

Computer Metadata File

Electronic Evidence Workbook 2026

Craig Ball

University of Texas Schools of Law, Computer Science and Information Science

LAW386N, INF385T, CS395T: Digital Evidence & Discovery

Spring 2026 Ver. 26.0218 ©Craig Ball, All Rights Reserved

Name:

THIS PAGE INTENTIONALLY BLANK

Course Workbook Reading Assignments: Classes 1-14

NOTE: This table is for your convenience; but the timing and scope of your responsibilities in this course are established by the latest Syllabus, not this table. ***Always* consult the latest Syllabus!!** The following Workbook exercises must be submitted, and readings completed, **prior to the start of class** on the stated dates. Late submissions will not be credited without advance permission.

Class 1: Wednesday, January 14
Workbook: Read pp. 7-50 Complete Exercise 1, p. 63
Canvas: *Powell-Hatcher v. Wal-Mart* and Preface**;** 2026 E-Discovery Predictions and Practice; and Preface

Class 2: Wednesday, January 21
Workbook: Read pp. 51-62, 64-84**;** Complete Exercises 2, 3 and 4
Begin familiarizing yourself with FRCP Rules 1, 26 and 34 including 2006 and 2015 Committee comments (pp. 599-634)
Canvas: *Green v. Blitz* w/ Preface (14 pp.); *Domus-v.-Arch-Insurance with Introduction.pdf (18 pp.)*

Class 3: Wednesday, January 28
Workbook: Read pp. 105-136**;** Complete Exercises 5 and 6
Canvas: *In Re: Weekley Homes;* THE SEDONA PRINCIPLES: THIRD EDITION, *pages 51-54*

Class 4: Wednesday, February 4
Workbook: Read pp. 137-192**;** Complete Exercises 7 and 8.
You will submit screenshots for Ex. 7 and a table of file signatures for Exercise 8. **No submissions sought for Exercises 9 and 10.**

Class 5: Wednesday, February 11
Workbook: Read pp. 193 -232; Complete Exercise 11, pp. 253-257 PICK UP EVIDENCE DRIVE!

Class 6: Wednesday, February 18 :MEET ON ZOOM TODAY
Workbook: Read pp. 258-292; Complete Exercise 12, pp. 293-306

Class 7: Wednesday, February 25
Workbook: Read pp. 307-318**;** Complete Exercises 13, pp. 307-318
Canvas: *Zubulake* Cases I-V with Preface, *focus on decisions I and III*

Class 8: Wednesday, March 4
Workbook: Read pp. 319-353, Complete Exercises 14 and 15, pp. 341-353
Canvas:*Mid Central OEHWF v Hoosiervac LLC*_2025.pdf

Class 9: Wednesday, March 11

Workbook: Read pp. 354-394, Complete Exercise 16, pp. 391-394; DO NOT DO Exercise 17.

Canvas: *Columbia Pictures v. Bunnell* with Preface; *In Re: NTL Securities Litigation* (pp. 17-20 only re: control); *Hynix v. Rambus* w/ Preface; The Sedona Conference Commentary on Rule 34 and Rule 45 "Possession, Custody, or Control" (pp. 475-527 only)

SPRING BREAK March 14-22

Class 10: Wednesday, March 25

Workbook: Read pp. 395-456, Complete Exercises 18 (all parts) and 19A

Canvas: *Victor Stanley I;* Fordham JOLT article re keyword search issues*; Maurer v. Sysco Albany LLC* w/Preface

Class 11: Wednesday, April 1

Workbook: Read pp. 457-471, Complete Exercises 19B and 20

Canvas: *In re: State Farm Lloyds;* Kessler, The Myth of Native Production with Ball Preface*;* Relativity, *From Pocket to Courtroom: The Legal Significance of Mobile Data*

Class 12: Wednesday, April 8

Workbook: Read pp. 472-493, Complete Exercise 21

Canvas: *Fast v. GoDaddy.com; Brookshire Brothers* decision (TxS.Ct.)*;* Rodriguez_Brookshire Bros. Article

Class 13: Wednesday, April 15

Workbook: Read pp. 494-554, Complete Exercise 22

Canvas: Grimm & Grossman*, Judicial Approaches to Acknowledged and Unacknowledged AI-Generated Evidence*; Ball, Tailor FRE 502(d) Orders to the Case

Class 14: Wednesday, April 223

Workbook: Read pp. 555-557, Complete Exercise 23 (AI-Assisted Review)

Canvas: Ball, Linked (Cloud) Attachments in E-Discovery-The Overlooked Evidence; Grimm, Capra, Joseph, *Authenticating Digital Evidence,* pp. 1-55

Contents

Goals for this Workbook 7

Introduction to Discovery in U.S. Civil Litigation 9

Why E-Discovery and Digital Evidence? 13

What Lawyers Must Know About Electronic Discovery and Digital Evidence 21

The Electronic Discovery Reference Model (EDRM) 25

Introduction to Data Storage Media 27

Exercise 1: Identifying Digital Storage Media 63

Introduction to Metadata 64

Exercise 2: Metadata and Hashing 85

Exercise 3: Exploring DISCO, an E-Discovery Platform Tool 93

Exercise 4: Metadata: Geolocation in EXIF 99

Processing in E-Discovery 105

Exercise 5: Encoding: Decimal, Binary, Hexadecimal and Base64 121

Exercise 6: Encoding: Running the Bases 127

Encoding and Steganography 137

Processing Part II 141

Exercise 7: Encoding: File Extensions 145

Exercise 8: Encoding: Binary Signatures 160

Exercise 9: Encoding: Unicode 163

Processing Glossary 188

Exercise 10: Metadata: File Table Data 193

Introduction to Digital Forensics and Examination Protocols 197

The Perfect Preservation Letter 234

Exercise 11: Compiling a Checklist of Sources 253
Computer Backup Systems 258
TEN PRACTICE TIPS FOR BACKUPS IN CIVIL DISCOVERY (2026) 270
Databases in E-Discovery 271
Exercise 12: Forensic Imaging 293
Exercise 13: Data Mapping 307
Mastering E-Mail in Discovery 319
Exercise 14: E-Mail Anatomy 341
Exercise 15: Encoded Time Values in Message Boundaries 345
The Next Arc of Exercises 354
Custodial Hold: Trust but Verify 356
The Leery Lawyer's Guide to AI and LLMs in Trial Practice 363
Exercise 16: Legal Hold 391
Mobile Data in E-Discovery 395
Simple, Scalable Solutions to iOS-Device Preservation 405
Exercise 17: Preserving a Mobile Device 412
Search Science and The Streetlight Effect in e-Discovery 424
Exercise 18: Honing Your Search Skills 428
Exercise 19A: Negotiating Search Protocols 455
Exercise 19B: AI Search Terms Analysis 457
Exercise 20: Processing, Culling, Search and Export 458
Exercise 21: Culling, Search and Export Part II 472
Forms that Function 473

Exercise 22: Forms of Production and Cost 494

The Annotated E-Discovery Protocol: A Primer on ESI Protocols 508

Preparing for Meet and Confer.....543

Exercise 23: AI-Assisted Review.....555

Exercise 24: Meet and Confer.....558

APPENDIX A: Materials for use with Exercises 16 and 24 561

The "E-Discovery Rules" (1,16,26,34 & 45) of the Federal Rules of Civil Procedure 590

Federal Rule of Evidence 502.....633

FRE Rule 902. Evidence That Is Self-Authenticating.....635

Exemplar FRE 902(14) Certification.....637

TRCP Rule 196.4 Electronic or Magnetic Data (enacted 1999) 638

Goals for this Workbook

The goal of this course is to change the way you think about electronically stored information and digital evidence. Despite its complexity, all digital content—photos, music, documents, spreadsheets, databases, social media, and communications—exist in one common form: as faint electric charges or impossibly tiny reversals of magnetic polarity. These minute polar fluctuations are read by a detector flying above the surface of a spinning disk on a cushion of air one-thousandth the width of a human hair in an operation akin to a jet fighter flying at more than 800 times the speed of sound less than a millimeter above the ground, precisely counting every blade of grass it passes!

That's astonishing, but what should astound you more is that there are *no pages, paragraphs, or markers of any kind* to define the data stream. It's the history, knowledge, and creativity of humankind distilled to two different states (on/off, one/zero) as a continuous, featureless expanse. It's a data stream that carries not only the information we store, but all the instructions needed to make sense of the data as well—all the data *about* the data required to play it, display it, transmit it, or otherwise put it to work. It's a reductive feat that will make your head spin and make you want to buy a computer scientist a drink.

Yet, it should comfort you to know that no matter the volume or variety of digital electronic evidence, electronic evidence is more alike than different. Digital evidence is daunting, but it's far easier to identify, preserve, collect, search, process, review, authenticate, challenge and produce once you divine its common threads. That's why information technologists are prone to dismiss overblown claims of burden with the observation, "It's just data."

In these exercises and readings, we will delve into the fascinating journey that data takes from its binary notation as ones and zeroes to the vast array of documents, communications, records, recordings, and formats that lawyers encounter in litigation. As you engage with the material, I encourage you to ask yourself "how" and "why?" Specifically, "How does it work?" and (for law students) "Why will understanding this benefit me in my practice?"

Through these exercises and readings, you will gain insights into key concepts such as:

1. How computers store and retrieve data
2. The differences between common storage media
3. The distinction between system and application metadata
4. The use and application of cryptographic hashing
5. The ways in which computers encode and decode data
6. The use of binary signatures and file extensions to identify file types
7. The encoding of foreign languages compared to English
8. The visible and invisible elements of an e-mail message

9. The location of deleted data and methods for data recovery
10. The preservation, processing, and presentation of data for attorney review
11. Techniques for reducing data volumes and isolating relevant evidence
12. The functioning of search tools and review platforms
13. The concept of forensic imaging
14. The use and challenges presented by load files
15. The impact of alternate forms of production on cost and utility
16. The growing role of artificial intelligence in discovery and review

While some of these concepts may seem removed from the day-to-day practice of law, they are essential building blocks—a foundation for developing practical skills. In any discipline, the absence of a solid foundation ultimately limits how high you can go.

And hey! *This is exciting stuff!* The amount of data that we create and store in today's digital world is staggering. Our lives are increasingly lived online through digital devices, and we are more connected, tracked, and monitored than ever before. We are more telemetered today than the Apollo astronauts of a generation ago! This presents both opportunities and challenges for the law, as the amount of probative and reliable evidence available to us has never been greater. We are truly blessed to have access to such powerful tools for finding the truth as we enter the age of AI.

I hope you will find this class to be enriching and engaging. Together, we will explore the thrilling world of data and its impact on the practice of law. Thank you for enrolling in this class.

Craig Ball, January 7, 2026

About the Author

I'm a Texas trial lawyer, 40-odd years in practice, and a certified computer forensic examiner. I was lead trial counsel in a personal injury, products liability and mass torts practice for 25+ years and I've taught digital evidence at UT for 13 years. My professional interests include digital forensics, emerging technologies, visual persuasion, e-discovery, and trial tactics. I limit my law practice to service as a court-appointed Special Master and consultant in Electronically Stored Information. I've spoken thousands of times at educational programs for the bench and bar and served as an instructor in computer forensics and electronic evidence to multiple law enforcement and security agencies around the world. For nine years, I wrote a syndicated column on computer forensics and e-discovery for American Lawyer Media called "*Ball in your Court*" and my writings still appear in the national media. You'll find my blog posts at ballinyourcourt.com, other writings at craigball.com and me in the flesh in New Orleans when I'm not flying, driving, training, hiking or sailing somewhere (travel is my true passion). You can reach me as craig@ball.net or call or text me at 713-320-6066. ***I'm here to help you, so reach out!***

Introduction to Discovery in U.S. Civil Litigation

Until the mid-20th century, civil trials in the United States could fairly be characterized as "trial by ambush." Each side typically learned the other's claims, defenses, and evidence only once the case was in trial. Lawyers had to find witnesses on their own; witnesses had no duty to talk to counsel, and what they volunteered informally could change substantially once they took the stand. Parties generally had no right to review an opponent's documents before trial.

Evidence scholar John Henry Wigmore captured this mindset in his 1904 treatise, *A Treatise on the System of Evidence in Trial at Common Law*. Quoting the Latin maxim *nemo tenetur armare adversarium suum contra se* ("no one is bound to arm his adversary against himself"), Wigmore wrote:

> To require the disclosure to an adversary of the evidence that is to be produced would be repugnant to all sportsmanlike instincts. Rather permit you to preserve the secret of your tactics, to lock up your documents in the vault, to send your witness to board in some obscure village, and then, reserving your evidential resources until the final moment, to marshal them at the trial before your surprised and dismayed antagonist, and thus overwhelm him. Such was the spirit of the common law; and such in part it still is. It did not defend or condone trickery and deception; but it did regard the concealment of one's evidential resources and the preservation of the opponent's defenseless ignorance as a fair and irreproachable accompaniment of the game of litigation. *Id*. Vol. III, § 1845, p. 2402.

Our forebears at common law,[1] feared that disclosure of evidence would facilitate unscrupulous efforts to tamper with witnesses and promote the forging of false evidence. The element of surprise was seen as a safeguard against skullduggery.

Legal reformers hated "trial by ambush" and, in the late-1930's, they sought to eliminate surprise and chicanery in U.S. courts by letting litigants obtain information about an opponent's case before trial in a process dubbed "discovery."[2] The reformer's goal was to streamline the trial process and enable litigants to better assess the merits of the dispute and settle their differences without need of a trial.

---

[1] "Common law" refers to law developed by judges in court decisions ("*precedents")*, as opposed to "Code law" (statutes enacted by legislatures).

[2] Discovery was not entirely new. Some jurisdictions allowed a Bill of Discovery, a separate equity action to obtain testimony or documents to support one's own case, but it generally did not permit probing the opponent's claims, defenses, or evidence—the information most needed. Some states also experimented with procedural codes allowing discovery of documents and testimony (for example, the Field Code in New York). For a comprehensive overview, *see* George Ragland, Jr., Discovery Before Trial (1932).

After three years of drafting and debate, the first **Federal Rules of Civil Procedure** took effect on September 16, 1938. Though amended many times since, the core discovery tools introduced then remain in place today:

- Depositions, oral and written (Rules 30 and 31)
- Interrogatories (Rule 33)
- Requests to inspect and copy documents and inspect tangible and real property (Rule 34)
- Physical and mental examinations (Rule 35)
- Requests for admissions (Rule 36)
- Subpoenas for witnesses and records (Rule 45)

Tools of Civil Discovery Explained[3]

Depositions

A deposition is sworn testimony taken before trial from a party or witness (the "deponent"). It may be conducted face-to-face (an oral deposition) or through written questions answered under oath (a written deposition). A court reporter typically prepares a transcript, and depositions are frequently video recorded. Deposition testimony can be used later if a witness is unavailable or to challenge a witness who offers inconsistent testimony at a hearing or trial.

Interrogatories

Interrogatories are written questions one party serves on another, to be answered in writing under oath. They function like a deposition on paper, without live testimony or a court reporter. The answers bind the responding party.

Requests for Production

Requests for Production seek documents, electronically stored information (**ESI**), tangible things, or access to land or property for inspection. This is the primary way parties obtain ESI.

Requests for Physical and Mental Examination

If a party's physical or mental condition is in dispute (for example, in a personal injury case), the opposing party may ask the court to order an examination by a qualified professional.

[3] This discussion concerns *civil* discovery in U.S. civil litigation. It differs in key respects from *criminal* discovery and from procedures in non-U.S. legal systems. In U.S. criminal cases, constitutional and statutory rules govern disclosure, with distinct obligations such as the prosecution's duty to disclose exculpatory evidence; the scope, timing, and tools differ markedly from civil discovery. Outside the United States, many civil law (inquisitorial) systems provide far more limited *party-driven* discovery; judges typically control evidence-gathering, and broad American-style discovery is uncommon. Even among common-law jurisdictions, the breadth of discovery varies significantly.

Requests for Admission

Requests for Admission ask a party to admit, under oath, that certain facts are true or that a document is genuine. Securing admissions narrows the issues and saves time and cost.

Subpoenas

A subpoena directs someone to act—usually to appear and testify or to produce documents or other evidence. Subpoenas are often used to obtain evidence from people or organizations that are not parties to the lawsuit. Strictly speaking, the Federal Rules of Civil Procedure do not characterize subpoenas as a discovery mechanism because their use is ancillary to depositions and proceedings. Still, they are employed so frequently and powerfully in discovery as to warrant mention here.

Scope of Discovery

Rule 26(b)(1) of the Federal Rules of Civil Procedure defines the scope of discovery:

> Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the case, considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit. Information within this scope of discovery need not be admissible in evidence to be discoverable.

In plain terms, discovery covers nonprivileged information that bears on the claims or defenses and is proportional to what the case reasonably requires. "Relevant" generally means the information has any tendency to make a fact more or less likely. Information can be discoverable even if it is not admissible at trial—for example, hearsay or documents whose authenticity is uncertain.

"Proportionality" was expressly added to Rule 26(b)(1) on December 1, 2015, though courts have long had authority to limit discovery that is excessive or unfair.

Certain matters fall beyond the proper scope of discovery because they enjoy a privilege from disclosure. The most common examples of these privileged matters are confidential attorney-client communications and attorney trial preparation materials (also called "attorney work product"). Other privileged communications include confidential communications between spouses, between priest and penitent and communications protected by the Fifth Amendment of the U.S. Constitution.

Protection from Abuse and Oppression

The discovery rules are both sword and shield. Parties may object to improper requests and withhold responses based on valid objections. They may also ask the court for a **protective order.**

Under Rule 26(c), a court may, for good cause, protect a party or person from annoyance, embarrassment, oppression, or undue burden or expense by ordering one or more of the following:

(A) forbidding the disclosure or discovery;
(B) specifying terms, including time and place or the allocation of expenses, for the disclosure or discovery;
(C) prescribing a discovery method other than the one selected by the party seeking discovery;
(D) forbidding inquiry into certain matters, or limiting the scope of disclosure or discovery to certain matters;
(E) designating the persons who may be present while the discovery is conducted;
(F) requiring that a deposition be sealed and opened only on court order;
(G) requiring that a trade secret or other confidential research, development, or commercial information not be revealed or be revealed only in a specified way; and
(H) requiring that the parties simultaneously file specified documents or information in sealed envelopes, to be opened as the court directs.



Why E-Discovery and Digital Evidence?

As noted, discovery is the formal process by which parties to litigation exchange information about witnesses and evidence. That spare definition, often dispatched in a few hurried minutes during first-year Civil Procedure class, cannot prepare students for the extent to which discovery devours litigators' lives. For every hour spent in trial, attorneys and trial teams devote hundreds, sometimes *thousands*, of hours to discovery and its attendant disputes. Trials are the tip of the spear. Discovery is the shaft.

This course treats electronically stored information—ESI—as the day-to-day substrate of civil discovery and the lens through which we study the rules and practice. This course is structured accordingly: early modules teach the nuts-and-bolts of digital evidence, including digital storage, encoding, metadata and hashing, and then move to hands-on preservation, search, and production so you appreciate why discovery takes the time and care it does in practice.

Before we dive in, let's construct a mental map of why these things matter and how they connect. Electronic discovery is not a bag of tricks or a menu of tools; it is a *system*. This course is about learning that system well enough to navigate it competently, ethically, and with confidence.

So, this chapter is a brief orientation to the terrain we will traverse in depth throughout the Workbook and the semester. In later chapters and exercises, we will handle ESI directly, from identification and preservation through collection, processing, review, production, and presentation—all comprising the conceptual schematic called the **Electronic Discovery Reference Model** or **EDRM**.

Discovery is also the trial lawyer's most demanding ethical challenge. It requires lawyers to seek out and surrender information that may aid and comfort the enemy—often over client objection, irrespective of the merits, and despite deep distrust of the opponent. Is there another profession that so routinely obliges its practitioners to act against perceived self-interest in service of a larger system of justice?

We will confront the discipline required to preserve and produce electronic evidence even when human nature supplies the impulse is to hide it or "clean it up," and we will study the **legal hold** as the primary instrument by which counsel enforces that duty across an organization.

Electronic discovery is also challenging because it demands specialized knowledge that few practicing lawyers possess. E-discovery skill has long been a requisite component of trial lawyer competency yet is often denigrated, deferred or delegated. To close that gap, our Workbook's early exercises require you to decode digital data, calculate hashes, capture and read message source, and image evidence media, so both law students and computer science/information science graduate students may acquire a shared, practical vocabulary for authenticity, integrity, and defensible handling of electronic evidence.

Civil discovery is a high-stakes game of "Simon Says." Counsel must phrase demands for information with enough precision to target what is relevant, yet with sufficient breadth to forestall evasion. It is as confounding as it sounds, and it is something of a miracle that discovery works as well as it does. The forces that make it work are not elegant rules or perfect systems, but counsel's professional integrity and judges' willingness to enforce the rules when integrity falters.

Professional integrity in discovery is not mere altruism. Failure to preserve and produce relevant evidence carries consequences ranging from damaged reputations and costly remedial measures to so-called "death penalty" sanctions, where a discovery abuser forfeits the right to pursue or defend a claim. Lawyers themselves may face monetary sanctions, adverse inferences, and referral to disciplinary authorities. Discovery is where credibility is earned—or squandered.

The modern American system of civil discovery rests on a simple premise: just outcomes are more likely when parties have access to facts established by relevant evidence. Because relevant evidence often lies within the exclusive control of those whose interests are *not* served by disclosure, justice requires a mechanism to compel production—subject to limits grounded in privilege, privacy, and proportionality.

Once more, Rule 26(b)(1) of the Federal Rules of Civil Procedure articulates the scope of discovery as, "Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the case… Information within this scope of discovery need not be admissible in evidence to be discoverable." Rule 401 of the Federal Rules of Evidence[4] defines evidence as relevant if it has any tendency to make a fact more or less probable and the fact is of consequence in determining the action (*i.e.*, the fact is "material"). In practice, proportionality turns on what's at stake, the relative accessibility of the information and the burden and benefit of supplying it. Throughout the course, you will weigh proportionality, *e.g.*, when proposing forms of production, choosing between forensic imaging and targeted collection or when negotiating keyword search and file culling.

Relevant.
Proportional.
Nonprivileged.

Commit these touchstones to memory; we will return to them often. They undergird the course roadmap, which follows the lifecycle of ESI: identifying and mapping sources, preserving against routine loss, collecting defensibly from devices and cloud services, processing and indexing for search, reviewing with rigor, producing in forms that function, and presenting evidence that can be authenticated and trusted.

[4] Note the distinction here between the Federal Rules of Civil *Procedure* and those of Civil *Evidence.* Each state has rules, too. Fear not! The handful of Rules from each you'll need to know for this course are set out at the back of the Workbook.

The discovery of an opponent's ESI often begins with a Request for Production under Rule 34 of the Federal Rules of Civil Procedure or, for matters in state court, a similar state rule. Rule 34 lets a party request that any other party produce designated documents or electronically stored information—including writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations in the responding party's *possession, custody, or control*. The responding party must respond in writing within 30 days and may lodge specific objections and withhold production pursuant to those objections. Much modern evidence also resides with nonparties, reached by subpoena, which this course treats in turn when we tackle discovery planning and data mapping beyond the parties themselves.

The simplicity of the rule governing requests for production hardly hints at its complexity in practice. A multibillion-dollar industry of service providers and consultants exists to support discovery, and a crazy quilt of rulings lays bare the ignorance, obstinance, guile, and ingenuity of lawyers and clients grappling with the preservation and exchange of electronic evidence. Here, *method matters*. Early in the term we preview three habits that make complexity tractable and keep you out of trouble, each developed fully in later chapters and exercises:

First, form matters. Whether evidence is produced in native or near-native forms with metadata or reduced to static images profoundly affects utility, searchability, and cost. You will learn to preserve, specify and deliver forms that preserve content and functionality, and to recognize when images and load files are appropriate and when they are not.

Second, preservation precedes requests. The duty to preserve arises when litigation is reasonably anticipated—typically long before formal requests issue—and is implemented through legal holds tailored to people, systems, and risk. You will design and critique hold strategies that combine custodial instructions with objective, systemic measures, and you will practice calibrating those measures to proportional needs.

Third, search is a science, not a magic word list. You will experiment with indexing, culling, deduplication, and iterative keyword refinement, then compare conventional techniques with AI-assisted methods to understand strengths, pitfalls, and defensibility across tools.

This opening chapter is only a first taste. What follows in this Workbook are the concrete skills that make electronic discovery work in practice: reading and preserving metadata without corrupting it; computing and using hash values to establish integrity; distinguishing forensic images from targeted collections; negotiating modern ESI protocols and production specifications; conducting and documenting defensible search; and delivering productions in forms that function. As you progress, your appreciation for the rules and the ethics introduced here will deepen through repetition, testing, and the discipline of doing the work yourself.

Clean Up on Aisle Nineteen



To appreciate what competent counsel must know about digital discovery, consider the everyday case where a customer slips and falls in a grocery store. A store employee witnesses the fall, assists the customer, and escorts her to the manager, who prepares an incident report. The customer alleges she slipped on grease near a roasted-chicken display. Pain worsens, an emergency room visit follows, and months later physicians diagnose a spinal injury requiring surgery. When the grocer refuses to cover care, the customer retains counsel.



From a discovery perspective, the case stands on three legs: liability, causation, and damages. To establish liability, the plaintiff must show duty and breach. A store owes patrons reasonably safe premises and must act reasonably to correct or warn of hazards such as slippery floors. That obligation turns on awareness: *actual knowledge* of the condition or *constructive knowledge,* because the hazard existed long enough that reasonable inspection would have revealed it. Did the store know of the hazard or, acting reasonably, should it have known?

Defense counsel counters that there was no grease, or that another customer caused any spill, and despite reasonable care the store had no opportunity to discover and remedy it before the plaintiff fell. The defense also argues contributory negligence, contending plaintiff failed to watch where she walked, and challenges causation and damages by pointing to possible pre-existing health conditions or unrelated trauma.

As plaintiff's counsel ponders the potentially relevant evidence in the store's control, he wonders:

1. Who might have witnessed the fall or the conditions?
2. Were witness statements obtained?
3. How did the store clean up after the fall?
4. Were photographs taken?
5. Were video cameras monitoring the premises?
6. Is there a history of other falls?
7. Did the roasted chicken display leak?
8. How frequently are the floors inspected and cleaned?



Defense counsel has her own questions:

1. Did the plaintiff stage the fall to profit from a claim?
2. Did the plaintiff suffer from a pre-existing condition?
3. Has the plaintiff made other claims?

4. Was the plaintiff impaired by drink, drugs or disability?
5. Has the plaintiff behaved inconsistently with her claimed infirmities?

Both sides wonder whether the other side acted diligently to preserve relevant evidence and if anyone has altered or destroyed probative material. The court will only permit discovery in proportion to the needs of the case. In gauging proportionality, comparable cases have prompted damage awards ranging from one-half million to two million dollars; but, absent liability and causation, there will be no recovery.

Plaintiff's counsel inventories potential proof within the store's control: eyewitnesses and their statements, photographs, cleanup steps, surveillance video, inspection schedules and logs, prior similar incidents, the condition and maintenance of the roasted-chicken kiosk, and the store's written logs, policies and procedures, to name a few.

Defense counsel likewise considers proof within plaintiff's control: medical history and prior claims, earnings history, and activities and statements inconsistent with the claimed disability.

Both sides, rightly, worry about preservation—what was kept, what was altered, and what was lost—because the court will confine discovery to only what is *proportional to the needs of the case.* In gauging proportionality, comparable cases have prompted damage awards ranging from one-half million to two million dollars; but, absent liability and causation, there can be no award of damages.

Now, place this dispute in its everyday *digital* setting. The store is part of a national chain with policies for monitoring hazards and documenting incidents. Digital video cameras record the store, warehouse, and parking lot. A point-of-sale database logs transactions; self-checkout kiosks include cameras. Employees clock in and out electronically. Third-party suppliers and contractors come and go. Nearly everyone carries a mobile device recording geolocation and activity. A corporate claims database tracks incidents and dispositions. Thus, even a simple slip on some schmaltz casts a long shadow of electronic artifacts.

Video is decisive evidence. Store policy requires prompt review and preservation of surveillance video before routine overwriting every 14 days. The manager saves footage from the deli-area feed beginning one minute before the fall until five minutes afterwards—all before the cleanup occurs—yet a kiosk blocks the view of the floor from that one camera. The manager also preserved video of the plaintiff arriving and leaving the premises. In one, the plaintiff is looking at her phone. The

surveillance system overwrote other video recordings two weeks later. The only contemporaneous photographs of the area were taken after employees mop and place caution cones. The store maintains they mopped grease spilled by the plaintiff when she fell, not pre-existing grease.

The parties engage in discovery seeking the customary complement of medical records and expenses, lost earnings documentation, store policies and procedures, similar prior incidents, and incident investigations.

To identify eyewitnesses or customers likely to have spilled grease, plaintiff requests the store "*produce for a period one hour before and after the fall, any photographic or transaction record (including credit- and loyalty-card identifying data) of any persons on the premises*." Plaintiff makes the same request for "*any persons who purchased roast chicken*." Plaintiff also demands the names, addresses, and phone numbers of employees or contractors on the premises within one hour on either side of her fall.

In its discovery, the store asks that plaintiff "*produce any texts, call records, application data or other evidence of phone usage for one hour before and after the alleged fall and the contents of any social networking posts for six months prior to the alleged injury to the present where any content, comment, or imagery in the post touches or concerns the Plaintiff's state of mind, physical activity, or consumption of drugs or alcohol*." The store also demands that plaintiff produce "*data from any devices (including, but not limited to, phones, apps, fitness equipment, fitness monitors, and smart watches) that record or report information about the plaintiff's sleep, vital signs, activity, location, movement, or exertion from six months prior to the alleged fall to the present date*."

At this point, it is common for both sides to balk. It emerges that neither party preserved all such ESI when the duty arose. The parties will meet and confer to narrow scope and discuss sources and forms. If agreement proves elusive, they move to compel and for protective relief.

The parties will object on various grounds, alleging that the information isn't relevant, doesn't exist, or is not reasonably accessible. Lawyers will point to undue burden and cost, oppression, excessive inroads into private matters, and even claim the data requested is privileged or a trade secret. Requests will be challenged as "disproportionate to the needs of the case."

One side assures the judge it's just a few keystrokes to gather the data sought. With equal confidence, the other side counters that the task requires teams of expensive experts and months of programming and review.

Case in point: Plaintiff's counsel notes that every roast chicken bears a Universal Product Code (UPC) scanned at checkout, updating inventory and recording transactions. Purchasers who use credit, debit, loyalty, or assistance cards can be identified through those transactions. "*It's on the receipts*," counsel urges. "*Just print them out*." *"It's the same for every employee," he adds, "they scan people in and out like roast chickens."*

Plaintiff's counsel is less sanguine about phone, social, and wearable data, uncertain how to collect, review, and produce whatever's not been lost to the passage of time. "*It will take forever*," he protests. "*And who knows what's relevant? It's disproportionate!*" The store concedes it tracks purchases and card usage, but not in the same system. Pairing identifiers may require custom queries or vendor assistance, and even then, purchasers are not necessarily eyewitnesses. "*Judge, this is a fishing expedition*," defense counsel says.

Meanwhile, time and technology erode evidence. Legacy logs purge. Devices are upgraded. Video overwrites. Only a minute of pre-event footage remains when the first request arrives—too little to show how long the grease was present. Plaintiff moves for sanctions based on failure to preserve. The defense counters that a minute seemed sufficient and points to plaintiff's deletion of a social media account, arguing prejudice from missing posts that might show activity inconsistent with claimed injury and damages.

So, the outcome of even this humble slip-and-fall dispute hinges on a panoply of relevant and probative sources of electronic evidence. Scale the same issues to a class action for 100,000 employees, a mass data-breach for one hundred million customers, or a patent war between tech titans, and the volume, velocity, and variety of ESI multiply. We cannot throw up our hands and say, "*Too much, too hard, too expensive!*" The task, rather, is to balance access to relevant evidence against burden, expense and intrusion, guided by competence. Parity matters: when *both* sides understand information systems and the mechanics of digital evidence, consensus follows more readily.

What does competence look like here? It includes knowing what ESI is, where it lives, and how it should be preserved and produced in forms that function; understanding metadata's role in authenticity and search; discerning when collection can be targeted and when forensic preservation is warranted; and communicating obligations in plain language to technical personnel. It is the practical ability to seek, cull, search, sort, review, and produce ESI defensibly, and to estimate time and cost credibly. We will address these skills in turn across the Workbook, beginning with data storage and metadata, then hashing and chain of custody, then collection, processing,

search, and forms of production, culminating in meet-and-confer strategy and protocol drafting. Along the way, we will examine sanctions for spoliation[5] and proportional remedies, and we will test, with real tools, how "just a few clicks" sometimes is, and often is not, the whole truth.

We expect banking lawyers to understand banking and real-estate lawyers to understand closings. Trial lawyers should understand electronic evidence and discovery. That starts with fundamentals. This course proceeds on the premise that while one can be a fine physician without biochemistry, you would rather your doctor had taken and passed it. Here, you will master the argot of ESI: storage, collection, encoding, processing, metadata, search, forms, and the vocabulary of IT and forensics. If you believe trial lawyers need only know the law, this course will challenge that view. Here, we celebrate the "e" in e-discovery and e-evidence. You will work with data, use modern tools, and learn to speak enough "geek" to lead. Law students will emerge as more tech-savvy litigators; computer science and information students will see ESI through the lens of relevance, privilege, privacy, proportionality, and admissibility in the civil justice system. All will become better equipped to manage the intersection of information technology and litigation.

[5] **Spoliation** is the destruction, alteration, loss, or failure to preserve evidence—particularly electronically stored information—when a party knows or should know that the evidence is relevant to pending or reasonably anticipated litigation, and the evidence cannot be restored or replaced through additional discovery.

What Lawyers Must Know About Electronic Discovery and Digital Evidence

Discovery sometimes entails gaining access to physical objects—real property, defective products, or people (e.g., medical examinations). But in modern litigation, discovery overwhelmingly concerns information: human recollection elicited through testimony and records memorialized either as ink on paper or as electronically stored information. Discovery becomes *electronic discovery* when the relevant matter consists of data created, transmitted, or stored in digital form.

Requests for production were born of simpler times. One party asked another (or a third party) to produce documents or records, either by specifying particular items or by describing categories of information sought. The responding party had roughly thirty days to locate responsive materials and make them available for inspection or copying. Privileged material could be withheld or redacted, provided the responding party supplied a log sufficient to permit evaluation of the claim. Courts served as referees—shielding parties from undue burden and compelling production when responses fell short.

For centuries, this system operated against a backdrop of paper-centric recordkeeping. Until the mass adoption of personal computing and the internet in the 1990s, nearly all personal and business communications existed as paper documents or as ephemeral conversations—literally vibrations in the air. Paper discovery was expensive, contentious, and time-consuming, but it was manageable. People were schooled from childhood in reading, organizing, and understanding paper records. Documents felt finite. You could see the pile and estimate the work.

Then everything changed.

Today, virtually all personal and business communications involve the movement of electrons. Conversations once fleeting are now persistent texts. Paper documents have become pixels on screens, often designed to mimic the look and feel of paper while behaving very differently. More importantly, digital communications and transactions are accompanied by information describing the who, what, when, where, and how of their creation and use. This data about data—*metadata*—often conveys more probative value than the content itself.

In the span of a generation, civilization stopped doing things the way it had for millennia. The legal system experienced digital whiplash.

For many senior lawyers, paper remains the sole enduring frame of reference for discovery. Paper discovery was tangible, finite, and familiar. Electronic discovery feels infinite—replicated, distributed, fragmented, and intangible. People store ESI wherever it is convenient, without structure or discipline, assuming that when the time comes, the skills honed on paper will suffice. They do not.

Electronic discovery is more complex than paper discovery, but so are electric lights compared to candles and automobiles compared to wagons. It is a worthwhile complexity. ESI is inherently searchable. It can be sorted, filtered, deduplicated, and analyzed at scale. Metadata affords ways to assess authenticity, integrity, and context that paper never could. The variety, ubiquity, and richness of ESI blaze new trails to the truth of events and transactions.

Yet more information does not inevitably yield better information. Volume can obscure signal. Accordingly, competence in electronic discovery requires not only the ability to locate relevant information, but also the discipline to manage noise—separating what matters from what merely exists.

Character and Competence in Discovery

We live in a golden age of evidence. Never before has so much probative information been generated, and never has that information been so precise or objective. Yet many lawyers regard electronic evidence as a burden rather than an asset—like farmers complaining of oil beneath their land. The problem is not the evidence; it is unfamiliarity with its value and handling.

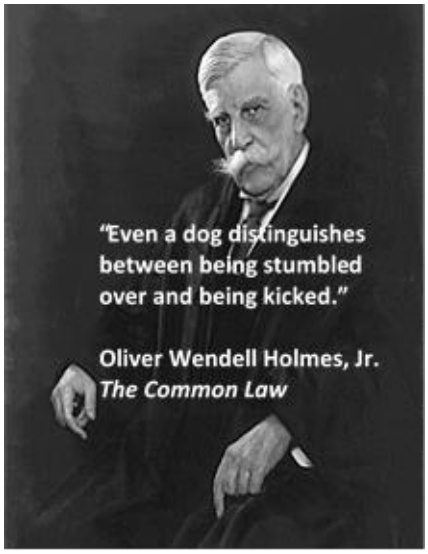
That unfamiliarity is hardly surprising. Few lawyers practicing today received practical instruction in electronic evidence. Law schools historically ignored it. Continuing legal education often skims the surface. Apprenticeship fails where mentors themselves lack technical fluency. Competence in electronic discovery remains rare, and there is little momentum to change that beyond vague exhortations to stay abreast of "relevant technology." [6]

Discovery is much maligned as a too costly, too burdensome, too intrusive fishing expedition.[7] As digital evidence proliferates—across cloud services, mobile devices, collaboration platforms, and the Internet of Things—the gap between competent and incompetent counsel widens. Discovery abuse and obstruction flourish where ignorance prevails.

[6] Rule 1.1 of the American Bar Association Model Rules of Professional Conduct provides that, "[a] lawyer shall provide competent representation to a client. Competent representation requires the legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation." Comment 8 to Rule 1.1 adds, "[t]o maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology…." Emphasis added. Many states have added the same language to their disciplinary rules.

[7] Such concerns are not new. Well before the original Rules went into effect, the Chairman of the Rules Advisory Committee exclaimed, "We are going to have an outburst against this discovery business unless we can hedge it about with some appearance of safety against fishing expeditions." Proceedings of the Advisory Committee (Feb. 22, 1935), at CI-209-60-0-209.61. Many still curse "this discovery business," particularly those most likely to benefit from the return of trial by ambush and those who would do away with trials altogether.

Character in discovery is hard to instill and harder still to measure. Competence is not. Competence can be taught, required, and assessed. And when competence is present, character becomes easier to discern. As Justice Holmes observed, even a dog knows the difference between being stumbled over and being kicked.



Recognizing this, some regulators have articulated concrete expectations for lawyer competence in electronic discovery. A leading example is the advisory opinion issued by the State Bar of California's Standing Committee on Professional Responsibility and Conduct in 2015. The Committee began with a simple premise: when representation is likely to implicate electronic discovery, competent counsel must *assess* that reality at the outset, evaluate their own knowledge and resources, and, if necessary, acquire the requisite skill, associate competent assistance, or decline the engagement.

From that premise flows a practical mandate: ***learn it, get help, or get out.***

The Committee identified nine core capabilities that competent counsel must be able to perform—personally or through appropriately supervised assistance—including:



- assessing electronic discovery needs;
- implementing appropriate ESI preservation procedures;
- understanding a client's ESI systems and storage;
- identifying custodians of relevant ESI;
- performing appropriate searches;
- collecting ESI in a manner that preserves integrity;
- advising clients on preservation and collection options;
- engaging in meaningful meet-and-confer discussions; and
- producing ESI in recognized and appropriate forms.

These are baseline competencies, not aspirational ideals. Each requires a blend of legal judgment and technical understanding that cannot be satisfied by outsourcing decisions wholesale to vendors.

Competent preservation demands more than issuing a generic legal hold. Counsel must understand where data resides, which data is fragile or ephemeral, how routine systems purge information, and how preservation can occur without corrupting evidence or metadata. Identifying custodians likewise requires understanding modern information systems and communicating effectively with technical personnel.

Search is not intuition or incantation. It is a discipline; *a science*. Tools differ. Queries must be tested to balance precision (how often your hits are actually relevant) and recall (how much of the relevant universe you managed to find). Errors at the search stage cascade into downstream cost and risk. Counsel who cannot understand how ESI is indexed, culled, and filtered cannot meaningfully negotiate protocols or assess proportionality.

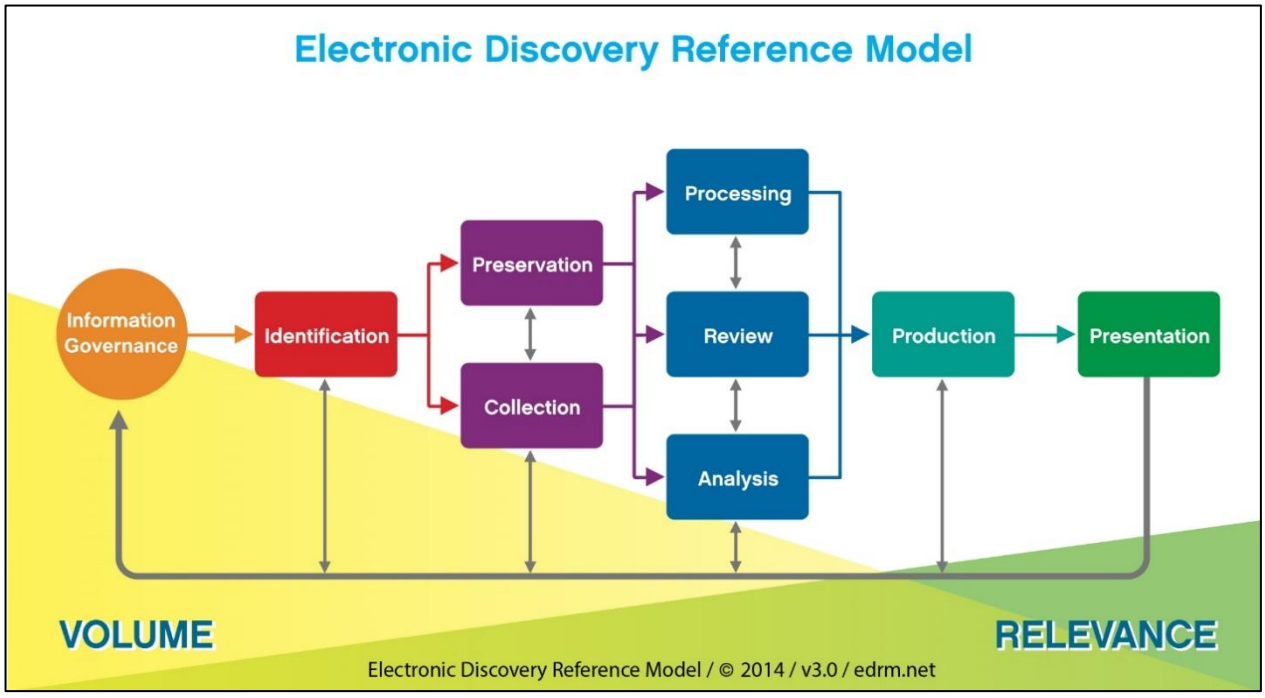
Competent counsel also understands that not all ESI should be handled alike. Some data may be preserved and searched in place; other data must be collected and processed. Forms of production matter. Native production may preserve functionality and metadata; image-based production may be appropriate in other circumstances. Each choice carries strategic and economic consequences.

Big firm lawyers and corporations have grown comfortable outsourcing electronic discovery to service providers whose expanding roles blur the boundary between technical support and substantive law practice. Delegating e-discovery duties to vendors has allowed lawyers to muddle through, but at great expense and at peril of error, delay and miscommunication. Delegating electronic discovery to vendors does not absolve lawyers of responsibility. Vendors provide tools and labor; lawyers bear ethical and strategic accountability. If trial lawyers intend to remain the captains of litigation, they must be competent in *doing*, not merely *buying*, e-discovery. If you want to be at the helm, it's wise to know how to steer.

Virtually all evidence today is born digitally. Every trial lawyer need not be a computer scientist or forensic examiner, but every trial lawyer must be minimally competent in the fundamentals of electronic evidence: where it lives, how it behaves, how it is preserved, how it is searched, and how it is produced in forms that function. That is the premise of this course.

The Electronic Discovery Reference Model (EDRM)

In 2005, when e-discovery had hardly entered lawyer lexicons, Minnesotans George Socha and Tom Gelbmann saw we weren't speaking the same language. George and Tom proposed a conceptual view of the electronic discovery process, expressed as a diagram to serve as a basis for discussion and analysis. The now-iconic EDRM diagram helped the legal and vendor communities communicate more clearly and reinforced the iterative nature of a successful e-discovery model; one that vanquishes volume and enriches relevance.



Information Governance

Getting your electronic house in order to mitigate risk & expenses should e-discovery become an issue, from initial creation of ESI (electronically stored information) through its final disposition.

Identification

Locating potential sources of ESI & determining its scope, breadth & depth.

Preservation

Ensuring that ESI is protected against inappropriate alteration or destruction.

Collection

Gathering ESI for further use in the e-discovery process (processing, review, etc.).

Processing

Reducing the volume of ESI and converting it, if necessary, to forms more suitable for review & analysis.

Review

Evaluating ESI for relevance & privilege.

Analysis

Evaluating ESI for content & context, including key patterns, topics, people & discussion.

Production

Delivering ESI to others in appropriate forms & using appropriate delivery mechanisms.

Presentation

Displaying ESI before audiences (at depositions, hearings, trials, etc.), especially in native & near-native forms, to elicit further information, validate existing facts or positions, or persuade an audience.

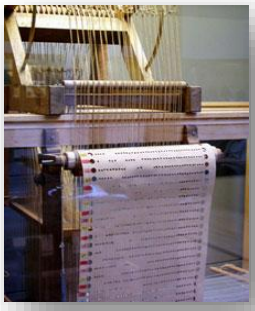
In this course, the EDRM serves as our conceptual framework for understanding the functional stages of electronic discovery from identification through presentation.

Too, we must peer past the colorful confines of the EDRM and weigh the evidentiary and societal aspects of electronic evidence. In a world without notarized pen-and-ink signatures—where simply flipping a bit value changes everything—how do we prove authenticity or attack the integrity of digital evidence? How will we trust, train and perhaps restrain Artificial Intelligence (AI) systems serving as proxies for human judgment? How do we balance personal privacy and the need to safeguard trade secrets against the obligation to bring the best evidence into court?

Introduction to Data Storage Media

Mankind has been storing data for thousands of years, on stone, bone, clay, wood, metal, glass, skin, papyrus, wax, paper, plastic and film. In fact, people were storing data in binary formats long before the emergence of modern digital computers. Records from 9th century Persia describe an organ playing interchangeable cylinders. Eighteenth century textile manufacturers employed perforated rolls of paper to control Jacquard looms, and Swiss and German music box makers used metal drums or platters to store tunes. At the dawn of the Jazz Age, no self-respecting American family of means lacked a player piano capable (more-or-less) of reproducing the works of the world's greatest pianists.

Whether you store data as a perforation or a pin, you're storing binary data: two data states: hole or no hole, pin or no pin, one or zero.



Jacquard Loom



Music Box



Player Piano

Punched Cards

In 1889, U.S. inventor Herman Hollerith (1860-1929) was granted a patent for his system for storing data on perforated paper cards, a system that revolutionized the 1890 U.S. census. Using 43 Hollerith machines and 62 million punched cards, the time to tabulate the census dropped from eight years to three.[8] In the 1930's, demand for mechanized data storage led to widespread adoption of Hollerith cards as a fast,



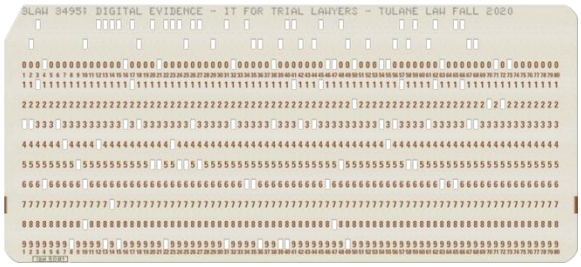
VITAL VOCABULARY

Actuator Arm
Analog
Areal Density
Bates Numbering
Behind the Firewall
Binary
Cloud
Clusters
Cylinders
De-Duplication
De-NISTing
Electromagnetic
Encoding
File System
Form Factor
Forensic Image
Hashing
HDD (Hard Disk Drive)
Head Crash
Interface
IT
Logical Block Address (LBA)
Master File Table
Metadata
Network Share
NTFS
RAID Array
Random Access
Read-Write Head
SAN and NAS
Sectors
Slack Space
Solid State
Solid-State Drive (SSD)
Storage Media
Tracks
Unallocated Space

[8] Hollerith founded The Tabulating Machine Company, based in the Georgetown neighborhood of Washington, D.C. Hollerith's company was later merged with others and renamed International Business Machines Company, now IBM.

practical and cost-effective binary storage media. These punched cards, initially made in a variety of sizes and formats, were ultimately standardized by IBM as the 80 column, 12 row (7.375" by 3.25") format (right) that dominated computing well into the 1970's. In the mid-1950s, punched card sales accounted for 20 percent of IBM's revenues! From 1975-79, this author spent many a night in the basement of a computer center at Rice University typing program instructions onto these unforgiving punch cards, cousins to the oily, yellow perforated paper tape (right) that Bill Gates and I used on opposite coasts to program mainframe computers via a teletype terminal in the early 1970s.

IBM 5081-style 80-column card





In the punch card era, storing programs and data sets meant storing stacks of IBM cards, often in motorized "tub files" (right).



The encoding schemes of these obsolete media differ from those we use today principally in speed and scale; but the binary fundamentals are unchanged and connect our toil in e-discovery and computer forensics to the likes of Charles Babbage, Alan Turing, Ada Lovelace, John von Neumann, Robert Noyce and both Steves (Wozniak and Jobs).

The IBM punched card held 80 columns of 12 punch positions, or 960 possible bit locations. Nominally, that's 120 eight-bit bytes; but because eight columns weren't always used for data storage, the effective capacity was closer to 864 bits, or 108 bytes. In practice, each column was typically dedicated to a single 7- or 8-bit character, so a card realistically stored about 80 characters—roughly 80 bytes.[9]

[9] After hours researching the capacity question, I couldn't arrive at a definitive answer because capacity varied according, *inter alia*, to the type of information being stored (binary versus ASCII) and a reluctance to punch out too many adjacent perforations lest it become a "lace card" too fragile to feed into a reader.

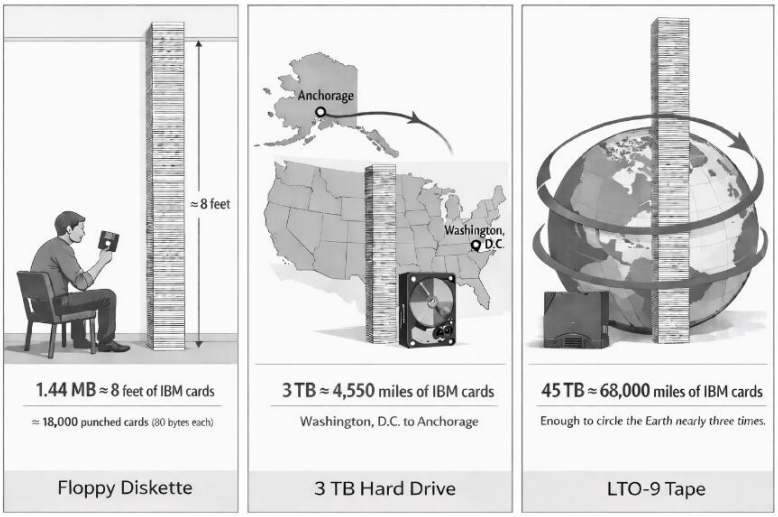
Using the 108-byte value, the formatted “1.44 MB” 3.5-inch floppy disks commonly used from the mid-1980s to early 2000s held about 1.4 million bytes. A single floppy therefore stored the equivalent of roughly 13,653 IBM cards—about seven 2,000-card boxes, or, at 143 cards to the inch, an eight-foot stack. That’s a common ceiling height and taller than anyone who ever played in the NBA.





The prim Eisenhower-era programmer in the photo below steadies 62,500 punched cards said to hold five megabytes of program instructions for the massive SAGE (Semi-Automatic Ground Environment) military computing network (at roughly 80 bytes per card)

Fast-forward to today’s capacious hard drives: a sixty-dollar, three-terabyte drive holds 3,298,534,883,328 bytes. That’s over *forty-one billion* IBM cards (41,231,686,041, to be precise). Stacked at 143 cards per inch, our tower of cards would rise about *4,550 miles*—roughly the distance between Washington, D.C., and Anchorage.



So, consider backup tape. The 45-terabyte (compressed) capacity of an LTO-9 cartridge corresponds to roughly 615 billion IBM cards. Stacked one atop another, they would span more than *68,000 miles*—enough to circle the Earth at the Equator nearly three times.

These are not real-world metrics because much storage capacity is lost to file-system overhead. If you used a warehouse for physical storage, you would sacrifice space to shelving and aisles, and you would find that not everything fits wall-to-wall and floor-to-ceiling. Similarly, digital storage sacrifices capacity to file tables and wastes space by using fixed cluster sizes. If a file is smaller than the clusters (i.e., storage blocks)

allocated to it, the bytes between the end of the file and the end of the cluster are wasted "slack space."

Electromagnetic Storage: Turning Polarity into Proof

Punched cards made data tangible, but they were slow, bulky, and unforgiving. The next great leap in storage came from exploiting a different physical phenomenon: **magnetism**. Magnetic storage did not change the binary nature of data; it found a denser, faster, and reusable way to express it. Instead of holes and cards, information is recorded as tiny regions of magnetization—**magnetic domains**—each oriented in one of two directions ("poles") that are read as a binary state.

At a conceptual level, nothing fundamentally changed. Data is still reduced to two states, still recorded on a physical medium, and still dependent on machinery to interpret it. What changed was scale. Magnetic media allowed enormous volumes of data to be packed into far smaller spaces and accessed far faster than any punched-card system could manage.

Early magnetic storage appeared first on tape, then on disks. Tape favored sequential access and long-term retention; disks enabled random access and speed. Both persist today, not as relics, but as workhorses of modern computing and electronic discovery.

Magnetic Tape: Sequential, Durable, and Still Very Much Alive

The earliest popular form of electromagnetic data storage was **magnetic tape**. Spinning reels of tape became a clichéd visual metaphor for computing in films and television from the 1950s through the 1970s. Though the miles of tape once visible on those reels now reside inside cartridges and cassettes, tape remains a surprisingly enduring medium for the backup and archival storage of electronically stored information.





Compact cassette tape (right) was the earliest data storage medium for personal computers, including the pioneering Apple II, the Radio Shack TRS-80, and IBM's first personal computer, the Model XT.

Magnetic tape predates hard disks, and—contrary to periodic predictions of its demise—continues to play a central role in modern data storage. Tape is not fast, and it is not elegant. It is, however, **dense, durable, inexpensive at scale, and remarkably reliable**. Those qualities explain its persistence and its recurring appearance in discovery disputes.

Unlike disks, which support random access, tape is a **sequential "("serial") medium**. To reach data recorded near the end of a tape, the drive must physically wind past everything that comes before it. That characteristic once made tape a poor choice for interactive computing but an excellent choice for backup, archiving, and long-term retention. Those use cases have not disappeared. If anything, they have grown with the explosion of data volumes.

How Magnetic Tape Stores Data

At a physical level, tape storage resembles disk storage more than its form factor[10] suggests. Data is recorded by altering the magnetic orientation of particles embedded in a thin coating on a flexible ribbon. As with disks, information is expressed as binary states—magnetized one way or the other—and interpreted by a read/write head as ones and zeros.

Where tape differs is in **layout and access**. Data is written in long linear tracks that run the length of the tape. Modern tape systems organize data into blocks and files, but the underlying constraint remains: the tape must move past the head to be read. There is no shortcut.

Tape Drives and Cartridges

A tape cartridge contains hundreds of meters of tape, precisely wound and tensioned. When inserted into a tape drive, the tape is drawn out of the cartridge and wrapped around a series of guides and rollers before passing across the read/write head. The process is mechanically complex but highly refined.

Modern tape drives are not crude devices. They employ sophisticated error correction, precise servo tracking, and constant monitoring to maintain alignment. Tape heads float above the tape surface at extremely small distances, and the systems compensate continuously for stretching, temperature changes, and wear.

The **LTO-9** format tapes, introduced in 2021, house **3,150 feet (960 meters)** of half-inch tape in a half-pound cartridge just four inches square and less than an inch thick. Yet each cartridge natively

[10] A form factor is the physical size and shape of a piece of computer hardware.

holds **18 terabytes of uncompressed data** and up to **45 terabytes of compressed data**, delivered at a transfer rate of approximately **400 megabytes per second**.[11]

LTO tapes use a **back-and-forth, or linear serpentine, recording scheme**—"linear" because data is stored in parallel tracks running the length of the tape, and "serpentine" because the recording path snakes back and forth, reversing direction on each pass. Thirty-two of the LTO-9 cartridge's **6,656 tracks** are read or written simultaneously as the tape moves past the recording heads. As a result, it takes **208 back-and-forth passes**, or "wraps," to read or write the full contents of a single LTO-9 cartridge.

LTO-9 Ultrium Tape



Sony AIT-3 Tape



SDLT-II Tape



That's **124 miles of tape passing the heads**—roughly the driving distance between **Austin and Houston**. Consequently, it takes hours to read or write each tape; for example, approximately **twelve and one-half hours** to write a full LTO-9 cartridge at maximum speed. While tape is not as fast as hard drives, it has proven to be **more durable and less costly** for long-term storage—so long as the data is being stored, not restored.

Capacity, Compression, and Marketing Numbers

Tape capacities are often advertised using **compressed figures**, typically assuming a 2.5:1 compression ratio. Compression may be achieved by eliminating redundancy in the data stream, but the actual result depends entirely on the nature of the data being stored. Text compresses well; already-compressed files, such as images and video, do not.

For discovery purposes, it is important to distinguish between **native capacity** and **compressed capacity**. Claims about how much data a tape "holds" can be misleading if compression assumptions are left unstated. As with hard drives, the physics is honest even when the marketing is optimistic.

[11] The LTO-10 format hitting the market in January of 2026 will offer a 40TB native capacity and 100TB compressed.

Why Tape Still Matters in Discovery

Tape remains the medium of choice for large-scale backups and long-term archives, particularly in enterprise and institutional settings. Cloud providers themselves rely heavily on tape behind the scenes, even when customers never see it. As a result, litigants often possess relevant data that exists **only on tape**, sometimes stored offsite and sometimes forgotten until litigation brings it to light.

From a discovery perspective, tape raises distinctive issues. Restoring data from tape is slower and more labor-intensive than copying files from disk. Older tapes may require legacy drives or software to read. Some tapes are overwritten on a schedule; others are retained indefinitely. These characteristics inform arguments about burden, proportionality, and feasibility.

Tape Is Not "Dead Data"

It is tempting to think of tape as a graveyard for obsolete information. That temptation should be resisted. Tape is not dead storage; it is **deferred access storage**. The data remains intact, often for decades, and can be restored if the necessary equipment and expertise are available.

Courts and practitioners who assume that tape data is inaccessible or irrelevant risk error. Tape has been the source of pivotal evidence in numerous cases, particularly where routine deletion on active systems was offset by long-term retention on backup media.

Continuity Across Media

The persistence of tape underscores a recurring theme: storage technologies accumulate rather than replace one another. Newer media change how data is accessed and used, but older media continue to shape what data exists and how long it survives. Electronic discovery operates across these layers, not above them.

As storage grows denser and systems more abstracted, the underlying physical realities remain decisive. Whether data resides on disk, tape, or solid-state memory, it is still subject to the constraints of the medium that holds it.

Floppy Disks

Today, the only place most computer users are likely to encounter a floppy disk is as the familiar "save" icon in Microsoft Office applications. But floppy disks played a central role in software distribution and data storage for personal computing for nearly **thirty years**.

Floppy disks are another form of electromagnetic storage. Each contains a spinning, flexible plastic disk coated with a magnetic oxide—essentially rust. In composition, a floppy disk is much like magnetic tape in disk form. Floppies were formatted—either by the user or pre-formatted by the manufacturer—by dividing the disk into concentric rings of data called **tracks**, with each track subdivided into small arcs called **sectors**. Formatting enables systems to locate data on physical storage media much as dividing a subdivision into streets and house numbers enables us to locate homes in a neighborhood.

8" Floppy Disk in Use

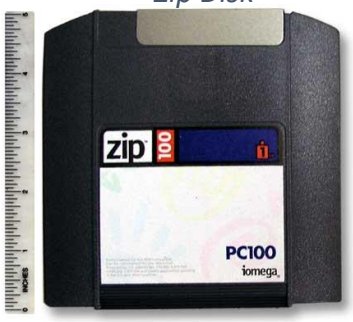




8", 5.25" and 3.5" Floppy

Although many competing floppy disk sizes and formats have been introduced since 1971, only a handful are likely to be encountered in e-discovery. These include the **8-inch**, **5.25-inch**, **3.5-inch standard-density**, **3.5-inch high-density**, and **Zip** formats. Of these, the **3.5-inch high-density (1.44-megabyte)** floppy disk is by far the most prevalent legacy format.

Zip Disk



The Zip disk was one of several proprietary "super-floppy" products that enjoyed brief success before the high capacity and low cost of recordable optical media (CD-R and DVD-R) and flash drives rendered them obsolete.

Optical Media

The most common forms of optical media for data storage are **CD**, **DVD**, and **Blu-ray** disks, each available in read-only, recordable, or rewritable formats. These media typically take the form of a **4.75-inch plastic disk** incorporating a metalized reflective coating and, in some formats, a dye layer. In contrast to tape, floppy disks, and mechanical hard drives, optical storage media do **not** rely on electromagnetism to store and retrieve data; instead, a focused laser beam alters or interrogates the metalized reflective coating to create or detect microscopic **pits and lands**. Variations in reflected laser light are interpreted as the ones and zeros of digital data storage.

The practical differences among CDs, DVDs, and Blu-ray disks lie in their **storage capacities** and the speed, or **throughput**, at which data can be read or written.



A **CD** (Compact Disc), or **CD-ROM** (CD Read-Only Memory), is read-only and not recordable by the end user. It is typically manufactured to distribute music or software. A **CD-R** is recordable by the end user, but once a recording session is closed, it cannot be altered in normal use. A **CD-RW** is rewritable and can be erased and recorded multiple times. A standard-size CD holds approximately **700 megabytes** of data.

A **DVD** (Digital Versatile Disc) likewise exists in read-only, recordable (DVD±R), and rewritable (DVD±RW) formats. The most common DVD has a native storage capacity of approximately **4.7 gigabytes**, meaning that a single DVD can hold roughly **six and one-half CDs** worth of data.



"I should have had him put into a more manageable format years ago."

By employing the shorter wavelength of a blue laser, a **dual-layer Blu-ray disk** can store up to approximately **50 gigabytes** of data—equivalent to about **ten and one-half DVDs**. Like its predecessors, Blu-ray media is available in recordable (BD-R) and rewritable (BD-RE) formats.

How Hard Drives Store Data

A traditional electromagnetic hard disk drive (HDD) stores data on one or more **spinning platters** coated with a ferrous ("magnetizable") material. Each platter surface is divided into concentric rings called **tracks**, which are further subdivided into **sectors**—the smallest addressable units on the disk. Modern drives abstract these physical details behind **logical block addressing**, allowing software to treat storage as a continuous array of numbered blocks rather than discrete tracks and sectors.

Hovering above each platter surface is a **read/write head**, mounted on an actuator arm. The head never touches the platter. Instead, it "flies" on a microscopic cushion of air generated by the spinning disk—at a height measured in **nanometers**, far thinner than a human hair. The tolerances involved are astonishing. A useful metaphor is a jet fighter screaming at many times the speed of sound, skimming just inches above the ground, while accurately counting every blade of grass it passes.

As the platter spins beneath the head, the drive writes data by altering the magnetic orientation of tiny domains on the surface. Reading data reverses the process: changes in magnetic polarity induce signals that the drive electronics interpret as “on” and “off,” notated as ones and zeroes. The drive has no sense of files, documents, or meaning—only magnetized regions and locations.

Computers, Hard Drives and Servers

Though ESI resides on a dizzying array of media and devices, by far the largest complement of same occurs within three closely related species of computing hardware: *computers, hard drives* and *servers.* A server is essentially a computer dedicated to a specialized task or tasks, and both servers and computers routinely employ hard drives for program and data storage.

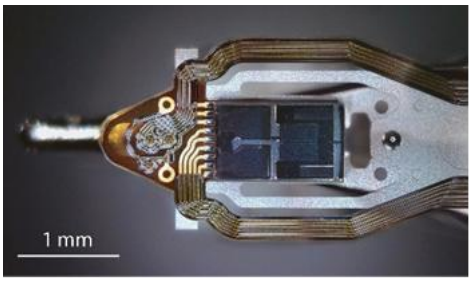
Electromagnetic Hard Drives

As noted, mankind has long stored information by translating it into physical manifestations: cave drawings, Gutenberg bibles, musical notes, Braille dots or undulating grooves in a phonograph record. Because it’s simply a long sequence of ones and zeros, binary data can be memorialized by many physical phenomena. You could build a computer that stored data as a row of beads (the abacus), holes punched in paper (a piano roll), black and white vertical lines (bar codes) or bottles of beer on the wall (still waiting for this one!).

If we build our computer to store data using bottles of beer on the wall, we’d better be thirsty because we’ll need a boatload of bottles and a whole lot of time to set them up, count them and replace them as data changes. Too, we would need something like the Great Wall of China to hold them. So, our beer bottle data storage system isn’t practical. Instead, we need something compact, lightweight and efficient --in short, a refrigerator magnet and some paper clips.

Okay, maybe not a refrigerator magnet *per se*, but the principles are the same. If you take a magnet off your refrigerator and rub it against a metal paperclip, you will transfer some magnetic properties to the paperclip. Suppose you lined up about a zillion paper clips and magnetized some but not others. You could go down the row with a piece of ferrous metal (or, better yet, a compass) and distinguish the magnetized clips from the non-magnetized clips. If you call the magnetized clips “ones” and the non-magnetized clips “zeroes,” you’ve got yourself a system that can record binary data. Were you to glue all those paper clips in concentric circles onto a spinning phonograph record and substitute an electromagnet for the refrigerator magnet, you wouldn’t be too far afield of what goes on inside the hard and floppy disk drives of a computer, albeit at a much smaller scale. In case you wondered, this is also how we recorded sound on magnetic tape, except that instead of determining that a spot on the tape is magnetized or not as it rolls by, we gauge varying degrees of magnetism which corresponding to variations in the recorded sounds. This is **analog** recording—the variations in the recording are analogous to the variations in the music.

Since computers process electrical signals much more effectively than magnetized paper clips jumping onto a knife blade, what is needed is a device that transforms magnetic signals to electrical signals and vice-versa—an energy converter. Inside every floppy and hard disk drive is a gadget called a read/write head. The read/write head is a tiny electromagnet that perform the conversion from electrical information to magnetic and back again. Each bit of data is written to the disk using an encoding method that translates zeros and ones into patterns of magnetic flux reversals. Don't be put off by Star Wars lingo like "magnetic flux reversal" --it just means flipping the magnet around to the other side or "pole."



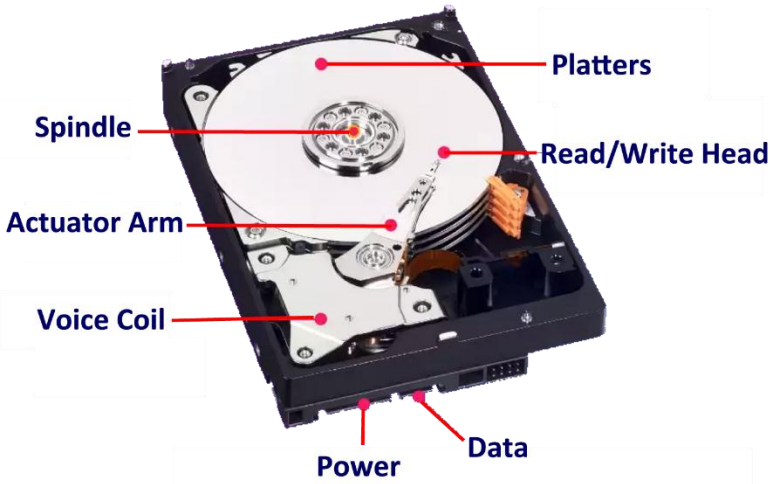
Hard Drive Read-Write Head

Older hard disk heads make use of the two main principles of electromagnetic force. The first is that applying an electrical current through a coil produces a magnetic field; this magnet field imparts magnetic properties—writes--to the disk. The direction of the magnetic field produced depends on the direction that the current is flowing through the coil. The converse principle is that moving a magnetic field alongside a coil of wire induces an electrical current to flow through the coil. That current corresponds to previously written magnetic information and so serves to "read" the disk. Newer disk heads use different physics and are more efficient, but the basic approach hasn't changed: electricity to magnetism and magnetism to electricity.

A hard drive is an immensely complex data storage device engineered to appear deceptively simple. When you connect a hard drive to your machine, and the operating system detects the drive, assigns it a drive letter and—presto! —you've got trillions of bytes of new storage! Microprocessor chips garner the glory, but the humdrum hard drive is every bit the paragon of ingenuity and technical prowess.

A conventional electromagnetic hard drive is a sealed aluminum box measuring (for a desktop system) roughly 4" x 6" x 1" in height. A hard drive can be located almost anywhere within the computer's case, customarily secured by several screws attached to any of ten pre-threaded mounting holes along the edges and base of the case. One face of the case is labeled to reflect the drive specifications, while a printed circuit board containing logic and controller circuits will cover the opposite face.

A conventional electromagnetic hard disk contains round, flat discs called **platters**, coated on both sides with a special material able to store data as magnetic patterns. Much like a record player, the platters have a hole in the center allowing multiple platters to be stacked on a spindle and read on both sides for greater storage capacity.



The platters rotate at high speed—typically 5,400, 7,200 or 10,000 rotations per minute—driven by an electric motor. Data is written to and read from the platters by **read/write heads** mounted on the end of a pivoting extension called an **actuator arm** that functions similarly to the tone arm that carries a phonograph cartridge and needle across the face of a vinyl audio record. Each platter has two read/write heads, one on the top of the platter and another on the bottom. So, a conventional hard disk with three platters typically sports six surfaces and six read/write heads.

Unlike a record player, the read/write head never touches the spinning platter but fly overhead. When the platters spin up to operating speed, *their rapid rotation causes a cushion of air to flow under the read/write heads and suspend them over the surface of the disk*—the same principle of lift that operates on aircraft wings and enables them to fly. The head then reads the magnetic patterns on the disc while flying just .5 millionths of an inch above the surface. At this speed, if the head bounced against the surface, there is a good chance that the head would burrow into the surface of the platter like a fighter jet flying into the ground, obliterating data, destroying both read/write heads and rendering the hard drive inoperable—a so-called "head crash."


IBM 350 Disk Storage Unit

The hard disk drive has been around for more than 60 years, but it was not until the 1980's that the physical size and cost of hard drives fell sufficiently for their use to be commonplace.

Introduced in 1956, the IBM 350 Disk Storage Unit pictured was the first commercial hard drive. It was 60 inches long, 68 inches high and 29 inches deep (so it could fit through a door). Called the **RAMAC** (for Random Access Method of Accounting and Control), it

held fifty 24” magnetic disks of 50,000 sectors, each storing 100 alphanumeric (7-bit) characters. Thus, it held about **3.75 megabytes**, or one or two cellphone snapshots today. It weighed a ton (literally), and users paid $3,200.00 per month to *rent* it. That’s about $30,000.00 in today’s dollars.



Now, you can buy a ten terrabyte hard drive storing *two million times* more information for a fraction of that monthly rental. That 10TB drive weighs less than two pounds, can hide behind a paperback book and costs less than $170.00.

Over time, hard drives took various shapes and sizes (or “form factors” as the standard dimensions of key system components are called in geek speak). Two electromagnetic drive form factors are still in use: 3.5” (desktop drive) and 2.5” (laptop drive). A third, the 1.8” (iPod and microsystem drive, is wholly supplanted by solid state storage.

Hard drives connect to computers by various mechanisms called “interfaces” that describe both how devices “talk” to one-another as well as the physical plugs and cabling required. The five most common hard drive interfaces are:

PATA for **P**arallel **A**dvanced **T**echnology **A**ttachment (sometimes called EIDE for **E**xtended **I**ntegrated **D**rive **E**lectronics) [long obsolete]

SATA for **S**erial **A**dvanced **T**echnology **A**ttachment [most common today]

SCSI for **S**mall **C**omputer **S**ystem Interface

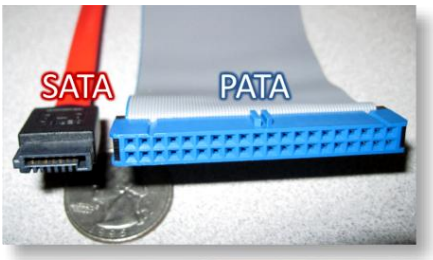
SAS for **S**erial **A**ttached **S**CSI

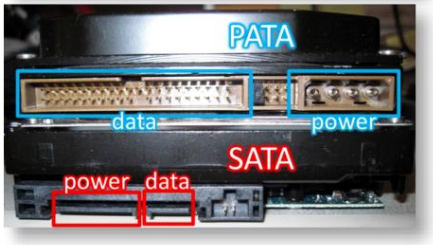
FC for **F**ibre **C**hannel

Though once dominant in personal computers, PATA drives largely disappeared after 2006. Today, virtually all laptop and desktop computers employ SATA drives for local storage. SCSI, SAS and FC drives tend to be seen exclusively in servers and other applications demanding high performance and reliability.

From the user’s perspective, PATA, SATA, SCSI, SAS and FC drives are indistinguishable; however, from the point of view of the technician tasked to connect to and image the contents of the drive, the difference implicates different tools and connectors.

The five drive interfaces divide into two employing parallel data paths (PATA and SCSI) and three employing serial data paths (SATA, SAS and FC). Parallel ATA interfaces route data over multiple simultaneous channels necessitating 40 wires where serial ATA interfaces route data through a single, high- speed data channel requiring only 7 wires. Accordingly, SATA cabling and connectors are smaller than their PATA counterparts (see photos, right).





Fibre Channel employs optical fiber (the spelling difference is intentional) and light waves to carry data at impressive speeds. The premium hardware required by FC dictates that it will be found in enterprise computing environments, typically in conjunction with a high capacity/high demand storage device called a **SAN** (for Storage Attached Network) or a **NAS** (for Network Attached Storage).

It's easy to become confused between hard drive interfaces and external data transfer interfaces like USB, Lightning, Thunderbolt (USB-C) or (the now obsolete) FireWire seen on external hard drives. The drive within the external hard drive housing will employ one of the interfaces described above (except FC); however, to facilitate external connection to a computer, a device called a **bridge** will convert data written to and from the hard drive to a form that can traverse a USB or Thunderbolt connection. In some compact, low-cost external drives, manufacturers dispense with the external bridge board altogether and build the USB interface right on the hard drive's circuit board.

Why Spinning Disks Still Matter

Despite the rise of solid-state storage, spinning disks remain ubiquitous. They dominate **bulk storage**, archives, backup systems, and much of the infrastructure behind cloud services. When litigants claim that data volumes are too large to preserve or search, they are often talking about data residing on magnetic disks, whether on-premises or in remote data centers.

For discovery purposes, it is a mistake to treat hard drives as obsolete. Their physical characteristics continue to shape what data exists, how it persists, and how it can be recovered.

Cloud Services as Everyday Evidence Repositories

Today, many of the most important categories of electronically stored information do not reside on local devices at all. Business email lives in Microsoft 365 or Google Workspace. Collaboration occurs in Teams, Slack, and Zoom. File repositories exist in OneDrive, Google Drive, and Dropbox.

Accounting runs in online platforms. Customer data resides in hosted CRM systems. Even litigation support tools themselves run in the cloud.

These systems blur the line between storage and application. They do not merely hold files; they also generate and retain logs, version histories, timestamps, and permissions—metadata that frequently becomes central to discovery disputes. A user may swear they did not alter a document, while the platform quietly preserves a full revision trail proving otherwise.

A practical takeaway for lawyers is this: **asking only for "documents" may miss the most probative evidence**. Counsel must learn to request the system-generated context as well. This includes activity logs, revision histories, message metadata, and export formats that preserve interoperability.

Cloud services also change how proportionality is assessed. Some data is easily searched and exported. Other data requires expensive specialist intervention. Competence includes understanding these differences well enough to negotiate realistic, proportional discovery plans rather than defaulting to either "everything" or "nothing."

File Systems, Clusters, and Wasted Space

Hard drives do not store files as neat, contiguous sequences. A **file system** keeps track of which disk locations belong to which files and which locations are free. To simplify management, file systems allocate space logically, in fixed-size units called **clusters** (or allocation units). If a file does not completely fill its last cluster, the unused portion remains allocated but empty. That unused space is known as **slack space**.

The warehouse analogy fits neatly. If you rent storage units of a fixed size, storing a few small boxes still consumes an entire unit, leaving empty space around it. Digital storage works the same way. In forensic analysis, "empty" slack space may contain remnants of previously stored data, fragments of older files, or nothing at all, depending on the drive's history.

Deletion Is Bookkeeping, Not Destruction

When a file is deleted from a hard drive, the data is usually not truly erased ("overwritten"). Instead, the file system marks the clusters previously assigned to the file as available for reuse and removes the directory entry that pointed to them. The underlying magnetic domains remain unchanged until new data overwrites them.

Whether deleted data persists is a function of time, system activity, and storage behavior—not user intent. This distinction underlies many disputes about spoliation, recovery, and proportionality. A lawyer who understands how deletion works is far better equipped to assess what can reasonably be found and what cannot.

As illustrated below, the physical organization of disks—platters, tracks, sectors, and clusters—governs what survives ordinary use and what is lost. Software may hide these details, but it does not eliminate them.

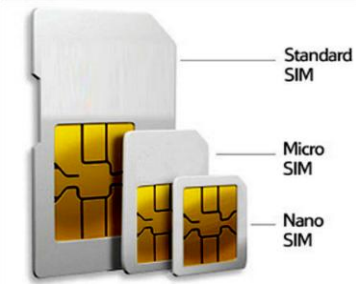
Flash Drives, Memory Cards, SIMs and Solid-State Drives

Many late-model laptops and nearly all portable computing devices and phones employ data storage devices with no moving parts where the data resides entirely within the solid semiconductor material which comprise the memory chips, hence the term, "solid-state." Historically, rewritable solid-state storage was volatile (in the sense that data disappeared when power was withdrawn) and expensive.



Solid State Media Cards

Beginning around 1995, a type of non-volatile memory called NAND flash became inexpensive enough to be used for removable storage in emerging applications like digital photography. Further leaps in the capacity and dips in the cost of NAND flash led to the near eradication of film for photography and the extinction of the floppy disk, replaced by simple, inexpensive and reusable USB storage devices called, variously, Smart Media, Compact Flash media, SD cards, flash drives, thumb drives, pen drives and memory sticks or keys.



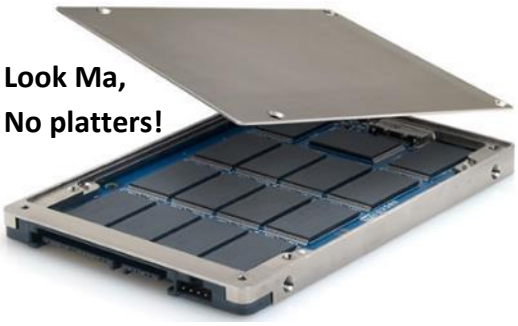
SIM Cards

A specialized form of solid-state memory seen in cell phones is the **Subscriber Identification Module** or **SIM card**. SIM cards serve both to authenticate and identify a communications device on a cellular network and to store a modest number of SMS messages and phone book contacts. A SIM's storage capacity is small, typically 256 KB (262,144 bytes).



USB Flash Drives

As the storage capacity of NAND flash has gone up and its cost has come down, the conventional electromagnetic hard drive is rapidly being replaced by **solid-state drives** in standard hard drive form factors. Solid-state drives are significantly faster, lighter and more energy efficient than conventional drives, but they currently cost more per gigabyte stored than their mechanical counterparts. All signs point to the ultimate obsolescence of mechanical drives by solid-state drives, and some products (notably laptops and tablets like the iPad and Microsoft Surface) have eliminated hard drives altogether in favor of solid-state storage.



Look Ma,
No platters!

Until recently, solid state drives assumed the size and shape of mechanical drives to facilitate compatibility with existing devices. However, the size and shape of mechanical hard drives was driven by the size and operation of the platters they contained. Because solid state storage devices have no moving parts, they can assume virtually any shape. So, slavish adherence to 2.5" and 3.5" rectangular form factors is fading in favor of shapes and sizes uniquely suited to the devices that employ them.

Today, it's common for solid state drives to take one of three forms. The first is the familiar 2.5" enclosure with a conventional SATA interface (below left). Modern laptops employ either NVMe SSD cards like the one lower right top or M.2 SSD "sticks" seen at lower right bottom. M.2 cards are 22mm wide but come in lengths from 42mm to 110mm, with 80mm (3.1") most common.







With respect to e-discovery, the shift from electromagnetic to solid-state drives is inconsequential. However, the move to solid-state drives will significantly impact computer forensic analysis. Because the NAND memory cells that comprise solid-state drives wear out rapidly with use, solid-

state drive controllers must constantly reposition data to ensure usage is distributed across all cells. Such “wear leveling” (also called “TRIM”) hampers techniques that forensic examiners have long employed to recover deleted data from conventional hard drives.

Traditional hard drives often preserved deleted data for long periods, allowing forensic analysts to recover remnants. That expectation is far less reliable today. Solid-state drives constantly reorganize data internally to extend their lifespan. Features such as **TRIM** instruct the drive to purge deleted content so that storage cells may be reused. Once overwritten, the data is typically unrecoverable.

Cloud storage further reduces the likelihood of recovering deleted content. Data may be divided across systems and overwritten rapidly according to provider-managed policies. In many environments there is **no physical artifact** to examine.

The result is not that parties may delete freely, but that **restoration limits are real and unavoidable.** Analyses of intent and prejudice must recognize that modern storage behaves differently than the magnetic-disk world of the past.

Mobile-Resident ESI

Mobile devices are now the primary computing platforms for many users. Phones and tablets store vast quantities of electronically stored information, including text messages, messaging-app content, call logs, photos, videos, notes, location history, authentication tokens, and cloud-sync data. Much of this information resides inside individual apps rather than visible “files.” Each application may maintain its own database or cloud connection, with its own retention settings.

Because modern phones synchronize with cloud services, data may exist in multiple locations at once. A message may appear on a device, a laptop, and a hosted platform. Deletion on one device may or may not remove the record elsewhere. Counsel must therefore think in terms of **ecosystems rather than devices.**

Preservation usually occurs through account-level action rather than physically seizing the phone. However, some content is device-resident only, and forensic examination may still be appropriate when proportional. Competent handling of mobile ESI requires awareness of privacy concerns, authentication challenges, and the increasing use of encrypted services.

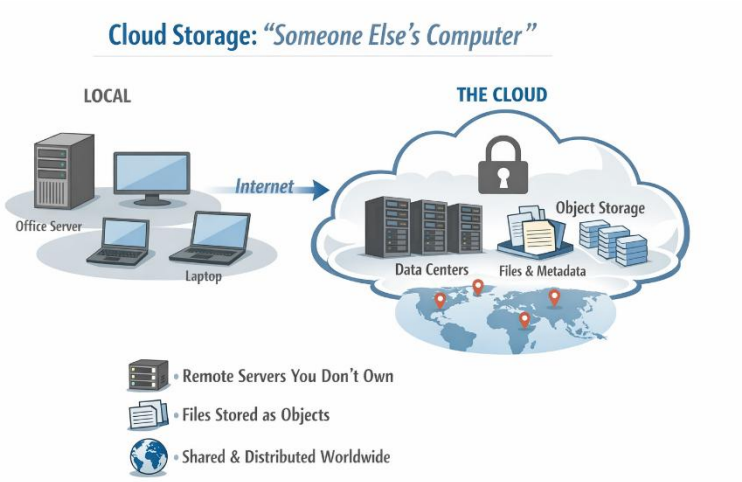
What matters for discovery is not technical mastery of mobile architecture, but recognition that *the most responsive conversational data may live in a user’s pocket.*

Cloud-Native Storage: “Someone Else’s Computer,” But Still Your Data

For most of computing history, data lived on machines you could see and touch—a tape cartridge on a shelf, a floppy on your desk, or a hard drive humming inside a server room. Today, an ever-larger share of electronically stored information never sets foot inside your building. Instead, it resides "in the cloud." Cloud servers are simply remote, provider-managed computing and storage resources accessed over the Internet.

Cloud storage is not magic. It is simply storage delivered as a service. Your files may physically reside on spinning disks or solid-state drives in vast data centers scattered across the country—or across continents. The cloud provider handles hardware, power, cooling, networking, redundancy, and security. You supply the data; they supply the infrastructure. Lawyers sometimes imagine the cloud as vaporous or intangible. It is neither. Cloud data is real data, on real machines, somewhere on Earth. You just don't own the machines.

Cloud storage commonly takes the form of **object storage**, where files are treated not as documents in folders, but as "objects" tagged with metadata and distributed across multiple systems for speed and resilience. If one server fails, your data remains intact because it is mirrored or divided across others. Cloud services also layer sharing, collaboration, and versioning features atop storage, meaning that a single "document" may have dozens of prior versions, access logs, comments, and embedded history.



The e-discovery implications are profound. First, **possession, custody, and control** no longer align neatly with physical ownership. Your client may not own the hardware, but courts routinely hold that they control data they can access and direct, even when it lives in a vendor's system. Second, preservation changes character. You cannot simply unplug a server and label it "Do Not Touch." Instead, you work with administrators—or the service provider—to suspend deletion rules, retain accounts, and preserve audit logs. Third, exports from cloud systems may not resemble "files" at all. You may receive JSON data, structured tables, or archives containing embedded metadata essential to meaning.

Cloud services—from email to chat platforms to shared drives—now represent the predominant repositories of responsive data in civil discovery. Competence requires understanding that the cloud is less a place than a model of computing. Your duty is not to chase molecules of silicon, but

to know where client data resides, how long it persists, how it can be preserved, and in what form it can be produced.

Cloud Collaboration as Evidence Ecosystems

Modern cloud platforms are not merely storage locations. They are **evidence ecosystems.** Business communication increasingly occurs inside hosted collaboration services like Microsoft 365, Google Workspace, Slack, Zoom, and Teams. These systems do far more than store documents. They also retain conversations, shared links, comments, reactions, approvals, version histories, and audit logs. In effect, each platform becomes its own universe of data and metadata documenting not only what exists, but who created it, how it changed over time, and how others interacted with it.

This shift means discovery rarely seeks only "documents." Instead, lawyers must consider whether the **context** surrounding a document is probative. A message thread may reveal negotiation, consent, instruction, or authorization. A revision history may show when text was added or removed. An access log may demonstrate who knew what and when. Cloud-resident systems quietly collect these historical traces as a matter of course.

Practically, exports from these systems may not resemble traditional files. Counsel may receive structured data formats designed for computers, not humans. Sometimes the native environment supplies meaning that is lost when data is flattened into static form. Competence requires understanding that the platform is often the true repository—not merely the files downloaded from it.

The essential takeaway is that **cloud applications generate evidence as they operate.** In modern litigation, appreciating this reality is often as important as knowing where a document is stored.

Ephemeral and Auto-Deleting Storage

Many modern communication platforms automatically delete messages or retain them only briefly unless a retention policy extends their lifespan. Some tools allow messages to "expire" after they are read. Others overwrite chat history after a set interval. These features were designed for privacy and risk management but have become common in routine business communications.

For lawyers, this creates a tension between ordinary business practice and preservation obligations. The fact that a system deletes data automatically does not excuse the duty to preserve when litigation is reasonably anticipated. Preservation may require disabling expiration, retaining user accounts, or coordinating with administrators so that routine deletion does not destroy evidence.

It is also important to understand that **metadata and logs may disappear sooner than content.** A message may remain visible while its corresponding audit trail does not. Lawyers should not assume that system-generated history exists indefinitely.

Courts recognize that some loss occurs through normal system operation. But where parties fail to intervene after a duty to preserve arises, sanctions risk increases sharply. The key is not technical brilliance, but **timely, documented communication and defensible action**.

Distributed and Network Storage: Many Drives Acting as One

As data volumes grew, organizations discovered that a single hard drive—or even a single server—was rarely enough. The solution was to network storage so that multiple users, machines, and even entire organizations could access shared data simultaneously. At the simplest level, this takes the form of **Network-Attached Storage (NAS)** or a **Storage Area Network (SAN)**—centralized repositories holding vast amounts of electronically stored information that users tap into from laptops, desktops, and mobile devices.

Network storage systems often behave like a single enormous drive, but behind the curtain they rely on **distributed storage architectures**—clusters of drives working together across multiple machines or locations. Data may be sliced, mirrored, striped, or duplicated so that performance improves and failure of one drive—sometimes even one server—does not result in data loss. In effect, the storage system becomes a living organism: as components fail or are added, the system reshapes itself to continue serving data.

For discovery, distributed storage presents both opportunity and complexity. On the positive side, such systems maintain logs, permissions, and metadata about who accessed what and when—evidence that may be more probative than the documents themselves. On the challenging side, **no single box contains "the file."** A document may exist in fragments across drives, replicated in multiple data centers, and presented as a seamless whole to the end user.

Preservation in distributed environments requires planning rather than panic. You cannot simply "image the server," because there may be no single server to image. Instead, you preserve at the logical layer—retention settings, legal holds, user accounts, and shared folders—working with IT and vendors who understand how the system stores and manages data. Competence here means speaking enough "geek" to ask the right questions, recognize limits, and document what was done so that your preservation efforts are defensible.

The key takeaway is this: **the physical reality of storage has become abstracted from the user's experience of it.** Lawyers must understand the abstraction well enough to reason about relevance, proportionality, and authenticity—even when the underlying bytes may live nowhere near the courthouse or the client.

Custody, Control, and Preservation in the Cloud Era

One of the most persistent misconceptions about cloud storage is the belief that data stored on third-party systems somehow falls outside a party's obligation to preserve and produce it. Courts have routinely rejected that argument. If a client has the practical ability to access, manage, or export information—even if the hardware belongs to a vendor—the data remains within the client's **control** for discovery purposes.

Preservation in the cloud therefore requires thoughtful, coordinated action. It may include suspending automatic deletion policies, retaining inactive user accounts, preserving shared workspaces, and ensuring that logs and version histories are not lost to routine system cycles. The process is often more about **policy and communication** than plugging in forensics tools.

This shift reinforces a recurring theme of this Workbook: discovery success depends less on mastering every technical nuance than on developing the fluency to collaborate productively with technical custodians. The days when discovery meant "pull paper from a file room" are gone. Today's lawyer must be comfortable engaging with administrators, IT staff, and vendors to ensure that cloud-resident evidence is preserved and produced in forms that remain useful and trustworthy.

RAID Arrays (Hard Drives Working Together)

Whether local to a user or in the Cloud, storage systems account for nearly all the electronically stored information attendant to e-discovery. In network server and Cloud applications, hard drives rarely operate singly. Instead, hard drives are ganged together to achieve greater capacity, speed and reliability in so-called **R**edundant **A**rrays of **I**ndependent **D**isks or **RAID**s. In the Storage Area Network (SAN) device pictured at right, the 16 hard drives housed in trays *could* be accessed as **J**ust a **B**unch **o**f **D**isks or **JBOD**, but it's far more likely they are working together as a RAID.



RAIDs have two advantages over single drives: **redundancy** and **performance.** The redundancy aspect is obvious—two mirrored drives holding identical data safeguard against data loss due to mechanical failure of either drive—but how can multiple drives improve performance? The answer lies in dividing the data across multiple drives using a technique called **striping**. Physical movement of disks and heads ("atoms") in a mechanical hard drive is a glacially slow process compared to the lightspeed pace of electrons in a circuit; however, you can speed the mechanical transfer by reading and writing data to and from several drives *at the same time*

A RAID improves performance by allocating data across more than one physical drive, supporting simultaneous reads and writes. Each split swath of data in an array is called a "stripe" and the

method of depositing the data across drives is called “striping.” If you imagine the drives lined up alongside one another, you can see why moving back-and-forth between them to store data is akin to painting a stripe across the drives. By striping data, each drive can deliver its share of the data simultaneously, increasing the amount of information handed off to the computer’s microprocessor, *i.e.*, faster *throughput* supporting speedier performance.

But, when you stripe data across drives, you lose Information if *any* drive holding striped data fails. You gain performance at the expense of security.

This type of RAID configuration is called a **RAID 0**. It wrings maximum performance from a storage system; but it's risky.

If RAID 0 is for gamblers, **RAID 1** is for the risk averse. A RAID 1 configuration duplicates everything from one drive to an identical twin, so that a failure of one drive won't lead to data loss. RAID 1 doesn't improve performance, and it requires twice the hardware to store the same information.

A helpful way to remember which RAID is which: ***When a drive fails using RAID 1, you’ve still got one copy of the data; when a drive fails using RAID 0, you’ve got nothing—zip, ZERO!.***

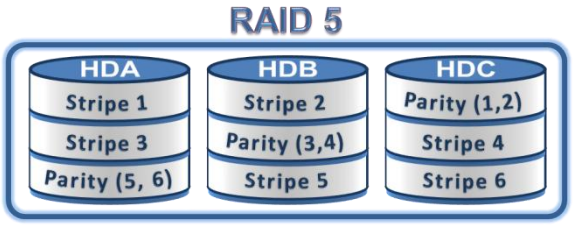
Other RAID configurations blend the *performance* features of RAID 0 with the *protection* of RAID 1.

Thus, a "RAID 0+1" mirrors two striped drives, but demands four hard drives delivering only half their total storage capacity. Safe and fast, but not cost-efficient. The safety flows from a concept called ***parity,*** key to a range of other numbered RAID configurations. Of those other configurations, the ones most often seen are **RAID 5 and RAID 7.**

To understand parity, consider the simple equation 5 + 2 = 7. If you didn't know one of the three values in this equation, you could easily solve for the missing value, *i.e.,* presented with "5 + __ = 7," you know the missing value is 2. In this example, "7" is the ***parity value*** or **checksum** for "5" and "2."

A similar process is used in RAID configurations to gain increased performance by striping data across multiple drives while using parity values to permit the calculation of any missing values lost to drive failure. In a three-drive array, any one of the drives can fail, and we can use the remaining two to recreate the third (just as we solved for 2 in the equation above).

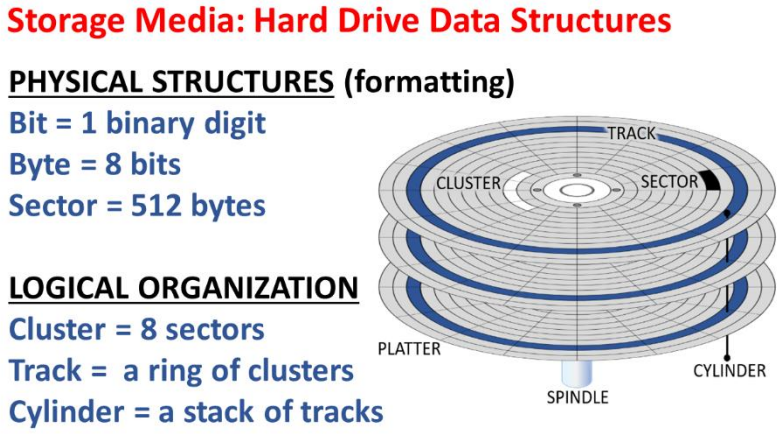
In this illustration, data is striped across three hard drives, HDA, HDB and HDC. HDC holds the parity values for data stripe 1 on HDA and stripe 2 on HDB. It's shown as "Parity (1, 2)." The parity values for the other stripes are distributed on the other drives. Again, any *one* of the three drives can fail, and *all* data is recoverable. This configuration is RAID 5 and, though it requires a minimum of three drives, it scales to dozens or hundreds of disks.



Knowing a little about RAID arrays helps lawyers gauge the burden and cost of preserving a server. Preservation may be as simple as swapping out a single drive from a RAID 1 or entail the duplication of 3 or more drives to preserve a RAID 5.

Sectors, and Clusters and Tracks, Oh My!

Now, we will shift gears and briefly touch on how data resides physically and logically on hard drives. Recall the earlier discussion of electromagnetic hard drives. At the factory, a hard drive's platters are organized to enable the storage and retrieval of data. This is **low level formatting,** divides each platter into tens of thousands of densely packed concentric circles called **tracks**. If you could see them (and you can't because they are nothing more than microscopic magnetic traces), they would resemble the growth rings of the world's oldest tree. It's tempting to compare platter tracks to a phonograph record, but a phonograph record's track is a single spiraling groove, not concentric circles. A track holds far too much information to serve as the smallest unit of storage on a disk, so each track is broken down into physical **sectors**. "Physical" because it resides in a fixed location on the media. A sector is normally the smallest individually addressable unit of information stored on a hard disk and historically held **512 bytes** of information (through about 2010). Today, sector sizes tend to be **4,096 bytes,** but emulate the 512-byte sector size for backward compatibility. **So, when we speak of hard drives, we still speak of 512-byte sectors**. The figure right shows a simplified representation of three platters depicting tracks, cylinders and sectors. The number of tracks, cylinder and sectors is far, far greater that the illustration suggests. Each platter is formatted on both sides for double the information storage, much as a phonograph album was recorded on both sides. Each platter has two read/write heads, one on the top of the platter and another on the bottom. So, a conventional hard disk with three platters uses



six surfaces and six read/write heads. Tracks aligned on both sides of the same platter and with tracks on other platters form so-called "cylinders" of data storage.

When electromagnetic hard drives held far less data than they do now, file storage locations were based on the physical geometry of the platters, addressed by Cylinder, Head and Sector tuples, so-called **CHS addressing**. Any specific sector could be located by specifying its cylinder, read-write head and sector number. Early hard drives were limited to a maximum of 1024 cylinders, 16 heads (two sides of eight platters) and 63 sectors per track. That meant the maximum addressable capacity of a CHS hard drive formatted in 512-byte sectors was a measly 528 MB[12] (512 x 63 x 16 x 1024)—absurdly small by modern standards but vast thirty years ago. As drive capacities grew, computer companies resorted to a series of schemes to deploy larger drives using software running on the circuit boards of the drives. This "firmware" remapped fake disk geometries to real data locations. Over time, even these workarounds couldn't keep up with burgeoning drive sizes and were abandoned.

The challenge computer scientists faced wasn't increasing the areal density of the drives.[13] New materials and recording technologies ensured that drive capacity was growing exponentially! Instead, the impediment was coming up with a way to *address* the vast volume of storage without making older systems obsolete. Today, hard drives employ **LBA** for **Logical Block Addressing**, numbering each sector sequentially and allocating many more bits to catalog the locations of the sectors. Modern computers can address up to 144 petabytes of 512-byte sectors. A petabyte is 1,000 terabytes or one million gigabytes. That should hold us for a while (but isn't that what we always assume?).

Decimal Byte Unit Order

Byte	8 bits	
Kilobyte	1,000 bytes	(thousand)
Megabyte	1,000,000 bytes	(million)
Gigabyte	1,000,000,000 bytes	(billion)
Terabyte	1,000,000,000,000 bytes	(trillion)
Petabyte	1,000,000,000,000,000 bytes	(quadrillion)
Exabyte	10^{18}	
Zettabyte	10^{21}	
Yottabyte	10^{24}	

To this point, we have described only *physical* units of storage. Platters, cylinders, tracks, sectors and even bits and bytes exist as discrete *physical* manifestations written to the media. Computers manage data not only physically but also *logically*. As it's impractical to manage and gather the data by assembling it from individual sectors, operating systems speed the process by grouping sectors into contiguous chunks of data called

[12] 504 MiB

[13] Areal Density describes the quantity of data (in bits) that can be stored on a given surface area of a computer storage medium.

clusters. Just as we don't buy single eggs at the grocery, clusters are akin to the 6-, 12- or 18-egg cartons that have become standard.

A cluster is the smallest amount of disk space that can be allocated to hold a file. Computers organize hard disks based on clusters, which consist of one or more contiguous sectors. *The smaller the cluster size, the more efficiently a disk stores information. Conversely, the fewer the number of clusters, the less space consumed by the table required to track their content and locations.*

To recap, data is stored in logical units called **clusters**, made up of multiple physical storage units termed **sectors**. A series of logical clusters, in turn, comprise **tracks** (concentric circles or "tree rings" of data) on **platters**, one or more disks of rotating electromagnetic storage media within the enclosure of a mechanical hard drive. Tracks that overlie one-another on both sides of a platter and across multiple platters is termed a **Cylinder** (although "cylinder" is an archaic term from the days when hard drive storage was tied to the physical geometry of the formatted disks)**.** In order of data capacity: **Bits > Bytes > Sectors > Clusters > Tracks > Cylinders > Platters > Drive>Array**

Operating Systems and File Systems

As hard disks have grown, using them efficiently is increasingly difficult. A library with thirty books operates much differently than one with 30 billion. The **file system** is the name given to the logical structures and software routines used to control access to the storage on a hard disk system and the overall structure in which files are named, stored and organized. An **operating system** is a large and complex collection of functions, including the user interface and control of peripherals like printers. Operating systems are built on file systems. If the operating system is the car, then the file system is its chassis. Operating systems are known by familiar household names, like **MS-DOS**, **Windows or MacOS**. In contrast, file systems go by obscure monikers like **FAT, FAT32 (DOS), ext2 (Linux), NTFS (Windows) and HFS+ and APFS (Apple)**.

NTFS File Systems

The Microsoft Windows environment, in particular the **NTFS file system** at the heart of Windows NT, 2000, XP, Vista and Windows 7-11, accounts for most personal computers in the world; however, there are many non-Microsoft operating systems out there, such as Unix, Linux and, MacOS. Though similarities abound, these other operating systems use different file systems, and the Unix or Linux operating systems often lie at the heart of corporate and web file servers—today's "big iron" systems and Cloud computing. As well, MacOS usage has grown markedly as Apple products have kicked down the doors of business computing and captivated consumers.

NTFS uses a powerful and complex file system database called the **Master File Table** or **MFT** to manage file storage. Understanding the file system is key to appreciating the evidentiary potential

of computer forensics*, viz.*, why deleted data doesn't necessarily go away and where probative artifacts reside. It's the file system that marks a file as deleted though it leaves the data on the drive. It's the file system that enables the creation of multiple partitions where data can be hidden from prying eyes. Finally, it's the file system that determines the size of a disk cluster with the attendant persistence of data within the slack space.

Formatting and Partitioning

Partitioning divides drives into **volumes,** which users see as drive letters (*e.g.,* C:, E:, F: and so on). **Formatting** defines the logical structures on the partition and places necessary operating system files at the start of the disk to facilitate booting. For most users, their computer comes with their hard drive partitioned as a single volume (universally called C:). Windows machines may also come with a hidden recovery partition holding files needed to repair the operating system. Some users will find (or will cause) their hard drive to be partitioned into multiple volumes, each appearing to the user as if it were an independent disk drive. Partitions can be designated "**active"** and **"inactive.** Only one partition may be designated as active at any given time, and that partition is the one that boots the computer. The significance in computer forensics is that inactive partitions are invisible to anyone using the computer unless they know to look for them and how to find them. Inactive partitions are a place where users with something to conceal from prying eyes may choose to hide it.

Computers



Historically, all sorts of devices—even people—were "computers." During World War II, human computers—women for the most part—were instrumental in calculating artillery trajectories and assisting with the challenging number-crunching needed by the Manhattan Project. Today, laptop and desktop personal computers spring to mind when we hear the term "computer;" yet smart phones, tablet devices, global positioning systems, video gaming platforms, televisions and a host of other intelligent tools and toys are also computers. More precisely, the **central processing unit** (CPU) or **microprocessor** of a system is the "computer," and the various input and output devices that permit humans to interact with the processor are termed **peripherals.** The key distinction between a mere calculator and a computer is the latter's ability to be programmed and its use of **memory** and **storage**. The physical electronic and mechanical components of a computer are its **hardware,** and the instruction sets used to program a computer are its **software.**

In 1774, a Swiss watchmaker named Pierre Jaquet-Droz built an ingenious mechanical doll resembling a barefoot boy. Constructed of 6,000 handcrafted parts and dubbed "*L'Ecrivain*" ("The Writer"), Jaquet-Droz' automaton uses quill and ink to handwrite messages in cursive, up to 40 letters long, with the content controlled by interchangeable cams. The Writer is a charming example of an early programmable computer.





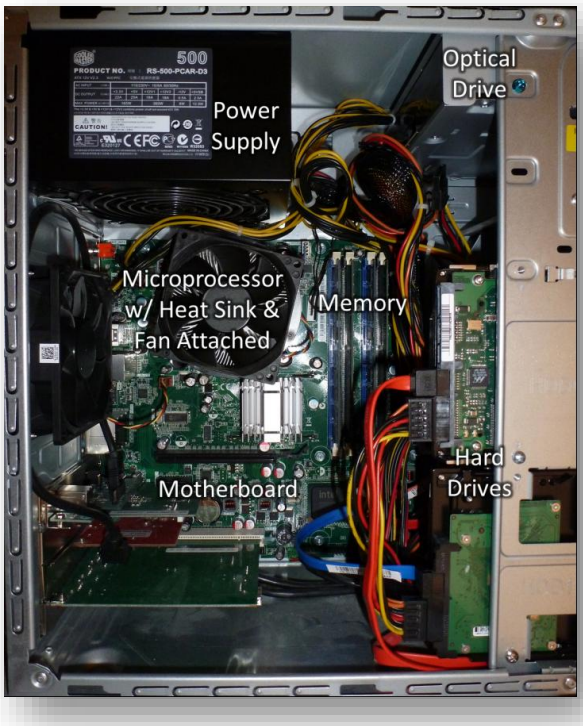
When you push the power button on your computer, you trigger an extraordinary, expedited education that takes the machine from insensible illiterate to worldly savant in seconds. The process starts with a snippet of data on a chip called the ROM BIOS storing just enough information in its **R**ead **O**nly **M**emory to grope around for the **B**asic **I**nput and **O**utput **S**ystem peripherals (like the keyboard, screen and, most importantly, the hard drive). The ROM BIOS also holds the instructions needed to permit the processor to access more and more data from the hard drive in a widening gyre, "teaching" itself to be a modern, capable computer.

Unlike the interchangeable cams of Pierre Jaquet-Droz' mechanical doll, modern electronic computers receive their instructions in the form of data retrieved from the same electronic storage medium as the digital information upon which the computer performs its computational wizardry.

This rapid, self-sustaining self-education is as magical as if you lifted yourself into the air by pulling on the straps of your boots, which is truly why it's called "bootstrapping" or just "booting" a computer.

Computer hardware shares certain common characteristics. Within the CPU, a microprocessor chip is the computational "brain" of system and resides in a socket on the **motherboard**, a rigid surface etched with metallic patterns serving as the wiring between the components on the board. The microprocessor generates considerable heat necessitating the attachment of a heat dissipation device called a **heat sink**, often abetted by a small fan. The motherboard also serves as the attachment point for memory boards (grouped as modules or "sticks") called **RAM** for Random Access Memory. RAM serves as the working memory of the processor while it performs calculations; accordingly, the more memory present, the more information can be processed at once, enhancing overall system performance.

Other chips comprise a **Graphics Processor Unit (GPU)** residing on the motherboard or on a separate expansion board called a **video card** or **graphics adapter**. The GPU supports the display of information from the processor onto a monitor or projector and has its own complement of memory dedicated to superior graphics performance. Likewise, specialized chips on the motherboard or an expansion board called a **sound card** support the reproduction of audio to speakers or a headphone. Video and sound processing capabilities may even be fully integrated into the microprocessor chip.



The processor communicates with networks through an interface device called a **network adapter** which connects to the network physically, through a Local Access Network or **LAN Port,** or wirelessly using a Wi-Fi or Bluetooth connection.

Users convey information and instructions to computers using tactile devices like a keyboard, mouse or track pad, but may also employ voice or gestural recognition mechanisms.

Persistent storage of data is a task delegated to other peripherals: previously **optical drives** (CD-ROM and DVD-ROM devices) and **floppy disk drives**, now **solid-state media** (*i.e.,* thumb drives) and, most commonly, **hard drives**.

All the components just described require electricity, supplied by batteries in portable devices or by a **power supply** converting AC current to the lower DC voltages required by electronics. To guard against data loss from power failure, computing systems may employ battery-powered **Uninterruptible Power Supplies (UPS)** that supply electricity until power is restored.

From the standpoint of digital evidence, it's less important to define these devices than it is to comprehend the information they hold, the places it resides and the forms it takes. Parties and lawyers have been punished for their failure to inquire into and understand the roles computers, hard drives and servers play as repositories of electronic evidence. Moreover, much money spent on electronic discovery today is wasted through parties' efforts to convert ESI to paper-like forms instead of learning to work with ESI in the forms in which it customarily resides on computers, hard drives and servers.

Servers

We defined servers as computers dedicated to a specialized task or tasks. But that definition doesn't begin to encompass the profound impact upon society of the so-called **client-server** computing model. The ability to connect local "client" applications to servers via a network, particularly to **database servers**, is central to the operation of most businesses and to all telecommunications and social networking. Google and Facebook are just enormous groupings of servers, and the Internet merely a vast, global array or "cloud" of shared servers.

Local, Cloud and Peer-to-Peer Servers

For e-discovery, let's divide the world of servers into three realms: **Local**, **Cloud** and **Peer-to-Peer** server environments.

"**Local**" or "**on-prem**" servers employ hardware that's physically available to the party that owns or leases the servers. Local servers reside in a computer room on a business' premises or in leased equipment "lockers" accessed at a co-located data center where a lessor furnishes, *e.g.,* premises security, power and cooling. Local servers are often easier to deal with in e-discovery because physical access to the hardware supports more and faster options when it comes to preservation and collection of potentially responsive ESI. Because on-premises ("on prem") computers and networks exist within a party's physical dominion, protected by the party's security protocols, they're often termed "**behind-the-firewall**."[14] The worldwide pandemic forced corporations and e-discovery service providers to broaden use of remote acquisition tools to search and collect potentially responsive ESI and prompted even more custodians to shift data to the Cloud.

"**Cloud**" servers typically reside in facilities not physically accessible to persons using the servers, and servers are not typically dedicated to a single user. Instead, the Cloud computing consumer is buying services via the Internet that emulate the operation of a single machine or a room full of machines, all according to the needs of the Cloud consumer. Web mail is the most familiar form of Cloud computing, in a variant called SaaS (for Software as a Service). Webmail providers like Google, Yahoo and Microsoft make e-mail accounts available on their servers across many massive data centers, and the data on those servers is available via the Internet, with no user having the right to gain physical access to the machines storing their messaging.

"**Peer-to-Peer**" (P2P) networks exploit the fact that any computer connected to a network has the potential to serve data across the network. Accordingly, P2P networks are decentralized; that is, each computer or "node" on a P2P network acts as client and server, sharing storage space,

[14] A firewall is a network security device facing the Internet that monitors incoming and outgoing network traffic in order to block or allow data transfers based on a set of rules for safe "white list" or unsafe "black list" transactions.

communication bandwidth and/or processor time with other nodes. P2P networking may be employed to share a printer in the home, where the computer physically connected to the printer acts as a print server for other machines on the network. On a global scale, P2P networking is the technology behind file sharing applications like Bit Torrent that have garnered headlines for their facilitation of illegal sharing of copyrighted content. When users install P2P applications to gain access to shared files, they simultaneously (and often unwittingly) dedicate their machine to serving up such content to a multitude of other nodes.

Virtual Servers

Though we've so far spoken of server hardware, *i.e.,* physical devices, servers can be deployed *virtually*, through software that *emulates* the functions of a physical device. Such "hardware virtualization" allows for more efficient deployment of computing resources by enabling a single physical server to host multiple virtual servers.

Virtualization is the key enabling technology behind many Cloud services. If a company needs powerful servers to launch a new social networking site, it can raise capital and invest in the hardware, software, physical plant and personnel needed to support a data center, with the attendant risk that it will be over-provisioned or under-provisioned as demand fluctuates. Alternatively, the startup can secure the computing resources it needs by using virtual servers hosted by a Cloud service provider like Amazon AWS or Microsoft Azure. Virtualization permits adding and paring back computing resources commensurate with demand and, being pay-as-you-go, requires minimal capital investment. Thus, a computing platform or infrastructure can be virtualized and leased, *i.e.,* offered as a service via the internet. Accordingly, Cloud Computing may be termed **PaaS** *(Platform as a Service)* or **IaaS** *(Infrastructure as a Service)*. Web-based applications (like Gmail) are **SaaS** *(Software as a Service)*.

It's helpful to be aware of the role of **virtual machines** (VMs) because the ease and speed with which VMs are deployed and retired as well as their isolation within the operating system can pose unique risks and challenges in e-discovery, especially with respect to implementing a proper legal hold and when identifying and collecting potentially responsive ESI.

Server Applications

Computers dedicated to server roles typically run operating systems optimized for server tasks and applications specially designed to run in a server environment. In turn, servers often support dedicated tasks such as serving web pages (*Web Server*), retaining and delivering files from shared storage allocations (*File Server*), organizing voluminous data (*Database Server*), facilitating the use of shared printers (*Print Server*), running programs (*Application Server*) or handling messages (*Mail Server*). These various server applications may run physically, virtually or as a mix of the two.

Network Shares

Eventually, all electronic storage devices fail. Even the RAID storage arrays previously discussed do not forestall failure, but instead afford a measure of redundancy to allow for replacement of failed drives before data loss. Redundancy is the sole means by which data can be reliably protected against loss; consequently, companies routinely back up data stored on server NAS and SAN storage devices to backup media like magnetic tape or online (*i.e.,* Cloud) storage services. However, individual users often fail to back up data stored on local drives. Accordingly, enterprises allocate a "share" of network-accessible storage to individual users and "map" the allocation to the user's machine, allowing use of the share as if it were a local hard drive. When the user stores data to the mapped drive, that data is backed up along with the contents of the file server. Although **network shares** are not local to the user's computer, they are typically addressed using drive letters (*e.g.,* M: or T:) as if they were local hard drives. Local network shares have their counterparts in the Cloud, including **Cloud Storage and File-Sharing Services** like Dropbox, Box, Google Drive, Apple iCloud and Microsoft OneDrive. So, again, users are encouraged to store their work on file shares to ensure that work is reliably backed up as a means of **disaster recovery**.

Structured and Unstructured Data in Discovery

Not all electronically stored information exists as free-standing documents. Discovery often involves two broad categories of data. **Unstructured data** consists of familiar files such as word-processing documents, PDFs, email messages, and images. **Structured data** resides inside databases and cloud platforms, where information is stored in fields and tables designed for rapid querying rather than reading.

Examples include accounting systems, customer-relationship databases, inventory tools, messaging archives, and cloud-based business platforms. In these environments, there may be no "document" to produce. Instead, counsel works with administrators to extract reports or exports reflecting the requested information.

Because the way a query is written affects the output, discovery of structured data requires careful collaboration and documentation. Parties must define fields, time ranges, filters, and formats so that the resulting production is accurate and reproducible. Proportionality considerations loom large; extracting all data from a large database may be unduly burdensome when a focused query would suffice.

Understanding the difference between structured and unstructured ESI helps lawyers craft requests that are realistic, proportional, and technically sound.

Hashing Data

Because all digital data are numbers, the arithmetic around parity values helps guard against data loss. More advanced math called "hashing" makes it possible to authenticate, deduplicate and cull digital data. Hashing is the use of mathematical algorithms to calculate a unique sequence of letters and numbers to serve as a reliable digital "fingerprint" for electronic data. These sequences are called "message digests" or, more commonly, "hash values."[15] Hashing is an invaluable tool in both computer forensics and electronic discovery and deployed by courts with growing frequency.[16]

Using a hash algorithm, any amount of data—from a tiny file to the contents of entire hard drives and beyond—can be expressed as an alphanumeric sequence of fixed length. The most common forms of hashing are MD5 and SHA-1. MD-5 is a 128-bit (16 byte) value typically expressed as 32 hexadecimal (Base16) characters.

A hash value is just a big, big, BIG number calculated on the contents of the file. A 128-bit number can be as large as 2^{128} – if you start doing the 2 x 2 x 2 x 2, etc. on that, you'll see how fast the values mount.

To say 128 bits or 2^{128} is a "big, big, BIG number" doesn't begin to convey its unfathomable, astronomic scale. In decimal terms, it's about 340 *billion billion billion billion* (a/k/a 340 undecillion). *That's 4 quadrillion times the number of stars in the observable universe!*

A SHA-1 hash value is an even larger 160-bit (20 byte) value typically expressed as 40 hex characters. So, a SHA-1 value is a WAY bigger number—*4.3 billion times bigger*.

The MD5 hash value of the plain text of Lincoln's Gettysburg Address is E7753A4E97B962B36F0B2A7C0D0DB8E8. ***Anyone, anywhere performing the same hash calculation on the same data will get the same unique value*** in a fraction of a second. But change "Four score" to "Five score" and the hash becomes 8A5EF7E9186DCD9CF618343ECF7BD00A. However subtle the alteration—an omitted period or extra space—the hash value changes markedly. The chance of an altered electronic document having the same MD5 hash—a "hash collision" in cryptographic parlance—is one in 340 *trillion, trillion, trillion*. Though supercomputers have fabricated collisions, it's still a level of reliability far exceeding that of fingerprint and DNA evidence.

[15] Please don't say "hash marks," unless you are speaking of insignia denoting military rank or the yard markers on a football field. The one-way cryptographic calculations used to digitally fingerprint blocks of data are "hash values," "hashes" or "message digests."

[16] In 2017, Federal Rule of Evidence 902 was amended to support self-authentication of digital evidence when supported by a process of digital identification like hashing. Fed. R. Evid. 902(14).

Hashing sounds like rocket science—and it's a miraculous achievement—but it's very much a routine operation, and the programs used to generate digital fingerprints are freely available and easy to use. Hashing lies invisibly at the heart of everyone's computer and Internet activities[17] and supports processes vitally important to electronic evidence, including identification, filtering, Bates numbering, authentication, de-duplication and blockchain authentication.

Identification

Knowing a file's hash value enables you to find its identical counterpart within a large volume of data without examining the contents of each file. The government uses this capability to ferret out child pornography, but you might use it to track down company secrets that flew the coop when an employee joined the competition.

Filtering and De-NISTing

A common e-discovery process is to cull data collected from computers that couldn't be evidence because it isn't a custodian's work product. It's done by matching hash values of collected data files to hash values on the National Software Reference Library's (NSRL's) freely published list of hash values corresponding to common retail software and operating systems. The NSRL is part of the National Institute for Standards and Technology (NIST), so this process is commonly called **"de-NISTing**" a data set. For more information on the NSRL, visit http://www.nsrl.nist.gov/.



Mechanical Bates Stamp

Bates Numbering

Hashing's ability to uniquely identify e-documents makes it a candidate to supplement, though not supplant, traditional **Bates numbering**[18] in electronic production in discovery. Though hash values don't fulfill the sequencing function of Bates numbers, they're excellent unique identifiers and enjoy an advantage over Bates numbers because they eliminate the possibility that the same number might be applied to different documents. An electronic document's hash value derives from its contents, so will never conflict with that of

[17] For example, many web services store the hash value of your password, but not the password itself. This enables them to authenticate a user by comparing the hash of the password entered to the hash value on file; however, the password cannot be reversed engineered from the hash value. A remarkable feature of hash values is that they are one-way calculations meaning that although the hash value identifies just one sequence of data, it reveals nothing *about* the data, much as a fingerprint uniquely identifies an individual but reveals nothing about their appearance or personality —it's computationally infeasible to derive the source data from the hash of the source data.

[18] Bates numbering has historically been employed as an organizational method to label and identify legal documents, especially those produced in discovery. "Bates" is capitalized because the name derives from the Bates Manufacturing Company, which patented and sold auto-incrementing, consecutive-numbering stamping devices. Bates stamping served the dual functions of sequencing and uniquely identifying documents.

another document unless the two documents are identical. Similarly, because two identical documents from different custodians (sources) will hash identically, the documents' hash values won't serve to distinguish between the two despite their different origins.

Authentication

Forensic examiners extensively use hashing to establish that a forensically sound duplicate of a hard drive faithfully reflects every byte of the source evidence drive and to prove that their activities haven't altered the original evidence. As e-discovery gravitates to production of native file formats instead of static page images, concern about intentional or inadvertent alteration requires lawyers to have a fast, reliable method to authenticate electronic documents. Hashing neatly fills the bill. In practice, a producing party calculates and records the hash values of all items produced in native format. Were there suspicion, alteration is apparent merely by hashing the file.

De-duplication

In e-discovery, manually reviewing vast volumes of identical data is burdensome and poses a significant risk of conflicting relevance and privilege assessments. Hashing serves to flag identical documents, permitting a single, consistent assessment of an item that might otherwise have cropped up hundreds of times and been differently characterized. This is **hash de-duplication**, and it drastically cuts the cost of reviewing data for responsiveness and privilege. But because even the slightest difference triggers different hash values, insignificant variations between files (e.g., different Internet paths taken by otherwise identical e-mail messages) may frustrate hash de-duplication. An alternative is to hash relevant *segments* of e-documents to assess their relative identicality, a practice sometimes called "**near de-duplication**."

In practice, each file processed, and each constituent item extracted, is hashed and their hash values compared to the hash values of items previously processed and extracted to determine if the file or item has been seen before. Thereafter, files and items with matching hashes are suppressed as duplicates, and instances of each duplicate and associated metadata are noted in a deduplication or occurrence log.

Takeaways on Hashing

The most important things to know about hashing:

1. Electronically stored information of any type or size can be hashed;
2. The algorithms used to hash data are not proprietary, and thus cost nothing to use;
3. No matter the size of the file that's hashed, its hash value is *always* a fixed length;
4. The two most common hash algorithms are called MD5 and SHA-1.

5. In a random population of hashed data, no one can reverse engineer a file's hash value to reveal anything about the file

6. The chance of two different files accidentally having matching MD5 hash values (a so-called *hash collision*) is one in 340 *trillion trillion trillion (i.e.*, 340 undecillion). So, it is highly improbable that two files with matching hash values are not identical

NOTE TO STUDENTS: We are seeking to lay a solid foundation in terms of your grasp of the fundamentals of information technology and the jargon used in e-discovery. Looking back over the preceding material, please list any topics and terms you don't understand and share your list with the instructor in order that we might go over those topics in class. Don't be shy!



Exercise 1: Identifying Digital Storage Media

All answers should reference lettered items in the image below.



I. Insert the letter(s) of the media described in the adjacent blank:

1. Backup Tape _____
2. USB Thumb drive _____
3. SD media card _____
4. SIM Card _____
5. RAID array _____

II. Identify three items that record data *magnetically*: ____ ____ ____

III. Identify four *solid state* digital storage devices: ____ ____ ____ ____

IV. Identify the three pictured items with the most meager (i.e., least) capacity to store digital information (capacity may be zero):

____ ____ ____

Introduction to Metadata

In the old joke, a balloonist descends through the fog to get directions. "Where am I?" she calls out to a man on the ground, who answers, "You're in a yellow hot air balloon about sixty-seven feet above the ground." The frustrated balloonist replies, "Thanks for nothing, Counselor." Taken aback, the man on the ground asks, "How did you know I'm a lawyer?" "Simple," says the balloonist, "your answer was 100% accurate and totally useless."

If you ask a tech-savvy lawyer, "What's metadata?" there's a good chance you'll hear, "Metadata is data about data." Another answer that's 100% accurate and totally useless.

It's time to move past "data about data" and embrace more useful ways to describe metadata—ways that enable counsel to rationally assess relevance and burden attendant to metadata in modern, networked, and cloud-based information systems.

It's time to get past defining metadata as a slogan rather than an explanation.

Metadata may be the most misunderstood topic in electronic discovery. Requesting parties demand discovery of "the metadata" without specifying what metadata is sought, and producing parties fail to preserve or produce metadata of genuine value and relevance —often because they do not understand what metadata exists, where it resides, or how easily it may be altered, lost, or generated as a byproduct of ordinary system operation.

It's Information *and* Evidence

Metadata is information that helps us use and make sense of other information. More particularly, metadata is information stored electronically that describes the characteristics, origins, usage, structure, alteration and validity of other electronic information—including when, where, how, and by whom electronic information was created, accessed, modified, transmitted, or deleted.

Many instances of metadata in many forms occur in many locations within and without digital files and systems, including operating systems, applications, databases, networks, mobile devices, cloud platforms, and collaboration services. Some are supplied by the user, but most metadata are generated by systems and software automatically, often without user awareness or control.

Some is crucial evidence, and some is merely digital clutter. Appreciating the difference—knowing what metadata exists and understanding its evidentiary significance—and recognizing when metadata is incomplete, misleading, or absent altogether is a skill essential to electronic evidence and discovery.

Metadata is Evidence!

If evidence is anything that tends to prove or refute an assertion as fact, then clearly metadata is evidence. Metadata sheds light on the origins, context, authenticity, reliability and distribution of electronic evidence, as well as provides clues to human behavior—including authorship, chronology, access patterns, collaboration, and intent.

It's the electronic equivalent of DNA, ballistics and fingerprint evidence, not because it is infallible, but because, when properly preserved, interpreted, and contextualized, it can be highly probative, with a comparable power to exonerate and incriminate or to mislead when misunderstood, incomplete, or taken out of context.

In *Williams v. Sprint/United Mgmt. Co.*, 230 F.R.D. 640 (D. Kan. 2005), the federal court ruled **in a dispute over Excel spreadsheets that had been altered prior to production**:

> [W]hen a party is ordered to produce electronic documents as they are maintained in the ordinary course of business, the producing party should produce the electronic documents with their metadata intact, unless that party timely objects to production of metadata, the parties agree that the metadata should not be produced, or the producing party requests a protective order.

Sprint had produced spreadsheets with cells locked and metadata removed, impairing the recipients' ability to evaluate formulas, relationships among cells, and the integrity of the data as maintained in the ordinary course. The court made clear that stripping or disabling such attributes was not ordinary-course production, but an alteration requiring justification.

Within the realm of metadata lies discoverable evidence that litigants are obliged to preserve and produce. There's as much or more metadata extant as there is information—often far more—and, like information, you don't deal with every bit of it. You choose wisely based on relevance, proportionality, and evidentiary value.

Often, files have *more* metadata than content.

A lawyer's ability to advise a client about how to find, preserve and produce metadata, or to object to its production and discuss or forge agreements about metadata, hinges upon how well he or she understands metadata —including how certain forms of ESI, such as spreadsheets and databases, collapse the distinction between "data" and "metadata" altogether.

'It's Just Ones and Zeroes'

Understanding metadata and its importance in e-discovery begins with awareness that electronic data is, fundamentally, just numbers. Though you've heard that before, you may not have

considered the implications of information being expressed so severely. There are no *words*. There are no spaces or punctuation. *There is no delineation of any kind*. *Solely binary numbers.*

How, then, do computers convert this unbroken sequence notated as ones and zeroes into information that makes sense to human beings? There must be some *key*, some *coherent structure* imposed to divine their meaning. But where does it come from? We can't derive meaning *from* the data if we can't first make sense *of* the data.

It's Encoded

Consider that written English conveys all information using fifty-two upper- and lowercase letters of the alphabet, ten numerical digits (0-9), some punctuation marks and a few formatting conventions, like spaces, lines, pages, etc. You can think of these collectively as a seventy- or eighty-character "code." Alternatively, the same information could be communicated or stored in Morse code, where a three-signal code composed of dot, dash and pause serves as the entire "alphabet."

We've all seen movies where a tapping sound is heard and someone says, "Listen! It's Morse code!" Suddenly, the tapping is an encoded *message* because someone has furnished metadata ("It's Morse code!") *about* the data (tap, tap, pause, tap). Likewise, all those 'ones and zeroes' on a computer only make sense when other ones and zeroes—*metadata*—reveal a framework for parsing and interpreting the data.

All those 'ones and zeroes' on a computer only make sense when *other* ones and zeroes—*metadata*—reveal a framework for parsing and interpreting the data.

So, we *need* data *about* the data. We need information that tells us the encoding scheme. We need to know when information of one sort concludes, and different information begins. We need the name, date, context, purpose and origin of information to support its utility and integrity. We need its ***metadata***.

The Metadata Continuum

Sometimes metadata is elemental, like the contents of a computer's file system structures detailing where recorded data blocks begin and end and how they are organized. This metadata is invisible to a user without special tools or forensic utilities capable of peering through the façade of the user interface into the utilitarian plumbing of the file system. Without file location metadata, each time a user sought to access a file or program, the operating system would have to peruse the entire drive to find it. It'd be like looking for someone by knocking on every door in town!

At other times, metadata supports enhanced functionality not essential to the operation of the system. The metadata that tracks a file's name or the dates a file was created or last modified may only occasionally be probative of a claim or defense in a lawsuit, but that information *always* makes it easier to locate, sort and segregate files and to manage information at scale.

Metadata is often instrumental to the intelligibility of information, helping us make sense of it. "Sunny and 70 degrees" aren't a very useful forecast without metadata indicating *when* and *where* it's the weather. Similarly, understanding information on a website or within a database, a cloud-based collaboration platform like Microsoft SharePoint or Teams, or a social network like Facebook depends on metadata that defines its location, origin, timing and structure. It's even common for computerized information to comprise more metadata than data, in the same way that making sense of the two data points "sunny" and "70 degrees" requires three metadata points: location, date and time of day.

There's No Such Thing as "The Metadata"

As we move up the evolutionary ladder for metadata, some metadata is recorded in case it's needed to support a specialized task for the operating system or an application. Standard **System Metadata** fields like "Camera Model" or "Copyright" may seem an utter backwater to a lawyer concerned with spreadsheets and word-processed documents; but, if the issue is the authenticity of a photograph or the origins of pirated music, these fields can make or break a case. ***It's all about relevance and utility in context.***

The point is, there's really no such thing as "the metadata" for a file or document. Instead, there's a continuum of **System** and **Application Metadata** that enlightens many aspects of ESI. The metadata that matters depends upon the issues presented in the case and the task to be accomplished; consequently, the metadata preserved for litigation should reasonably reflect the issues that can be reasonably anticipated, and it must also address the file management and integrity needs attendant to identification, culling, processing, review and presentation of electronic evidence. Again, relevance and utility.

File Systems and Relative Addressing

Most of those ones and zeroes[36] on a hard drive are files that, like library books, are written, read, revised and referenced. Computers use file systems to keep track of files just as libraries once used card catalogues and the Dewey Decimal system to track books.

Imagine you own a thousand books without covers that you store on one very long shelf. You also own a robot named Robby that can't read, but Robby can count books very accurately. How would you instruct Robby to get a particular book?

If you track the order in which the books are stored, you might say, "Robby, bring me the 412th book." If it was a 24-volume set of encyclopedias, you might add: "…and the next 23 books." The

[36] I cringe every time I refer to digital information as being stored as "ones and zeroes" because (as we should now know) that's just a convenient way to *notate* the data, not how it's stored on digital media.

books don't "know" where they're shelved. Each book's location is metadata *about* the book, but it's not stored within the book. The system tracks that metadata. It's *System Metadata.*

Locating something by specifying that it's so many units from a particular point is called *relative addressing* or *offset addressing*. The number of units the destination is set off from the specified point is called the *offset*. Computers use offset values to indicate the locations of files on storage devices as well as to locate information inside files.

Computers use various units to store and track information, so offsets aren't always expressed in the same units. As previously explained, a "bit" stores a one or zero, eight bits is a "byte," (sufficient to hold a letter in the Latin alphabet), 512 bytes is often a *sector or block* (see **Appendix A**) and (typically) eight contiguous sectors or blocks is a *cluster*. The cluster is the most common unit of logical storage, and modern computers tend to store files in as many of these 4,096-byte (4 KB) clusters, or "data baskets," as needed. Offset values are couched in bytes when specifying the location of information within files and as sectors when specifying the location of files on storage media.[37]

Application Metadata

To the extent lawyers are familiar with metadata, it's likely just the type called **application metadata** with the fearsome potential to inadvertently reveal confidential or privileged information embedded within electronic documents. Computer programs or "applications" store data in files "native" to them, meaning that the data is structured and encoded to uniquely support the application. As these applications added features--like a word processor's ability to redline changes or collaborate on a document--the files used to store documents necessarily retained those tracked changes and collaborative comments.

Microsoft Word was once notorious for its potential to store information unseen by users, and a cottage industry grew up offering utilities to strip embedded information, like comments and tracked changes, from Word documents. Because of its potential to embarrass lawyers or compromise privilege, metadata acquired an unsavory reputation.[38] But metadata is much more than embedded *application* metadata affording those who know how to find it the ability to dredge up a document's non-obvious content; it is an integral byproduct of modern software functionality.

By definition, **application metadata is embedded in the file it describes and moves with the file when copied.** But not all metadata is embedded for the same reason that cards in a library card

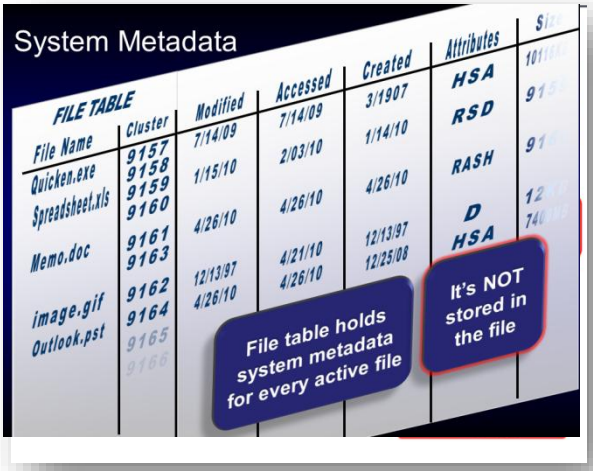
[37] On modern storage systems, including solid-state drives and virtualized storage, these locations are *logical* rather than fixed physical positions, and may change over time without the file system's awareness.

[38] Once, a few states' Bar disciplinary authorities (*e.g.,* NY, FL) deemed it unethical for lawyers to *look* at metadata in e-documents received from opponents! Happily, that notion quickly lost favor.

catalog aren't stored between the pages of the books: *You need to know where information resides to reach it*.

System Metadata

Unlike books, computer files aren't neatly bound tomes with names embossed on spines and covers. Typically, files don't internally reflect the name they've been given or other information about their location, history or ownership. The information about the file that's *not* embedded in the file it describes but stored apart from the file is its **system metadata**. The computer's file management system uses *system* metadata to track file locations and store attributes like each file's name, size, and dates of creation, modification and usage.



System metadata is crucial to electronic discovery because so much of our ability to identify, find, sort, cull and authenticate information depends on its system metadata. For example, system metadata helps identify the custodians of files, what the files are named, when files were created or modified and the folders in which they are stored. System metadata stores much of the *who, what, when, where* and *how* of electronic evidence.

Every computer employs one or more databases to keep track of system metadata. In computers running the Windows operating system, the principal "card catalog" holding system metadata is called the Master File Table or "MFT." In the predecessor DOS operating system, it was called the File Allocation Table or "FAT." The more sophisticated and secure the operating system, the greater the richness and complexity of the system metadata in its file table.

Windows Shell Items and Properties

In the Microsoft Windows ecosystem, Microsoft refers to discrete units of content—such as files, folders, messages, and contacts—as "Shell items." Each discrete attribute associated with a Shell item is called a "property." Windows maintains hundreds of such properties, organized across 34 categories, many of which may exist outside the file's content and yet still bear significant evidentiary value.

Shell item properties illustrate an important point for discovery: metadata is not confined to what applications embed in files, nor is it limited to the familiar created/modified/accessed dates. Operating systems routinely generate and maintain rich descriptive information that may never be

visible during ordinary use, yet may prove highly probative of authorship, chronology, provenance, access, or use.

Examining even a small subset of Shell item properties demonstrates the breadth of potentially relevant metadata that can exist within and without files, messages, and photographs, including document authorship and revision history, message routing and handling details, camera identifiers, and system-level attributes reflecting file handling and ownership.

Application vs. System Metadata
Application Metadata = CONTENT *If it changes what the document says or does, it's application metadata.* • Embedded **inside** the file • Travels **with** the file • Enables functionality (comments, formulas, revisions) • Can be **substantive evidence** or a **privilege trap** **Ask:** *Does this information belong to the document itself?* **System Metadata = CONTEXT** *If it describes where, when, or how the document existed, it's system metadata.* • Stored **outside** the file • Managed by the system, not the application • Used to locate, sort, and authenticate files • Essential to discovery logistics and provenance **Ask:** *Does this information describe the document rather than form part of it?*

Much More Metadata

The hundreds of Windows Shell item properties are by no means an exhaustive list of metadata. Software applications deploy their own complements of metadata geared to supporting features unique to each application. E-mail software, word processing applications and spreadsheets, databases, web browsers and presentation software collectively employ thousands of additional fields of metadata across modern platforms.

For example, digital photographs can carry dozens of embedded fields of metadata called **EXIF data** detailing information about the date and time the photo was taken, the camera, settings, exposure, lighting, and, where enabled,

Photos taken with cell phones routinely hold detailed EXIF information about where the photo was taken.

precise geolocation data. Cell phone photos contain detailed information about where the photo was taken often to within a few meters, depending on sensor and settings.

The popular Microsoft Outlook e-mail client application provides for over 180 standard application metadata fields which users may select to customize their view—and many more that are generated and maintained without user visibility.

But even this broad swath of metadata is only part of the probative information about activity recorded by computers. Within the Master File Table and index records used by Windows to track all files, still more attributes are encoded as structured system records not ordinarily visible to users. In fact, an ironic aspect of Windows is that the record used to track information about a file may be larger than the file itself!

Stored within the hives of the System Registry—the database that tracks attributes covering almost any aspect of the system—are thousands upon thousands of attribute values called "registry keys." Other records and logs track network activity and journal system, application, and user activity at a granular level.

Matryoshka Metadata

Matryoshka are carved, cylindrical Russian dolls that nest inside one another. Metadata works the same way. If the evidence of interest is a Word document attached to an e-mail, the Word document carries its own embedded application metadata; but once attached, its *system* metadata is reduced to what the transporting message conveys—typically little more than file name and type, plus limited contextual attributes supplied by the sending system.



The e-mail message, in turn, carries its own metadata concerning addressing, routing, structure, and encoding. That message is managed by an e-mail application such as Outlook, which maintains additional metadata about the message and its configuration. Depending on configuration, those messages and their metadata may reside in local container files (such as PST or OST files) or in server-side mailboxes. Those containers or mailboxes then exist within a file system or service that maintains still more system metadata about location, ownership, timestamps, size, and related attributes.

Within this Matryoshka maelstrom of metadata, some information is readily accessible while other layers are opaque, technical, and unintelligible without specialized tools and experience.

Forms of Metadata

As if the variety of metadata weren't enough, metadata also varies in *form*. It is not uniformly human-readable or self-explanatory. Some metadata fields are simple bit flags indicating "true" or "false." Others encode numeric values whose meaning depends entirely on context. Still others reuse the same numeric value to signify different states in different fields.

The form of metadata matters when deciding how to preserve and produce it. A response like "item type 0x0029 was set to 0x00" is meaningless unless translated into the functional question it answers—such as whether a read receipt was requested. Because many metadata values only make sense within the application that interprets them, preservation and production require more than mechanical extraction. Context is essential.

The challenge is not that metadata cannot be located or interpreted, but that counsel must know whether the firm, client, or service provider has tools and workflows that do so accurately and repeatably. Before committing to produce metadata—or objecting to its production—counsel must understand what metadata is routinely collected, how it is processed, and which fields require specialized handling to avoid alteration or misrepresentation.

Relevance and Utility

How much of this metadata is relevant and discoverable? Would I be any kind of lawyer if I didn't answer, "It depends?" In truth, it does depend upon the issues the data bears upon, its utility, and the cost and burden of preservation and review.

Metadata is unlike almost any other evidence in that its utility may flow from its probative value (its relevance as evidence) or from its utility—its ability to support searching, sorting, and interpretation of ESI—or both. If the origin, use, distribution, destruction, or integrity of electronic evidence is at issue, the "digital DNA" of metadata is essential probative evidence that needs to be preserved and produced for its *relevance*. Likewise, if metadata materially facilitates the searching, sorting, and management of electronic evidence, it should be preserved and produced for its *utility*.[39]

[39] This duality of metadata—relevance *and* utility—is sometimes overlooked by readers who focus narrowly on the text of the rules and ignore the Committee Notes. Rule 26(b)(1) permits discovery of nonprivileged matter that is relevant to a party's claims or defenses *and* proportional to the needs of the case. The 2015 Committee Notes explain that examples formerly enumerated in the rule—such as information concerning the existence, location, custody, and condition of documents—were removed not to narrow discovery, but because such inquiries are "deeply entrenched in practice" and remain discoverable when relevant and proportional. The Notes further observe that framing intelligent requests for electronically stored information may require detailed information about another party's information systems. Although the Committee did not use the term "metadata," the import is clear: discoverability is not limited to evidentiary relevance alone. Metadata may be discoverable for its utility as well as its probative value.

Put simply, metadata is an indispensable feature of ESI and should be considered for preservation and production in every case. Too often, much of what is dismissed as "mere metadata" is truly substantive content, such as embedded comments between collaborators in documents, speaker notes in presentations, and formulas in spreadsheets.

Does this then mean that every computer system and data device in every case must be forensically imaged and analyzed by experts? Absolutely not! Once we understand what metadata exists and what it signifies, a continuum of reasonableness will inform our actions. A police officer making a traffic stop routinely collects relevant "dog tag" data, *e.g.*, driver's name, address, vehicle license number, driver's license number, and the date, time, and location of the offense. We wouldn't expect a traffic cop to collect a DNA sample or fingerprint the driver; but make it a murder investigation and the calculus changes.

The crucial factors are burden and cost, balanced against utility and relevance. The goal should be a level playing field between the parties in terms of their ability to see and use relevant electronic evidence, including its metadata.

So where do we draw the line? Begin by recognizing that the advent of electronic evidence hasn't changed the fundamental dynamics of discovery: Litigants are entitled to discover relevant, non-privileged information, and relevance depends on the issues before the court. Relevance assessments aren't static but change as new evidence emerges and new issues arise. Metadata deemed irrelevant at the start of a case may become decisive when, *e.g.,* allegations of data tampering or spoliation emerge. Parties must periodically re-assess the adequacy of preservation and production of metadata and act to meet changed circumstances consistent with proportionality and evolving needs.

Periodically re-assess the adequacy of preservation and production of metadata, and act to meet changed circumstances.

Metadata "Musts"

There are readily accessible, frequently valuable metadata that, like the dog-tag information collected by a traffic cop, we should expect to routinely preserve and produce. Examples of essential system metadata fields for any file produced are:

- Custodian(s)
- Source Device
- Originating Path (file path as it resided in its original environment)
- Filename (including extension)
- Last Modified Date
- Last Modified Time

Any party producing or receiving ESI should be able to state something akin to, "This spreadsheet named *Financial Forecast.xlsx* came from the Documents folder on Sarah Smith's Dell laptop and was last modified on January 14, 2026 at 2:07 PM CST."

Another metadata "must" is the **UTC time-zone offset** applicable to each time value (unless all times have been normalized to a common zone). UTC—*Coordinated Universal Time*—is the modern reference standard for timestamps and avoids ambiguity caused by local time zones and daylight-saving rules. Without the applicable offset, time values may be misleading or uninterpretable.[40]

Application metadata is, by definition, embedded within native files; thus, native production ordinarily preserves application metadata without special handling. But when ESI is converted to other forms, the parties must assess what metadata will be lost or altered and identify, preserve, and extract relevant application and system metadata fields for production in ancillary files commonly called **load files**.

For e-mail messages, this is generally straightforward notwithstanding the many metadata values generated by client and server applications. The metadata "musts" for e-mail messages are, as available:

- Custodian
- To
- From
- CC
- BCC
- Subject
- Date Sent (or Received)
- Time Sent (or Received)
- Attachments (names or unique identifiers)
- Mail Folder Path
- Message ID

E-mail messages that traverse the Internet contain header data detailing routing and delivery. Whether header data should be preserved and produced depends on the reasonable anticipation that issues of authenticity, receipt, or timing may arise—but because headers are integral to the message, discarding them absent good cause is difficult to justify.

[40] UTC stands for both *Temps Universel Coordonné* and Coordinated Universal Time. It's a fraction of a second off the better-known Greenwich Mean Time (GMT) and identical to Zulu time in military and aviation circles. Why UTC instead of TUC or CUT? It's a diplomatic compromise, for neither French nor English speakers were willing to concede the acronym.

Metadata "musts" also include values generated during e-discovery processing, review, and production, such as Bates numbers, attachment ranges, hash values, production paths, duplicate identification, and family relationships.

When ESI other than e-mail is converted to non-native forms, preserving application metadata without impairing its utility or intelligibility can be difficult. Tracked changes and comments, for example, may be incomprehensible without context, yet producing them in static images can confuse recipients and degrade searchable text. Where native production is not permitted, an equitable alternative may be dual production—once with tracked changes hidden and once revealed.[41]

For certain ESI, there is no practical substitute for native production with metadata intact. Spreadsheets are the classic example. Similar losses of functionality occur with audio, video, animated presentations, databases, and collaborative environments, where structure and relationships—reflected in metadata—define utility. Production formats are addressed elsewhere in this book; suffice it to say that native production's principal strength lies in its ability to preserve and exploit metadata.

The Path to Production of Metadata

Producing metadata is not a single act but a sequence of judgments. The steps typically include:

- gauging spoliation risk,
- identifying potentially relevant metadata,
- balancing relevance, utility, and burden,
- considering authentication and admissibility,
- determining preservation methods,
- collecting metadata, and
- resolving privilege and production issues.

Gauge Spoliation Risks Before You Begin

German physicist Werner Heisenberg taught that observation can alter what is observed. The analogy applies neatly to metadata: opening or handling electronic files without care can change metadata values and destroy prior information. When you open any document in Office applications without first employing specialized hardware or software, metadata often changes,

[41] But the viability of this clunky "solution" must be weighed against the greatly increased cost to load and host alternate versions of documents considering that vendors typically charge for services by the gigabyte. Two sets of static images substantially inflate the cost of discovery for the parties receiving such a double-whammy production.

and prior metadata values may be lost. Such alteration can impair chronology, complicate review, and invite spoliation claims.[42] The threshold question, then, is how much risk the case warrants. Not every matter is a crime scene. Few cases require full forensic preservation, but those that do demand care proportionate to the issues. In many instances, modest precautions suffice to protect metadata likely to matter.

On Windows systems, three familiar timestamps—Created, Last Accessed, and Last Modified—illustrate the point. Last Accessed dates are fragile and, today, rarely updated; they are seldom useful or reliable values.

Created dates are frequently misunderstood: they reflect creation *on a particular storage medium*, not authorship, and change when files are copied while sometimes persisting when templates are reused. That is, when you copy a file to new media, you've "created" it on the new media as of the date of copying, and the created date changes accordingly. Once more, *created dates may or may not coincide with authorship*; so, it's a mistake to assume they do.

The most stable and useful timestamp is Last Modified, because it is not changed by copying, previewing or virus scans. It changes only when a file is opened and saved—though not necessarily when visible content changes because other changes—including subtle, automatic changes to application metadata—may trigger an update to the last modified date when the file is re-saved by a user.

Application metadata generally changes only when a file is opened. Accordingly, the simplest preservation strategy is to maintain a pristine copy and conduct review on working copies. Preserving hash values for the pristine set provides reliable proof of integrity. More rigorous methods—such as write blocking or forensic imaging—are available when warranted, and modern review tools are designed to avoid metadata alteration. Finally, containerized copies[43] effectively preserve system metadata values.

Identify Potential Forms of Metadata

You cannot preserve or assess metadata you do not know exists. For each principal file type, identify metadata fields of potential evidentiary or functional significance. Early collaboration with experts, IT personnel, or opposing counsel often narrows disputes and clarifies expectations. Knowing what metadata exists, where it resides, and what it signifies is foundational.

Assess Relevance, Utility, and Burden

[42] Spoliation is the negligent or intentional loss or destruction of evidence. Spoliation of ESI often flows from a failure to preserve relevant data promptly or properly. When spoliation is intentional, it may prompt significant sanctions (*i.e.*, punishments) assessed against the spoliating party.

[43] Typically a compressed .Zip file. The zip format replicates a broad range of system metadata values.

Producing every metadata field for every item is neither required nor sensible. Preservation may be broad, but production should be guided by relevance and utility. Modern tools routinely extract extensive metadata, and producing additional fields is often trivial if requested before production. Claims of undue burden are weak when production amounts to exporting another column.

Relevance is fluid. Metadata that seems immaterial at the outset may become decisive later. Consider two commonplace fields in Adobe PDF files: **PDF Producer** and **PDF Version**. These document properties identify the source application and the release of Acrobat used to create the file. They may appear esoteric until a dispute turns on the authenticity of a purportedly old contract. If the metadata shows the PDF was created using a scanner introduced last year and a recent version of Acrobat, it supports a claim of fabrication. If it reflects use of an early scanner and an older release, it bolsters the claim that the document was scanned years ago. Neither is conclusive, but both are relevant and warrant preservation and production.

Dialogue helps. Many disputes evaporate when an opponent concedes, "I don't need that." Others sharpen when metadata becomes the battleground.

Consider Authentication and Admissibility

Electronic evidence lacks paper's physical cues to authenticity, like signatures, handwriting and physical watermarks. Computer user accounts may be shared or compromised, and software and AI tools enable seamless alteration. Dates may be system-generated conveniences rather than reliable markers of authorship or timing.

When authenticity may be contested, preservation of original system metadata is critical. Relevant sources may include user identifiers, system and network logs, version histories, and evidence of contemporaneous activity. Preservation choices should balance burden against anticipated need—but selective preservation is risky. If you preserve metadata supporting your case, it is difficult to justify discarding metadata that may support the opposition or undermine your own proof.

Chain of Custody

An important role of metadata is establishing and maintaining a defensible chain of custody for ESI. Through every stage of e-discovery--collection, processing, review, and production—metadata should facilitate a clear, verifiable path back to the source ESI, device and custodian.

"Chain of custody" describes the processes used to track and document the acquisition, storage and handling of evidence so as to demonstrate that the integrity of the evidence has not been compromised. From movies and television, we're familiar with the signed and sealed evidence bags

in police property rooms and the sign in/out logs and other steps law enforcement agencies use to safeguard physical evidence. But what are the corollary steps required for digital evidence?

As a rule, counsel should be able to trace any item of digital evidence back to its origin. So, there must be a means to identify the device, repository, path, container file and custodian of the data. When electronic evidence is collected, or media imaged for preservation, collections and images should be hashed ("digitally fingerprinted") upon acquisition and those hash values recorded and preserved.

Digital evidence is unique in that its ability to be duplicated and authenticated without compromising any iteration deemed to be "original."[44] Nonetheless, it remains sound practice to protect data and interdict or log any actions that may alter the evidence or its hash values.

Evaluate Need and Methods for Preservation

Not every bit of metadata is important in every case, *so what factors should drive preservation*? The case law, rulings of the presiding judge and regulatory obligations are paramount concerns, along with obvious issues of authenticity and relevance. Another critical consideration is stability. As discussed, essential metadata fields, like Last Modified Date, change when a file is used and saved. If you don't preserve dynamic data, you lose it. Where a preservation duty has attached, by, *e.g.*, issuance of a preservation demand or by operation of law, the loss of essential metadata may prompt costly remedial measures. Worse, it may constitute spoliation subject to sanctions.

If you fail to preserve metadata at the earliest opportunity, you may never be able to replicate what was lost.

How, then, do you avoid altering metadata by review and collection? What methods will preserve the integrity and intelligibility of metadata? Poorly executed collection efforts can corrupt metadata. For example, when a custodian or reviewer opens files in native applications, copies responsive files to new media, prints documents or forwards e-mail as a means of collection, metadata is altered or lost. Consequently, metadata preservation must be part of a *defensible preservation protocol* and addressed in preservation directives, so-called "legal hold notices" sent to custodians of evidence when litigation is anticipated. Be certain to document what was done and why. Courts expect a modicum of transparency concerning data preservation, so consider sharing proposed protocols with opposing counsel in sufficient time to allow adversaries to object, seek court intervention or agree to alternate approaches.

Collect Metadata

[44] In the world of digital forensics, the notion of "original" or "best" evidence no longer means much in that one hash validated copy of ESI is indistinguishable from another.

Because metadata is stored both within and without files, simply duplicating a file without capturing its system metadata is insufficient. Not all metadata preservation efforts demand complex and costly solutions; methods should be tailored proportionally to the case. As feasible, record and preserve system metadata values before use or collection. This can be achieved using software that archives basic system metadata values to a table, spreadsheet or CSV file. Then, if there's corruption of metadata, the original values can be ascertained. Even just archiving files ("zipping" them) may prove a sufficient method to preserve associated metadata in small cases. Optimally, you (or your service providers) will use purpose-built tools for e-discovery and forensically sound collection.

Whatever the method chosen, safeguard the association between the data and metadata. For example, if data is the audio component of a voice mail message, recordings may be of little use unless correlated with metadata detailing the date and time of the call and the identity of the voice mailbox user. Similarly, email attachments must tie back to their transmittals. These efforts are termed "preserving family relationships."

When copying file metadata, know the limitations of the environment and medium in which you're working.

I learned that lesson the hard way many years ago while experimenting with recordable CDs to store evidence. Each time I copied a file with distinct MAC dates (modified, accessed, and created) from a hard drive to a CD, all three dates collapsed into a single value when read back. I was unwittingly corrupting the very metadata I meant to preserve!

The explanation was simple and sobering: optical media like CD-Rs are not formatted to store multiple timestamps the way magnetic drives are. With only one date field available, two of the three MAC values were discarded. When the files were later copied back to a hard drive, the operating system repopulated all three timestamp fields with that single surviving date—silently misrepresenting the file's history. The broader lesson remains: different storage media, operating systems, and applications support different metadata structures and limits. Test your processes, or risk altering, truncating, or losing metadata without realizing it.

Plan for Privilege and Production Review

The idea of reviewing metadata for privilege may seem odd unless you consider that application metadata may reveal deleted content, comments, or prior versions. The industry standard has long been to simply suppress the metadata content of evidence, functionally deleting it from review and production. This occurred without legal justification (*i.e.,* privilege). Producing parties didn't want to review metadata so simply, *incredibly,* purged it.

That's indefensible. Metadata must be assessed like any other potentially responsive ESI and produced when tied to responsive, non-privileged content.

When the time comes to review metadata for production and privilege, the risks of spoliation faced in collection may reappear during review. Counsel should consider:

- How will metadata be efficiently accessed?
- Will it exist in a form that can be interpreted?
- Will review alter the metadata?
- How will metadata be tagged for production?
- How will privileged or confidential metadata be redacted?

Fortunately, modern e-discovery review platforms are designed to address these concerns. What remains perilous is the use of native applications as review tools. *Don't do that!*

Application Metadata and Review

As noted above, many lawyers deal with metadata by pretending it does not exist. They employ review methods that suppress application metadata—such as comments, tracked changes, formulas, and speaker notes—reviewing only what "prints" instead of all the information contained in the document. Rather than adapt workflows to the evidence, they suppress application metadata out of fear that privileged or confidential content may be inadvertently produced, or simply from unfamiliarity. The usual defense is burden: reviewing application metadata is said to cost more than it is worth.

To ensure that requesting parties cannot access metadata the producing party never examined, counsel often strip metadata wholesale—by converting ESI to static images, typically TIFF. While effective at removing metadata, these practices impair the utility, integrity, and searchability of the evidence.

Producing parties then attempt to reintroduce fragments of stripped metadata and searchable text through ancillary load files, resulting in so-called "**TIFF-plus**" productions we will discuss later in the semester. This approach is costly, fragile, and ill-suited to modern ESI. Spreadsheets become unreadable. Multimedia disappears. Interactive, animated, and structured information breaks. As a rule, the richer the information, the less likely it is to survive conversion to static images.

This persistence raises a more fundamental question: why does any lawyer assume the right to unilaterally suppress—without review or privilege disclosure—integral parts of discoverable evidence? Stripping metadata because it *might* contain privileged material is little different from erasing handwritten notes in medical records because the handwriting is difficult to read.

Courts have repeatedly rejected speculative privilege as a justification for wholesale metadata removal. In *Aguilar v. Immigration & Customs Enforcement Div. of U.S. Dep't of Homeland Sec., 255 F.R.D. 350 (S.D.N.Y. 2008)*, the court emphasized that metadata associated with native files is discoverable when relevant and that disputes over metadata must be resolved field-by-field, not by categorical suppression. In *Covad Communications Co. v. Revonet, Inc., 258 F.R.D. 5 (D.D.C. 2009),* the court ordered production of native spreadsheets with metadata intact, recognizing that metadata in spreadsheets is often inseparable from the evidence itself. Decided after the 2015 amendments to Rule 26(b)(1), *Heraeus Kulzer GmbH v. Biomet, Inc., 633 F.3d 591 (7th Cir. 2011)* reaffirmed that proportionality does not permit the silent alteration or concealment of metadata necessary to assess authenticity and sequencing.

Against this backdrop, *Williams v. Sprint/United Management Co.* remains instructive. There, the producing party stripped metadata from native spreadsheet files based on generalized privilege concerns. The court rejected that approach, distinguishing targeted redaction—following review and accompanied by a privilege log—from blanket excision undertaken without examination or disclosure. Privilege, the court made clear, does not excuse non-review.

The upshot is straightforward: requesting parties are entitled to the same metadata benefits available to producing parties. Metadata may be redacted when privileged, but it may not be vandalized, suppressed, or silently altered. The rules governing privilege review and logging apply to metadata no less than to the face of a paper document. Issues of production format and load files are addressed elsewhere in this book.

The requesting party is entitled to the metadata benefits that are available to the producing party.

Resolve Production Issues

Metadata can be produced in many forms: as a database, a delimited load file,[45] embedded within native files, displayed through an online review platform, or—even now—rendered on paper. Each method carries risks because metadata presents production challenges unlike those posed by conventional documents.

One challenge is intelligibility. Metadata is often encoded and meaningless outside its native or processed environment. An unlabeled value may represent a creation date, a modification date, or something else entirely. Without decoding and labeling, metadata becomes inscrutable or, worse, misleading.

[45] Load files are commonly used to convey searchable text and metadata in electronic productions. Delimited load files consist of structured text in a predetermined sequence, with individual values separated by characters such as commas (as in CSV files), tabs, or quotation marks that serve as delimiters (i.e., separators). We will explore the use and structure of load files later in the semester.

Another challenge is form. Metadata is not always textual. It may be a numeric value or a single bit flag—true or false—without meaning unless one knows what the flag signifies. A third challenge lies in preserving the relationship between metadata and the data it describes and, where required, ensuring that both remain electronically searchable.

When data is separated from its metadata, much of its evidentiary value is lost. Consider a collaborative cloud document, such as a Google Docs file maintained in a shared environment. A printed or imaged copy may reveal what the document says, but it conceals who authored particular passages, when edits were made, what was deleted, and how the document evolved over time. Without version history, authorship, and timestamp metadata, allegations of fabrication, backdating, selective revision, or bad-faith editing cannot be meaningfully evaluated. What remains may resemble evidence, but it no longer reflects how the information was created, used, or understood.

Sometimes, producing a well-constructed load file preserving key originating metadata values will suffice. Other times, only native production can convey relevant metadata in a usable and complete way. Determining the method of metadata production suited to the case demands planning, technical competence, and cooperation with the other side.

Beyond "Data About Data"

The relentless march of digital technology has only heightened the evidentiary and functional importance of metadata. Today, nearly all information is born digitally and defined by its metadata. Authorship, timing, location, versioning, and integrity increasingly turn not on what evidence says, but on what its metadata reveals.

It is time to retire the glib definition of metadata as merely "data about data." Metadata is context, structure, and history. It is the means by which electronic information is understood, managed, and trusted. For lawyers engaged in electronic discovery, metadata is not a technical curiosity or a procedural nuisance; it is an indispensable feature of the evidence itself.

Crucial Distinctions: System versus Application Metadata:

- System metadata describes the *file as an object* (*e.g.*, name, location, dates). It is **CONTEXT** and typically **resides outside the file** in file-system tables.
- Application metadata describes the *information within the file* (*e.g.*, comments, formulas, geolocation). It is **CONTENT** and is typically **embedded in the file**.

System Metadata — Examples:
File name and extension; file size; file path; custodian; Modified, Accessed, and Created (MAC) dates.

Application Metadata — Examples:
Comments; tracked changes; formulas; revision history; editing time; last printed date; EXIF data in photos.

Production Implications:

- System metadata is commonly produced in structured load files (*e.g.,* CSV, DAT) or other tabular formats.
- Application metadata is ordinarily produced with the native file. When files are converted to non-native formats, relevant application metadata must be extracted and produced separately to avoid loss.

Appendix A: Just Ones and Zeros



The illustration above shows a single ASCII-encoded sector holding the text below and notated as binary data (excerpted from *David Copperfield* by Charles Dickens):

I was born with a caul, which was advertised for sale, in the newspapers, at the low price of fifteen guineas. Whether sea-going people were short of money about that time, or were short of faith and preferred cork jackets, I don't know; all I know is, that there was but one solitary bidding, and that was from an attorney connected with the bill-broking business, who offered two pounds in cash, and the balance in sherry, but declined to be guaranteed from drowning on any higher bargain. Consequently the advertisement was withdrawn at a dead loss--for as to sherry, my poor dear mother's own sherry was in the market then--and ten years afterwards, the caul was put up in a raffle down in our part of the country, to fifty members at half-a-crown a head, the winner to spend five shillings. I was present myself, and I remember to have felt quite uncomfortable and confused, at a part of myself being disposed of in that way. The caul was won, I recollect, by an old lady with a hand-basket, who, very reluctantly, pr *[end of sector]*



Exercise 2: Metadata and Hashing

GOALS: The goals of this exercise are for the student to:

1. Use cryptographic hashing to match duplicates and flag altered files.

2. Assess which actions alter file metadata and hash values

I've furnished an answer sheet at pp. 91, if that's easier for you.

Step 1: Download the compressed file http://www.craigball.com/MD5.zip and **extract** [46] the four files it contains to a convenient location on your computer:
MD5-1.xlsx
MD5-2.xlsx
MD5-3.xlsx
MD5-4.xlsx

DO NOT OPEN THESE FILES IN EXCEL before completing the exercise or you risk altering the evidence!

Step 2: Collecting System Metadata *Without opening the files*, examine the **properties** of each file and record the values in the table below.

FILE NAME	LAST MODIFIED DATE AND TIME	FILE SIZE IN BYTES
MD5-1.xlsx		
MD5-2.xlsx		
MD5-3.xlsx		
MD5-4.xlsx		

Step 3: Apart from having different file names, are they otherwise "identical" in terms of their system metadata values (date and time values and file size)? (yes or no): _____

Step 4: Find an online or local (standalone) MD5 hashing (AKA "checksum") tool of your choice and acquaint yourself with its interface. There are dozens of free, online hash calculators that can be found by searching Google for "online MD5 hash calculator." Examples:

http://hash.urih.com/
https://defuse.ca/checksums.htm
http://www.fileformat.info/tool/md5sum.htm
http://hash.online-convert.com/md5-generator

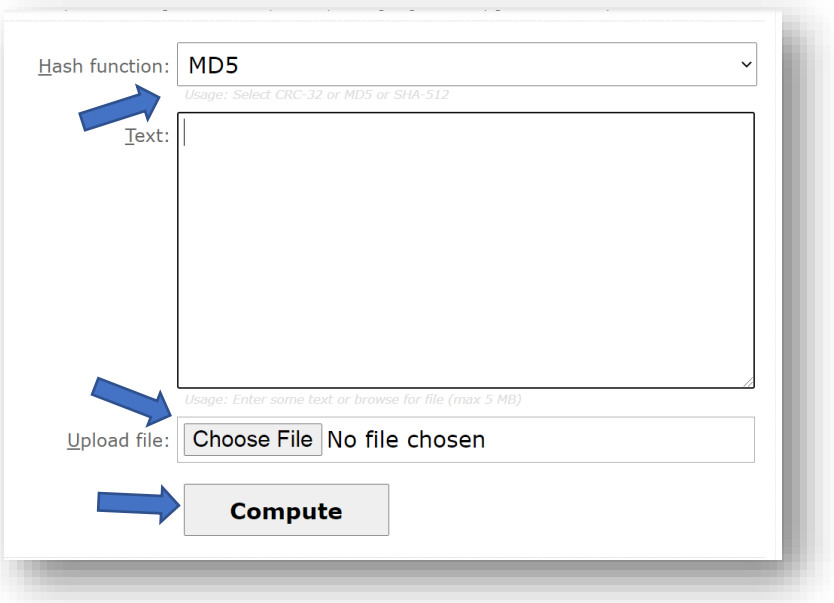
[46] I'm assuming all students have previously extracted and saved the contents of a Zip-compressed container. If you haven't, no worries. All Mac and Windows operating systems have the capability to view and extract the contents of Zip files or you may prefer to download and install a free Zip utility like 7-Zip for Windows at https://www.7-zip.org or the Unarchiver for Mac at https://theunarchiver.com/

https://www.pelock.com/products/hash-calculator

If you use an online hash calculator, be sure to use one that will allow you to browse your machine for a file to hash (all the listed sites do). Should you elect to use a hash calculator that you install as a local application, know that there is no need to purchase software for this purpose as there are many freeware options out there.

(MD5 is used here for instructional purposes. Although cryptographically broken for security applications, it remains sufficient to demonstrate hashing concepts such as content integrity and deduplication).

Below right is the interface for the online hash calculator found at https://hash.urih.com. Note that all that's required is to ensure that MD5 has been selected as the hash function, then click **Choose File**, navigate to and select the file you wish to hash, then click **Compute** and the 32-digit MD5 hash value of the file you selected should appear.



The online MD5 checksum calculators are the easiest method to use in this Exercise and support both Mac and Windows systems; nonetheless, if you're a Mac user and prefer to calculate MD5 hash values offline, you can follow these steps:

1. Open Terminal.
2. Type md5 and hit the SPACE button.
3. Drag the file you have downloaded into the Terminal Window. ...
4. Hit ENTER. You will see the MD5 Checksum

Step 5: Computing Hash Values Hash the named files and record the MD5 hash values for each in the table below.

FILE NAME	MD5 HASH VALUE (first eight characters only will suffice)
MD5-1.xlsx	
MD5-2.xlsx	
MD5-3.xlsx	
MD5-4.xlsx	

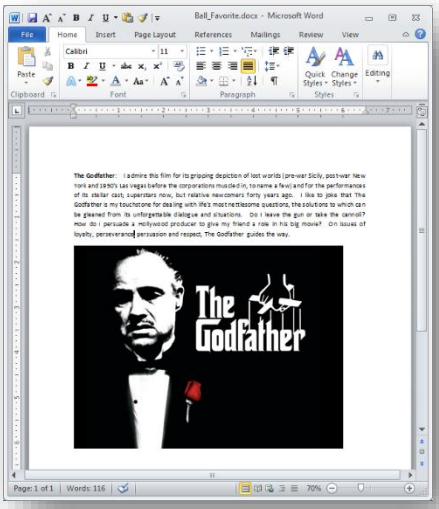
Step 6: Based on their hash values, are any of the four files *truly* identical in terms of their content? (Which ones, if any?): _____

Step 7: Create an Evidence File

Using the word processing application of your choice, create a document that identifies by title one of your favorite books or films, followed by one sentence saying why you like it. What you choose to write is of no consequence; you're merely creating a unique file to serve as evidence for the exercise. Feel free to embellish as you wish, *e.g.,* adding a still image from the film or of the cover of the book (as from Amazon.com) to paste into the document; but **you can stick with plain text if you prefer.** *Five minutes on this step, max!*

Name the document with your surname, an underscore and then "Favorite" (*i.e.,* YOURNAME_Favorite.docx). **Save** the document on your computer (not in the Cloud) and **close** your word processor. **This saved file is your original "evidence file" for purposes of this exercise**.

Step 8: Gather Baseline Metadata Values for the Evidence



To begin, establish the "true" or "baseline" system metadata for your original evidence file.

In Windows OS: Using Windows Explorer, determine the following metadata values for your original evidence file:

Filename: _____

Created Date and Time: _____

Modified Date and Time: _____

File size and size on disk: _____ _____

OR

In Mac OS: Use Get Info to determine the following metadata values for your original evidence file:

Filename: _____

Created Date and Time: _____

Modified Date and Time: _____

File size: _____

Record these values above.

Step 9: Establish Baseline Hash Values

Now, you need to establish the baseline hash value of your original evidence file.

Using the local or online hashing tool of your choice, determine the MD-5 hash value for your original evidence file.

Record the value you get for your evidence file here for ready reference.

MD5 HASH	

Step 10: Let's Identify Actions that Do and Don't Alter Metadata and Hash Values

Instructions: After completing each task below, determine the metadata and hash value for the resulting file and record them in the spaces provided (first eight characters of the hash will suffice):

Step 10a. E-mail a copy of your original evidence file to yourself as an attachment. ***Be sure to send the file to yourself using an alternate e-mail address than the transmitting account***.[47] When received, save the e-mailed attachment (*just* the attachment, not the whole message) from your e-mail client to disk under a different name (so you don't overwrite your original evidence file)[48] and record its metadata and hash value below:

Created Date and Time: (mailed) _____

Modified Date and Time (mailed): _____

File size (mailed): _____

MD5 HASH (mailed)	

Step 10b. Copy (not Move) your *original* evidence file to another storage device than your computer (such as a thumb drive, external hard drive or your online storage within Canvas). Determine the metadata and hash values of the copy:

Created Date and Time (copied): _____

[47] The reason I ask you to use an alternate address is because some e-mail systems won't transmit an attachment across the Internet when the sender and addressee are the same. So, *e.g.,* send from your personal to school account.

[48] In Windows, when you save files of the same name to the same folder, the operating system adds an incrementing number to the filename; e.g., YOURNAME_Favorite(1).doc. In Mac, the OS may add the word "copy" to the name. Let's just use a slightly different name.

Modified Date and Time (copied): _____

File size (copied): _____

MD5 HASH (copied)	

Step 10c. Rename your **<u>original</u>** evidence file using the file system,[49] *but make no other change to the file.* Rename it to something like "YOURNAME_Favorite_2.docx." Determine the metadata and hash values of the <u>renamed</u> document:

Created Date and Time (renamed): _____

Modified Date and Time (renamed): _____

File size and size on disk (renamed): _____

MD5 HASH (renamed)	

Step 10d. Edit your original evidence file *using the application you used to create it* and add a single space somewhere in the document. Save the modified file by a different name (e.g., YOURNAME_Favorite_3.docx). Determine the metadata and hash values of the edited document:

Created Date and Time (altered): _____

Modified Date and Time (altered): _____

File size and size on disk (altered): _____

MD5 HASH (edited)	

WHAT TO TURN IN: In Canvas, SUBMIT your answers to all subparts of Steps 2-10a-d, above AND submit the modified file you saved in Step 10d,

What does this exercise seek to demonstrate?

1. It supplies hands on experience in using a hash algorithm to digitally fingerprint files for reliable comparison, identification and deduplication. Imagine being tasked to compare two printed 100-page documents to determine if they are identical in every respect. You'd

[49] To rename a file using the file system, DO NOT open the file in the application you used to create it. *Doing so will likely alter the hash value.* Instead, in Windows OS, right click on the file and select "rename." In MacOS, change the file's name in the "Name and Extension" field of the Get Info screen.

need to compare them word-for word and line-by-line (or use a tool that makes the comparison). Note how use of hashing permits ready comparison of just 32-character values no matter the length or complexity of the source document.

2. It underscores the essential difference between system metadata and application metadata. Applications embed application metadata inside the file, so application metadata are part of the calculation when the file is hashed. Accordingly, **a change in application metadata alters the hash value of the file**, impacting the ability to deNIST and deduplicate the file. By contrast, system metadata values are stored outside the file and so are not hashed along with the file. Accordingly, **system metadata can be changed without altering the hash value of the file**. If you can change the name of a file without changing the file's hash value, that tells us something about the location of a file's name record. That explains why we can Bates number files without changing their hash values used for duplicate identification.

3. The interplay between system and application metadata confuses many students, a circumstance exacerbated by the tendency of software to display a mix of application and system metadata values on a file's properties screen. Some properties MUST move with the file to be of use (e.g., Last Printed Dates in Word documents). Others must *not* move with the file because they would be misleading in other environments (*e.g.*, a physical storage address on disk).

4. Metadata is dynamic and may be lost or changed inadvertently when files are copied or transmitted. How we handle files in e-discovery (our "chain of custody") impacts our ability to authenticate those files as evidence and may prompt charges of spoliation.

Answer Sheet for Exercise 2

Step 2: Collecting System Metadata

FILE NAME	LAST MODIFIED DATE AND TIME	FILE SIZE IN BYTES
MD5-1.xlsx		
MD5-2.xlsx		
MD5-3.xlsx		
MD5-4.xlsx		

Step 3: Apart from having different file names, are they otherwise "identical" in terms of their system metadata values (date and time values and file size)? (yes or no): _____

Step 5: Computing Hash Values

FILE NAME	MD5 HASH VALUE (first eight characters only will suffice for the last three)
MD5-1.xlsx	
MD5-2.xlsx	
MD5-3.xlsx	
MD5-4.xlsx	

Step 6: Based on their hash values, are any of the four files truly identical in terms of their content? (Which ones, if any?): _____

Step 8: Gather Baseline Metadata Values for the Evidence

Filename: _____

Created Date and Time: _____

Modified Date and Time: _____

File size and size on disk: _____ _____

Step 9: Establish Baseline Hash Values

MD5 HASH	

Step 10a. E-mail

Created Date and Time: (mailed) _____

Modified Date and Time (mailed): _____

File size (mailed): _____

MD5 HASH (mailed)	

Step 10b. Copy

Created Date and Time (copied): _____

Modified Date and Time (copied): _____

File size (copied): _____

MD5 HASH (copied)	

Step 10c. Rename Determine the metadata and hash values of the renamed document:

Created Date and Time (renamed): _____

Modified Date and Time (renamed): _____

File size and size on disk (renamed): _____

MD5 HASH (renamed)	

Step 10d. Edit Determine the metadata and hash values of the edited document:

Created Date and Time (altered): _____

Modified Date and Time (altered): _____

File size and size on disk (altered): _____

MD5 HASH (edited)	



Exercise 3: Exploring DISCO, an E-Discovery Platform Tool

Time Required: ≈ 45 minutes

GOALS FOR THIS EXERCISE

The goals of this exercise are for the student to:

- Successfully log in and validate DISCO reviewer credentials
- Navigate the core components of a commercial e-discovery review platform
- Understand how stages of the EDRM are implemented in practice
- Perform basic search, filtering, and review functions using a real-world email dataset

Background: The Dataset

You will be working with the **John Podesta Email Collection**, a publicly available corpus of approximately **50,000 emails and attachments** (approximately **5 GB in native form**) published by WikiLeaks in 2016. This collection has already been processed and loaded into DISCO for your use.

John Podesta served as campaign chairman for Hillary Clinton's 2016 presidential campaign and previously as Counselor to President Obama and Chief of Staff to President Bill Clinton. The emails span political, policy, and administrative topics.

We use the Podesta Email Collection because there are few publicly available, contemporary corpora of emails and attachments available in native or near-native form. The collection is large without being unwieldy and concerns issues, events, and individuals with which many Americans have at least passing familiarity. [50]

You will access and explore the collection using the **DISCO online review platform**, focusing on how documents, metadata, search results, and relationships are presented to reviewers.

Accessing the Dataset

The processed and indexed collection is reached by navigating to:

https://login.csdisco.com/Account/Login

[50] By contrast, the massive Enron e-mail collection widely employed in e-discovery is decades old, concerns persons and events unfamiliar to most students and has been filtered, culled and sterilized so as to be almost unrecognizable as a useful native corpus.

Log in using the username and password you established when you accepted the DISCO invitation sent to your **utexas.edu** email address.

After logging in, locate the **Matter** titled:

UT Spring 2026 (Podesta Email is Here) [51]

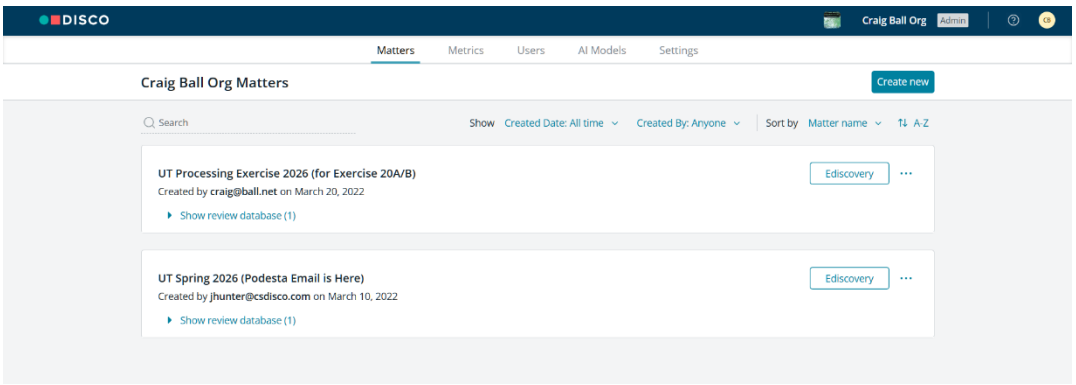
(A *Matter* is DISCO's term for a case or project.)

Step-by-step instructions follow.

Part 1: Access and Orientation (≈ 5 minutes)

Task 1.1: Log In and Validate Credentials

1. Navigate to the DISCO login page.
2. Log in using your assigned credentials.
3. Confirm that you can see your list of Matters.
4. **Deliverable 1.1:** Take a screenshot of your DISCO Matters screen (Mine looks like the image below. Yours may look a bit different).



Task 1.2: Orient Yourself to the Review Screen

Open the Matter **UT Spring 2026 (Podesta Email is Here)** by clicking the button alongside labeled **Ediscovery** to access the database named **Podesta Email**.

You should now see the **Review screen**, which consists of:

- A **search bar** across the top of the screen
- A **Filters panel** on the left

[51] Please do not use the matter called UT Processing Exercise 2026. That's for a different exercise.

• A **Review panel** containing a document list showing individual emails and attachments
• A document count is displayed at the Review panel, alongside the word “Review.”

The **Filters panel**, located to the left of the Review screen, allows you to refine and explore the document set using metadata, system-generated attributes, and document relationships. The Filters panel is scrollable and organized into expandable categories, including:

• Tags
• Annotations and Redactions
• Topics
• Fact Timeline
• Primary Languages
• Document Relationships, including family and attachment relationships
• Hidden & Supplemental Content
• Dates
• Document Types
• File Properties
• Chat Fields
• Email Fields
• Custodians
• Ingest Details

Filters operate in conjunction with the search bar and update dynamically as searches are run or refined. You may apply filters before running a search or use them to narrow an existing set of results.

Deliverable 1.2: Write down the total number of documents in the collection as shown at the top of the Review screen.

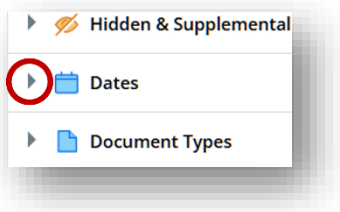
Part 2: Understanding the Dataset (≈ 15 minutes)

Task 2.1: Examine Dataset Characteristics

Using the information available through Filters:

Record the following values:

Deliverable 2.1.1: Approximate date range of the emails (expand the "Dates" filter by clicking on the arrow just left of the blue calendar icon) →



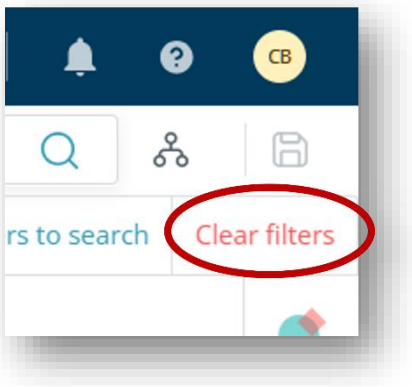
Deliverable 2.1.2: The three most frequent email recipients after John Podesta himself (expand "Email Fields" in Filters, then expand "Email To recipients").

Deliverable 2.1.3: Once more using the "Email To recipients" filter, what is the email count for recipient **jake.sullivan@gmail.com**?

Remember to write down and submit the values recorded for tasks 2.1.1 through 2.1.3.

Task 2.2: Explore Document Types and Families
Using the **Document Relationships → Families & Attachments** filter:
Identify and **record**:
Deliverable 2.2.1: How many documents are attachments
Deliverable 2.2.2: How many attachments are PDFs (here you will activate two filters simultaneously: Attachments and PDF (in Document Types)

IMPORTANT: CLEAR YOUR FILTERS → Note the link in the upper right corner of your screen. *Always remember to clear your filters* when you don't want them to impact on the outcome of your search or review!



Part 3: Lexical Search and Review Basics (≈ 20 minutes)

Task 3.1 Perform Keyword Searches

Use the Search bar to search for the term:

classified

Deliverable 3.1: Record the number of results returned for the term "classified."

Task 3.2: Refine Results Using Filters

Using the Filters panel, limit your "classified" search results to emails during **2015**.

Deliverable 3.2: Record the revised number of results.

Task 3.3: Review an Individual Email

Open the first email from your filtered search results. It should be Doc ID 53 from Gary Gensler. Simply click on the item to open it in the viewer.

Examine the metadata displayed for the email, including:

- From / To / CC
- Date and time sent
- Page count

Deliverable 3.3: Take a **screenshot** of the email with its metadata panel (right side of screen) visible and expanded.

IMPORTANT: Remember to **clear your filters** before tasks where those filters are not required.

Task 3.4: Practice Boolean Search

Conduct each of these searches and record the number of hits:

Deliverable 3.4.1: **"Benghazi"** (simple keyword)
Deliverable 3.4.2: **"Benghazi" AND "Libya"** (both terms must appear)
Deliverable 3.4.3: **"Benghazi" OR "Libya"** (either term appears)
Deliverable 3.4.4: **"Libya" NOT "Benghazi"** (first term but not second)

Submission Requirements for Deliverables

Using Canvas, submit a document containing (a template is supplied in Assignments):

1. **Screenshots** (2 total):
    - Deliverable 1.1: DISCO Matters screen after login
    - Deliverable 3.3: Email Doc ID 52 with metadata panel visible
2. **Data Collection** (12 total)
    - Deliverable 1.2: total number of documents in the collection
    - Deliverable 2.1.1: date range of the emails in the collection
    - Deliverable 2.1.2: three most frequent email recipients after John Podesta
    - Deliverable 2.1.3: email count for recipient jake.sullivan@gmail.com.
    - Deliverable 2.2.1: number of attachments in collection
    - Deliverable 2.2.2: number of attachments that are PDFs
    - Deliverable 3.1: number of items containing the term classified
    - Deliverable 3.2: number of 2015 items containing the term classified

- Deliverable 3.4.1: number of items hit by "Benghazi"
- Deliverable 3.4.2: number of items hit by "Benghazi" AND "Libya"
- Deliverable 3.4.3: number of items hit by "Benghazi" OR "Libya"
- Deliverable 3.4.4: number of items hit by "Libya" NOT "Benghazi"

Bonus Challenge (Optional!!!)

Search for emails discussing **"polling data"** in **swing states** (Hint: Do you find swing states by searching for “swing states?”).

- How many emails discuss these topics?

Learning Outcomes Check

After completing this exercise, you should be able to:

✓ Successfully access and navigate a commercial e-discovery platform
✓ Understand how digital evidence is organized and made searchable
✓ Perform basic keyword and Boolean searches
✓ Recognize metadata associated with email communications



Exercise 4: Metadata: Geolocation in EXIF

GOALS: The goals of this exercise are for the student to:

1. Grasp the fundamentals of global positioning
2. Explore the range and volume of metadata in and for everyday ESI; and
3. Identify and use geolocation information in EXIF data to track stolen funds.

OUTLINE: Students will examine various image files to extract embedded EXIF metadata and assess its value as evidence, then use that information and online research to follow the ill-gotten gains.

Background

Some cameras and virtually all modern mobile phones sold in the United States are GPS-enabled. For mobile phones, the latter capability is legally mandated to support 911 emergency services. In fact, modern cell phones facilitate geolocation in at least three distinct ways, by:

1. Communicating with nearby cell towers (cell-site location information)
2. Using internal GPS/GNSS receivers; and
3. Interacting with Wi-Fi access points and Bluetooth beacons.

Consider how many issues in civil and criminal litigation might be resolved by the availability of detailed and reliable geolocation evidence? But right now, how reliable *is* such data?

Global Positioning System (GPS)

Few who use GPS have more than a vague idea that GPS-enabled devices establish their location by listening to a network of satellites. In fact, your phone does not transmit to satellites; it listens for signals from a minimum of four orbiting transmitters equipped with atomic clocks, part of a constellation of approximately 31 operational U.S. GPS satellites that collectively provide global coverage.

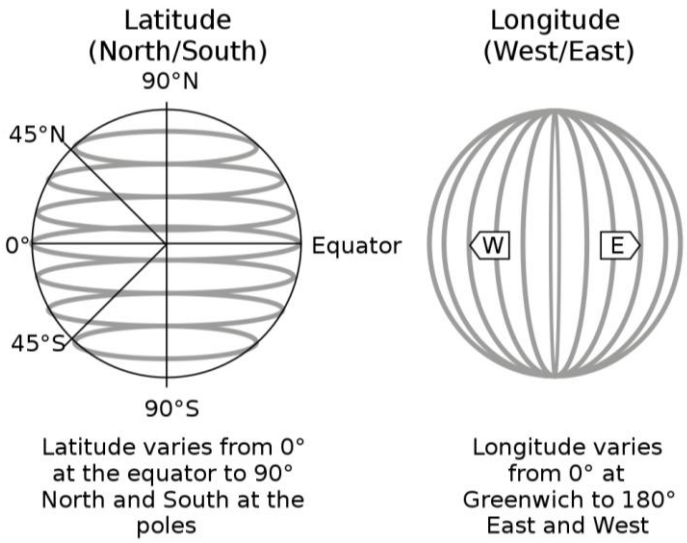
Each satellite continuously broadcasts a signal containing the precise time the signal was sent. GPS receivers compute their location by measuring the time difference between when the signal was transmitted and when it was received. Because the signal travels at a constant speed (~the speed of light) and the satellite positions are known, these time differences can be converted into distances. Using distances from four or more satellites, the receiver calculates a three-dimensional fix reflecting latitude, longitude, and elevation.

Phones commonly use **Assisted GPS (A-GPS)** to speed acquisition of a location fix. Rather than waiting to receive an almanac of satellite positions directly—which can take minutes—A-GPS allows devices to obtain that information quickly via Wi-Fi or cellular data.

Geographic Coordinate System: Latitude and Longitude

Global Positioning tells you *where you are*, but to be useful as evidence, geolocation data must be expressed in a standardized geographic coordinate system that ties where you are in relation to where everything else is, too. GPS systems employ the **World Geodetic System (WGS)**, most commonly **WGS 84**, which remains the prevailing reference system for GPS coordinates.

The WGS divides the planet vertically into **meridians of longitude** running from pole to pole, separating the Eastern and Western Hemispheres at the **Prime Meridian**, which passes through the Royal Observatory in Greenwich, England. Each hemisphere spans 180 degrees of longitude.



The Earth further divides into horizontal **parallels of latitude**—concentric circles beginning at the Equator and extending 90 degrees north and south to the poles. Latitude is the angular distance between a point on the Earth's surface and the Equator, measured either in degrees, minutes, and seconds (sexagesimal—Base 60—format) or as decimal degrees.

Together, the parallels of latitude and meridians of longitude form a cage-like grid called a **graticule**.



REMEMBER: The divisions from pole-to-pole are **meridians of longitude;** the concentric circumferential divisions are **parallels of latitude**. Keep them straight by noting that the Prime **Meridian** is a *north-south* demarcation and that parallel lines never touch, so **parallels** of latitude are belts *around* the Earth like the zero-degree parallel of latitude called" the Equator."

The Mission: *You've been hired to track down embezzled assets. The culprit is a globetrotting gastronome named Bernie who has stashed stolen cash and diamonds all over the world. Bernie is skilled at covering his tracks, destroying his phone and laptop before his arrest.*
But Bernie forgot something.



Each time he opened a new bank account or safety deposit box, Bernie treated himself to a celebratory meal and texted an iPhone snapshot of that meal to his wife in Palm Beach. Your client obtained copies of ***eleven*** *of these cellphone photos through discovery, all in their native* ***.jpg*** *format.*

Your first task is to determine ***where*** *in the world the cash and diamonds might be—and* ***when*** *Bernie visited those locations.*

Step 1: EXIF Geolocation Data

Download the Zip file at www.craigball.com/exercise4.zip and extract its contents to a folder on your desktop or any other convenient location. Locate eight numbered files called **YUM_1.jpg** through **YUM_8.jpg** and two other photos named **key.jpg** and **huh.jpg**.

Find the file called **YUM_1.jpg** and explore its properties:

In Windows:

Right-click the file and select **Properties → Details**. Note the **Date Taken** (under "Origin"). This reflects the date and time the photo was taken, though it may be adjusted to your computer's local time zone and daylight-saving settings. Also note the GPS values for **Latitude**, **Longitude**, and **Altitude**.

In MacOS: Open the file in **Preview**, select **Tools → Show Inspector**, then click the **More Info** tab (the "i" in a circle), followed by the **GPS** tab. Note the **Date Stamp** and **Time Stamp**, as well as the GPS values for **Latitude**, **Longitude**, and **Altitude**. Your Mac may also display a map showing the photo's location.

Step 2: Record the GPS Coordinates and Date Taken

For **YUM_1.jpg**, locate the GPS Latitude and Longitude values embedded in the EXIF data and the date the photo was taken.

You should see values substantially similar to the following:

Latitude: 25° 8' 29.13" N
Longitude: 55° 11' 6.23" E
Date Taken: December 30, 2019

Step 3: Where's Waldo, I mean Bernie?

Now let's determine *exactly* where this photo was taken. In Google, carefully enter the latitude and longitude values embedded within the photo as described below:

If the values you found were formatted as: Latitude AA; BB; CC.CCCCCC and Longitude XX; YY; ZZ.ZZZZZ, enter them in the Google search box in the following format:

AA BB**'** CC.CCCCCC**,** XX YY**'** ZZ.ZZZZZ

So, for yum1.jpg: the Google search is**: 25 8' 29.130000, 55 11' 6.239999** (Windows) or, using the Mac data: **25° 8' 29.13 N, 55° 11' 6.23 E.** Either way, all roads lead to Rome. I mean, Dubai.

That's right, if the map you retrieved points to the spectacular Burj al Arab hotel, you've done it!. If not, check the formatting of the coordinates carefully and try again.

Be sure you include the apostrophe after the second longitude and latitude values and the comma separating the values. By way of example, the coordinates for The University of Texas School of Law in a photo might appear as **Latitude 30° 17' 18.696" N, Longitude 97° 43' 49.846" W**. In Google, they must be entered as: **30 17' 18.696, -97 43' 49.846**.

Step 4: A Quicker Way
Since we must get through several photos, let's speed things up.

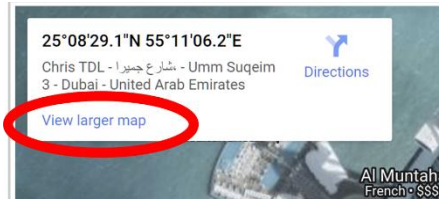
On macOS:
Click the **Locate** button at the bottom of the GPS pane to launch the location directly in your browser.

On Windows:
Open the photo in the default Photos app, click the **three dots** in the upper-right corner, and select **File info**. Location data (if present) will appear in the information panel.

If neither method works, you may use an online EXIF viewer, such as:

Jeffrey's EXIF Viewer – http://exif.regex.info
or FotoForensics – http://fotoforensics.com/

In FotoForensics, upload each image, then select **Metadata** from the Analysis menu. Scrolling down reveals GPS data and a map. You may click **View Larger Map** to open the location in Google Maps.



Determine **where and when** Bernie took each photo, and complete the following table:

Photo	**Location Taken**	**Date photo wastaken**
Yum_1.jpg	Dubai, United Arab Emirates	December 30, 2019
Yum_2.jpg		
Yum_3.jpg		
Yum_4.jpg		
Yum_5.jpg		
Yum_6.jpg		
Yum_7.jpg		
Yum_8.jpg		

Step 5: Follow the Money

You've learned that, traveling by private jet and using a foreign passport, Bernie ferried funds and diamonds to Europe. Based on large transfers, it appears Bernie made these trips during the ***summer months of 2018, 2019, and 2022****.*



Along with the food photos, you found this image: **key.jpg**.

You suspect Bernie hid money and diamonds in safety-deposit boxes at European branches of **Deutsche Bank**. With approximately **1,200 Deutsche Bank branches and offices worldwide**, where do you begin?

Used your list of **European cities** Bernie visited during **June, July, or August** of 2018, 2019, and 2022, locate the street address of the nearest Deutsche Bank outpost to the restaurant in each city.

City	**Closest Deutsche Bank to Restaurant**

Step 6: How Trustworthy is EXIF data?

One more.

Open **huh.jpg from** where you stored it.



Ah, oui—the Eiffel Tower. The Bateau-Mouche. The Seine. One can almost hear Edith Piaf singing, non? Can this be anywhere but the City of Lights and *amour*, Paris?

Check the GPS data and map it.

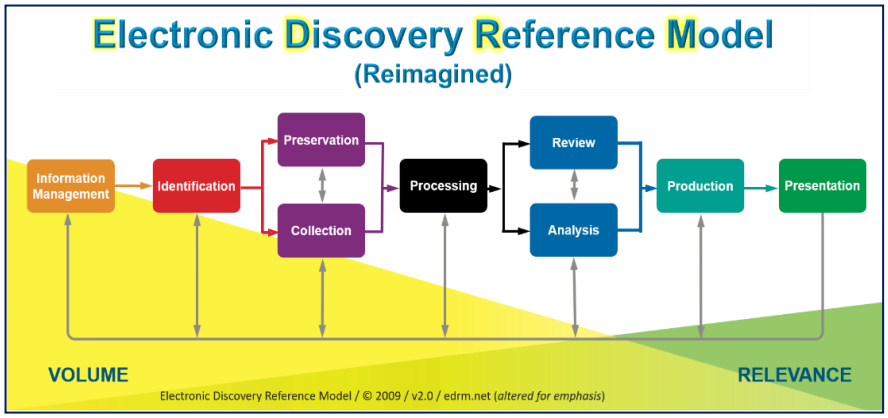
According to the EXIF geolocation data, where was this photo taken?

_____

_____

Processing in E-Discovery

Processing is the "black box" between preservation/collection and review/analysis. Though the iconic Electronic Discovery Reference Model (EDRM) positions Processing, Review and Analysis as parallel paths to Production, processing is an essential prerequisite— "the only road"—to Review, Analysis and Production.[52] Any way you approach e-discovery at scale, you must process ESI before you can review or analyze it with only rare and highly constrained exceptions that do not meaningfully alter this reality. If we recast the EDRM to reflect processing's centrality, the famed schematic would look like this:



VITAL VOCABULARY

ASCII
Base64
Binary
Bits
Bytes
Character Encoding
Code Page
Composite Character
Content-Transfer-Encoding
Content-Type
Decoding
Diacriticals
File Header
Hexadecimal
MIME E-mail
Mojibake
Multibyte Encoding
Normalization
Precomposed Character
Radix
Steganography
Tokenization
Unicode
Unitization

There are hundreds—perhaps thousands—of articles delving into the stages of e-discovery that flank processing in the EDRM. These are the stages where lawyers have had a job to do. But lawyers tend to cede processing decisions to technicians. When it comes to processing, lawyer competency and ken is often little more than "stuff goes in, stuff comes out," a dangerous abdication given how many downstream decisions are locked in at this stage.

Why process ESI in e-discovery? Isn't it "review ready?"

We process information in e-discovery to catalog and index contents for search and review. Unlike Google, e-discovery is the search for *all* responsive information in a collection, not just one information item deemed responsive. Though all electronically stored information is inherently electronically searchable, computers don't structure or search all ESI in the same way; so, we must process ESI to ***normalize*** it to achieve uniformity for indexing and search.

[52] That's not a flaw. The EDRM is a conceptual view, not a workflow.

Thus, "processing" in e-discovery might fairly be called normalized access: the set of steps by which heterogeneous files—created by different applications, operating systems, languages and users—are reduced to a common, searchable form. Processing extracts text and metadata, resolves character encodings, expands containers, identifies file types and organizes content so that search and review tools can treat unlike things alike.

ESI processing tools perform five common functions.[53] They must:

1) Decompress, unpack and fully explore ingested items, including recursively opening embedded containers (e.g., ZIPs inside ZIPs)
2) **Identify** and apply templates (filters) to **encoded** data to **parse** (interpret) contents and **extract text, embedded objects**, and **metadata**.
3) **Track** and **hash** items processed, **enumerating** and **unitizing** all items and tracking failures.
4) **Normalize** and **tokenize** text and data and create an **index** and **database** of extracted information.
5) **Cull** data by file type, date, lexical content, hash value and other criteria.

Files

If we polled lawyers asking what to call the electronic items preserved, collected and processed in discovery, most would answer, "documents." Others might opt for "data" or reflect on the initialization "ESI" and say, "information." None are wrong answers, but the ideal response would be the rarest: "*files*." Electronic documents are *files*. Electronically stored information resides in *files*. *Everything* we deal with *digitally* in electronic discovery comes from or goes to physical or logical data storage units called "data files" or just "files." Moreover, all programs run against data files are themselves files comprising instructions for tasks. These are "executable files" or simply "executables."

So, what is it we process in the processing stage of e-discovery? The answer is, "**we process *files***." Let's look at these all-important files and explore what's in them and how they work.

A Bit About and a Byte Out of Files

A colleague once defended her ignorance of the technical fundamentals of electronically stored information by analogizing that "she didn't need to know how planes stay aloft to fly on one." She had a point, but only for passengers. If you aspire to be a pilot or a rocket scientist—if you want to be at the controls or design the plane—you must understand the fundamentals of flight. If you

[53] While a processing tool may do considerably more than the listed functions, a tool that does less is unlikely to meet litigants' needs in e-discovery.

aspire to understand processing of ESI in e-discovery and manage e-discovery, you must understand the fundamentals of electronically stored information, including such topics as:

- *What's* stored electronically?
- *How* is it stored?
- *What forms* does it take?

The next few pages are a crash course in ESI storage, particularly the basics of encoding and recording textual information. If you can tough it out, you'll be undaunted by discussions of "Hex Magic Numbers" and "Unicode Normalization" yet to come.

Digital Encoding

All digital evidence is encoded, and how it's encoded bears upon how it's collected and processed, whether it can be searched and in what reasonably usable forms it can be produced. Understanding that electronically stored information is numerically encoded data helps us see the interplay and interchangeability between different forms of digital evidence. Saying "*it's all ones and zeroes*" means nothing if you don't grasp *how* those ones and zeros underpin the evidence.

Processing vs. Indexing (Not the Same Thing) Modern SaaS e-discovery platforms often blur the distinction between *processing* and *indexing*, encouraging the mistaken belief that the moment data is uploaded, it is "searchable." Processing is the upstream work of unpacking files, resolving character encodings, extracting text and metadata, normalizing dates and time zones, and identifying what each file actually is. Indexing comes later: it is the creation of data structures that allow processed text and metadata to be searched quickly and efficiently. Indexing cannot fix processing failures. If text was never correctly extracted, if characters were mis-encoded, or if metadata was misinterpreted during processing, the index will faithfully—and invisibly—preserve those errors. Searches that return nothing may be functioning perfectly; it is the processing that failed them.

Electronic evidence is just data, and data are just numbers; so, understanding the numbers helps us better understand electronic evidence. Those numbers only acquire meaning through the interpretive rules imposed by software, standards and human convention.

Decimal and Binary: Base 10 and Base Two

Understanding encoding requires we hearken back to those hazy days when we learned to tally and count by numbers. Long ago, we understood quantities (numeric *values*) without knowing the **numerals** we would later use to symbolize quantities. When we were three or four, "five" wasn't yet Arabic **5**, Roman **V** or even a symbolic tally like 𝍸.

More likely, five was this:



If you're from the Americas, Europe or Down Under, I'll wager you were taught to count using the **decimal system**, a **positional notation** system with a **base** of 10. Base 10 is so deeply ingrained in our psyches that it's hard to conceive of numeric values being written any other way. Decimal just *feels* like one, "true" way to count, but it's not. Writing numbers using an alternate base or "**radix**" is just as genuine, and it's advantageous when information is stored or transmitted digitally.

Think about it. Human beings count by tens because we evolved with ten digits on our hands. Were that not so, tasteless old jokes like this one would make no sense: "*Did you hear about the Aggie who was arrested for indecent exposure? He had to count to eleven.*"

Had our species evolved with eight fingers or twelve, we would have come to rely upon an octal or duodecimal counting system, and we would regard those systems as the "true" positional notation system for numeric values. Ten only feels natural because we built everything around ten. Again, if we'd evolved with eight or twelve fingers, it really wouldn't matter because you can express any number—and consequently any data—in any number system. So, it happens that computers use the base two or binary system, and computer programmers are partial to base sixteen or hexadecimal. Data is just numbers, and it's all just counting.

When we were children starting to count, we had to *learn* the decimal system. We had to *think* about what numbers *meant*. When our first-grade selves tackled big numbers like 9,465, we were overtly aware that each digit represented a decimal multiple. The nine was in the thousands place, the four in the hundreds, the six in the tens place and the five in the ones. We might even have parsed 9,465 as: *(9 x 1000) + (4 x 100) + (6 x 10) + (5 x 1)*.

But soon, it became second nature to us. We'd unconsciously process 9,465 as nine thousand four hundred sixty-five. As we matured, we learned about powers of ten and now saw 9,465 as: $(9 \times 10^3) + (4 \times 10^2) + (6 \times 10^1) + (5 \times 10^0)$. This was exponential or "base ten" notation. We flushed it from our adolescent brains as fast as life (and the SAT) allowed.

Computers don't have fingers; instead, computers count using a slew of electronic switches that can be "on" or "off." Having just two states (on/off) makes it natural to count using Base 2, a binary counting system. By convention, computer scientists *notate* the status of the switches using the numerals one and zero. So, we tend to say that computers store information as ones and zeroes—a useful shorthand, though not literally how storage media record data.

Computer storage devices like IBM cards, hard drives, tape, thumb drives and optical media store information as physical phenomena that can be reliably distinguished in either of two distinct states, *e.g.,* punched holes, changes in magnetic polar orientation, minute electric potentials or deflection of laser beams. We *symbolize* these two states as one or zero, but you could represent the status of binary data by, say, turning a light on or off. Early computing systems did just that, hence all those flashing lights.



You can express any numeric value in any base without changing its value, just as it doesn't change the numeric value of "five" to express it as Arabic "5" or Roman "V" or just by holding up five fingers.

In positional notation systems, the order of numerals determines their contribution to the value of the number; that is, their contribution is the value of the digit multiplied by a factor determined by the position of the digit and the base.

The base/radix describes the number of unique digits, starting from zero, that a positional numeral system uses to represent numbers. So, there are just two digits in base 2 (binary), ten in base 10 (decimal) and sixteen in base 16 (hexadecimal). E-mail attachments are encoded using a whopping *64 digits* in base 64.

We speak the decimal number 31,415 as "thirty-one thousand, four hundred and fifteen," but were we faithfully adhering to its base 10 structure, we might say, "three ten thousands, one thousand, four hundreds, one ten and five ones. The "base" ten means that there are ten characters used in the notation (0-9) and the value of each position is ten times the value of the position to its right.

10,000	1,000	100	10	1
3	1	4	1	5

The same *decimal* number 31,415 can be written as a *binary* number this way: **111101010110111**

16,384	8,192	4,096	2,048	1,024	512	256	128	64	32	16	8	4	2	1
1	1	1	1	0	1	0	1	0	1	1	0	1	1	1

In base 2, two characters are used in the notation (0 and 1) and each position is twice the value of the position to its right. If you multiply each digit times its position value and add the products, you'll get a total equal in value to the decimal number 31,415.

A value written as five characters in base 10 requires 15 characters in base 2. That seems inefficient until you recall that computers count using on-off switches and thrive on binary numbers.

The decimal value 31,415 can be written as a base 16 or hexadecimal number this way: **7AB7**

4,096	256	16	1
7	A	B	7

In base 16, sixteen characters are used in the notation (0-9 and A-F) and each position is sixteen times the value of the position to its right. If you multiply each digit times its position value and add the products, you'll get a total equal in value to the decimal number 31,415. But how do you multiply letters like A, B, C, D, E and F? You do it by knowing the letters are used to denote values greater than 9, so A=10, B=11, C=12, D=13, E=14 and F=15. Zero through nine plus the six values represented as letters comprise the sixteen characters needed to express numeric values in hexadecimal.

Once more, if you multiply each digit/character times its position value and add the products, you'll get a total equal in value to the decimal number 31,415:

	7 x 4,096	=	28,672
The A means 10 so:	10 x 256	=	2,560
The B means 11 so:	11 x 16	=	176
Finally, we add:	7 x 1	=	7

			31,415

Bits

To recap, computers use **binary digits** in place of decimal digits. The word **bit** is even a shortening of the words "**B**inary dig**IT**." Unlike the decimal system, where we represent any number as a combination of *ten* possible digits (0-9), the binary system uses only *two* possible values: zero or

one. This is not as limiting as one might expect when you consider that a digital circuit—essentially an unfathomably complex array of switches—hasn't got any fingers to count on but is very good and very fast at being "on" or "off."

In the binary system, each binary digit—each "bit"—holds the value of a power of two. Therefore, a binary number is composed of only zeroes and ones, like this: 10101. How do you figure out what the value of the binary number 10101 is? You do it in the same way we did it above for 9,465, but you use a base of 2 instead of a base of 10. Hence: $(1 \times 2^4) + (0 \times 2^3) + (1 \times 2^2) + (0 \times 2^1) + (1 \times 2^0) = 16 + 0 + 4 + 0 + 1 =$ **21**.

Moving from right to left, each bit you encounter represents the value of increasing powers of 2, standing in for zero, two, four, eight, sixteen, thirty-two, sixty-four and so on. That makes counting in binary easy. Starting at zero and going through 21, decimal and binary equivalents look like the table at right.

DEC = BIN	DEC = BIN
0 = 00000	11 = 01011
1 = 00001	12 = 01100
2 = 00010	13 = 01101
3 = 00011	14 = 01110
4 = 00100	15 = 01111
5 = 00101	16 = 10000
6 = 00110	17 = 10001
7 = 00111	18 = 10010
8 = 01000	19 = 10011
9 = 01001	20 = 10100
10 = 01010	21 = 10101

Bytes

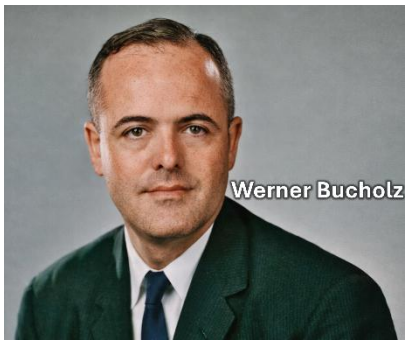
Computers also work with binary data in eight-character sequences called **bytes**. A byte is a string (sequence) of eight bits. The biggest number that can be stored as one byte of information is 11111111, equal to 255 in the decimal system. The smallest number is zero or 00000000. Thus, you can store 256 different numbers as one byte of information (0-255). So, what do you do if you need to store a number larger than 256? Simple! You use a *second* byte. This affords you all the combinations that can be achieved with 16 bits, being the product of all of the variations of the first byte and all of the second byte (256 x 256 or 65,536). So, using bytes to express values, we express any number greater than 256 using at least two bytes (called a "word" in geek speak), and any number above 65,536 requires three bytes or more.



"01101001, 00111011, 00011010, but, but!"

Why are eight-bit sequences the fundamental building blocks of computing? It just happened that way. In these times of cheap memory, expansive storage and lightning-fast processors, it's easy to forget how scarce and costly such resources were at the dawn of the computing era. Seven bits (with a leading bit reserved) was the smallest block of data that would suffice to represent the minimum complement of alphabetical characters, decimal digits, punctuation and control instructions needed by the pioneers in computer engineering. It was, in a sense, all the data early

processors could bite off at a time, perhaps explaining the name "byte" (coined in 1956 by IBM scientist Dr. Werner Buchholz).



Hexadecimal

Once more, a binary sequence of eight ones and zeros ("bits") can be arranged in 256 unique ways. Long sequences of ones and zeroes are hard for humans to follow, so happily, two hexadecimal characters can also be arranged in 256 unique ways, meaning that just two base-16 characters can replace the eight characters of a binary byte (*i.e.*, a binary value of 11111111 can be written in hex as FF). Using hexadecimal characters allows programmers to write data in just 25% of the space required to write the same data in binary, and it's easier for humans to follow.

Let's take a quick look at why this is so. A single binary byte can range from 0 to 255 (being 00000000 to 11111111). ***Computers count from zero***, so that range spans 25**6** unique values. The following table demonstrates why the largest value of an eight-character binary byte (11111111) equals the largest value of just two hexadecimal characters (FF):

Binary Byte (0-255)

128	64	32	16	8	4	2	1
1	1	1	1	1	1	1	1

The decimal value of this binary byte is the sum of one multiplied by each of the positional values in the shaded area; thus, **128+64+32+16+8+4+2+1 = 255**

Hexadecimal Byte (0-255)

16	1
F	F

The decimal value of this hexadecimal byte is the sum of 15 (recall F=15) multiplied by each of the positional values in the shaded area; thus, **(15x16) + (15x1) = 255**

Hexadecimal values are everywhere in computing. Litigation professionals encounter hexadecimal values as MD5 hash values expressed in hexadecimal notation and may run into them as IP addresses, Globally Unique Identifiers (GUIDs) and even color references.



The Magic Decoder Ring called ASCII

Back in 1935, American kids who listened to the Little Orphan Annie radio show and drank lots of Ovaltine could join the Radio Orphan Annie

Secret Society and obtain a Magic Decoder Ring, a device with rotating disks that allowed them to read and write numerically encoded messages.[54]

Similarly, computers encode words as numbers. Binary data stand in for the upper- and lower-case English alphabet, as well as punctuation marks, special characters and machine instructions (like carriage return and line feed).

Encoding Text

So far, I've described ways to encode the same numeric value in different bases. Now, let's shift gears to describe how computers use those numeric values to signify intelligible alphanumeric information like the letters of an alphabet, punctuation marks and emoji. Again, data are just numbers, and those numbers signify something *in the context of the application using that data*, just as gesturing with two fingers may signify the number two, a peace sign, the V for Victory or a request that a blackjack dealer split a pair. What numbers mean depends upon the **encoding scheme** applied to the values in the application; that is, the encoding scheme supplies the essential context needed to make the data intelligible. If the number is used to describe an RGB color, then the hex value 7F00FF means violet. Why? Because each of the three values that make up the number (7F 00 FF) denote how much of the colors red, green and blue to mix to create the desired RGB color. In other contexts, the same hex value could mean the decimal number 8,323,327, the binary string 111111100000000011111111 or the characters 缀ÿ.

ASCII

When the context is text, there are a host of standard ways, called **Character Encodings** or **Code Pages**, in which the numbers denote letters, punctuation and symbols. Now nearly sixty years old, the **American Standard Code for Information Interchange** (ASCII, "ask-key") is the basis for most modern character encoding schemes (though both Morse code and Baudot code are older). Born in an era of teletypes and 7-bit bytes, ASCII's original 128 codes included 33 non-printable codes for controlling machines (*e.g.*, carriage return, ring bell) and 95 printable characters. By limiting the ASCII character set to just 128 characters (0-127), we can express any character in just seven bits (2^7 or 128) and so occupy only one byte in the computer's storage and memory. The ASCII character set follows:

[54] A similar toy, a secret decoder *pin*, was featured in the movie, *A Christmas Story*.

ASCII	Hex	Symbol
0	00	NUL
1	01	SOH
2	02	STX
3	03	ETX
4	04	EOT
5	05	ENQ
6	06	ACK
7	07	BEL
8	08	BS
9	09	TAB
10	0A	LF
11	0B	VT
12	0C	FF
13	0D	CR
14	0E	SO
15	0F	SI

ASCII	Hex	Symbol
16	10	DLE
17	11	DC1
18	12	DC2
19	13	DC3
20	14	DC4
21	15	NAK
22	16	SYN
23	17	ETB
24	18	CAN
25	19	EM
26	1A	SUB
27	1B	ESC
28	1C	FS
29	1D	GS
30	1E	RS
31	1F	US

ASCII	Hex	Symbol
32	20	(space)
33	21	!
34	22	"
35	23	#
36	24	$
37	25	%
38	26	&
39	27	'
40	28	(
41	29	)
42	2A	*
43	2B	+
44	2C	,
45	2D	-
46	2E	.
47	2F	/

ASCII	Hex	Symbol
48	30	0
49	31	1
50	32	2
51	33	3
52	34	4
53	35	5
54	36	6
55	37	7
56	38	8
57	39	9
58	3A	:
59	3B	;
60	3C	<
61	3D	=
62	3E	>
63	3F	?

ASCII	Hex	Symbol
64	40	@
65	41	A
66	42	B
67	43	C
68	44	D
69	45	E
70	46	F
71	47	G
72	48	H
73	49	I
74	4A	J
75	4B	K
76	4C	L
77	4D	M
78	4E	N
79	4F	O

ASCII	Hex	Symbol
80	50	P
81	51	Q
82	52	R
83	53	S
84	54	T
85	55	U
86	56	V
87	57	W
88	58	X
89	59	Y
90	5A	Z
91	5B	[
92	5C	\
93	5D	]
94	5E	^
95	5F	_

ASCII	Hex	Symbol
96	60	`
97	61	a
98	62	b
99	63	c
100	64	d
101	65	e
102	66	f
103	67	g
104	68	h
105	69	i
106	6A	j
107	6B	k
108	6C	l
109	6D	m
110	6E	n
111	6F	o

ASCII	Hex	Symbol
112	70	p
113	71	q
114	72	r
115	73	s
116	74	t
117	75	u
118	76	v
119	77	w
120	78	x
121	79	y
122	7A	z
123	7B	{
124	7C	|
125	7D	}
126	7E	~
127	7F	DEL

Here's the same ASCII character set expressed in binary values:

Binary	Decimal	Character	Binary	Decimal	Character	Binary	Decimal	Character
00000000	000	NUL	00101011	043	+	01010110	086	V
00000001	001	SOH	00101100	044	,	01010111	087	W
00000010	002	STX	00101101	045	-	01011000	088	X
00000011	003	ETX	00101110	046	.	01011001	089	Y
00000100	004	EOT	00101111	047	/	01011010	090	Z
00000101	005	ENQ	00110000	048	0	01011011	091	[
00000110	006	ACK	00110001	049	1	01011100	092	\
00000111	007	BEL	00110010	050	2	01011101	093	]
00001000	008	BS	00110011	051	3	01011110	094	^
00001001	009	HT	00110100	052	4	01011111	095	_
00001010	010	LF	00110101	053	5	01100000	096	`
00001011	011	VT	00110110	054	6	01100001	097	a
00001100	012	FF	00110111	055	7	01100010	098	b
00001101	013	CR	00111000	056	8	01100011	099	c
00001110	014	SO	00111001	057	9	01100100	100	d
00001111	015	SI	00111010	058	:	01100101	101	e
00010000	016	DLE	00111011	059	;	01100110	102	f
00010001	017	DC1	00111100	060	<	01100111	103	g
00010010	018	DC2	00111101	061	=	01101000	104	h
00010011	019	DC3	00111110	062	>	01101001	105	i
00010100	020	DC4	00111111	063	?	01101010	106	j
00010101	021	NAK	01000000	064	@	01101011	107	k
00010110	022	SYN	01000001	065	A	01101100	108	l
00010111	023	ETB	01000010	066	B	01101101	109	m

Binary	Decimal	Character		Binary	Decimal	Character		Binary	Decimal	Character
00011000	024	CAN		01000011	067	C		01101110	110	n
00011001	025	EM		01000100	068	D		01101111	111	o
00011010	026	SUB		01000101	069	E		01110000	112	p
00011011	027	ESC		01000110	070	F		01110001	113	q
00011100	028	FS		01000111	071	G		01110010	114	r
00011101	029	GS		01001000	072	H		01110011	115	s
00011110	030	RS		01001001	073	I		01110100	116	t
00011111	031	US		01001010	074	J		01110101	117	u
00100000	032	SP		01001011	075	K		01110110	118	v
00100001	033	!		01001100	076	L		01110111	119	w
00100010	034	"		01001101	077	M		01111000	120	x
00100011	035	#		01001110	078	N		01111001	121	y
00100100	036	$		01001111	079	O		01111010	122	z
00100101	037	%		01010000	080	P		01111011	123	{
00100110	038	&		01010001	081	Q		01111100	124	|
00100111	039	'		01010010	082	R		01111101	125	}
00101000	040	(		01010011	083	S		01111110	126	~
00101001	041	)		01010100	084	T		01111111	127	DEL
00101010	042	*		01010101	085	U		**Note: 0-127 is 128 values**		

So, “E-Discovery” would be written in a binary ASCII sequence as:

0100010100101101**01000100**01101001**01110011**01100011**01101111**01110110**01100101**01110010**01111001**

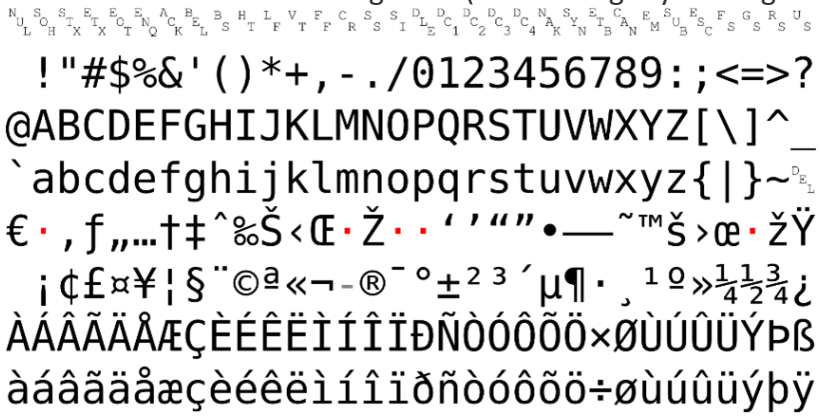
It would be tough to remember your own name written in this manner! *Hi, I’m Craig, but my computer calls me **01000011**01110010**01100001**01101001**01100111.***

Windows-1252

Note that each leading bit (i.e., the first character) of each byte in the binary table above is a zero. It wasn't used to convey any encoding information; that is, they are really all *7-bit* bytes. Later, when the byte standardized from seven to eight bits, 128 additional characters could be added to the character set by simply changing the leading bit to a one, prompting the development of *extended* character encodings that include, *e.g.*, accented characters used in foreign languages and line drawing characters.

In the mid-1980s, international standards began to emerge for character encoding, ultimately resulting in various code sets issued by the International Standards Organization (ISO). These retained the first 128 American ASCII values and assigned the upper (extended) 128-byte values to characters suited to various languages (*e.g.,* Cyrillic, Greek, Arabic and Hebrew). ISO called these various character sets ISO-8859-*n,* where the *"n"* distinguished the sets for different languages. ISO-8859-1 was the set suited to Latin-derived alphabets (like English) and so the nickname for the most familiar code page to U.S. computer users became "**Latin 1**."

Arguably the most used single-byte character set in the world is the **Windows-1252 code page,** the characters of which are set out in the following table (red dots signify unassigned values).



Note that the first 128 control codes and characters (from NUL to DEL) match the ASCII encodings and the 128 characters that follow are the extended set. Each character and control code has a corresponding fixed byte value, *i.e.*, an upper-case B is hex 40 and the section sign, §, is hex A7. To see the entire code page character set and the corresponding hexadecimal encodings on Wikipedia, click here. Again, ASCII and the Windows-1252 code page are *single byte* encodings so they are limited to a maximum of 256 characters.

Unfortunately, these extra characters weren't assigned in the same way by all computer systems. The emergence of different sets of characters mapped to the same high byte values prompted a need to identify these various **character encodings** or, as Microsoft calls them in Windows, these "**code pages**." If an application used the wrong code page, some information displayed as

gibberish. This is such a familiar phenomenon that it has its own name, **mojibake** (from the Japanese for "character changing"). If you've encountered a bunch of Asian language characters in an e-mail or document you know was written in English, you might have glimpsed mojibake. In e-discovery, mojibake can corrupt search results, break deduplication and quietly distort productions.

Note that we are speaking here of textual information, not typography; so, don't confuse character encodings with fonts. The former tells you whether the character is an A or b, not whether to display the character in Arial or Baskerville.

Unicode

ASCII dawned in the pre-Internet world of 1963—before the world was flat, when the West dominated commerce and personal computing was the stuff of science fiction. The Windows-1252 code page works reasonably well so long as you're writing in English and most European languages; but sporting only 256 characters, it won't suffice if you're writing in, say, Greek, Cyrillic, Arabic or Hebrew, and it's wholly unsuited to Asian languages like Chinese, Japanese and Korean.

Though programmers developed various *ad hoc* approaches to foreign language encodings, an increasingly interconnected world needed universal, systematic encoding mechanisms. These methods would use ***more than one byte*** to represent each character, and the most widely adopted such system is **Unicode**. In its latest incarnation (version 17.0, effective September 2025), Unicode standardizes the encoding of 172 written character sets called "scripts" comprising 159,801 characters, plus multiple symbol sets and emoji characters.

The Unicode Consortium crafted Unicode to co-exist with the longstanding ASCII and ANSI character sets by emulating the ASCII character set in corresponding byte values within the more extensible Unicode counterpart, UTF-8. UTF-8 can represent all 128 ASCII characters using a single byte and all other Unicode characters using two, three or four bytes. Because of its backward compatibility and multilingual adaptability, UTF-8 has become the most popular text encoding standard, especially on the Internet and within e-mail systems.

Mind the Gap!

Now, as we talk about all these bytes and encoding standards as a precursor to further discussion of hexadecimal notation, it will be helpful to revisit how this all fits together. A byte is eight ones or zeroes, which means a byte can represent 256 different decimal numbers from 0-255. So, two bytes can represent a much bigger range of decimal values (256 x 256 or 65,536). Character encodings (aka "code pages") like Latin 1 and UTF-8 are ways to map textual, graphical or machine instructions to numeric values expressed as bytes, enabling machines to store and communicate information in human languages. As we move forward, keep in mind that *hex, like binary and*

decimal, is just another way to write numbers. Hex is not a code page, although the numeric values it represents may correspond to values *within* code pages. [55]

Hex

Long sequences of ones and zeroes are very confusing for people, so **hexadecimal notation** emerged as more accessible shorthand for binary sequences. Considering the prior discussion of base 10 (decimal) and base 2 (binary) notation, it might be enough to say that hexadecimal is base 16. In hexadecimal notation (**hex** for short), each digit can be any value from zero to fifteen. Accordingly, we can replace four binary digits with just a single hexadecimal digit, and more to the point, we can express a byte as just two hex characters.

The decimal system supplies only 10 symbols (0-9) to represent numbers. Hexadecimal notation demands 16 symbols, leaving us without enough single character *numeric* values to stand in for all the values in each column. So, how do we cram 16 values into each column? The solution was to substitute the letters A through F for the numbers 10 through 15. So, we can represent 10110101 (the decimal number 181) as "B5" in hexadecimal notation. Using hex, we can notate values from 0-255 as 00 to FF (using either lower- or upper-case letters; it doesn't matter).

It's hard to tell if a number is decimal or hexadecimal just by looking at it: if you see "37", does that equate to 37 ("37" in decimal) or a decimal 55 ("37" in hexadecimal)? To get around this problem, two common notations are used to indicate hexadecimal numbers. The first is the suffix of a lower-case "h." The second is the prefix of "0x." So "37 in hexadecimal," "37h" and "0x37" all mean the same thing.

The ASCII Code Chart below can be used to express ASCII characters in hex. The capital letter "G" is encoded as the hex value of 47 (i.e., row 4, column 7), so "**E-Discovery**" in hex encodes as:
0x45 2D 44 69 73 63 6F 76 65 72 79

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

[55] Don't be put off by the "math." The biggest impediment to getting through the encoding basics is the voice in your head screaming, "I SHOULDN'T HAVE TO KNOW ANY OF THIS!!" Ignore that voice. It's wrong.

That's easier than:

0100010100101101**01000100**01101001**01110011**01100011**01101111**01110110**01100101**01110010**01111001?**

Why Lawyers Should Care

Encoding errors are not abstract computer-science mishaps; they are discovery failures with real litigation consequences. When text is mis-encoded, characters drop, substitute, or render as gibberish, searches fail silently. Responsive documents go undiscovered not because counsel chose bad search terms, but because the underlying text was never correctly extracted or indexed in the first place. Courts have sanctioned parties where keyword searches missed responsive material that was later shown to exist, and "the computer did it" is no defense when counsel never understood—or bothered to ask—whether processing settings handled character sets, Unicode normalization, or legacy encodings correctly. A lawyer who abdicates processing details may believe discovery is complete, while critical evidence sits unreadable in plain sight. Competence in e-discovery today requires at least enough technical literacy to recognize when encoding choices can determine what is found, what is missed, and who pays for it when the miss comes to light.

NOTE TO STUDENTS:

For most students of this course—law students particularly—the last twelve pages on digital encoding constitute perhaps the most challenging material we cover. If your eyes glazed over on first reading, I understand. But it's foundational stuff, so give it a second chance when you're fresh and be sure to ask questions about the material until it's clear to you. You'll be glad you did when you start taking the quizzes.



Exercise 5: Encoding: Decimal, Binary, Hexadecimal and Base64

All digital evidence is encoded, and its encoding bears upon how it's collected, whether it can be searched and in what reasonably usable forms it can be produced. Understanding that electronically stored information is, in essence, numerically encoded data helps students to see the interplay and interchangeability between different forms of digital evidence. Simply saying, "*it's all ones and zeroes*" means nothing if you know nothing of how those ones and zeros underpin evidence you must authenticate or undercut.

GOALS: The goals of this exercise are for the student to:

1. Understand the correspondence between binary data and hexadecimal; and
2. Understand the correspondence between data in hex and encoded text and content.

OUTLINE: We will examine evidence data in Text and Hex modes, noting the correspondence between text and its hexadecimal equivalents. We will then examine the role of Base64 as an encoding scheme for e-mail attachments.

Exercise 5A: Notate ASCII as Hex

Please write your surname in ASCII/hex:

____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____

Exercise 5B: Viewing data in Hex

In this exercise, you will use online data viewer tools to examine common file types in hexadecimal. Remember that hexadecimal is just a method to notate numeric values. Such values can be expressed in any notation, e.g., base 2 (binary) or base 10 (decimal) or any other base. *It's all just numbers that are written differently but mean the same thing*. Still, be mindful of the distinction between the notation employed to *record* the information (the "numbers") and the encoding scheme used to *express* the information (e.g., ASCII, ANSI, Unicode, etc.). The notation is akin to an alphabet (e.g., Roman, Cyrillic, etc.) and the encoding scheme is like the language (e.g., English, French, Russian, etc.).

In the preceding exercise, the encoding scheme was ASCII and the notation was hexadecimal. Put another way, ASCII supplied the translation table, and hex served to record the locations within that table.

Step 1: View the File Structure of a Compressed Container (Compound) File

Download the compressed archive file called GBA.zip from http://www.craigball.com/gba.zip. Save the file to your desktop or any other convenient location on your computer.

Using your web browser or the hex viewer of your choice,[56] go to the Online HexDump Utility at http://www.fileformat.info/tool/hexdump.htm and click "choose File." Using the selection box that will appear, navigate to the file **gba.zip** you just saved and select it. Click "Open." Now click the blue "Dump" button on the Online HexDump Utility page. You should see this:

```
file name: GBA.zip
mime type:

0000-0010:  50 4b 03 04-14 00 00 00-08 00 30 66-bf 3e 67 aa  PK...... ..0f.>g.
0000-0020:  d4 42 c3 02-00 00 c9 05-00 00 07 00-00 00 47 42  .B...... ......GB
0000-0030:  41 2e 74 78-74 65 54 4b-8e db 30 0c-dd 0f 30 77  A.txteTK ..0...0w
0000-0040:  e0 01 3c 39-44 17 2d 0a-14 5d 15 e8-9a b6 68 5b  ..<9D.-. .]....h[
0000-0050:  18 59 4c 25-d9 86 6f df-47 c9 4e 52-74 13 44 96  .YL%..o. G.NRt.D.
0000-0060:  44 be 1f f5-55 d7 44 79-d0 24 c4 d1-51 96 4d 22  D...U.Dy .$..Q.M"
0000-0070:  1d c2 29 13-4f 4a b6 3b-72 99 05 eb-3e e9 3a cd  ..).OJ.; r...>.:.
0000-0080:  85 46 4d 65-26 8d 54 66-9f 69 d0 58-7c 94 58 3a  .FMe&.Tf .i.X|.X:
0000-0090:  62 8a b2 53-e4 e2 35 76-b6 31 88 df-c4 91 8f f4  b..S..5v .1......
0000-00a0:  c3 f7 92 ca-d1 d5 26 4e-9c 1f b8 60-a7 28 8a 08  ......&N ...`.(..
0000-00b0:  dd 93 de 35-7b bb 87 35-17 e2 10 68-01 10 06 ac  ...5{..5 ...h....
0000-00c0:  21 49 3d 2b-7f 56 0e 37-7a 7f 7b 7f-fb a9 3b ed  !I=+.V.7 z.{...;.
0000-00d0:  52 77 25 4e-3c b5 1e 4c-93 1d a5 c1-6f 3e d0 ce  Rw%N<..L ....o>..
0000-00e0:  a9 a3 22 19-e0 26 da 67-31 0e ad f8-05 50 13 d0  ..".&.g 1....P..
0000-00f0:  1c e7 92 b2-be 20 ae 52-e8 13 28 d8-70 a4 a0 28  ....... .R..(.p..(
0000-0100:  25 d1 ad 49-6e f4 bb f5-5f a4 98 14-57 ef 9e 4b  %..In... _...W..K
0000-0110:  09 f2 31 7a-09 8e 74 6c-0d 01 a5 9e-9f 79 03 1d  ..1z..tl .....y..
0000-0120:  5d c4 78 5f-b5 71 f5 0e-41 0d c2 75-be de 86 54  ].x_.q.. A..u...T
0000-0130:  b0 00 ff 23-07 4a 27 8f-7b e0 41 cc-00 1c d4 2c  ...#.J'. {.A....,
0000-0140:  e0 a5 04 5e-42 93 95 06-45 9f 28 80-41 6e 85 5e  ...^B... E.(.An.^
0000-0150:  e8 d2 e2 cd-3b db bc d1-f7 42 70 8e-43 d1 a9 e9  ....;... .Bp.C...
0000-0160:  32 fa 52 eb-1b 6f 33 e3-92 0a 22 e7-59 57 70 71  2.R..o3. .."YWpq
0000-0170:  5a ed 6e fa-7f 59 e1 76-55 3c 70 9a-70 38 4b cc  Z.n..Y.v U<p.p8K.
0000-0180:  d2 d9 71 d3-29 6a 79 d2-fb f8 78 fd-0c 89 b3 0c  ..q.)jy. ..x.....
0000-0190:  e9 ff 8d 19-96 c3 55 7c-ac a9 9a 90-b5 e8 6e f4  ......U| ......n.
0000-01a0:  0b e9 e8 93-91 43 1c 3a-83 7f c1 74-c2 d0 c8 04  .....C.: ...t....
(A) OFFSET                    (B) HEX                    (C) ASCII TEXT
```

The three columns of information represent, from left to right, (A) the byte offset (location) of the hex value within the file, expressed in hexadecimal notation, (B) the contents of the file in hexadecimal notation and (C) the hexadecimal content expressed as ASCII text (to the extent the hex values *have* a corresponding ASCII value).

Note that the first two ASCII characters in the file are PK and the first two hex values are 50 4b.

If you check the ASCII Code Chart, you'll see that everything matches up: 50 4B = PK. That PK at the start of the file serves an important purpose. It's the file's **binary header signature**. In computing, a file's header refers to data occurring at or near the start of the file that serves to identify the type of data contained in the file and may also furnish information about the file's length, structure or other characteristics. [Don't confuse *file* headers with *mail* headers, which carry information about, *e.g.*, sender, addressee(s), routing, subject, etc. for e-mail messages.] That PK means that the file data that follows is encoded and compressed with Zip compression. In other words, as a file header," PK" signals to the operating system that the data will only make sense if it is interpreted as Zip-compressed content.

[56] a capable alternate online hex editor can be found at https://hexed.it/?hl=en

Why PK? Because the unfortunate fellow who came up with the Zip compression algorithm was named Phil Katz! Phil insured his place in history by using his initials as the binary header signature for Zip files. So long as it's not already used to identify another file type, a binary header signature can be almost anything, and the person or entity that originates the file structure/type gets to choose it. More on that to come.

Step 2: Unpack the Archive

Open the zip file and extract (unzip) its contents to a convenient location on your machine.

The zip file should hold the seven files listed below:

	Name	File Type
1.	**GBA.doc**	Word Document
2.	**GBA.docx**	Word Document
3.	**GBA.htm**	Web Page
4.	**GBA.pdf**	Adobe PDF
5.	**GBA.rtf**	Rich Text Format
6.	**GBA.txt**	Text
7.	**GBA.eml**	E-mail

Remember where you stored these extracted files.

Step 3: Exploring the Contents of the Archive

Six of these files hold precisely the same famous text, but each in their own unique ***encoded*** way. The seventh, an e-mail, also holds the text, but encoded as *both* an image and attachment.

One-by-one, load each file except GBA.eml into the Online HexDump Utility, http://www.fileformat.info/tool/hexdump.htm, (or the hex viewer of your choice) and explore each file's hex and ASCII content. Now, please answer the following questions about the files:

Exercise 5C: Encoding Anomalies

1. Who is the famous author of the text? _____

2. As you go through each file, can you identify any **date or time values** in the body of the file (e.g., application metadata values like Creation Date, Last Modified Date, Date Last Printed or the like)?

   _____

3. Which, if any, of these files do not show the famous text as human-readable text anywhere in the ASCII Text column? _____

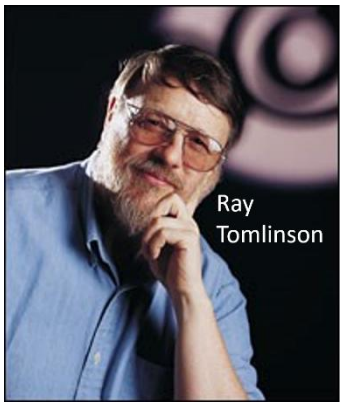
4. What are the first four **hex** values seen in the file GBA.doc? _____

5. Do you note anything striking about the file GBA.txt in terms of its file signature?

_____

Base64

Internet e-mail was born in 1971, when a researcher named Ray Tomlinson sent a message to himself using the "@" sign to distinguish the addressee from the machine. Tomlinson didn't remember the message transmitted in that historic first e-mail but speculated that it was probably something like "qwertyuiop." So, not exactly, "*Mr. Watson, come here. I need you*," but then, Tomlinson didn't *know* he was changing the world. He was just killing time.



Also, back when the nascent Internet consisted of just four university research computers, UCLA student Stephen Crocker originated the practice of circulating proposed technical standards (or "protocols" in geek speak) as publications called "Requests for Comments" or RFCs. They went via U.S. postal mail because there was no such thing as e-mail. Ever after, proposed standards establishing the format of e-mail were promulgated as numbered RFCs. So, when you hear an e-discovery vendor mention "RFC5322 content," fear not, it just means plain ol' e-mail.

An e-mail is as simple as a postcard. Like the back-left side of a postcard, an e-mail has an area called the message body reserved for the user's text message. Like a postcard's back right side, we devote another area called the message header to information needed to get the card where it's supposed to go and to transmittal data akin to a postmark.

We can liken the picture or drawing on the front of our postcard to an e-mail's attachment. Unlike a postcard, we must convert e-mail attachments to letters and numbers for transmission, enabling an e-mail to carry any type of electronic data — audio, documents, software, video —not just pretty pictures.

The key point is that *everything in any e-mail is plain text*, no matter what's attached.

And by plain text, I mean the plainest English text, 7-bit ASCII, lacking even the diacritical characters required for accented words in French or Spanish or any formatting capability. No **bold**. No underline. No *italics*. It is text so simple that you can store a letter as a single byte of data.

The dogged adherence to plain English text stems in part from the universal use of the Simple Mail Transfer Protocol or SMTP to transmit e-mail. SMTP only supports 7-bit ASCII characters, so sticking with SMTP maintained compatibility with older, simpler systems. Because it's just text, it's compatible with any e-mail system invented in the last 50 years. Think about that the next time you come across a floppy disk or CD and wonder how you're going to read it.

How do you encode a world of complex digital content into plain text without losing anything?

The answer is an encoding scheme called Base64, which substitutes 64 printable ASCII characters (A–Z, a–z, 0–9, + and /) for any binary data or for foreign characters, like Cyrillic or Chinese, that can be represented by the Latin alphabet.

Base64 is brilliant and amazingly simple. Since all digital data is stored as bits, and six bits can be arranged in 64 separate ways, you need just 64 alphanumeric characters to stand in for any six bits of data. The 26 lower case letters, 26 upper case letters and the numbers 0-9 give you 62 stand-ins. Throw in a couple of punctuation marks—say the forward slash and plus sign—and you have all the printable ASCII characters you need to represent any binary content in six-bit chunks. Though the encoded data takes up roughly a third more space than its binary source, now any mail system can hand off any attachment. Once again, *it's all just numbers.*

Exercise 5D: Exploring Base64

In this exercise, we will open the e-mail GBA.eml in a plain text viewer and locate its Base64-encoded content. If you are using a Windows machine, you can use Notepad as your text viewer; else, you can use the free application at **http://www.rapidtables.com/tools/notepad.htm.**

Step 1: Open the File in the Text Viewer

Returning to the seven files you extracted from the GBA.zip archive, find the file named GBA.eml and, using the commands, File>Open, open the file in your preferred plain text viewer. Once visible, scroll down until you see this block of data:

--089e0118431ed478e705164be95e--
--089e0118431ed4790705164be960
Content-Type: image/gif; name="GBA.gif"
Content-Disposition: attachment; filename="GBA.gif"
Content-Transfer-Encoding: base64
X-Attachment-Id: f_i9suf1i90

This snippet tells the recipient system that the attachment data is encoded in base64 and that it should be interpreted as a GIF image file named GBA when decoded. The first two lines are boundaries signaling where sections of the message begin and end.

Now, look at the gibberish text that follows, all the way until the end of the message. What you are seeing is a .gif image file—a drawing--that's been numerically encoded to be able to traverse

the network as an e-mail attachment. Note that the entirety of the data preceding the end boundary of the message: **==--089e0118431ed4790705164be960—** are composed of 64 characters: the 26 lower case letters, 26 upper case letters, the numbers 0-9 and the forward slash and plus sign.[57]

The Base64-encoded content you see should begin:

```
R0lGODlhGQNOAXAAACH5BAEAAMcALAAAAAAZA04BhwAAAKy3wtDOzlJ7ocnHx8/Nza+np8nGxs7K
yj1qk4mFhcHAwM7MzLSysjszM46KisTBwcbDw6ekpDhhh4WCgri2ts3Ly7y6uo6MjDIyMqCensjG
xp6cnC4uLjhhhjNafBUSEoaEhMrIyK+srF5eXnt5eba0tMzKyq6rq1BOTjNZezVcf0E+PsbExKin
p1RUVGNgYLCtrcC9vZKQkDVbfjtmjYN/f8vJybe1tVdXV7q4uGRfXzpliz9tlrSwsMvIyNDNzc7L
y8PDw83MzMzLy0BwmOBAQM/MzNDMzMvLy8jIyGlaWkFxnJCQkNHNzdLMzN+/v9HOzq6urgoKCtTJ
yc/Pz8XCwru7uzs6OjAvL2tpaQ4ODjk4OHd2drq1tW9fXxISEgcHBykKF1bWwQEBBUVFTQzM6qh
oScmJiAfH83KyiUkJB0dHU1NTTo5OYSCgtHMzA8PD8vKyrOzs8O/v8fExAgICFdUVElSG5ubpCP
jz08PAkJCU5NTRYWFlJRUREQEL+/v4yKioF/fxoaGjExMRAQECEhIaCfn2FfXwICAoaFhaupqXt6
elhWVh8fH3JxcVtaWp6dnaCgoD89PQMDA8C8vFhXV4SDg3Z0dA0NDScnJ0dHRykpKR4eHiUlJUNC
Qjg2Njc2Ni8vLzg3NxkZGW9ubmlpaYiGhkZFRSckJGxra1xbW0FAQCUiIhEREQ8MDAYGBp+ennNz
c3JwcEVDQwEBAaimpo+OjjIxMRsbGyQjIz8/P19dXTw6Ojw8PJ2ammZlZW9tbTc3N2hoaBQUFJGP
jwAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

…and continue for another 333 lines of base64-encoded information.

Takeaway: The different ways in which data is encoded—and the different ways in which collection, processing and search tools identify the multifarious encoding schemes—lie at the heart of significant challenges and costly errors in e-discovery. It's easy to dismiss these fundamentals of information technology as too removed from litigation to be worth the effort to explore them; but, understanding encoding, binary signatures and the role they play in collection, indexing and search will help you realize the capabilities and limits of the tools you, your clients, vendors and opponents use.

To that end, we're just getting started.

[57] Turning all that binary data into alphanumeric data has a dark side in e-discovery reminiscent of those apocryphal monkeys at typewriters who will, in the fullness of infinite time, type Shakespeare's sonnets. Trillions of seemingly random alphabetic characters necessarily form words by happenstance--including keywords searched in discovery. Very short keywords occur in Base64 with alarming frequency. If the tools employed in e-discovery treat encoded base64 attachments as text, or if your search tool doesn't decode base64 content before searching it, false or "noise" hits may prove a significant problem.



Exercise 6: Encoding: Running the Bases

All ESI is encoded, so success in electronic discovery often depends upon the ability to extract intelligible text from encoded forms to facilitate search and review. Each of the following questions for you to answer is encoded in a common digital format. The first question is in binary (base2), the second is in hex (base16) and the third is in base64. Decode each to find the three questions you must answer in this exercise. Again, decode each to find the question, then answer the three questions.

GOALS: The goals of this exercise are for the student to:

1. Decode each of three questions encoded in various ways: and

2. Answer the questions posed in the decoded text. You will submit these answers in Canvas.

You are encouraged to use free online tools[58] and solicit help as needed, with the proviso that you must be prepared to demonstrate your solutions to the problems. The tasks should take no more than about 15-30 minutes. The exercise has **five** parts (Questions 1-5). Answer and submit responses to all five.

IMPORTANT: Please don't try to decode these by hand (an electronic version of the text can be found at http://www.craigball.com/runningbases2021.txt). And remember: Google and Wikipedia *are* your friends!

Question 1: Easy

```
01010100 01101000 01100101 00100000 01010011 01110101 01101101 01100101 01110010 01101001 01100001
01101110 00100000 01100001 01100010 01100001 01100011 01110101 01110011 00100000 01101001 01110011
00100000 01110100 01101000 01100101 00100000 01100110 01101001 01110010 01110011 01110100 00100000
01101011 01101110 01101111 01110111 01101110 00100000 01110000 01101000 01111001 01110011 01101001
01100011 01100001 01101100 00100000 01101001 01101110 01110011 01110100 01110010 01110101 01101101
```

[58] Examples:
Binary decoders:
http://www.roubaixinteractive.com/PlayGround/Binary_Conversion/Binary_To_Text.asp
https://paulschou.com/tools/xlate/
http://nickciske.com/tools/binary.php
Hex decoders:
http://www.convertstring.com/EncodeDecode/HexDecode
http://www.unit-conversion.info/texttools/hexadecimal/
http://bin-hex-converter.online-domain-tools.com/
Base64 decoders:
http://codebeautify.org/base64-to-image-converter
https://onlineimagetools.com/convert-base64-to-image
http://www.freeformatter.com/base64-encoder.html

01100101 01101110 01110100 00100000 01100110 01101111 01110010 00100000 01110100 01101000 01100101
00100000 01110000 01110101 01110010 01110000 01101111 01110011 01100101 00100000 01101111 01100110
00100000 01100011 01100001 01110010 01110010 01111001 01101001 01101110 01100111 00100000 01101111
01110101 01110100 00100000 01100011 01101111 01101101 01110000 01110101 01110100 01100001 01110100
01101001 01101111 01101110 00101110 00100000 00100000 01001001 01110100 00100000 01100010 01101111
01110010 01100101 00100000 01101100 01101001 01110100 01110100 01101100 01100101 00100000 01110010
01100101 01110011 01100101 01101101 01100010 01101100 01100001 01101110 01100011 01100101 00100000
01110100 01101111 00100000 01110100 01101000 01100101 00100000 01101101 01101111 01100100 01100101
01110010 01101110 00100000 01000011 01101000 01101001 01101110 01100101 01110011 01100101 00100000
01100001 01100010 01100001 01100011 01110101 01110011 00101100 00100000 01110100 01100001 01101011
01101001 01101110 01100111 00100000 01110100 01101000 01100101 00100000 01100110 01101111 01110010
01101101 00100000 01101111 01100110 00100000 01100001 00100000 01100110 01101100 01100001 01110100
00100000 01110011 01110101 01110010 01100110 01100001 01100011 01100101 00101100 00100000 01110011
01110101 01100011 01101000 00100000 01100001 01110011 00100000 01100001 00100000 01110011 01110100
01101111 01101110 01100101 00100000 01110100 01100001 01100010 01101100 01100101 01110100 00100000
01101001 01101110 01100011 01101001 01110011 01100101 01100100 00100000 01110111 01101001 01110100
01101000 00100000 01110000 01100001 01110010 01100001 01101100 01101100 01100101 01101100 00100000
01101100 01101001 01101110 01100101 01110011 00101100 00100000 01100001 01101110 01100100 00100000
01110000 01100101 01100010 01100010 01101100 01100101 01110011 00100000 01110101 01110011 01100101
01100100 00100000 01100001 01110011 00100000 01100011 01101111 01110101 01101110 01110100 01100101
01110010 01110011 00100000 01110100 01101111 00100000 01110100 01110010 01100001 01100011 01101011
00100000 01110001 01110101 01100001 01101110 01110100 01101001 01110100 01101001 01100101 01110011
00101110 00100000 00100000 01010100 01101000 01100101 00100000 01010011 01110101 01101101 01100101
01110010 01101001 01100001 01101110 01110011 00100000 01101101 01100001 01100100 01100101 00100000
01110011 01101001 01100111 01101110 01101001 01100110 01101001 01100011 01100001 01101110 01110100
00100000 01100011 01101111 01101110 01110100 01110010 01101001 01100010 01110101 01110100 01101001
01101111 01101110 01110011 00100000 01110100 01101111 00100000 01101101 01100001 01110100 01101000
01100101 01101101 01100001 01110100 01101001 01100011 01110011 00100000 01100001 01101110 01100100
00100000 01100001 01110010 01100101 00100000 01100011 01110010 01100101 01100100 01101001 01110100
01100101 01100100 00100000 01110111 01101001 01110100 01101000 00100000 01100010 01100101 01101001
01101110 01100111 00100000 01110100 01101000 01100101 00100000 01100110 01101001 01110010 01110011
01110100 00100000 01110100 01101111 00100000 01110101 01110011 01100101 00100000 01110011 01111001
01101101 01100010 01101111 01101100 01110011 00100000 01110100 01101111 00100000 01110010 01100101
01110000 01110010 01100101 01110011 01100101 01101110 01110100 00100000 01100111 01110010 01101111
01110101 01110000 01110011 00100000 01101111 01100110 00100000 01101111 01100010 01101010 01100101
01100011 01110100 01110011 00100000 01110100 01101111 00100000 01100011 01101111 01101101 01101101
01110101 01101110 01101001 01100011 01100001 01110100 01100101 00100000 01101100 01100001 01110010
01100111 01100101 00100000 01101110 01110101 01101101 01100010 01100101 01110010 01110011 00101110
00100000 00100000 01010100 01101000 01100101 01111001 00100000 01100001 01100100 01101111 01110000
01110100 01100101 01100100 00100000 01100001 00100000 01100011 01101111 01110101 01101110 01110100
01101001 01101110 01100111 00100000 01110011 01111001 01110011 01110100 01100101 01101101 00100000
01100010 01100001 01110011 01100101 01100100 00100000 01101111 01101110 00100000 01110100 01101000
01100101 00100000 01101110 01110101 01101101 01100010 01100101 01110010 00100000 00110110 00110000
00101100 00100000 01110111 01101000 01101001 01100011 01101000 00100000 01100111 01100001 01110110
01100101 00100000 01110101 01110011 00100000 01110100 01101000 01100101 00100000 01110011 01101001
01111000 01110100 01111001 00101101 01101101 01101001 01101110 01110101 01110100 01100101 00100000
01101000 01101111 01110101 01110010 00101100 00100000 01110011 01101001 01111000 01110100 01111001
00101101 01110011 01100101 01100011 01101111 01101110 01100100 00100000 01101101 01101001 01101110
01110101 01110100 01100101 00100000 01100001 01101110 01100100 00100000 01110100 01101000 01100101
00100000 01101110 01101111 01110100 01100001 01110100 01101001 01101111 01101110 00100000 01110011
01111001 01110011 01110100 01100101 01101101 00100000 01110101 01110011 01100101 01100100 00100000
01110100 01101111 00100000 01110000 01101100 01101111 01110100 00100000 01110000 01101111 01101001
01101110 01110100 01110011 00100000 01101111 01101110 00100000 01110100 01101000 01100101 00100000
01100111 01101100 01101111 01100010 01100101 00101100 00100000 01100001 01101100 01110011 01101111
00100000 01110101 01110011 01100101 01100100 00100000 01110100 01101111 00100000 01100100 01100101
01101110 01101111 01110100 01100101 01100100 00100000 01101100 01100001 01110100 01100101 01110101
01100100 01100101 00100000 01100001 01101110 01100100 00100000 01101100 01101111 01101110 01100111
01101001 01110100 01110101 01100100 01100101 00100000 01110110 01100001 01101100 01110101 01100101
01110011 00100000 01100101 01101101 01100010 01100101 01100100 01100100 01100101 01100100 00100000
01101001 01101110 00100000 01110100 01101000 01100101 00100000 01000101 01011000 01001001 01000110
00100000 01100100 01100001 01110100 01100001 00100000 01101111 01100110 00100000 01110011 01101101
01100001 01110010 01110100 01110000 01101000 01101111 01101110 01100101 00100000 01110000 01101000
01101111 01110100 01101111 01100111 01110010 01100001 01110100 01101000 01100111 01100001 00101110
00100000 01010100 01101000 01100001 01110100 00100000 01100111 01100101 01101111 01101100 01101111
01100011 01100001 01110100 01101001 01101111 01101110 00100000 01100100 01100001 01110100 01100001
00100000 01101101 01100001 01101011 01100101 01110011 00100000 01000111 01010010 01000101 01000001
01010100 00100000 01100101 01110110 01101001 01100100 01100101 01101110 01100011 01100101 00101110
00100000 00100000 01010100 01101000 01100001 01101110 01101011 00100000 01111001 01101111 01110101
00101100 00100000 01010011 01110101 01101101 01100101 01110010 01101001 01100001 01101110 01110011
00100001 00001101 00001010 00001101 00001010 01010001 01110101 01100101 01110011 01110100 01101001

```
01101111 01101110 00111010 00100000 01010111 01101000 01100001 01110100 00100000 01101001 01110011
00100000 01110100 01101000 01100101 00100000 01100010 01100001 01110011 01100101 00101101 00110110
00110000 00100000 01110011 01111001 01110011 01110100 01100101 01101101 00100000 01101111 01110010
01101001 01100111 01101001 01101110 01100001 01110100 01100101 01100100 00100000 01100010 01111001
00100000 01110100 01101000 01100101 00100000 01010011 01110101 01101101 01100101 01110010 01101001
01100001 01101110 01110011 00100000 01100011 01100001 01101100 01101100 01100101 01100100 00111111
```

ANSWER:_____

Question 2: Harder

```
54 68 65 20 66 61 6d 65 64 20 45 6e 67 6c 69 73 68 20 70 6f 65 74 20 61 6e 64 20 6e 6f 74 6f 72 69
6f 75 73 20 70 6c 61 79 62 6f 79 2c 20 4c 6f 72 64 20 42 79 72 6f 6e 2c 20 68 61 64 20 61 20 64 61
75 67 68 74 65 72 20 68 65 20 61 62 61 6e 64 6f 6e 65 64 20 61 66 74 65 72 20 66 69 76 65 20 77 65
65 6b 73 2e 20 20 53 68 65 20 77 61 73 20 72 69 67 6f 72 6f 75 73 6c 79 20 65 64 75 63 61 74 65 64
20 69 6e 20 6d 61 74 68 20 61 6e 64 20 73 63 69 65 6e 63 65 2c 20 65 78 74 72 61 6f 72 64 69 6e 61
72 79 20 66 6f 72 20 61 20 77 6f 6d 61 6e 20 69 6e 20 74 68 65 20 65 61 72 6c 79 20 31 39 74 68 20
63 65 6e 74 75 72 79 20 62 75 74 20 69 6e 74 65 6e 64 65 64 20 74 6f 20 64 69 73 74 61 6e 63 65 20
68 65 72 20 66 72 6f 6d 20 74 68 65 20 6c 69 62 65 72 74 69 6e 65 20 69 6e 63 6c 69 6e 61 74 69 6f
6e 73 20 6f 66 20 68 65 72 20 70 61 70 61 2e 20 20 53 68 65 20 6d 65 74 20 6d 61 74 68 65 6d 61 74
69 63 69 61 6e 20 61 6e 64 20 69 6e 76 65 6e 74 6f 72 20 43 68 61 72 6c 65 73 20 42 61 62 62 61 67
65 20 61 74 20 61 20 64 69 6e 6e 65 72 20 70 61 72 74 79 20 69 6e 20 31 38 33 33 20 77 68 65 72 65
20 42 61 62 62 61 67 65 20 73 68 6f 77 65 64 20 68 65 72 20 61 20 70 72 6f 74 6f 74 79 70 65 20 6f
66 20 68 69 73 20 64 69 66 66 65 72 65 6e 63 65 20 65 6e 67 69 6e 65 2c 20 74 68 65 20 66 69 72 73
74 20 61 75 74 6f 6d 61 74 69 63 20 63 61 6c 63 75 6c 61 74 6f 72 2e 20 20 49 6e 20 31 38 34 33 2c
20 73 68 65 20 70 75 62 6c 69 73 68 65 64 20 61 6e 20 61 72 74 69 63 6c 65 20 66 6f 72 20 53 63 69
65 6e 74 69 66 69 63 20 4d 65 6d 6f 69 72 73 20 69 6e 20 77 68 69 63 68 20 73 68 65 20 69 6e 63 6c
75 64 65 64 20 64 65 74 61 69 6c 65 64 20 69 6e 73 74 72 75 63 74 69 6f 6e 73 20 66 6f 72 20 73 65
74 74 69 6e 67 20 74 68 65 20 42 61 62 62 61 67 65 20 64 69 66 66 65 72 65 6e 63 65 20 65 6e 67 69
6e 65 20 74 6f 20 63 6f 6d 70 75 74 65 20 42 65 72 6e 6f 75 6c 6c 69 20 6e 75 6d 62 65 72 73 2e 20
20 48 65 72 20 61 6c 67 6f 72 69 74 68 6d 20 61 6e 64 20 69 6e 73 74 72 75 63 74 69 6f 6e 20 73 65
74 20 61 72 65 20 67 65 6e 65 72 61 6c 6c 79 20 72 65 67 61 72 64 65 64 20 61 73 20 6f 6e 65 20 6f
66 20 74 68 65 20 66 69 72 73 74 20 70 75 62 6c 69 73 68 65 64 20 63 6f 6d 70 75 74 65 72 20 70 72
6f 67 72 61 6d 73 2e 0a 51 75 65 73 74 69 6f 6e 3a 20 57 68 6f 20 77 61 73 20 74 68 69 73 20 72 65
6d 61 72 6b 61 62 6c 65 20 65 61 72 6c 79 20 70 72 6f 67 72 61 6d 6d 65 72 3f 0a
```

ANSWER:_____

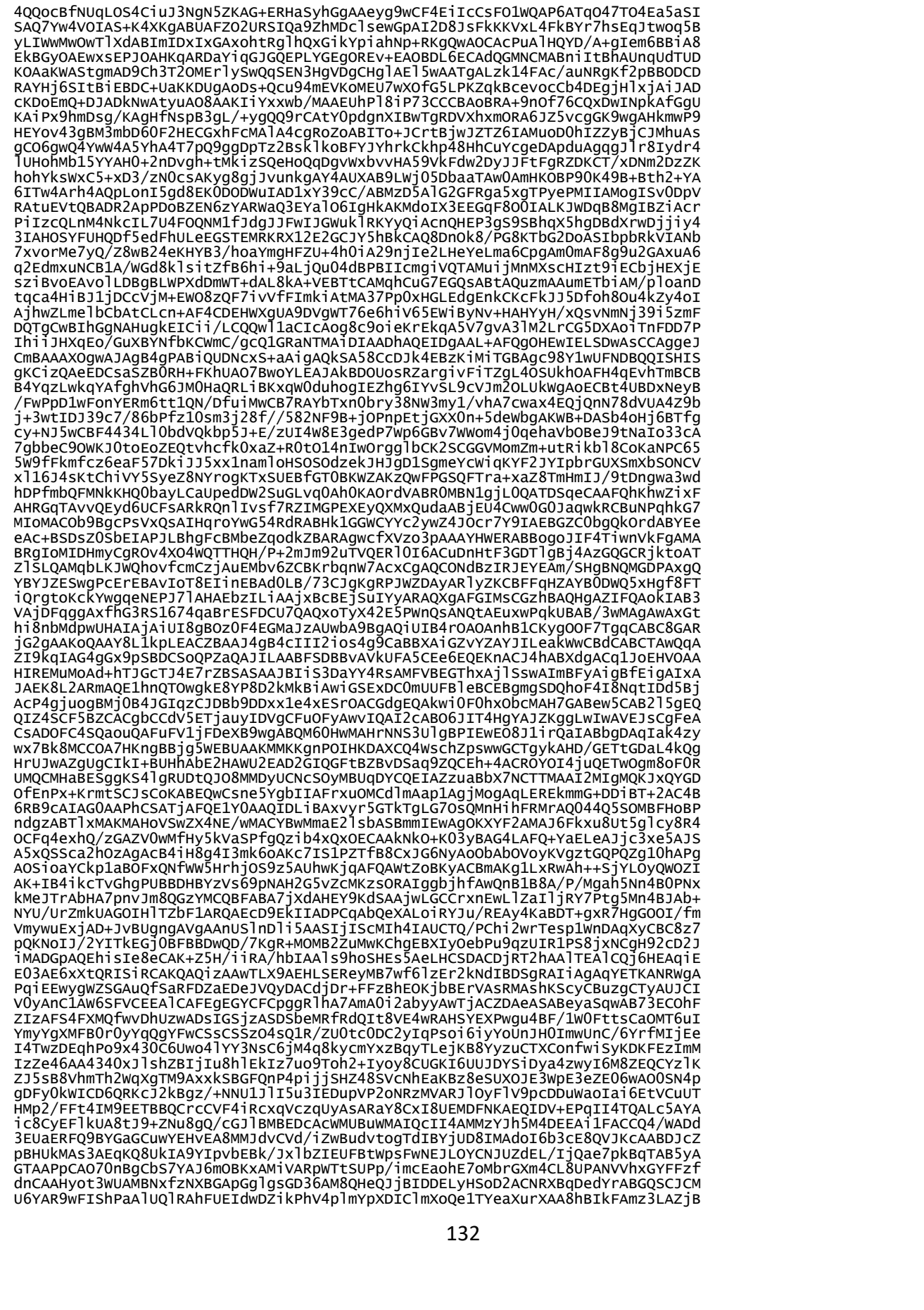
Question 3: Hardest

```
Content-Type: image/gif; name="unicode.gif"
Content-Disposition: attachment; filename="unicode.gif"
Content-Transfer-Encoding: base64
```

```
R0lGODlhgAG7AHAAACH5BAEAAP8ALAAAAACAAbsAh////+/v74yUlPfv94SEjGtrY+bm5lpjY729
vc7W1vf/98XOzmtrc8w9xffv5t7e3nuEhJyclFpSWpScpaWlpUpCSsW9tUpSSntze2tze97v5kJC
OlpaSrWttbWtpQgZEObm1vf3/zoxMc7e3q2trQAAALW1vZytpYyEhN7Fve/v3sXOvQicpTpCSjoQ
SnuEa3u9a6U6UnM6GaUQUnMQGZzexa1a5t6c5kpapd5C5pxjQq0Z5q1atd6ctd5CtaOZtXveOnuc
OmtjGRBaWnveEHucEHtzc0pazkreEEoZhEoZzkqcEKWllGs6St7e1t7m5iEhGUJaGdbWxWve5ine
5imc5hlaMRlapRla7xneMRkZpRkZ7xmcMRnvYxmtYOre5gje5gic5hnOYxmMY6WUlObF3rW9rXtC
hHsQhKU6GXvva2sQUq29a6UQGUpaWt7Wva2tta3eOjopGa2cOq3eEK2cEJyUzjoxQjEIEEpa7Orv
YOreMUoZpUoZ7OqcMUqtYxAISkrOYOqMY2taWqWMa1Jrc5xrhK1ChKOQhM7O3nOElk3va4yEc973
Ut73Gd73jDExMUpSOpyU7+8ZjO8ZOu8ZY+8ZEM4ZjM4ZOs4ZY84ZEJyMjHtrYxkpMd73va3mnHtr
5u9r5u8Q5q1jEHsp5ntrte9rte8QtXsptYzmnHtK5s5r5s4Q5oxjEHsI5ntKtc5rtc4QtXsIta2t
nK3m7+/OjO9ajO9aOu/OOu+UjO+UOu/OY+9aY+9aEO/OEO+UY++UEBlrEBlrhBlrzhnvEBkphBkp
zhmtEFKtpUpSWozm787OjM5ajM5aOs7OOs6UjM6UOs7OY85aY85aEM7OEM6UY86UEBlKEBlKhBlK
zhnOEBkIhBkIzhmMEFKMpZS15qW1vbW91mut3mvvpSnvpSmtpUqt3krvpQjvpWuM3mvOpSnOpSmM
pUqM3krOpQjOpcXF1ntzhFprYxApWqVjY5TFnO/m5nOtpebe3kpKezpjWq2Mpeb3/3uMhP/v/wAA
AAj/AAEIHEiwoMGDCBMqXMiwocOHECNKnEixosWLGDNq3Mixo8ePIEOKHEmypMmTKFOqXMmypcuX
MGPKnEmzps2bOHPq3MmzJOQFFYFWnAUgAL+HARIm9cmUoAIHJI5WRPCA5IMFSzkqUIEg6OqhAPgp
ILEQKFiCCxYMdOD1oZQNkO6ckAglAIEIDhEcQBgAA9mKCxAsfPD3YARIkCS0lUjYpooKDiwWEHww
AN6MCDDg+/hAggaKD+BcVIAAEhQoIlZOYujg8JOPAgZuIjBwwSB+JKo+tIBhaYAEC8oAxbcgQdGO
SVUscLJaBT8VV9UqiP55YICqVzcTfKAA39UH2QGA/l+g2OnaolelMGAQIEDazfhGWPi8qcUDsQVn
VUgq1h5y8cCpkAhWO/33QFJO2LWfeOQVJQVw3zmFT1IHKrCAFGJJkYh21kUHgANPoDfdgwJihU8L
aalwHIH4JJKAA+QtddYgETQYQIgHIiBCImepMEKDAJinW1pVQXceApx8dmFbKuBolIYIArndCmWs
IIGB5wXghBRPuIfVcSvgQOAyARRY1W8QSqkQARMAoIAcT3zpnhQyOsVBIm4CQAAGwH2ogQOQkFBd
```

```
QyscIABZZkBRABQLqCDBBZyMwEkFFRSgAAYVDAKFBn6RUYEEkCBg2yAlnOBPnghI8FgBUVk3GQbJ
gP9awKcacCKBBHdc+GkFJJjx6SYiQEHCArYqNqsEFWjAARQERFbQIGOKFEABHFTAgAa3ipBMARsQ
4MSttHICDgBQOFEBFAKsUICthEFxZQUY0DbQpXNVgEAEEkBQBgnrHpCVAmZ0UoAIuVXATwMcdEZp
pZBE4IDAlSqwyboYaDBIC3dJgIFaBh3wqX17AsDJCS/gkQGeAl16xyCQpCVBAQU8YMEFyIajH7IR
rNAbATBj4BQBtNnrQQUELOBBARhIUJVZAdw6WQvIFgBABJwE+gAkFVDA78t6VsABCedGgMAgST8w
i7srbIBBbAxhsIlAUGQAic8TF5DBYmQUkFUEGxT/AAkcFkiAABT27iYBAxAoQIYEADAiAL4CeXAA
Pw5IwDN3UPCDAThkZADALBmQEXSdZjDwgAhdpVzU5oxAILIiACRDLG2OD0JWqiR4NnWbDyBAwqac
tAnJA1YixIFoAuV9lBy2gQOoexU8wIFxIiywMQAtaGABAyJLAMcBGcwySFEbkHBWAIyMeyUjMStQ
QQYkdEKZm88kM/XGjNv2QDIjKJA5CRhwAiT48QAaVUAAE2DUIMygpOGMwCkpk8BcOPE4121OegVx
wNpEJoBBFAAcf+OAYCJQABICYAH24oRqKECADTxQIA7YRGwGgYATKAOAHGAA2OaSMg/YL0gtuM4G
/34DDk5k4AEVeKAUEECACtjmgQrwgOdgljsImDAAkADHWRLygjY5IHNSKMACOhGB+nBsXgVoEwDv
JTUEOHF8V4KIzpZCAu4x4i6e+1w8HHApDHAiRJBAnxnIEBsyFMJXGdAbQcxgqQRgIB4yUgAnwsGI
EyiAAHNJhhPEBYAXCKAC4QCAlUgAgaREIDazMQMkRrA5AGwAHqmaFOE44brk5XED4NiY+7QkgQS8
oCoVMAMGBHOHB5ihEACwGwLA4YQTxEMgC4CAIt1ELwBcIJQRyEAEBkEABCAgREI55v2KtwClFUAK
COAEAIRpLrZgAALQ8qYTrqeAE4ysIApQwGSmZv9F140MiVJJGSPaRAB4xsN3/JDDAOmgzdg84A46
M8Md4GCG1AmFADNEwFgKYMNuguOFQGECMoPkLyQugBERIAADEnAbADAgAiewV8wEUkcASCAeZgDH
AyJQSylAgBP/QsgGAaDOBUzmDiTwZlsCV5Vw1CilAEiAE/fCgTM2JHAnoAoJ9uI41SBgAghowSkv
kIADEIARmdscVGdRgA5AgqNSAQoCCjBPRhxgUCpg3QT8MVAAVCABk8odOPBFgkH0IzNA2USyOICB
RD2hlatcASQSgQD25OmEkDjUCWYhgcdVQAVifIwCVvDZufGNU4b6QABUudNBgGMWiYiAz5xwl2T/
zYsREogAowiQuwi8AAMrEMC/LCC1bKoAEicYxAWwdSA5AACACYACBAjgmc7eq1Zq2UTuCOAB4c7r
YiRYJdh+B4cHlG9QkuRVy2x4r74cwAO5jR/VMqAzDVQgAnBo1UA8dQI8cG67jGAAAmahGzdJARIe
iAB8CWivZPCLAd/6DME2sQGLQYACCPDAHTTAswGPYBPPlIIAxmTUAh8EA3gJgHPDuL9unqBOl6ql
bALV2RWMj2gmXogKZLuxBIjGmwDIDAm0BLSqRBMOmwgAAkawAguckAQRcHBuAXCqIMEhALMoGkEc
QBULqEVUz+UUBCBgvrG8YC4JSJOKDvoAAWwi/wJKfmAE8MGPTZxgp215ANAEswKg2UMBOp0aP/Ax
CydkYLoPVAFvN6EAf/BWySgWpWAcMAtGhJIgAUBgBDTwgEqGJQLx8EBAASAF0ViAgeHgrQcCUN6p
BckMATgBCRhhHCcIYGOrrQpt8TLUgZiBBNwtCrBhOxYB6AYojMgAAcyggFj7hY8TSB8AyLA24RrT
AWIiwAoKEoAJRCACD+AwnOvpl6M464TxyI35LKNkRpAADhqYxamkgFG86JkAGlCO+UgpAH+sgDKi
W4A97vwhhJiBY7HRwJAV3U1uk0AKBVkBlCNAQLIs4AVWVcipRq0QsGzRIAFO1bhkORFOqg6CHf9P
eT6peZCPE2tQly1JlfH58ZQT5Cgrb7OJa47ytWCyIYTkOD5JTnSE+EMocQ2KVspCEp5fRMk1Okij
Tg4TGC1GIk4vekSEsnKVOCHHCHmAE3iedZZ43CFlh0naMcL1oVO95wOR+kTOsvaFyN0kbUeIWWLe
FI7c/e1DT3vdT7L3nQy+IWT3SdYPjxLGdz3vIkn82xfvkJn/RCNOP3niay75s6fsLHE9VXcyDsEA
rODqBhGKlvAElJk/5bnz+/xBku4mpcLd6UcveFhYzncIelz0vTeLUFTQgIR0ne+Fn5frUz8Qjm+x
8J53kwooQKdhwX3oarK55kfrFahEABzL8Uf/QBv9cbOIvspXkdG5r/8QD1Kd67mXZTvdhHOtz4sB
igrAg/ix/I8fLebxJxD4UAAgFSSgUkyXNXN3IXsEMV4x53JSUxC9dXx5EoADEYCrdUIFwCFUpnmi
RwZtOnpuYgH2oHt85yxS4C/zUn6Wd4EwFHORIYKX5Sg+AwAWsAF3gBflJALRw1OgwAAx+HZ+5FSF
c4HRNxCnpDpjET2iVAFy4DMDiDVLAOMG4ytQOCZKtgk6c2zIJxAtSBBbeBAx+IUO8S2Z8yHXOR4T
IhDQkRQBoALtsTT44BzYRmfNNiFJIXpDAwAjcG8qOoZFgW1JsRX2YEjoMYjQQTnu4UADgQ8V/zAX
BwIA3jGI3tFXAbCGn8cJBHAUk4iGmLhOjBMAGrCGnPACRwGIb9glb3gg1wEUaWiDyTJoAriGcegV
6IOFE6IAYYQ8eFgUKmAPlGMPAZACycCKReEdaKgC2nEdGpAUvbiC+KAB5vaGR5F7C6AonxEJ8FYG
j2EGz6EC8CMFP6R6OUg5vNIOcyGKSxMA9qAB/tAkRiGJisY2QOEBHFCMANAA0KFQjPBMaUQQHuAz
C4APvpIUX+VXVsUWE6KIbhIAbJGG12GDkdIevhgiRdEeqfh3B1FJ/3gYFSACciMBpDEIEuAvjJAY
pAUAm3AHLUMC1bIJjlIBbOUQB2YG/uCIUP9gOiyjNOkjhebSAu8TAJOyH78jAXOGCRuQRAORAOoO
ELOwAVCjPUj5iAVEkgW2AHLQCbPgK4nxBIcRRwJRPMgiAgegAq+RVRUAlHqGLCQQK3JwAKACZ5iy
SpzwAbZTACpAAi2wAYrhQVjzQm7CJm4kAcJCBlDQCQmQOxoDAAdwBOozCJ3QWFADCbRBABwACeGw
AJCgKUM2KR+wUoMAl1Y1NImhaFATe2TACBggGpxgOs9VAIOCARMgBRJQfw4wMA2DAB9AmVDQAr4j
ARxAAI5yBOVTALdCPJQiAvTYkBKAMkUBCU9gBgeQKs4pMhMgFG4UjZxwFJxwGUCRNDJJGBj/wA8e
YDp8qSjJIAcIsALJ4ISlhC8V0CaEkzE31HuBJwEkoB8AgAGeEwkTMC1WAgELUD4MADvmMgK2ZYML
QFOB4ERvMmpDA5uEVBTeVAGJ8ALPhAEnsEERwB4LsAJyAA+Q4wR3IB5MGDkVMC9QECLBE0eDsKEF
sAIXwDYCMUz4QEM2dQKMMFIDETgKcAGCIQJ1hhfV0zhWJDWOmBbJYgH2swJSpVMpGkZPoFDJlFyw
I5spEOMTEDgagEJ98UByIGsXQFqiYUKidAf8QFoWQiwQYAEGI0Ukup+ZkQFzlUeS6EQA8KLJ5hQB
UFUY4DPTwiuyFVec8HCMg4RSg0SghSec/7Ac8ambOjMurTlX8RA81lkQBQRFDSSJkHlXBBEJl6ZP
DnUZPDVLnuM4AblOR5REbkSehUCb/cMocuBUxQgFD8QJvSFH1gIBmyIAM4QV4qIx37agZIFEbzEQ
KFY2B/BAq1QQ/NBZWSZK3YkinoZitrNODGA9J9AyoGOD/qICjOgmTLkWq9FJn2Qcj2OcUGZV/8SI
ayNDESeScRQJI4BRACAHSTEB8UJx4PoEUlAI+leMGEAG5YNBZXAA9gAFApEBExCjlOKjKjkBwqR/
QdSo/uNtmxZHdyQQtElqEkBvE9OmUiM5VyMABOhB33ZpHzJELpVSEGshzOIxL2QGyWBXWf8hm50R
UBsEruW0bYMQDlZCABSwCZKCJx7kbTY2LhEAOzO4rAJhBoqhJ/7OTFUWCYKhAB3aFoQwngPhaaBm
Bp5TOt9ier1ROtbzGaECBScAU4AiEBpgsqTXcvsZQLMDNDiEANhFSwnQAQ+goUD0BPc1LFDQAJbJ
XFgOKB5ABlZyAhQgONS1ALnit2sTAdNZAfGALwe2AJODAMcFbJCAVxqzAuCWg3BwJYPwAiSwAScg
ABmwAHAAmCLzYkizAtWzCUw7EKmiABIwAg7Qg7RBNqliBgtYQHEyCNDjK1J1ZyIQDtLZoALgRsOj
STR6ixO7AnegT/HgPhMQGE2jFkODABT/VZso5CtGFQ/lBAAeMAhOIAfJQBskFBhnFACDcFBOxCZO
oYliVwHgOE0EcABI1kbDMjId26P6S12xoxaDcHqDIAALwLeDYBzvSxUYMBkbALF5WT6rlVkjUCsE
WACwIxTXSgYE8yXICrEbxLoL8DUvMwIcwA9gu1qDcGDfFj3B1ADhoGIO5VaFsRCUxjFtBmQRMAIB
QAJCfGuzsFqhpGgnFC8ECTRDNgEhgm8DoWga6g92sWkoRQD2YAbbVlEAYMQnoAKzQQBx8heZ8Wbg
5LZ7IlyO1UoB8AIEQAbh4A+y5VODYX16FmpBFnvigRfgdklbLBgBOA9+scdFAWf48J9i/4y1d7QA
/pBlEncdEDBMU8Mjy7RI4eAEZCEmqgoeqkkWxuYmspUbeIEPcAZAAgAHzXRCWYVUBUAAlhEvXqFn
L+AB67Rt87IJ1ZFUs6ZleRwBKgAODZBpTgFABFAV4PY5apEAe9IVBHbIa9MeBSXHaMEz8UAYr2y+
OxUvn7Fy1MYPbLUnJnYBVkUC2+Y7ALChMGUPDqPJO/Gfp4QBeKIz2qYAtcREdtwQrjdz/ReP5ScU
QeiCTnFuzUZlQCEWoucbzXducYVOXsgOawEW+GF/HifReQJ5Cah67OcQ4ict87J+ExOWFEhNr4dp
zfd28Vh/DSkQlENNYGMbriMW5Ud7l/9FOK7IbS+odw7NOtTEH+ezgifoAPgRGeOBObJklh9ydF7h
BDNFcmCRFVVGOFT2L3XiJuu3doWnOgzodjWXe2TIdPbXgcmHOZOXeZyndyjXdlx2fA9tOAyxd9C3
ghbYhc4KeGXBRLD2gOz3EE53hBtt15OHeJAnVzBHkwtAOyaocWTNeIEH1ErogW5iect3OReokYjt
dpRdgZlHOY/de/EngmdneWeXeZ9ddFsEfKUNFOsn2ZTtLI2meW59eZeVd86y1iKd1vaZdEIx17O9
OcJ3e1Um2lttn17ofFUmd/jxeH+NOVgt1q5XeOenhApw2WXNclLBR4CtfNm91cAHgOb/KIPPN2pd
Z4HC3dt7PXsah9bJp3OuuHK/nXK27XvSLXfTzdfG197x/dq4Ld2ydNbsTX9lAd5eSNxvnaXywtb7
nXqvDct5QgaAiQ+5sdeuHdho7d0LsMOYTXRBzHxlsX6YTXZc5gG+o7kQJxDmUeHEXXMOQAYcg+By
WxYeUHwGsQL3dN4UsdmXJXFOjb4L4OF0DdYdl3eb3Ro7JxFOAGfkghAO4C2Id3JyUGD4ORBMXZ3s
p9xE19d3fKhuMkj3PRAbEHuCJy3cQ+AMQW2UWU6FYQEk8NVkzeFFcSUWcAITnuFGZxCRpjqnssBU
ftpcbeOzYD5k2LgIQUGy3cdgF9bXBpdPx6y7fP8cEaOxN8e1AJtQFaYMZx4ACa6ZMhZAAE5GW0es
OJwrGJ87Ft87ECkQPKE8xJPu1iuguUO2CZPWAZugFhZwL4IhBR1ABuZzQrAuEAggOnbqiHiK12L9
FN/mALYjNh/iAQSwGbU+ZJs+aUwk6XgQxjtVyr6DAGIyZDp9NQIRPBHAQIHxXAws5bcuGBbAjd9m
```

```
kLCODxbQAXCmM/pOB9FydFCRZA7wb7P+IUzkZOfxoehLALDejycg42GJBxDQjCm1GWZgAf/ZiBFA
BkkhBdrlAOwMZ/deFcOihU3ziKTG6SPIY5g2tH3xOILhD2Yw67pIAhJPW7YcZC4mB/HwJ8ceADb/
s2rfVoIFp2cSP23o7CYeIPPq4m1qwWXHvHsNoR/99gFIcwAKEA8TszEVUEpOmVveckCDwB5FhZcb
5ikC4F8DAQItYCiBpJr4ogFCsQ8NExcvIAHMCzT7UQDOzjxkQEaBIgUVOAH4KVknIAfc44qDUIz4
EgHB+XlkUzSDkIO5YgbxwAhkcjEUdRcbUAaDHwFnk5nQwp8CVABSJTaFXRBILTINkyuiA7UI4GQC
eAcalKLJOGcdtDFfswCJ8IQHAEk16j2zBAFkEESDcEB3EABkkAwEC++xkVJz9TsLSkZUKhAPIAcC
kAhIO0RP0PvcHpj4Ukp34KtzxQACsOnapCoV/9C+TmQr4YD3uAO5EyAHdnopZpUZ2V8BgHMXkGAP
GJCDD1BCEgAOvdVYB+QoADGBQIEHFQqA21QgXCIADZ1UiFAAQwAoByw0BCBFhIAHEiDK4ZfQDCQV
GE2exLgCgwYFkAAsIKghHAEJCgowVNBw0ImGsyQAUNHigcsFK0U8OQggacM3BUYAqEBCggAEQjEy
IgAgggAAGE4oWDGLwwIGCAAwEhAhKxkIjA7AnMCJBAASDDAqiAAB6goACCo0VOAACkYJZrqScLCA
BKQAnMABUJAC3IETGeYCKAhg1iASJCokGKQBACcIDFHexSeiYQE4ZyeQIbBiEBzRGCsgGJSMRP+B
J/iKZnAi4UQA3PxW/GxIhhE/jMEj+F3AybAIJ2MBbCKA9jqBCJs0VNCAYQIACY8hQ9XgxCWADQsG
mcf4wAKGeBEwNERQIAAAfnLgwDkgkUgSUYCTcLZSgJBkYLIvArrsaggfDh5oSDsC9FrBgwqKcvCl
/CK67AGnVoiKjALyY8gnvnJSwETMKhiKOYwcGGQEBySABwAO8oMgnA1PAxKmB1p6qQASI2AkmQAG
WcCkQcbzR6oXhwJgBf3uGMEx8i4DwIkCmtzAMwIQgGO/htQCQAABFMCABEMYiKCCRFjrKq14OoyH
JgRIeIITriK6C81IHvMLowBAamjBriZgpAD/Ejpp7DGfZknmBE7GU0ADl2aJCpwFrHwCgAA8QyCA
/fxBCZ+/AJjNtRMYCcCJDJA784Cd7hCAnw3sO0CFBwioQLaXlOwJgpwaWoADAsBBQIMXwgEAkmQf
wOoAvQgQgIwD7HNNRy6jfUC9hjYcBNoZd+UkAwEywE8CCgGQg6oOFPgsgALM2MRBAW7j84WsIoAQ
qB+7+koA7hwVYQHuXhJBqgkK4Ckj3qSAhARwGkiEASlEPRIjnzATaj1DJSDwLajAGQSDPc0E8i4A
8JGg18FgQkCCgZIB4AB3MZIKAwIewuCArCQoIDcVINGAjAoYgQI+KQxiQIIABBiEgBNk7Cor/0YU
GQ2cdAlor4AKIIAxgk7AXmCBCgggQL6xK2AXIzKk3USCoLlUYBAGMFhAAgQKFIARBgiQw97HyLC5
AnB6juDo5wrIYExQu3pBgqIinrGhBwZjdS4M7Cwq6Luc+GDMEszaaJAvCQj6OH4sqEBGvwg4lsOC
uIsU3hEQrzztDRipIC1I1g7PwagAA9feAjjRbOeT/NlVAuCSuV2lA/geyGp/RHhABUfNCL4gCOJ5
QG1GWrivIQfqxmCFS7OOYBABpP28SoiS2QSBZPy0h+gyrmcEOJxwgDJoJR7CwcjTgiYAipzEAUU7
DpgWcAIJYKVlJznWAibgAGgljjFm8ECTFP9DoWMh4ASGSUAEzLAfFUzADAt4HT8UAAcSlGpGC4BD
BCikAhKcwAIsAgAInqIBCi1AAwEwAwnCoYIMQIAETYqfGf5WpQiQYD97QkCTDDWLJmHRZQE4AQl8
I5oEPEEDNXzMAkqiAARIUQVIPAGOFBMAKcziBA9IDHNWUMOXaPE0WlwAhZwwggdOr4csywkc3mhD
J3RmBKOaxSObpAI/stGK5+FhDRWwgCvKEAET0CIWNcCJMd3uAaFaQahKmJMH9JBCK2AZskhghgS8
pDMBaCUZAuAAEjwHABbYTyAVwMcFVNEeAEhiFk3Cwwg8QQpDcoITkFjDWL2rjXCgEAJmIcf/CWhA
BVVcwCb384ATIABrAEjhCvuCkkaWbz/AJM5zjnVBl9ETg/bE5zzpqc/T2MQskNkE1/qJT4IW1CT8
NGhCUYJQhRaUoQ3ZBPUs11CKVnSfFsXoBR+aUSBtlKAebagDsBZLjpbOPCZF6UU5qgBTkTSlL9Uo
TBOqT5DG9KQyxWjt7vnSmiKPnzO9CapwulOUAjWnJT2nSY26ONMI9TRJRV5UDSpUhEL1pkNl6kGB
mNKHzhOhCnDqefxRO6WW1Ks8bZlHw1rP85T1oFKtqE5PCtSzwhUlVs3qTF22OZ/G5ykoQRVzGOrR
s1qAQEQFjBT+SdWrXrWuDnwASbcq1sfm/wRVvpHqLBCZE8HaVKovjKpO5cpW0eJnY+pzyJgA89WU
aJFFx3LASTT71mPJ8KOQyUnagDTWnOCDAN9C1QOm6NODxnaubyXucUXFScQ6AAH4QK5BjStVfm6C
hGPFCAIokNecjKpJO5XRTy2Lkfq1zAPQklJD+CFUoTqApilBwHSXKYAFOAEcZpIvQ/FhxbA6YQIk
ldZJY6vP2GItvzcdhIMIgKqcHBgw2G0Ig0/DgE2ghBNluYttTRLCrtwJMmsV1ROj1SSwvvc8eG2s
VjBQW8gkFXGPgTAypfbhecoXSDqcK4SdiqoA4PguLCLDkCRgrrC69IJSiEADXiZP51rAiv+kI0At
c+KBCGhACtBaQQQiEA4QNIAEK/olJFbSlwj4kTgr84BZFIDkJjFiOKUKABnmElwOSCAslpviN81p
qFvCeQFSYEAFMGeSB3ACARgQNJ/JwA/nRmA/ZpBCLyEBAVP5OgnxRYAUKADdSCygA/jZDxIx0yR8
XAQBjsbMcwKwAAtQIADzGdUHSEBCMlNICl4mcZab9AC+SKFJSC4zBiIAh39iJl5PCACVj5mCOXZJ
BF+yDwL+6QQCnPaXBglAsCZwWguUOSNNckI4u/0UXb8k0uFQQGRI8Dk2+hgoFCCBBiTklPNkuQxm
YMDiIhQBC/AjAA2AAydvCZQIdGABcsj/QJO6LUgnx/OvydZAIL2NKjPIQREpgNisIdPtUBW0AVVj
gAYkEBEy4KMTDJDIE+RAAC0CbNawic4EoHBqSBAgYQOhUchpwgi/aQ4KnJBAzRNWhtvBaBOBM9bI
B9KQAHDAIggonQg6kOOT7ITpGGAAJ9yzAAf8FnjjQYn3wsGIC4BdKZGjss+aF5U4EaAxNDnBCuIR
AEignAH+6HS58BGglw3iARFoQY8FYAaJHCDbz3lIynC2HxJAAdXXgZME7EGAmkPiAeCLwB06cBwA
wKYjcYIDI2quOM2JOQEYKMABNSC2CezQICmEwvzSJgGpmItEhQjABtYmgSdQcBMVgIMH/+4zC5bT
fgGHJgAHAoCBmm+gDBa4AOHiQSIRiyplBfhSBRigRfAFDgFitpm94pTpTqhMLeFPWTw8sIGEw4Ez
MGLEHU4Aon/igwGFrgAGJsowCARnaSLgBxCpACMDEp/ZOBEYAdYhkb4oAKh4l+uoEX4gg2zJijt4
gE24k1Z5Ga0DAChgDkbwMCfoBBU7C67gtRZYgPg7AGR6DxJoAfJyEBIYBADAFk4pCkrLiAFsiLFY
PgTwAAjAGqx5AL4xg/T5JYIQlUiYANgZgdt4CU4YIveBAAvIALpbNd7gnrwQvgZjBIH4gETgAAIp
AxcsCM0qCG+CBBlRgTu4C4gIh/d4H//O8IrOwwAz+Am1IIEqVBm2KUHbcIIH2ID9uIOCMLvRmAsQ
rJNNeIsEawgVwIDH6DTvqRmzsMN167zYSJixaiXL4QSe8AoBuI8I6AcS8aMVYMMAWBC5wAir64tB
4AczWMHyGbkFBIoU3D6RYZLGiIcmHDu9mMRjUQEJGAFHKZCIoZERUICJChZOUIQFuAM/sid/aYhl
6AQKKKfMsAAJOJW/ggwEEIEJmIUpHLkCwAuuKYAOYSKzgIL9IIAQ/AsyeIEAHLwKmIVOGAGeO0YN
OTIuEbYHAQBCiIfF6CUtkoINkJENiRQzuBOWOgnvSQASwMFWpJAAEIHf+ooKeIoUyID/yBKBCJAA
CFiBKgTA4CgIWaQMngmeBLsDBaAJAbgDFQgPvzkAZNuAd3GAEQQMEYiHWYi3NmEVAeAEO9CMDNAf
rYgNiCBHDJgFPLGNQAJAHfEa+BiN1jCYs1AE4Oml v1IBDMPIxvCMLkqGWQhFY1mA6CMDDhAAy5lD
tLgQrVCEACCRrKgSNkzGyyGM1gAAr+GHmimIOJmAFVhBe+GE1lkVqOCkAjiAXvo7TDmOuMGHm3AM
NuEKoGCSuoQKOFieXqoNfPKAkROAMhAOMkgEEZSxvOGbhqCyc4yAgBSBZBAAe9ga8ugQ5bEXYTOZ
uTQgqySBHYGCROiHtJCA4BAAPkoO/8uhIKtMALYzDRVIGR3iuTgBBxI5gTIYhFgSkU+5gIjxnvoI
h8EkATjAO4YAgetJAEiomSnMgKOpEt7jHhrZAG6UAiiYPihgFO54gROwPALYhAywwyewB6RRHOgg
g9ogjvuCGFY5AbGgIDNQDO6BgPVLBgLgBz85i7gBgE5rDGFLhkIbNEbQBykggBcYDaoQjmZSHwwI
De5hTnAYiAg4gOicCg7AACnYhM8ZzAi4AJjwxMEbuQp4gQYAGLlUgeDBgPTsRLnZls4oAH5YhgUx
T4iwAJwZIFAOvOBhoHRZAW0hgf/Sjm4jgBV8GWF8j7wBOxrBgOAAkxUYiC1NKIXkiv8FiAdGSIAA
aA18MAyFIUi24Qdfc1Gp+JkmMQNuLIoRUAG2gY9R6aMqmVMtY6AVIFQrUph4KDYVcKJLewm+UAkC
4EZqg4AhYRuNUwsFYCCTIA7RODqMWAABeIHIAsW5iDfIAIcXYCYBSAR8MKdZ8LdSuaSONAlegqKn
uMAIGA4SeIFSgQPmWDSeEYCOm5r6YCOG8ABktA8SYA6FjIAUEg6JWABy8yMSOCZqY7m+eECgWJYy
aJIVANTP6biXwIBWCoBXHJLVvAziIwNOYptHwkMNQIACQtdhahM1YpsHdAKseArDChTuEC7MAE8z
OM9HwksLyBZwUIDTEwB8oLZSUQv/R1sAhsCH1XwXBVghMxgSBDCNlwBJBKAQK+qx+iBAnGq7F8Wq
mPOxmSXVA6VZgqKULx2Om+XZnvVZrFIAJygDKXArs/rZlGqpoOOrc3UCEFPap4XaqJVaiiraqbXa
q8XarNXajqJal9qlekq3AHittMKg2pmsrd0nU7ktru2qlrUnt6UotbWosEXbknqmmEOMsbWhgZoF
e1gAJXueKhmOk2gbqIozaFwttf2qBTgmDIqttEGVTJrIzhChzhDZhiKOcaOoJxA2FBsoyCCBkqCn
U7OrfXKAuYMMKWAI4kixi+LcwbUoJIPblVIBRrukgtJchcIADHBalGLOBSCDwLug/7bZk9PAt83c
wWJbJgzAS35agcm8qhtRXpOAHR3EPqLpigPgBBghAVIqAbxMKO/iqIgYkoSKrRX4AAttGRehKKvQ
CqmJgGRIt4YSACRsqHmiwnXFqUJlx9y9IFeOKD9hyapBJoDhBGTbBAIOnYnovNsJgPDgmx47R6ES
kfilEMIjiAIZCCiCS3MBAA+4HvmoQu5AoOvLCinQP3BwgKApgAoLgDWVkLM5DXsJBwtICNSFjBTE
AH5QT6RQG7MAOJchluX1SS27HcwYCIUQWRLIHpYKUlGJiAh4HQHYPjLwisuqAEgIBzKwj8srmkfC
DjmkN4zghHjgABlSC0ZcN9IAgP8TsAuxZOCG2A3e+LugUTiJYAz1GQSyGRJ8W4kV9orlq5oWWQEN
2IT9wIcNkJbGEBq6IACtywvOeIGbQKbdFQ0ZxD4zGcLC4A8AcICIiIfXOZs2SwhX3dOuKYDyGCbW
```



```
BSwAPshZjJbBmyGABcQgq7UL1ChAAQzZQ2DQXDHEhFUEL22SrI1AU9nDCGRVhqTSGalYBByAClTA
VOCdlUaAPgAAEygSBvCD/4LmV8zED8rEwwRdi5aWqqmeqnOgmQRAwvlB2Kp22AJgGdQEgAQkAzY9
wB1UACRQFySIAL4QABSYwQF01hMQ2ms4ga1CqWwNDCOcAAXgSwFAwjQJACRgEwDMAhRIQAs4wYcS
q6IwQgsAwAEwzAKIwLAuAIZWwATKGCRY28/1KvU4GhMpD/kAF6reK77mK3ykmy9q45AYRymu3M+c
Y9voq8EerJaWCWwQoRdBqMhIio9wCzkeY8AirMVe7E0CikNWIzvKpIN0imzOi7gMbMcmioQQCkM6
LMauLMtiZHmorDXOyS/Sy8DCbMveLM7SyoNCJDymYzrabL+gY84OLdHy4ivE9gpDXuQ2cgrTTkvS
9uyjAG3RTi3V7qLUVi3WZq3Wbi3Xdq3Xfi3Yhq3Y/84kqcaHl6wM1Jril5AsQXyHj5Rtl2CjlzRK
fLSkhNaI3XIt3JbqyLqtOS7terAtbAjuLlbFz1RsbOwthSru1jLuljiufcRkRIKscRCuOlYkcyAu
wmru13KufPCDiJAIU3qjP3ZHf/jH3eAMQhAIPgTsJb1IQajEokSEExBAT/jk6XZH6o5I29JJRfhr
myhkbLBhecwCU3psL4UHh5RjdzytuyhtbGAX6t1JfvTHiNBE7sbtiWRv9l4vSXRv98ZG7lovg5iB
QCCJfXjAKeFFTtFuHnOkSt7EjDFEVVnPaXTXtX7NBPjHl/wcsvjIempve/gDBMBKbGxFBaSQAuBD
Bf9UgBy4lO58G6PtrwC1BwODhOF2ZZMObcjurGyICAxFxL5UhAjcxZUYrm1kjpV85HioCr+ApIPM
bvz+q/T9ybcWwvGCiI6sXG8UhBxAhBQMAiRoZ8XIwa6GLPo9h1F8SQRU61N8QAPj1SZAcTwogAdU
q3M5RLsuywxbhAN/wB2c5I7saeJ6MJXsZO54ZC8lggPkBbeRgEA4OVSIiAKQwWNhBFx5DYaZBWOt
gAjEiwvNigM4AWMZhErYjWp0Uh4ugBmMpyH/BQkwwCyJhX/ygxTIwfvp1AE8nOAsUZR5wEYpBOyk
yGC9sRS4hiGDBUikhgcghAstRWbkykSoRy81Fbj/JTAAsFJKNEQ4LJFOkcAI9BgczARXzEJkqEA4
EEbhoAtKfIUIQEBIGERFOIFSdMdEPEANIgC4WAQH6ckCeIDeFE5aSMRftpDWxRUHoYZmIEAiTIUG
Qx3QhcSgRUDs8UMHFAJN4Fs8jOUYLkZcEdVAeKQ/Qk9EIGBlrRP0xhWQJUQJ/gU4SAC4GMQiZuGk
NIcHMEAak+ljEQY5TdReKcDz3UEL+AcW6UUC225VEFYLYIAI7JEEHFOO+ar+2dV6VUABiIAGfJIA
LAAHOCcnEcAgNCAGDMIpJdNe6d+HcoIyUQAeeJ9wBEAeE4RH49JhRQAUZMDCZFxVmule7ZW82HTq
/1TAscGN6TEQBxxUGC3AQoncAljAHWCAHFip6TCmBkRCAQzCmpHByxxADIKyQFFABWSAMuFDo1WA
E2ArBlhA6pz1OXDELMhBIYSDXkPAYg1CCzydRETC9pgBJDAAB0TQ4ahr7HEEPoTRdNWGBcjBAbj1
AdDOBTCApFrAWFUA3EFC5EXE/QTIBoADchSAaEkGYO6EWCdFkLUA9NBgBdTWi8nYQzREQX2N3FWS
f1yIjQ4EBvSTBNSfwIvAV7TcNDXfe8YMo1x0GuNSGCUEJFTf30ABjGpA8T3JCunFg/1T4gzEC/xq
PESybFXEgv5GAAzCCABdq1nntSoPhp5AYA+G5/89ACTIGQNsDD9sgnPBEXSRiII272VtU6JlzxNE
qlI2hAKQ04MDgFO2GqOdjQbIwfpAwhMYhAR0hCoBAAGE8Xwrj64lBPoAgAgsi1txgEAZd0S422OU
wE4cmEFAgYtLxgkMQiEgNKulpk+RwQdwklzjRQVAzILF3giswASIaPDUaHcmAOAnBCMsyx3kRAOZ
jQWEa6NJ2mrJ+CTL2GwwggQoO2tIQGQIOdxURQSE8YtbROyx6ECkoUc8QOQMwmIdgCIhTUX/RrdF
KkIElDONmhnISwROwhNAgACYjXMk41LlxAqEK3YxQJsRlkqApvWdJ2zkpl3CUi7PSnMHz37SakX/
ONr6ogaA4xkA5FZZhJNGBlQZrMAI9BYcjQA5ldMsOQS+rMABrNiieVAuRvCQTZdqSYHaWITIifAd
pKiOJ8CoDcQmtEsCtHQD2lW0fM25M0IGSEEDPEEDhXBA25CTmUTNjBxcDEQkRGPm2NoCbFIF9PLZ
rBcJbAJnyg2raSAxlfsLTC9B4MPLzJVFEBMn9NOEwM1AyFgCPKEI75YDXEB/xGjQWEA+yLjySPBv
NGuNrpmt2codvMeBvjoTMQKf+1i4RYQYxQYv6YSkXtYJhB+eeMDIaxlCyNGPShp61tGz+OqQ3ccg
MABg8wQMwRI5sRcAoBaiY4DYNIQTYEAykIBD/5heZXFZJ8wCJPChjyUD6L2MRyUEgE/Zru8ZGXC3
B7QA5ynTTjPCdLlXy126/q3AOTmodtpGD2V9QwBcnyfAJkwAPyOc9LRvQrBU1un4BcVR6ijPz6nO
FSL5BTB+BhTA1oHYA5iBc8a1CHtWfw3gxAfUx0wgAHzSBMwTBOBhMmQYyWlGcXTaXDGAglKdBnhW
AmAAA7xNxXCTCGFABpgBFIhRBkwYOGwUHE3ACzD4BDgFzrzAPbk6E/zSBqTEpRICvlyWothoIYyA
FEBCBnTAAqTAOV13d3jWBVUARv2NZIBTOz2ExFkb3WxXxWBQMowAQFioEEECCXAFAvgDsJChmf8D
CyUUiMBhAYZBEaCAY1SABJQnDjZU0BCggjOAChimTIlSQYATBFYs9EBAgwIKABBoAEBmhIiXFR4s
VOCBkaIIAFREEKCTYQoEmxAAGAEOgAUCAcwICHrSgYcACIKaefBAQFQFEUgAWECABL8AESKYCcAP
7opEDBWsYLtwFr+F/hQ4SbvQb15wBEYAWLFArQB8JGgydIkzKIkAASA/WLwwKYGgCzaRWPA2QgCc
bBX4MxMBTmK/agnYI6Hia+JZBKKmvL1ASgMAD+AAIBGBjIYHVKviDr7WDD8QBBIoIENgQd6tCx2Q
WIpAwokJCyGbuYsANwAnBIpLBquq7Ukz1AH/4OuglrFdyZsI4H6AX8ByAqVTegCx7OBLSa+OHJiO
MScgMY+AE0x7QKnEAHgvKikQeIOhB+LL7SwB7nIpgnAeCKA/CjjEIB6cJHgNJZVgjFFGzgY5IYIM
TJtRxx157NFHGF/8UcghiSzSyCMXsgADJFdi0skZRTMyyCP54YSxCKR4csYHSLBMyy+n/FLMMcks
U6UA7jJTTbx4DHNNGR1YQMM36awTSDCJdNPOPJnU8yQh/VQzOCa1HPRNQ//ckO82U1IIRg4TRZIl
odok1EyFEK1UxzDdDDTTO8ucMlNHixT1SJRI9fEBBEzLjlIAAvCAMUv/1MueSSUDUKVJIzhB/6hU
ZXQUOkg3uQklBJZ89TcvYWUknL9e9LNToVJDa8cyCkgJAUgIYGjOSH9MLUguc2QU2hmDtBJcDv1y
gFMZA/NM2V3xErbJTxHIzVwHJNjwVQwWS0uBIMMcVkfAcsVqVh7zBQDhH/3yGAHTpLgDu4W+YswB
fjA1GAAPMNApAIPfMm2uPxXgJC0MjnLRNH5Q2hjlRPnhUFwAWMaOhGxjZugBMl7DhwGzyhXKAYUG
MUOhYUO22oEXvoMVY1gLW6ECaGvEOFqZHVZgYw6HbYkrqf/MUYqouYbVZkoDeLFtvBSg6KRyHwbA
gQggMJpeDlfghCkOc4QZ0qPxMi2Cbu22N//HojFGYNYcUTJZDgA05FAhCcAajOS0p8a0ZsqjzVGF
TQKAI6iR0XUAkga5+hZdnAoI6oEKMIsA4I+deAsDBBxAVl4PIOAERwEJSPOkTfhZS1cGK1iAkQz+
cxcyMoYlXcIn8ImAH5Bj68yMeI6aDoOho4ogEcwQIMEBfAhg5IBw7ibAg5SkiOeEBSApSAPbo9rE
DAzSEgD8FOA7tvMKAgSwCRV0gBEEuNpCFiCHAizgPBiwwFc2EQF8cOY/1MnPZUgAQAWE43CkS8FL
PEMu1nDiWVKAAAbkdSwMKLCCZvgNIzYRiXBgRoROAJ9JkFKBO4AjOzBkCngAaIECEAADjFn/CyMe
4AAClAUfAmgibGBSgRakhQQQOMpaPAOZwWhrAekTACMS8r/SQGJ7jEGeCv4DCceZrldXJA8DdPUV
LwZAQId7SreKKICREYAfJHiAFKbIuJRYIAMSoEoZbAc5HnmgADpxgkjuwLI7aGB3GMiAGSZghgog
oADxYFAHBnEAfFRgOBJ4kQpE8IBBgOhF+JAABALAiCOmAw4ewFwBBMCQApzAAh8IB8jMcIcVDCIe
UvhABB5gv1lIQJQZIAMu8QCBBXACDpW8ZQQgkQhOQMAMEmCM1yqgFVlCYFUTwEAlRYBHDiSiABkg
ASTgsAC0QKFKFYCHKEkwiAcqAAR3uN8g/whghk7IMhnweJEFoHACDkCCBN2RQmiggA9dIqAC4CBD
J8xQSgQUAh9yiIcABqGBOyDgBf1aCIPAcb+BzKJ5zJwFFPQ5CPwtgAOeYUoAOHEAAWKOEwxoyEpJ
4IEKeJF2XZLAUkqgQRKcgBEVUAFCZ9EA4e1Ppx39KAIsAIkFLGxuANwERZ/KHe1oQAQBwABV5KAB
95kBCunLgBTkMAEGhEwECxhE3hYiSp4SM6KD6AQJNhAOmnIHAiqARAAmOJDzmWshnEBASAEwiAwk
QJEzQgADnmAekXAimVAAajlzw4i8XfIE/UJAPMwAidsUgEIkeQtHTIMSBgSHAN/BwASKGv9CRTDE
ApjrBBkksILfKoYBAlnIb7lZoQzgY5jyYCUGnFCIJOzOAU2EhABsFxy+MCAtg6CKYNxHkhJxAhzm
BEAEJqACCnREAy/wFd4AkIAHnmQQiSClaWrECF8xZAX9ms57BXkCOMghEZsIpmpnsSKQzYIBGpCA
Biz8gAO8YAIZEIovkzQIfvCDEe+8C3sZUADUUoAjGorAcDlwXtYBgBHBVEwGHqCATlTlBKUs6ELO
NotOIEDEC4EvAApwABJwIgKzWJICMLC7XT1IKQCYRQYikC27PfatALjDCAbxLPN+bwEH8EeBNRyA
DETtLKzEowWSAYD/4Mx230HzHdyqT5b/FS1aCoBClPkZUl3tSCFSkMBdCKIB0gIACvUNhyNhxcSP
HUCpVp6gWBfwgNfAMgAqSAQktqIAjtDYV5w4wTtHkGnJyAEDAuBEC+S6JIkt4IEvAAceASCAbA1C
BCvgwAHg4IQCPOGxDzCnCEQzGuw8oLgLKABVxkmCA2gAKIoGBweiUiMCxM4jBKAKGbJlhhbghSLF
Da0IwsEI5KCkuO8lgAIcvIkCfHoE8cgbMGcRZwMyYHYBcIIEErCB/1CoKg8UyMiAOUEAbABZ/B7N
AxZQEOy8OLNHESVDnpekkDlADiPAHAYgoIEG8SsCCLhDAi58khoDoKgjSMD+VkQbEuBZ/9l2YwQc
8PNeBswiW24RgajTEtdRj9ID8fArPwrMIPRAjgwiWcAT2p1zU+fnvTj+eQFYtYBkgGMsMCLDAZTS
HX+MgJVCikBIJDACJ2SgtJy87AUqgAEQwL2jCpBABSqAEAxUgAObYIg/NuxAkTDkrHCAGstpIgEJ
6DQlg2iB3fulAbqDZdx0fsAdOFABo27iA3LtRAAi/pZOZD4copQABoBI4wqkfRaQwM/qL1yBEUD2
sndY/QlIoPnOY8BXKkgG3i+QqEEOQAEEaMEFMODWMbLI3SYSgCgLgFNw/p5LkPh9OFZQCBWIxAnJ
cIIIOnEHAThKAY6XwHlaEHe65A/TsP/XAAE24N4kQSECBa7ABfRFEs2HNMeNHsTdM8Cxnmvx2IgR
IKF5EAAKBCABkkHxJGRJnGAQLiB2GGBWpOgEHEznAuDuLmxyTuAOBqHztuUA6orM+gXO8AEKIKEF
```

```
xAsAQOAAVm8BViDOqM7ULgnvFmD7FC3dMq8iECclYgw8JCACLmADyIpH+GEsVMBu3ia05kIB8KFE
kOIBmEIFqDBHxoIpUEInrLBcWqJETuYyfmMstiYAuFB0RkAFYGZx/MIKR4JyTAaIWqVEVGBgsjAl
nkAK+ahtrjA+MCZCcqdt8MEeTMItMAYMU0IN7WYsYEZsNsRkIBEAnuAJ7CGXcMk0yCD/AINiY/xh
CR0A9KyKlNLkC0+nDxVgLN7GOE7HCZZwQx4gtFBxTtxQDBXAJAIgDedicYwjt8jwFE1CFWuGF4MC
ZFIiFyPxFoPCFqnQJPwhFj9RalqinN4CRxjCCnPMH3IrG6fQknZsYNwQKRQJdU5iJB4g9TQFVOBm
SDAFWBiiIqLsHV8gHuLxBWDoOlLCL1IlWr7EdQQFXsylMGakVxoCsODl/9KJWs4xRuomVIrECRKA
VqgkUVRA8SSuSFRA4AgjRj4FUFQCZhCySRoRXNCRcijFCeDiBMjARiZgFiYgJdEit+glUvKRVADy
HmFkIVMDIQkmJnHlL3xSJ2llYDSS/1GEskemxCP9EVyK8iQeZlgAo1PepSMTZSchciKbhSRTwmCC
hB+9BVR6hl5cZGZE8kfq5kWQEiLB5EXycVo2Eh0HRVoy0m7G0lDO8U/IL0Y0pCdjskgUgh3pBVGo
Ui/Hcl4cRixfhycpRU/aUiUQBldQgh8NJS/Rkk+mRUoQUx2H0kmE0msysy6VUjF3pC5NJR0TkyhN
kzIPMzPR5S4/cjDjkkhYs2CC5TRlBCDV8u+UxTMzJUwWUi7PUTQpK1xOoieDBFMYwh7MAB8QYINe
MyF3pAGO0DkfhQQ8hDZHsjGDMyE9MzsThZOwcleM0yqNMlGM01KoEoIGTTqDJH0uxv8195Ihsico
BkZeXuc8d3NMKlNIIMVR2LEuf+YtjohmfHM0MwUzZgU062UqE2Vw4iFqcnMo71Iti5MlwuZeuuZX
/IWytHIhUqDyXidVNpRN9BE4Y9IvlehlWOIuvZBNWHMTCJJGbLPP5oVDpGAQaktuztNcXgk+ZPRP
PAB57NI8kg83m+TmmBJVgsQJJmYl/FJHnCctsgcu9oMMFoI8gEYFSMAeTsAe4AhAguQtPKMrMMAD
/KGMMkhASEABFgAO4MLf2NQCEGD2NmETPpECCEAK8JEhZsEqEEC5GCMAyOCK3mITGMBX3mN+pCAC
NiF73oOsBiZ9CABON4ExOiPHOq3/AxQAWaICWQIoUDGgRBQ1KKxiflRCL8iAj9gUQKwiAvarHf/j
MojDAhRgAhjBeIQCTLMkhNJCCjogVBkhLeCUACgAkaLCAf5nFgAAQwhgEzIVgKLiMYqmjBQDfhQk
OQQGMjhhuFKCDD7AM7LniphyIbJHXsiCAEirS9OrSIqIMWaCSstIUb8if3LncB4AH1BkAUIIEiLA
XRgCIzQxPx4yQPAuLSzkvcbiRzsBA/BBSarTR8xgEOCgOxDgAyhABO6AVReApirqTico7f7q6SSy
qEqOAAqClW6KVS+2r8CJIJ4PCiagAXbnY+1h25QqtBJl9SIAD/YKIU6PILYJjyAh6xzAjaMWgAxc
NgCgYK8GoWjwgQOG8AOUFgAGArl4SjSkAKcOYxNOoK82odU44AkQKsvwx3FUAh8wILYQwAPo75YK
igQyr2zlZwMM6QMEoAwyYHceKys5QSLSjZoGggLOVYs2QQLS7WIPcBNkr+mubZ+yDAKUCxyuZprY
hAyMigwYQbDsgTsetiUloPVsjCTDKhz+jRHIABJslivg4pYCQF854Q6MDQMYAQoIzm42IRm6ZBM4
YGMtoBPkjQEuSZDC9gAegAMogHAZgXDD0gPuwJDSiQA4wB5VgAMwYASQCwGM8OvwA6CcQA5E12P/
eOR4DyNm5ewoBCxnVKATsMXY9Glo5iQCKaR53gt8usXJACC5quxGrsxhbM12uE4CcMN1VSISRiMD
cjDtoGAhOGIDnqD4AgqAMIACTuOkduwBFO5FHADrAKBBnAOBkqFXCkC5MOB8hjCKOid3/+O/oACU
dKoAJhVGVoCjTuCzAOAE4oEMXkCOdkUKykkAnMBiXpAt5A872GghRCAxEOAADoziXoAEGOE7eqt+
aY0xMvFwAKC91i4AjJD3VCJ/mUARIo4CqqQA4KCBcqaLVeIiq6LCcOY7xgVZdudqHiADTiClMusu
3GYQcqOjKqQAIkwtKkDUniUjsiwRnClbQKYn/51AqJxgAxYisVTC1KAM6zAHAOSgBTrtwgIAoWa3
Ukh2MZygnGhMAEjs1IJpAVpAlowtDTnKLAqN8RbOAuQMAl6i6qQWNDzMdvLXrcAhUyNhAvy3jMJS
ahNBBVgIBAwYge+pRFSsACqQVWznJERALWzUN1WgAJ6lQZQLoMLhglSg6ahwfLaNo5iYASBgh6Ag
n57OHgSAdlJiAfqqAnCvcocJA/iBoOAHyO7gYRdw3B7gAiaVd4rRdBuZMWahEMggmKj4BXDuKOYs
GXiqfOwnbzzgAFYgAzSACVZvRcGNxiAgMJ5qb8MBLAgIZ4IpSGAJJ+iYmnxy+/AmA1bAYvzKDP+S
4QlI4pwwZhAGI3eFgyMUIVP/eBBWwM9kxTeQq1+crCYB4A0kAPSiuZa7kptS5gASgFU07NCMzTQ4
CSESYSP9qgAmiDtImsYmQOskIMaEjQEF5P/0ZSHWGXNyrwDMiQmW5EakdgUioAUEChzMwMNUAICy
rAJWYJqWiHH6CrJWAB+Swa0UDyHuunXBobgK4FPpmtHMI4hhBYMP2K84KSIigJNSydkkQMBOTwQM
DwoErvg4uyLgWpIWmQDkIJ7fK3YU720TZXgxQOAYhK0j4geFwoHwOPf+LxFs5yzypkFXborqtwGg
z5zoN9+2SQMcAAo8TCWk4AIOYAMggJQ6OOL/CoC0EOACBOpFYUUCDkAKMGCxJan84LqS/s9pR4AT
FM8IXOXckFiUDgBzPABHXhrJ7uA8UtucDkAApGADhjRaWkACKFjxhlQ3xG8GmVkDLuIEpu0CzOn/
PjUZmDNTnAAB4IeP4iNCTGMBLMCJUqpENONAq7GMaqKMnAAAMAxWZMcSN4OPYMbEyRb0yFYlcuwU
+RBm+tQJV4BEltC/qqPrFCBL/G1OHEAK//R0XvguVLxMc0ItVsAJtukEHgDUAiDIHODK7XERw2E0
PNwK88uTX0UBymAx2uZxXjg6HeCFQ6uMHpLMqdDF9/A3dILJyVYFuDDHNE0B5BtGygABXi5j/0wj
AbAvxBHACXSCVE5xNDTghV8GZVQgzIMCH7BPOzJGCovxhT9jxGG8JeY8JxygAYS8DJAiMcZqM1Fi
wzutT8NxrNTCccoUVqTAH6Z6y1cAMzJAMJ/EAUQgWb6kyvBTUXzEgSKis5LS2FXTSMpgA7jSTBZz
SAYl2glz2YVELJidI9/TMrmzNkMTT77TSILZPbddI8WzTyQl3MO9Naf93PGyUEpzJD3l2ovE2odT
THISSb4lMEe0NYVz3JtdOhdlV+DS4HX0ULrdRwIlT3194QlzSrx3QglFBczReyHzMnHFw8MVSR8U
OBEULN2k0xx+4JMyP8sz4lVzNCdzMlHC4v+DEzjxobOWUi8fPh23OzAXUzwHBjDwwXv38t+BxDGj
6EcQoDSEHlVOQkOOvlz4VSgBwx9AILRmAkbacKspxWYsOOvzvTA5REnhCB9MZevFUul9EyryfWmq
4ylVIk4CAwNacTAJpig9/jZ98+49Xi7ZnjWT9So+9aWSzO2eXul7PlLQwuPR4jWcYCsq9BSDwloW
lEMJqT3ZRz4dxQq1vjVfxAkwwB+kgHG8d21I9EX2gzm7iOIgRShfozH9clJGM+V4RCEgLlVIJfV1
PDrpxUh3hkirVNiB/kdoTS0Yp95pZQW+/DcGYejjo19WoJIe9CzbE+9b3t53BB+q44I4Ixz/cnRe
tjMRNjlnnsuGGPMkZiGYmExGIA5GtpxxPOBzz+UnkYIxFO5Cef8vwRMpGmBxXBnbGwVXAMIfgIEE
CXYKUDChQjMRNByQAClCAEYSCmjA8EAFgwCcDhSIQEaDBIgkABCQAEUCQimQoBDw0ElChQcBCsjB
gK/ChwMCASwYuUnByTucADwY1GmTAO6QKiwAgOHOBgkPBHDEAKmAPykSREBCQKACAYQASHCoUCEA
vgKQMEwMt6JCCwQDC3ygGsGrxIGDShzwMFIOOABmZJYcGBfiggdoSTDOCmCWzAr4AoQtYA/BgQol
FSjAUAEziQJoI2QNEGBQ0wdOIHB8MNAJ/wYHYccCUCASEoHRFSIQeCpBhZQKkGaRIVETEid+jAQM
ZAnFdSE5Ep4CAAGpBAkMB1rM9PdiZAALBC5WcAsAX3lOWDHqHMT8jkUHgypAQUCxwgoMCD5AIlOB
URgkcltckJTEiQQMEITPIBJgIAUDsBVAwlhmbFDBBok8QBoB/JDRVQGoSXBTABRRRRAZG9yxgmzE
UUAYJINAoYFCNg5EBgYqQLGABdVV589RGjhBFQeKAOAbPlCMMNoCg1R15EAYSORBWh9pEE4CByBg
hlYDKVABAgtw4CQECkAyAgYQAKDBJgtGQOEBKkRQgRMWcQABh04i8EALUhSwAD8EYSCBAv+cQKBC
OClIMAIn4HgASY2IVcCPjxqU0QJsALzRgk8VhFNYAE51qSmkAUSAwQpQPBGABOBwdMImKr0AQagF
CEDGHWQBQIaIGBAQgQQaDHJAqwv8ys8mGaxQwBMVjDDQTxrIoYECAznwqwIBkNFCAA9wsoACd3B4
AmpTqiBmAeEwMgFBnDhHgFsvxHPtJvQyUiwGAsySwQMvKIJAAQ9AEuhzAlcggJ0IzBIPP4UB8JFv
HI5QwZorSACABHSROZqgCljwwAkFJLABTe5GkAAGJwwCjgpoAqofSylbJcKxBQDg1CYY8HOqhALc
PBAC+Ej2wAYrSHHHE4Mg4EQn/PROYOH/HuR5BwBSHPBAsBEcNQI+wg7yFBnlQeJTAYUtwO6g7ZKw
IAEQ4PMCUwggsKCgKkxVgAQrOIozCXxDfAIAFmDACQEAOJGMExkovYADGcDxEQIYnwSHA7f5FpmE
L2DQyQKcIODBggSVwQEATGDgGcuxBcisBoyNgAcDeT8xkAc3W5DMAgESCRsZGWAu+SadZCDBBPwW
xEloEkQQgQD8NA/AARZE8pTtmtnDAbSHqOQnCYICsPRAs0DAjwqA4vyACLRDNQE+EMi9iXNSSqQ6
3etHsOaUSEKwCSQZDCIAH3FIU0YZxBPIVBM4eIAAChAeaCKgrIAFgEyeAgAHwsGrAhDA/znW2k8F
5HQS7w0EEhWRABxm8SsBSIETkHuCP8y2mQyQKW0cAMcBzDAQTrSgAAXIAEFQmAwJKO4BCuhEAgwI
AEgQMGo/xIAGyLanB6zgAyNwVZWG9ZQkdUJwg4hLTL43EdfM4gBhDFYAkhEO3FULAA4Ik7dqMpgw
McI1D8CAiHR2ggOMYBN1sohT4IiVFiwgAFKYxR0kRRENcEIRBBiESDwHBxJgjCAke4JkEoGAC0jK
CZAYj4hWkKkNjCAAT/ieB3TFiGV1SgUXOAG5ZKVICIxRAU7Ylg8JQogM8OMBzBFAA9ckgXAUzkTx
WAEDnKW9FSQjAAsgwUxuc6gHZG1N5v+bwAo6cZQJ0IQTJ4jAMl0VvOGtBwMBwMALvhcsqETgeRnI
0RttN4IDrM9qhXhdAkM3glnAh4hk6ATCcPYUCwRIAidQwAPkwAB4DCQAGyABPJ1AgkjxpVy8BEEF
qOU53aVxA8KMACGzO64NgCOVT3hAvFwnKSU9gASJK8AIHACFB7hqAVAAIxOD5lEy4gMDC4iXW3pl
x3ISiCE1OdxuNCgAwyFGAiRAgHNIYM0MNHIBKpCAUsuSAX0FQABPAdcCDoABCmigABjIAAIMZUcn
jgUDLxUAAgYxpXiARV+WQ9IdIOBIZp3EpAtwakHKmScCZIATexmIBAI4lrXwYxZkpUD/Xc3QCQgc
YEO3HAOGUGcaAsizJpcdwQrkNxANMOCyD5gFCfxBgsARYAUcuuxiVMKJ9dXxKlz5DlhJQIK9kOAA
```

```
FAlAwMwagQ4sQINgm0XggkahB8iDcAS6zU8QEAEcksCXl33BAnoaANcwiAAqYCttWKvBq7CVER+J
R8oScDi3/KwBCihAMgiiAFSpKWW/IstwfwWb8ACgDB6Ckx2lSNaxnAAOAABXa0NyWQxg8Da/KkA8
VFCejAFAZIxIy02/BADUBOk2CFGBdfhRyxqpACGoAQDtHOAAZ4KFbASRpj9IjBBpakBQ+FjiAwYJ
AHuQ2FrMJDGOMzzKAGjLKAt9wAUE/xDRQVpnIASIR4mNQhMF4CPD81xoLQGQgCVqKxEl5seUATCC
JQPAA4Pw1kAI+ICXhvHJzITNkwkSgCxnOAA8g3GGBymUAvAMvlPu8V8H+WaaSOrGcO7PAzTw4l1B
+c6KNihNSGwPIMPGAah5M6X9oQKBlFiaAODHYjLGnyPX1aDP4YcDGFBYNo7AZCpIgLXgTNWB7Ll8
ApEmndmE4xLrGscKWMCXrzUCFdSILLBRQBk0oOgLKztqUMuwbyIgZgwr+9XLrjYAwLKbhFiLS9bu
NhMXOK6bIqAD3lb2TlcwkGYv26bqZna5rW3TqEkhqRNI9l83gRBqv3vf1da3rJfdbv9+F9pGddU2
tf1dbmv5W+EF13bABW4jaj/8RvG2tsJvY3D4YnzfCLcRPzpOkInfaOIVx7CQL2xvhYDc2iIviL/V
vfKovfrVKYd4QqA2c5fbfNkN1/iNau5zbROk5zeNecKlvfOkb1zpTL+R0Zmo76c3PehST4jliu1t
Sje86suOutB/rYARC/zNEi/I1gke9HczPAA4n3ra3b5zrkNd7XDvtgNMlnCMIOYwNyXiqQjkAL13
PeIYFhFBfmOjNyo7Ah4QeouTDfKtBn7nWSNgzuuOeZsb/fKZ77xCslbljxP933DewPoaQJebLgU1
BUi9sluebsffCAoFCVPigaZzhRT/INw2MqmkkL5OZxMe6V5XyOAs7/nkd/vpgvd28ZUPXwx8IC0I
aAxBbHee2ghOJoOswBM2QYIVwEEFBOgeXSRTASdHZvokIF4yNpEeBixPATnKfm8I8qfenIoOqKpA
4JglAR0QAJtQAA3lDxdje/7zTxoDFVZFFhhQAi5lFhXAe0/AQ4ajP41ETp8RGhEABU7xAnDlDxVG
FalBGk/hFj9RAXBgLRxSHeXkYFhWH/UWAYMgIhXGHwWwGa43etDngz8IhAD3MYMQDi6TAIWhPRiQ
OBsAAQuAUQsQARshFgegAfjzAFAQDgKAEaJUAGtiFCIADw+AJyMQkQRAADnRKAcQ/xe/sUWC0gDX
1UkFwACqcgLOZD4koDti6FYVoAEWIAU6owBQ0FQSkADVgSrk8hzqOwAsExea8hHDdSwn4AEZIBsC
EAGcMAJv6CpFFAFPcALCYk4P8AGJwAhO1EmEJWcOACe08yt2QgIJsAIroCscoEtkoEd95X/OFIS7
yIu9eFP7IUUBAj6dEU4MOBOeUwHLJActQGBkkCcBOiUJOEkCAFp86ADqAgAigABngQE8wgiB03oA
AAWvJgUEUABLggEEljRVVXOVAQnhcD5KdCwVUCid1F5nwwHJUAB3EG6iEhkZgBCqAwAYQhrgkI6c
IB8XYI4EZhScQES6MhGGkwB1Uv81XeIE45I9BeEE5zFBoYEwsnGOijQY+oIY/RIANeWLKamSu5gC
8rAAnMRMB/AUDsAIR6I6HPBWTVUsHHASGjA+AyM4IwMJMvJ7G+BrBYBuNcNNPMMPKgMxBLZFhjUW
cvAE6ZhE/MAVXwVoigQPgZg+C8AvKtA5fxIO1fEuPPNq6RMAo7FqFYBeOAMOe8ZNZqgAkXYr3nIU
Z3VI/hAv9EeJZINPd6ACgwAHAZBSVSUAg1AuCMEBVnEQPDUR5BQAZsAJGiCWRlfK6mZmzl1MMgR
g2UbvBI4/AE+IxAvnLARlWmJuSUSPsEIceEbDDkRDAAWwPEAklOKczKSGFQBr4b/KmGRNYMhLGtx
KrxVHoinEmpCVqOih24BKERCANYla4eSMoOFHoQxCNEJYeMnAZuTCFeDAU72AgKjEofDWxdUVYRB
AA/wXraYjp6xHQUADitwAPHwGxEQD4xwBwrACBhEfr9iBgTAKlZSErDHmQeKoHNnFCsgKGLyV7Rj
MrzkACuwaliZYcfGKm72AB4wF3nzPQHAWibjBC2IADVSGVJmLUs2oYuxSzHmD6TWawggZPjgGU+h
AGIiZE5wYwhxZbdpefiAZytgHa/mBHzCJiSWJWdWohlWhESkpDVibBlGRNaxACtAFiDaXD5aIBuC
G2ShAXwCoO+BdwlKpmWakhCWH1Sut3xmupLNx6ZvCqeep3hxSqd1aqd3iqd5qqd7yqf/feqnfwqo
gSqog0qohWqoyeemh6qoi8qoApeojQqpkSqpnbdwQfioFpd7mPpyGHdxpNcTJddsIOcxBAGqnRZ7
SPep1zKpq+pznkGolwpwSRdzBsqqtSpzFwar5SYQ+iYQJRd8g6dyENepvxp8ohpyThesuCp0uWqr
byerxGpzJdeDlDp8yCp7wBd0L/pu34Nwr6atlXqrfzV4B3dwpwqtzTqu1jJqSmcttIqru+py8LqL
r7aut6EAPaGtO6qtBbGvKuemnfqtybp1D+eq7VquF3evBwc1HzdO6carF+er6AqE1BaxHGcjW9E4
82qtx1p0ChGq8BZ746EpUjet5oqt/wo6dOX6Y+4qsRiWoqARGtEmcMGipphaEBGVerJoG6mBERIA
kSh7rvvWIORkFodhLSchswThBDaYIrYHfDslMJi6CVBwXITHFfTCZE5BeCbCVAaDFqHRO9z4e2Zw
HgwCGslgG4yABx8ABQzwe8w6qZBSAsnAASXwAUbrGT2RsKM6quaAHigVJxQAhJBrtr6ogogKIGn
cPKKJBEoKIxQAlCAXk5QAmkBCR8gKT1RrzjTSWinqvdqr4druAU7ECrgH+VUAl44EBVQAoNBrAvw
AR/gLiXAQcGnAiIwihhmOaBrr8MKgVGiqsHbH1FpQSXAkIH3cS+6kSVQAlFiAbE7Ev8bEQEl4E8l
gDHY8gHUm291SxwlsDOp9RMlcFUt2234sAGtKz4lOAK0w2koQCUlm+WU2mEUQIZIGK8tFDzy0Z0
pnVf8mYFsQJ4sEVtZLeDQQapGwCXaw9omb/z6wBrOwIMy0Zz1r51xmOKBr+WY2qVAQAqgA/c2sFO
Cr+KlgIfQLyC2y6ooSOlBr8eU2kujBoFR2lQMbgNHG8eULkEwbqu2OY3Yy2bMBWOGzR4OALfgw9y
QCMBIAedkAhOIAKbULmC8gYfUDVOQMS7gsMSULHkeyMWoL5kcbslOMTWmx5QMMAeaLf0shZ2KwIr
wLzM+y0fEDgWcL6RS2CzUL9yUAL/kEAgbhwgo8pjF1ACI6AAayvEnPABjXMHd8G8hlNV2Zu6KsC6
zFsUlMS6OOfIMzII2TtanLwSEsC8UJDCS3wqQkwQdVuEm1y5ZLEVZjwoJTAB/MAJUKDKWhwAIgAF
I7oJ2fsBdIHGlQsbVbXHeuwcS5u9X1EQXtwpA1G3xwsJy1QQUOy8VCx2nHQQBVYCgeMZe5y42YEA
J5C6CjcCfREP3srFynrAenYtrDsClPteIAAFVUMClzsXvTzPB2CQ3WW94cARrXvEHzABEeAfDOAG
sSsADFC/ggMFZGQjggsOXnyOyaAAS+xlImC9As3HHOJQiwwXsTsBzdXBcvABBCC4/1rcF72hx3dg
GiUQLsQ8AXrcNUSsAZALWjjDOtPLAHBwSwNhAQI8u+OiuCpt0S0jAp3wBDh8Bx3AAA1gBiVwB2aA
OFpcOmgcy33BAATAtmDkxcJYvK+SW3JwByQAB4hkyj39AS3QApwgKrTrgR9gB9fi1N9jBpC8vrXD
vMmwROesbPtQAvqguXW7ACAwt+mxxIYSuXcABSUQD16sGwjhxreEyOAwz1gLgQpUAguyAHOLEAtA
O5yHxxkwteDQBIk8fR2MuwH5AYORNXSiHQCgxzYlBXZ7B5f7AQrQF4Nx2xNRAmIy2CZBuyd5ENJc
EKiM1A1REFIABYAwu+skuIEDxf8C4AAbYNS5zcwl4J+L/ARIDBsQKAAasLYicAdwTBDPWzXVDQ6w
m9ixSzADIdzXBgUfcIHXZh+dkB2doQAlsAEeYwgpkR3OIzh+Aberai1mgAf7+Rwi/QSxjTFKMg8O
ULfxOE0RAC3OtMd9mNDYHA4H3C782boHTC8pYLOlR21SUODSTcMMgAftgg+4a2qbrCjzAQl4ONrv
/TOCgdSxMgECuMlPORco2LpePEnP/QBE/BmwXHssTX+se0hKu7azCw6GjUGQW29LbA86sWTnSyCS
/AH2YMKPC8sPQL3MQwZwAEYgXt423TgP0ABNsSEEQdM8/QHJOGwBUAZimLtstMf/AzHXGOOEq8JG
jqHX3nY3qdtpgutD+KC+H5LnkOuFg8Qa4GO8ma3OggsHbtwtDoC7C+ABH+BDUhDnAWBRWNtzYGK3
GDMLhkwL+xUFgBBpOw7FECAS2ZxEJfAEk3cdTlOjOmTbqY3hugOODkC9T0Dqg2TGA2jK1c0nRATK
rnsb51vM2QEtDg0A0uwAd4BNkEsvZOA5Ca3MBDwBBGwVSEwgzBR7gSjrqlICZOEAz5wQz83TRAxG
CYAQCBDFX5LovDLYA5O7E7kXAb6qCCDj7828x0TAHyDSVXNQkcu2iTC973OQVxi5FvgBBNYXHyDj
zuHhANAAJQAfTb3MuSe4Rk5T/3meHpcbAA4gAR+QRrEr8BSfxtFnt+9dzhyg8pDu43QBgYa8M/hw
xW+uw9bd8CasaP2h8EJcE8ar21MOCU/A82w78Sqgx+8NBQqFw2yb2OOyvQrP1Z6x6ZFrx9EsQuvH
vFTcV9MHRn2R2GiSHiIg43brURYt8PKDxwwd6MCqABXCeKCc7qciAE8AB/1YIVyVYWLjUQNRBgKA
byuwnkGzQc0lBdDmbGbgDw7gVlHzAPEDpQuUeikWy9e2CVP2hxPQAAIAG/5AAYiXzJtAAIXp+bBB
HrCxAC+gKZ+F+qZ7SOUIOoLDXdsCLPbmAKe/CdojOAIALY+fCHdfb85GAAvGD/8kAAH7EGnPQQAU
8ASQbxSqPwGSQm32ODz7kJkJUY4EoPp9fyokcHFVuBvf4wAnYIaq3wB/Lv4F08GoX/f9lhAD2C15
+q8Qp4ohDhAABA5UMNDgQYQJFS4UyI8hw4IIHT6UCEBBRIoZNW7k2NHjR5AhMO5EiPFhQZIiD/pT
uXEEHkgkPJp8yNIiSJsMc1Lc2XJhT59BhQ4lWlQhOJsfaRpdeHHBA6YHIy5dGtXqVaxZtXKsujUh
Uo8pOyogibFrOpNRz2Zd69XtW4JSr64t2Dbu3bEGqeKtevYiWrwQIYINLPinz6lNBdKF29jxY8iR
JU+mXNnyZcyZNW/m3NnzZ9AUoUWPJl3a9GnUqVWvZt3a9WvYsWUHz6Zd23blgAA7
```


Hint: Be sure to decode *only* Base 64 content. ***Don't cut or paste content before the R0lGOD… or after the final …bLgAA7***

Note also that you are converting this encoded data into a specified content-type (*i.e.*, a GIF image); so, *some online converters may require you to save the decoded base64 data to a file with a .gif extension before you can view the image*. If one online converter doesn't work for you, another likely will.

ANSWER:_____

Question 4: Look at the base64 data in Question 3, above. Can you spot any combinations of letters within the data that form English words or names of three, four or even five letters? For example, I spotted "GOD," "PILE," "PUB," "NECK" and "MOM." List at least three others you found:

_____

_____

Think about the ways the occurrence of happenstance words and names may complicate electronic searches for relevant documents.

Question 5: Find the Secret Word

Question 5: Download the file at http://www.craigball.com/Find_the_Secret_Word.zip Extract the two files in the Zip and use what you've learned in encoding readings (along with your use of a hex viewer) to Find the Secret Word. Once you've found the secret word (by solving a three-part puzzle), Supply the secret word as your answer via Canvas.

The secret word is: _____

Hint 1: Not all data associated with a file is visible or intentional; some may exist in the slack space—the unused portion of a file's final storage block that can retain remnants of prior data.

A Note on Slack Space (You May Need This)

Files are stored on digital media in fixed-size blocks or clusters. When a file does not completely fill the last block allocated to it, the unused remainder is called **slack space**. That slack space is not automatically erased when a file is saved and may still contain fragments of data left over from earlier files or system activity. Although slack space is not part of the file's active content and is invisible to ordinary users, hex viewers can visualize it.

Hint 2: If you discover and decode the password, you must figure out how to use it to uncover the secret word. Zip files can be password-protected, so you may need to use a decompression app that supports password protection. Your OS-supplied app may not be the one to do that.

Encoding and Steganography

We record information every day using 26 symbols called "the alphabet," abetted by helpful signals called "punctuation." So, you could say—loosely and metaphorically—that we write in something like hexavigesimal (Base-26), though unlike true positional encodings, our letters do not map cleanly to numeric values.

"Binary" or Base2 encoding is notating information using just two symbols. It's often said that "computer data is stored as ones and zeroes;" but that's a fiction. In fact, binary data is stored physically, electronically, magnetically or optically using mechanisms that permit the detection of two clearly distinguishable "states," whether manifested as faint voltage potentials (*e.g.,* thumb drives), polar magnetic reversals (*e.g.,* spinning hard drives) or pits on a reflective disc deflecting a laser beam (*e.g.,* DVDs). Ones and zeroes are simply a convenient way to *notate* those states; the storage media themselves know nothing of numbers. You could use *any* two symbols as binary characters, or even two discrete characteristics of the *same* symbol—such as uppercase versus lowercase letters, or the presence or absence of a serif.



I free you from the trope of ones and zeroes to plumb the evolution of binary communication and explore an obscure coding cul-de-sac called Steganography, from the Greek, meaning "covered writing." But first, we need an aside of Bacon.

I mean, of course, lawyer and statesman Sir Francis Bacon (1561-1626). Among his many accomplishments, Bacon conceived a bilateral cipher (that is, one based on two distinguishable states; a "code" in modern parlance) enabling the hiding of messages *omnia per omnia,* or "anything by anything."

Bacon's cipher used the letters "A" and "B" to denote binary values; but if we use ones and zeros instead, we see the straight line from Bacon's clever cipher to modern ASCII and Unicode encoding.

As with modern computer encoding, we need multiple binary digits ("bits") to encode or "stand in for" the letters of the alphabet. Bacon chose the five-bit sets at right:

Letter	Code
A	AAAAA
B	AAAAB
C	AAABA
D	AAABB
E	AABAA
F	AABAB
G	AABBA
H	AABBB

Letter	Code
I/J	ABAAA
K	ABAAB
L	ABABA
M	ABABB
N	ABAAB
O	ABABA
P	ABBBA
Q	ABBBB

Letter	Code
R	BAAAA
S	BAAAB
T	BAABA
U/V	BAABB
W	BABAA
X	BABAB
Y	BABBA
Z	BABBB

If we substitute ones and zeroes (right), Bacon's cipher starts to look uncannily like contemporary binary encodings.

Letter	Code
A	00000
B	00001
C	00010
D	00011
E	00100
F	00101
G	00110
H	00111

Letter	Code
I/J	01000
K	01001
L	01010
M	01011
N	01100
O	01101
P	01110
Q	01111

Letter	Code
R	10000
S	10001
T	10010
U/V	10011
W	10100
X	10101
Y	10110
Z	10111

Why *five* bits and not three or four? The answer lies in binary math ("*Oh no! Not MATH!!*"). Wait, *wait*; it won't hurt. I promise!

If you have one binary digit (2^1), you have only two unique states (one or zero), so you can only encode two letters, say A and B. If you have two binary digits (2^2 or 2×2), you can encode four letters, say A, B, C and D. With three binary digits (2^3 or 2x2x2), you can encode eight letters. Finally, with four binary digits (2^4 or 2x2x2x2), you can encode just sixteen letters. *So, do you see the problem in trying to encode the letters of a 26-letter alphabet?* You must use at least *five* binary digits (2^5 or 32) unless you are content to forgo ten letters —a sacrifice few secret correspondents are willing to make.

Sir Francis Bacon wasn't especially interested in encoding text as bits. His goal was to *hide* messages in any medium, permitting a clued-in reader to distinguish between differences lurking in plain sight. Those differences—whatever they might be—serve to denote the "A" or "B" in Bacon's steganographic technique. Bacon's insight was that *any* observable difference could serve as a carrier of meaning. Consider these examples:

I hi**d** m**y** name in this s**e**nten**c**e as a se**r**ies of b**o**l**d**e**d** l**e**tters.

I hi*d* m*y* name in this sente*n*c*e* as ita*l*icized *c*ha*r*acte*r*s.

I hid my name in this sentence as sans serif characters.

That last one is quite subtle, right? Here's how it's done:

Using Bacon's cipher, CRAIG BALL is AAABA BAAAA AAAAA ABAAA AABBA AAAAB AAAAA ABABA ABABA
C R A I G B A L L

To conceal my name in each of the respective examples, every unbolded/unitalicized/serif character signifies an "A" in Bacon's cipher and every boldface/italicized/sans serif character signifies a "B" (ignore the spaces and punctuation; only the visual differences carry meaning). The bold and italic approaches look wonky and could arouse suspicion, but if the fonts are chosen carefully, the absence of serifs should go unnoticed. Take a closer look to see how it works:



I hid my name in this sentence as sans serif characters.
A AAB AB AAAA AA AAAA BAAAAABB AA AAAB AAAAA ABABAABABA.

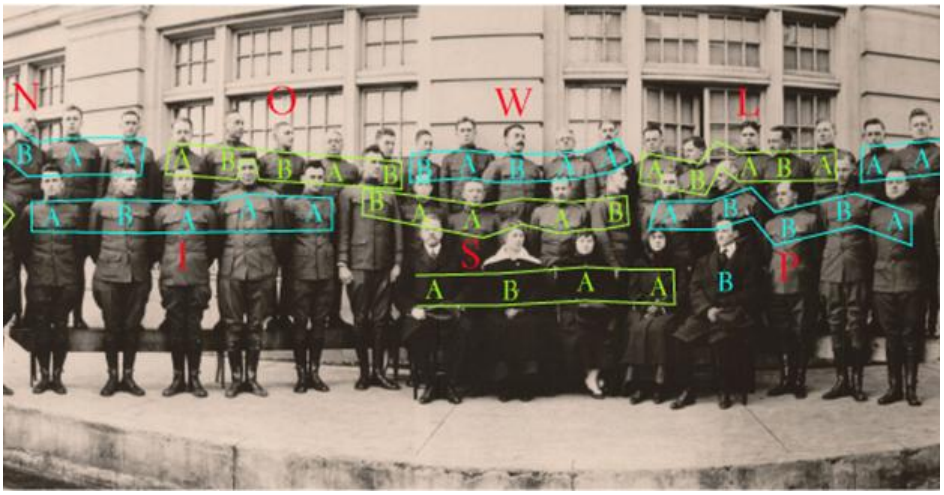
In my examples, I've used Bacon's cipher to hide text within text, but it can as easily hide messages in almost anything. My favorite example is the class photo of World War I cryptographers trained in Aurora, Illinois by pioneering American cryptographers William and Elizabeth Friedman.[59] Before they headed for France, the newly minted codebreakers lined up for the cameraman; but there's more going on here than meets the eye.



Taking to heart *omnia per omnia*, the Friedmans ingenuously encoded Sir Francis Bacon's maxim "knowledge is power" within the photograph using Bacon's cipher. The 71 soldiers and their instructors convey the cipher text by facing or looking away from the camera, a difference invisible to the untrained eye, but obvious once you know what to look for. Those facing denote an "A." Those looking away denote a "B." There weren't quite enough present to encode the entire maxim, so the decoded message actually reads, "KNOWLEDGE IS POWE." Here's the decoding:



A closer look:



[59] For this material, I'm indebted to "How to Make Anything Signify Anything" by William H. Sherman in Cabinet no. 40 (Winter 2010-2011).

Steganography is something most computer forensic examiners study but never encounter. Still, it's a fascinating discipline with a history reaching back to ancient Greece, where masters tattooed secret messages on servants' shaved scalps and hit "Send" once the hair grew back. Digital technology brought new and difficult-to-decipher steganographic techniques enabling images, sound and messages to hitch a hidden ride on a wide range of electronic media. The tools have changed, but the impulse—to hide meaning in plain sight—has not.

Processing Part II

Why Care about Encoding?

All this code page and Unicode stuff matters because of the vital role that electronic search plays in the identification and exploration of digital evidence. Modern evidence isn't sitting in a box; it must be teased out of databases, media, and vast, largely unorganized repositories of so-called **"unstructured" data."** If we index an information item for search using the wrong code page, the information in the item won't be extracted, won't become part of the index and, accordingly, won't be found even when the correct search terms are run. Many older search tools and electronic records are not Unicode-compliant. Worse, because **binary or compressed content** may include a smattering of ASCII text, a search tool may conclude it's encountered a document encoded with Latin-1 and won't flag the item as having failed to index. In short, encoding issues carry real-world consequences, particularly when foreign language content is in the collection. If you aren't aware of the limits of the processing tools you, your opponents or service providers use, you can't negotiate successful protocols. What you don't know *can* hurt you.

The many ways in which data is encoded—and the diverse ways in which collection, processing and search tools identify the multifarious encoding schemes—lie at the heart of significant challenges and costly errors in e-discovery. It's easy to dismiss these fundamentals of information technology as too removed from litigation to be worth the effort to explore them; but, understanding encoding and the role it plays in processing, indexing and search will help you realize the capabilities and limits of the tools you, your clients, vendors and opponents use.

Let's look at these issues and file formats through the lens of a programmer making decisions about how a file should function.

VITAL VOCABULARY

Bates Numbers
Binary File Signature
Chain of Custody
Compound Files
Compression
Container File
Concordance Index
Custodian
Decryption
Deduplication
De-NISTing
Entropy Testing
Exception Reporting
Family Group
Filtering
Hash Collision
Immaterial Items
Indexing
Ingestion
Load File
Near-Deduplication
Noise Words (Stop Words)
OCR
Production Processing
Recursion
Shingling
Text Stripping
Threading
Time Zone Normalization
Tokenization
Unicode Normalization
Unitization
UTF-8

Hypothetical: TaggedyAnn

Ann, a programmer, must create a tool to generate nametags for attendees at an upcoming law school reunion. Ann wants her tool to support different label stock and multiple printers, fonts, font sizes and salutations. It will accept lists of pre-registrants as well as create name tags for those who register onsite. Ann will call her tool *TaggedyAnn*. TaggedyAnn

has never existed, so no computer operating system "knows" what to do with TaggedyAnn data. Ann must design the necessary operations and associations.

Remembering our mantra, "ESI is just numbers," Ann must lay out those numbers in the data files so that the correct program—her TaggedyAnn program—will be executed against TaggedyAnn data files, and she must structure the data files so that the proper series of numbers will be pulled from the proper locations and interpreted in the intended way. Ann must develop the **file format**.

A file format establishes the way to encode and order data for storage within a file. The file format tells software not only *what* the data is, but *how to read it*. Ideally, Ann will document her file format as a published specification, a blueprint for the file's construction. That way, anyone trying to develop applications to work with Ann's files will know what goes where and how to interpret the data. But Ann may never get around to writing a formal file specification or, if she believes her file format to be a trade secret, Ann may decide not to reveal its structure publicly. In that event, others seeking to read TaggedyAnn files must reverse engineer the files to deduce their structure and fashion a document filter that can accurately parse and view the data.

In terms of complexity, file format specifications run the gamut. Ann's format will likely be simple—perhaps just a few pages—as TaggedyAnn supports few features and functions. By contrast, the published file specifications for the retired binary forms of Microsoft Excel, PowerPoint and Word files run to 1,125, 649 and 578 pages, respectively. Microsoft's latest file format specification for the .PST file that stores Outlook e-mail occupies 193 pages. Imagine trying to reverse engineer such complexity without a published specification! For a peek into the Byzantine structure of a Word .DOCX file, see Illustration 2: Anatomy of a Word DOCX File.

File Type Identification

In the context of the operating system, Ann must link her program and its data files through various means of **file type identification**.

File type identification using binary file signatures and file extensions is an essential early step in e-discovery processing. Determining file types is a necessary precursor to applying the appropriate document filter to extract contents and metadata for populating a database and indexing files for use in an e-discovery review platform.

File Extensions

Because Ann is coding her tool from scratch and her data files need only be intelligible to her program, she can structure the files any way she wishes, but she must nevertheless supply a way that computer operating systems can pair her data files with only her executable program. Else, there's no telling which application might be invoked, and the result may be unintelligible.

Filename extensions or just "file extensions" are a means to identify the contents and purpose of a data file. Though her executable files must carry the file extension .EXE, any ancillary files Ann creates can employ almost any three-letter or -number file extension Ann desires *so long as her choice doesn't conflict with one of the thousands of file extensions already in use*. That means that Ann can't use .TGA because it's already associated with Targa graphics files, and she can't use .TAG because that signals a DataFlex data file. So, Ann settles on .TGN as the file extension for TaggedyAnn files. That way, when a user loads a data file with the .TGN extension, Windows can direct its contents to the TaggedyAnn application and not to another program that won't correctly parse the data.

Binary File Signatures

But Ann needs to do more. Not all operating systems employ file extensions, and because extensions can be absent or altered, file extensions aren't the most reliable way to denote the purpose or content of a file. Too, file names and extensions are *system metadata*, so they are not stored inside the file. Instead, computers store file extensions within the *file table* of the storage medium's file system. Accordingly, Ann must fashion a unique **binary header signature** that precedes the contents of each TaggedyAnn data file in order that the contents are identifiable as TaggedyAnn data and directed solely to the TaggedyAnn program for execution.

A binary file signature (also called a **magic number**) will typically occupy the first few bytes of a file's contents. It will always be represented as hexadecimal values; however, the hex values chosen may correspond to an intelligible alphanumeric (ASCII) sequence. For example, PDF document files begin 0x 25 50 44 46 2D,[60] which is **%PDF-** in ASCII. Files holding Microsoft tape archive data begin 0x 54 41 50 45, translating to **TAPE** in ASCII. Sometimes, it's a convoluted correlation, *e.g.,* Adobe Shockwave Flash files have a file extension of SWF, but their file signature is 0x 46 57 53, corresponding to **FWS** in ASCII. In other instances, there is no intent to generate a meaningful ASCII word or phrase using the binary header signatures; they are simply numbers in hex. For example, the header signature for a JPG image is 0x FF D8 FF E0 which translates to the gibberish **ÿØÿà** in ASCII.

Binary file signatures often append characters that signal variants of the file type or versions of the associated software. For example, JPG images in the JPEG File Interchange Format (JFIF) use the binary signature 0x FF D8 FF E0 00 10 4A 46 49 46 or **ÿØÿà JFIF** in ASCII. A JPG image in the JPEG Exchangeable Image File Format (Exif, commonly used by digital cameras) will start with the binary signature 0x FF D8 FF E1 00 10 45 78 69 66 or **ÿØÿá Exif** in ASCII.

[60] The leading "0x" signals that the data that follows is notated in hexadecimal.

Ann peruses the lists of file signatures available online and initially thinks she will use 0x 54 41 47 21 as her binary header signature because no one else appears to be using that signature and it corresponds to **TAG!** in ASCII.

But, after reflecting on the volume of highly-compressible text that will comprise the program's data files, Ann instead decides to store the contents in a ZIP-compressed format, necessitating that the binary header signature for TaggedyAnn's data files be 0x 50 4B 03 04, **PK..** in ASCII. All files compressed with ZIP use the **PK..** header signature (again, because Phil Katz, the programmer who wrote the ZIP compression tool, chose to flag his file format with his initials).

How can Ann ensure that only TaggedyAnn opens these files and they are not misdirected to an unzipping (decompression) program? Simple. Ann will use the .TGN extension to prompt Windows to load the file in TaggedyAnn and TaggedyAnn will then read the **PK..** signature and unzip the contents to access the compressed contents.

But, what about when an e-discovery service provider processes the TaggedyAnn files with the mismatched file extensions and binary signatures? Won't that throw things off? It could—and often does. We'll return to that issue when we cover compound files and recursion; but first, an exercise to illustrate what you've read.



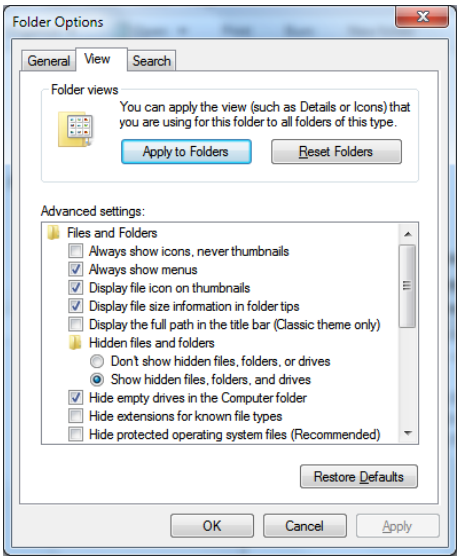
Exercise 7: Encoding: File Extensions

GOALS: The goals of this exercise are for the student to:

1. Through experimentation, understand how file extensions influence file appearance and application association in Windows or macOS; and
2. Assess the utility and reliability of file extensions as a filtering and identification tool in e-discovery.

OUTLINE: Section A: Students using Microsoft Windows operating systems will modify file extensions for a known file type to determine the impact on file appearance and application association. **Section B:** Mac users will investigate the treatment of file extensions and "Open With" option in the Info window. **Do Section 7(A) *or* Section 7(B), depending on your operating system—not both.**

7(A). File Extensions in Windows (Students with Windows machines do this exercise)



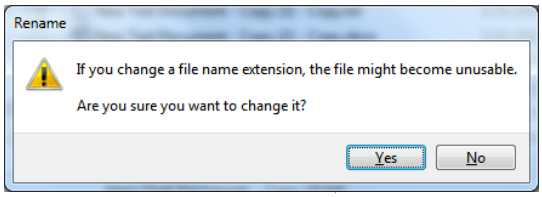
Step 1: Right click on an open area of your Desktop and select **New>Text Document**. Open the document "New Text Document.txt" that was created and type your name. Save the file.

NOTE: If you cannot see the file's .txt extension, your computer may be configured to hide file extensions. If so, open **File Explorer**, select **View** (or **View > Show**),[61] and ensure that **File name extensions** is checked. In older versions of Windows, this setting may appear under **Folder Options** as "Hide extensions for known file types."

Step 2: Locate the New Text Document.txt file you just created and, ***while depressing the Ctrl key,*** click and drag the file to create a duplicate of the file. **Do this four more times** to create five identical copies of the file (Windows will automatically adjust the filenames (e.g., "Copy," "Copy (2)") so they can coexist in the same folder).

[61] In earlier Windows versions, you can get there by typing 'folder options" at the Start screen to locate the Folder Options menu, then click the Folder Options menu and select the View tab and Advanced Options to continue to uncheck the option "hide extensions for known file types" option.

Step 3: Right click on any copy of New Text Document.txt and select Rename. Change only the extension of the file from .txt to .doc. Hit the Enter key to accept the change. You should get a warning message asking if you are sure you want to change the extension. Select "Yes."



Step 4: Change the file extensions for each of the four other copies of the file to .zip, .xls, .ppt and .htm. **Before opening the files, pause and note whether Windows has changed the file icons to reflect the new extensions.**

Step 5: Take a screenshot of the folder showing the icons for all six files (the original plus the five renamed copies). You will submit this screen shot as your work on the exercise. This is all you need to submit for Exercise 7.

Step 6: Try to launch each of the five copies with their new extensions. **Note whether the launched application successfully opens the file or merely attempts to do so.**

Step 7: Save one of the renamed files with **n**o file extension at all, then attempt to launch it. **What happens?**

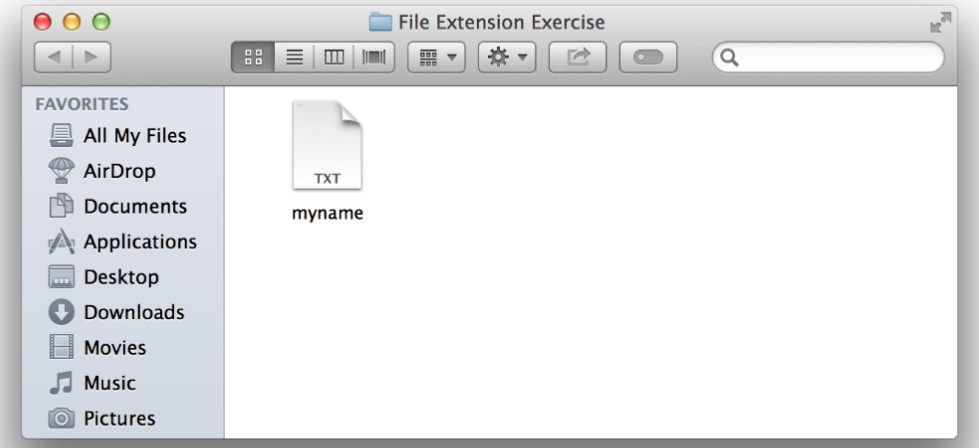
Points to Ponder:

1. How might a file acquire a file extension other than the correct one or lack a file extension?
2. How does Windows determine what application to run for each file extension?
3. How might you determine what application should be used to open a file with an unfamiliar file extension? Check out FileExt.com.
4. What risks arise if an e-discovery workflow relies solely on file extensions to identify file types?

B. File Extensions in MacOS (Students with MacOS machines do this exercise)

Step 1: Create an empty folder called "File Extension Exercise" on your Desktop. Run the text editor called TextEdit (found in your Applications folder). Create a new blank document by either clicking **File > New** or pressing **Command-N.** Type your name in this new blank document and save the file as "**myname.txt**" in your exercise folder.

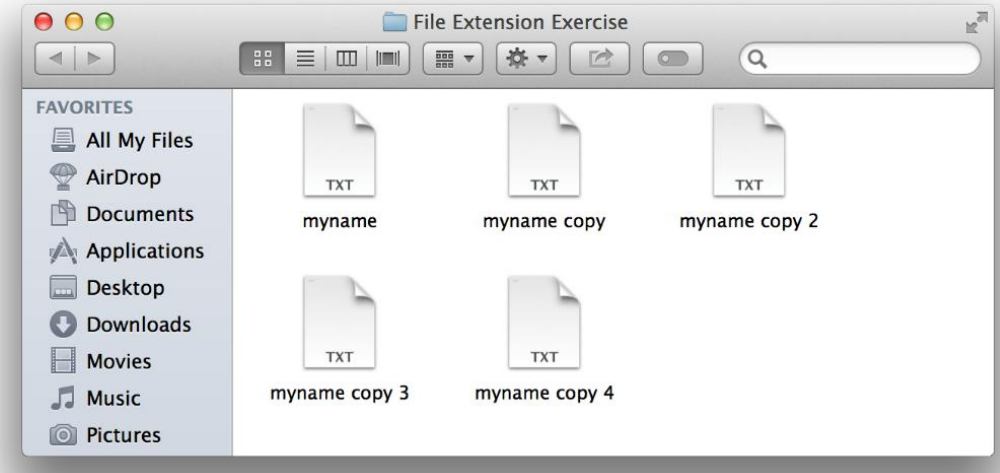
Be sure to save it in the new folder, like so:



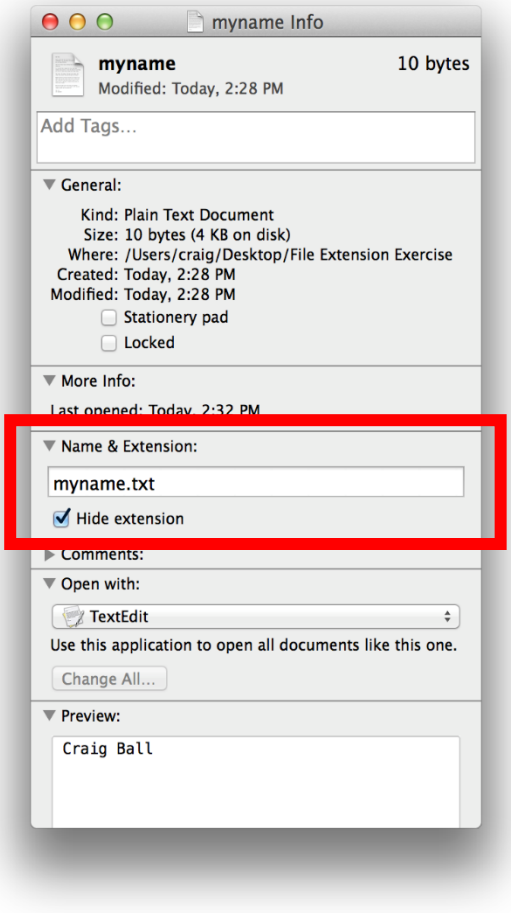
Step 2: Locate the file ***myname.txt*** you just created, select it, and press **Command-D (⌘-D)** four times to create four duplicates. You can also right-click and select **Duplicate**, repeating as needed.

The Apple Mac operating system hides file extensions by default; so, depending upon your system settings, you may or may not see file extensions for the five files.

In the next steps, you will change settings, if necessary, to show file extensions. Your folder should now look like this:



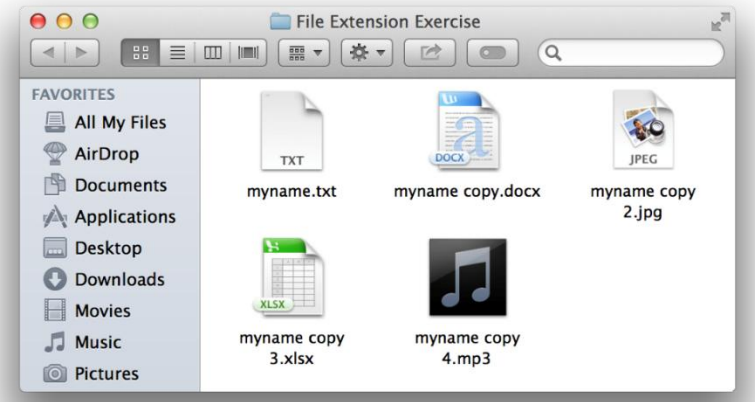
Step 3: Right click on the file **myname.txt** and select **Get Info**. As you can see in the figure at right, "Craig Ball" is nine letters and a space, so the file size (top right corner) is 10 bytes, reflecting the ten ASCII characters comprising its contents (your byte count may differ if you included additional characters).



Note in the **Name & Extensions** region that the option "Hide Extensions" is checked. This option can be turned on and off for individual files. If it's checked for your machine, uncheck the selection and note that the file icon now shows the name and extension, **myname.txt**.

Step 4: Select the first duplicate file you created called **myname copy** and Get Info**. I**n the **Name & Extensions** region, change the file's extension from .txt to .docx and uncheck "Hide extension." Hit the Enter key to accept the change. and when your system asks, "**Are you sure you want to change the extension from ".txt" to ".docx"?**" select **Use .docx.**

Step 5: Do the same thing for each of the three other copies of the file: uncheck "Hide extension" and change the file extensions of each to, respectively, *.jpg, .xlsx* and *.mp3*. Now, the contents of your exercise folder should look like mine at right.

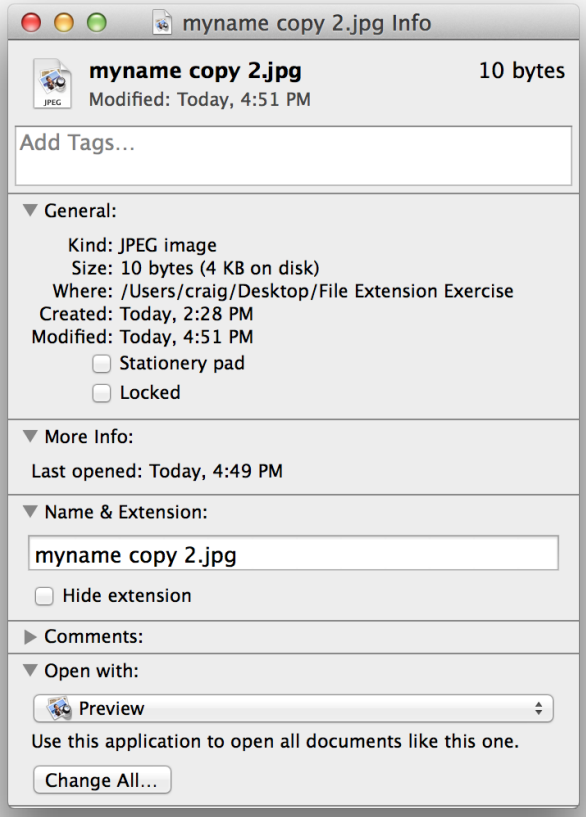


Step 6: Take a screenshot of your exercise folder showing the icons for all five files.[62] You will submit the screenshot to show your work on the

[62] On my Mac, the key combo **Shift-Command-4 plus space bar** generates a camera cursor that saves a screenshot of any folder or open window I select by clicking my mouse. A screenshot captured this way is stored on the desktop and features a white border around the window with a bit of a drop shadow.

exercise. This is all you need to submit for Exercise 7.

Step 7: Try to launch each of the altered icons and note what application opens (or fails to open). If you have Microsoft Word installed, clicking on **myname copy.docx** will launch Word instead of TextEdit. It works because Word knows how to parse text. Launching the next three files with altered extensions will prompt your system to try to launch the associated application for JPG images, spreadsheets and music, but all will fail because the applications are incompatible with the contents of the file. Still, that didn't stop the operating system from changing the appearance of the icons to show the incompatible applications or from trying to launch them for incompatible files.



Step 8: Unlike Microsoft Windows, Mac OS doesn't slavishly rely on file extensions to pair files with the proper applications. Pull up the Get Info screen for named **myname copy.jpg** and locate the "Open with:" panel (it probably displays Preview or another photo viewer app). Select the pull-down menu and select "Other…" In the following menu, change "**Enable: Recommended Applications**" to "**Enable: All Applications**." Scroll down the list of applications, select "**TextEdit**" and click Add. The file still displays a JPG icon; but now, launching the file starts TextEdit, *not* Preview. This illustrates that file extensions influence appearance and default behavior, but they do not determine the actual contents of a file.

Common File Extensions and Associated File Types[63]

Extension	Associated File Type
Text Files	
.doc	Microsoft Word Document
.docx	Microsoft Word Open XML Document
.pages	Pages Document
.rtf	Rich Text Format File
.txt	Plain Text File
Mail Files	

[63] No, of course I don't expect you to learn all these extensions. I expect you know many of them already.

.edb	Microsoft Exchange database
.eml	Outlook Express Mail Message
.mime	Multi-Purpose Internet Mail Message File
.msg	Outlook Mail Message
.nsf	IBM/Lotus Notes container file
.ost	Microsoft Outlook synchronization file
.pst	Microsoft Outlook container file

Data Files

.csv	Comma Separated Values File
.dat	Data File
.key	Keynote Presentation
.pps	PowerPoint Slide Show
.ppt	PowerPoint Presentation
.pptx	PowerPoint Open XML Presentation
.vcf	vCard File
.xml	XML File

Audio Files

.m3u	Media Playlist File
.mid	MIDI File
.mp3	MP3 Audio File
.mpa	MPEG-2 Audio File
.ra	Real Audio File
.wav	WAVE Audio File
.wma	Windows Media Audio File

Video Files

.asf	Advanced Systems Format File
.asx	Microsoft ASF Redirector File
.avi	Audio Video Interleave File
.mov	Apple QuickTime Movie
.mp4	MPEG-4 Video File
.mpg	MPEG Video File
.rm	Real Media File
.swf	Shockwave Flash Movie
.vob	DVD Video Object File
.wmv	Windows Media Video File

Raster Image Files

.bmp	Bitmap Image File
.gif	Graphical Interchange Format File

.jpg	JPEG Image File
.png	Portable Network Graphic
.psd	Adobe Photoshop Document
.tif	Tagged Image File

Vector Image Files

.ai	Adobe Illustrator File
.drw	Drawing File
.eps	Encapsulated PostScript File
.ps	PostScript File
.svg	Scalable Vector Graphics File

Page Layout Files

.pdf	Portable Document Format File

Spreadsheet Files

.wks	Works Spreadsheet
.xls	Excel Spreadsheet
.xlsx	Microsoft Excel Open XML Spreadsheet

Database Files

.db	Database File
.dbf	Database File
.mdb	Microsoft Access Database
.pdb	Program Database
.sql	Structured Query Language Data

Executable (Program) Files

.app	Mac OS X Application
.bat	DOS Batch File
.cgi	Common Gateway Interface Script
.com	DOS Command File
.exe	Windows Executable File
.gadget	Windows Gadget
.jar	Java Archive File
.pif	Program Information File
.vb	VBScript File
.wsf	Windows Script File

CAD Files

.dwg	AutoCAD Drawing Database File
.dxf	Drawing Exchange Format File

GIS Files

.mxd	Map Exchange Document
.kml	Keyhole Markup Language File

Web Files

.asp	Active Server Page
.cer	Internet Security Certificate
.csr	Certificate Signing Request File
.css	Cascading Style Sheet
.htm	Hypertext Markup Language File
.html	Hypertext Markup Language File
.js	JavaScript File
.jsp	Java Server Page
.php	Hypertext Preprocessor File
.rss	Rich Site Summary
.xhtml	Extensible Hypertext Markup Language File

Font Files

.fnt	Windows Font File
.fon	Generic Font File
.otf	OpenType Font
.ttf	TrueType Font

System Files

.cab	Windows Cabinet File
.cpl	Windows Control Panel Item
.cur	Windows Cursor
.dll	Dynamic Link Library
.dmp	Windows Memory Dump
.drv	Device Driver
.lnk	File Shortcut
.sys	Windows System File

Settings and Configuration Files

.cfg	Configuration File
.ini	Windows Initialization File
.keychain	Mac OS X Keychain File
.prf	Outlook Profile File

Encoded Files

.bin	Macbinary Encoded File
.hqx	BinHex 4.0 Encoded File
.uue	Uuencoded File

Compressed Files

.7z	7-Zip Compressed File
.deb	Debian Software Package
.gz	Gnu Zipped Archive
.pkg	Mac OS X Installer Package
.rar	WinRAR Compressed Archive
.sit	Stuffit Archive
.sitx	Stuffit X Archive
.tar.gz	Tarball File
.zip	Zipped File
.zipx	Extended Zip File

Disk Image Files

.dmg	Mac OS X Disk Image
.iso	Disc Image File
.toast	Toast Disc Image
.vcd	Virtual CD

Developer Files

.c	C/C++ Source Code File
.class	Java Class File
.cpp	C++ Source Code File
.cs	Visual C# Source Code File
.dtd	Document Type Definition File
.fla	Adobe Flash Animation
.java	Java Source Code File
.m	Objective-C Implementation File
.pl	Perl Script

File Structure

Now, Ann must decide how she will structure the data within her files. Once more, Ann's files need only be intelligible to her application, so she is unconstrained in her architecture. This point becomes important as the differences in file structure are what make processing and indexing essential to electronic search. There are thousands of different file types, each structured in idiosyncratic ways. Processing normalizes their contents to a common, searchable format despite their wildly different internal architectures

Some programmers dump data into files in so consistent a way that programs retrieve particular data by offset addressing; that is, by beginning retrieval at a specified number of bytes from the start of the file (an *offset*) and retrieving a specified extent of data from that point forward.

Offset addressing could make it hard for Ann to add new options and features, so she may prefer to implement a chunk- or directory-based structure. In the first approach, data is labeled within the file to indicate its beginning and ending, or it may be tagged ("marked up") for identification. The program accessing the data simply traverses the file seeking the tagged data it requires and grabs the data between tags.

There are many ways to implement a chunk-based structure, and it's probably the most common file structure because they allow data elements to grow, shrink, or be reordered without rewriting the entire file. A directory-based approach constructs a file as a small operating environment. The directory keeps track of what's in the file, what it's called and where it begins and ends. Examples of directory-based formats are ZIP archive files and Microsoft Office files after Office 2007. Ann elects to use a mix of both approaches. Using ZIP entails a directory and folder structure, and she will use tagged, chunked data within the compressed folder and file hierarchy.

Data Compression

Many common file formats and containers are compressed, necessitating that e-discovery processing tools be able to identify compressed files and apply the correct decompression algorithm to extract contents.

Compression is miraculous. It makes modern digital life possible. Without compression, we wouldn't have smart phones, digitized music, streaming video or digital photography. Without compression, the web would be a much different, much duller place.

Compression uses algorithms to reduce the space required to store and the bandwidth required to transmit electronic information. If the algorithm preserves all compressed data, it's termed "**lossless compression**." If the algorithm preserves *all* original data, it's termed **lossless compression**.

JPEG image compression is lossy compression, executing a tradeoff between image quality and file size. Not all the original photo's graphical information survives JPEG compression. Sharpness and color depth are sacrificed, and compression introduces distortion, often visible as blockiness or rough margins ("jaggies"). We likewise see a loss of fidelity when audio or video data is compressed for storage or transmission (*e.g.*, as MPEG or MP3 files). The offsetting benefit is that the smaller file sizes facilitate streaming video and storing thousands of songs in your pocket.

In contrast, file compression algorithms like ZIP are lossless; decompressed files perfectly match their counterparts before compression. Let's explore how that's done.

One simple approach is **Run-Length Encoding**. It works especially well for images containing consecutive, identical data elements, like the ample white space of a fax transmission. Consider a black and white graphic where each pixel is either B or W; for example, the image of the uppercase letter "E," below left:

WW**BBBBB**WW WW**BB**WWWW WW**BBBB**WW WW**BB**WWWW WW**BBBBB**WW	The image at left requires 45 characters (and thus 45 bytes when stored as ASCII) but we can write it in 15 fewer characters by adding a number describing each sequence or "run," *i.e.,* 2 white pixels, 5 black, 2 white.	**2W5B2W** **2W2B5W** **2W4B3W** **2W2B5W** **2W5B2W**

We've compressed the data by a third. Refining our run-length compression, we substitute a symbol (|) for each 2W and now need just 23 characters to describe the graphic, like so:	**|5B|** **|2B5W** **|4B3W** **|2B5W** **|5B|**	We've compressed the data by almost half but added overhead: we must now supply a dictionary defining |=2W.

Going a step further, we swap in symbols for 5B (\), 2B (/) and 5W (~), like so:	**|\\\|** **|/~** **|4B3W** **|/~** **|\\\|**	It takes just 17 characters to serve as a compressed version of the original 45 characters by adding three more symbols to our growing dictionary.

As we apply this run-length encoding to more and more data, we see improved compression ratios because we can apply symbols already in use and don't need to keep adding new symbols to our dictionary for each swap.

ZIP employs a lossless compression algorithm called **DEFLATE**, combining dictionary-based compression with entropy encoding and underpinning both ZIP archives and the PNG image format. Thanks to its efficiency and being free to use without license fees, DEFLATE remains the most widely used compression algorithm in the world.

Tools for processing files in e-discovery must identify compressed files and apply the correct algorithm to unpack the contents. The decompression algorithm locates the compression dictionary and coding tables, reconstructs the directory structure, and restores stored metadata for the contents.

Identification on Ingestion

Remember Programmer Ann and her struggle to select a TaggedyAnn file extension and signature? Now, those decisions play a vital role in how an e-discovery processing tool extracts text and metadata. If we hope to pull intelligible data from a file or container, we must first reliably ascertain the file's structure and encoding. For compressed files, we must apply the proper decompression algorithm. If it's an e-mail container format, we must apply the proper encoding schema to segregate messages and decode all manner of attachments. We must treat image files as images, sound files as sounds and so on. Misidentification of file types virtually guarantees failure.

The e-discovery industry relies upon various open source and commercial file identifier tools. These tend to look first to the file's binary header for file identification and then to the file's extension and name. If the file type cannot be determined from the signature and metadata, the identification tool may either flag the file as unknown (an "exception") or pursue other identification methods as varied as byte frequency analysis (BFA), which examines statistical patterns in the data, or the use of specialty file type detection tools designed to suss out characteristics unique to certain file types, especially container files. Identifiers will typically report both the apparent file type (from metadata, *i.e.*, the file's name and extension) and the actual file type. Inconsistencies between these may prompt special handling or signal spoofing with malicious intent.

The table below sets out header signatures aka "magic numbers" for common file types:

File Type	Extension	Hex Signature	ASCII	Notes
ZIP Archive	ZIP	50 4B 03 04	PK..	
MS Office	DOCX XLSX PPTX	50 4B 03 04	PK..	Compressed XML files
Outlook mail	PST	21 42 44 4E	!BDN	
Outlook message	MSG	D0 CF 11 E0 A1 B1 1A E1		
Executable file	EXE	4D 5A	MZ	For Mark Zbikowski, who said, "when you're writing the linker, you get to make the rules."
Adobe PDF	PDF	25 50 44 46	%PDF	
PNG Graphic	PNG	89 50 4E 47	.PNG	

File Type	Extension	Hex Signature	ASCII	Notes
VMWare Disk file	VMDK	4B 44 4D 56	KDMV	
WAV audio file	WAV	52 49 46 46	RIFF	
Plain Text file	TXT	none	none	Plain text files do not have fixed signatures, though some may begin with a Byte Order Mark (BOM).

Media (MIME) Type Detection

File extensions are central to Microsoft operating systems. Systems like Linux and MacOS X rely less on file extensions to identify file types. Instead, they employ a file identification mechanism called **Media (MIME) Type Detection**. MIME, which stands for **M**ultipurpose **I**nternet **M**ail **E**xtensions, is a seminal Internet standard that enables the grafting of text enhancements, foreign language character sets (Unicode) and multimedia content (*e.g.,* photos, video, sounds and machine code) onto plain text e-mails. **Virtually all e-mail travels in MIME format.**

The ability to transmit multiple file types via e-mail created a need to identify the content type transmitted. The **I**nternet **A**ssigned **N**umbers **A**uthority (IANA) oversees global Internet addressing and defines the hierarchy of media type designation. These hierarchical designations for e-mail attachments conforming to MIME encoding came to be known as **MIME types**. Though the use of MIME types started with e-mail, operating systems, tools and web browsers now employ MIME types to identify media, prompting the IANA to change the official name from MIME Types to **Media Types**.

Media types serve two important roles in e-discovery processing. First, they facilitate the identification of content based upon a media type declaration found in, *e.g.,* e-mail attachments and Internet publications. Second, media types serve as standard classifiers for files after identification of content. Classification of files within a media type taxonomy simplifies culling and filtering data in ways useful to e-discovery. While it's enough to specify "document," "spreadsheet" or "picture" in a Request for Production, e-discovery tools require a more granular breakdown of content. This granularity matters because search, culling, and analytics behave differently across media types. Tools must be able to distinguish a Word binary .DOC format from a Word XML .DOCX format, a Windows PowerPoint file from a Mac Keynote file and a GIF from a TIFF.

Media Type Tree Structure

Media types follow a path-like tree structure under one of the following standard types: **application, audio, image, text** and **video** (collectively called **discrete** media types) and **message**

and **multipart** (called **composite** media types). These top-level media types are further defined by subtype and, optionally, by a suffix and parameter(s), written in lowercase.

Examples of file type declarations for common file formats:
Note: File types prefixed by x- are not IANA. Those prefixed by vnd. are vendor-specific formats.

Application

Word .DOC:	application/msword
Word .DOCX:	application/vnd.openxmlformats-officedocument.wordprocessingml.document
Adobe PDF:	application/pdf (.pdf)
PowerPoint .PPT:	application/vnd.ms-powerpoint
PowerPoint .PPTX:	application/vnd.openxmlformats-officedocument.presentationml. presentation
Slack file (example):	application/x-atomist-slack-file+json
.TAR archive:	application/x-tar
Excel .XLS:	application/vnd.ms-excel
Excel .XLSX:	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
ZIP archive .ZIP:	application/zip

Audio

.MID:	audio/x-midi
.MP3:	audio/mpeg
.MP4:	audio/mp4
.WAV:	audio/x-wav

Image

.BMP:	image/bmp
.GIF:	image/gif
.JPG:	image/jpeg
.PNG:	image/png
.TIF:	image/tiff

Text (Typically accompanied by a **charset** parameter identifying the character encoding)

.CSS:	text/css
.CSV:	text/csv
.HTML:	text/html
.ICS:	text/calendar

.RTF:	text/richtext
.TXT	text/plain

Video

.AVI	video/x-msvideo
.MOV:	video/quicktime
.MP4:	video/mp4
.MPG:	video/mpeg

When All Else Fails: Octet Streams and Text Stripping

When a processing tool cannot identify a file, it may flag the file as an exception and discontinue processing contents; but the greater likelihood is that the tool will treat the unidentifiable file as an **octet stream** and harvest or "strip out" whatever text or metadata it *can* identify within the stream. An octet stream is simply an arbitrary sequence or "stream" of data presumed to be binary data stored as eight-bit bytes or "octets." So, an octet stream is anything the processor fails to recognize as a known file type.

The risk of treating a file as an octet stream and stripping identifiable text is that whatever plain text is stripped and indexed doesn't fairly mirror relevant content. However, because *some* text was stripped, the file may not be flagged as an exception requiring special handling; instead, the processing tool records the file as successfully processed notwithstanding the missing content. From a litigation standpoint, this is one of the most dangerous failure modes in e-discovery because it looks like success.



Exercise 8: Encoding: Binary Signatures

GOALS: The goals of this exercise are for the student to:

1. Identify and parse binary file signatures and their hexadecimal counterparts; and
2. Better appreciate the role data encoding plays in computing, e-discovery, and computer forensics.

OUTLINE: Students will examine the binary content of multiple files of various types to determine consistent binary and hex file signatures suitable for filtering, processing, and data carving in e-discovery and computer forensics.

Binary and Hex File Signatures

As we saw earlier, a file's header is data at or near the start of the file that serves to identify the type of data contained in the file (as well as information about the file's length, structure or other characteristics). File headers play a crucial role in the recovery of deleted data and the identification of hidden files. Computer forensic examiners often recover deleted files by scanning the recycled areas of magnetic hard drives—called **unallocated clusters**, space the operating system believes to be empty—for file signatures in a process called "file carving."

Step 1: Download the Zip file at www.craigball.com/filetypes.zip and extract its contents to your desktop or any other convenient location on your computer.

Step 2: The extracted contents will comprise nine folders (named BMP, DOC, DWG, GIF, JPG, PDF, TXT, WAV and XLS), each containing samples of file types commonly encountered in e-discovery.

Step 3: Identify file header signatures for common file types

Using your web browser, go to the Online HexDump Utility at http://www.fileformat.info/tool/hexdump.htm[64] and click "choose File." Using the selection box that will appear, navigate to the folder just extracted called BMP (you should see seven files) and select the file called TOC.bmp. Click "Open." Now click the blue "Dump" button on the Online HexDump Utility page. You should see this:

[64] If the Online HexDump Utility is unavailable, any hex viewer or forensic tool that displays raw bytes will suffice.

```
file name: TOC.bmp
mime type:

0000-0010:  42 4d 82 7d-04 00 00 00-00 00 36 00-00 00 28 00  BM.}.... ..6...(.
0000-0020:  00 00 39 01-00 00 eb 00-00 00 01 00-20 00 00 00  ..9..... ........
0000-0030:  00 00 00 00-00 00 c4 0e-00 00 c4 0e-00 00 00 00  ........ ........
0000-0040:  00 00 00 00-00 00 0b 0a-06 ff 0b 0a-06 ff 0b 0a  ........ ........
0000-0050:  06 ff 0b 0a-06 ff 0b 0a-06 ff 0a 09-05 ff 0a 09  ........ ........
0000-0060:  05 ff 09 08-04 ff 09 08-04 ff 09 08-04 ff 09 08  ........ ........
0000-0070:  04 ff 09 08-04 ff 09 08-04 ff 08 07-03 ff 08 07  ........ ........
0000-0080:  03 ff 08 07-03 ff 09 08-04 ff 09 08-04 ff 09 08  ........ ........
0000-0090:  04 ff 09 08-04 ff 09 08-04 ff 07 08-01 ff 08 08  ........ ........
0000-00a0:  02 ff 08 07-02 ff 08 06-03 ff 05 07-03 ff 06 07  ........ ........
0000-00b0:  03 ff 05 07-03 ff 03 07-05 ff 04 07-05 ff 04 07  ........ ........
0000-00c0:  05 ff 04 06-04 ff 05 06-04 ff 05 06-04 ff 05 06  ........ ........
0000-00d0:  04 ff 05 06-04 ff 05 06-04 ff 05 06-04 ff 05 06  ........ ........
0000-00e0:  04 ff 05 06-04 ff 05 06-04 ff 05 06-04 ff 05 06  ........ ........
0000-00f0:  03 ff 05 05-04 ff 05 05-05 ff 05 05-05 ff 04 04  ........ ........
0000-0100:  04 ff 07 06-08 ff 07 06-08 ff 08 07-09 ff 09 08  ........ ........
```

Note the first few bytes of the file. Load and peruse each of the remaining six bitmap files in BMP folder and identify text within the first few bytes of each that is **common to all** of the files in the BMP folder. Do you see that the first two characters of all the BMP files are **BM**, which corresponds to the hexadecimal signature **42 4D**? BM is the binary signature header identifying the content of each file as a bitmap image. Even if you changed the files' extensions to something other than BMP, that header signature gives away its content.

Now, use hexDump to view each of the six files in the folder called DWG (DWG is an extension typically denoting AutoCAD drawing files). Look at each of the six DWG files. **Can you identify a common binary header?** Note that all of the files begin "AC10" but the next two values vary from 15 to 18 to 24.

Header variation may indicate file formats that have changed over time. In the case of these DWG files, the headers AC1015, AC1018 and AC1024 indicating drawings created using different internal versions of AutoCAD. AC1015 indicates that the drawing was made using version 15 of AutoCAD, sold in the year 2000. Version 18 was released in 2004 and version 24 in 2010.

Step 4: Identify Binary Signatures for Common File Types

Because file headers can vary, as with the DWG files above, it's important to identify signatures that are common to all the files of a particular file type.

Examine the files in the DOC, GIF, PDF, TXT, WAV and XLS folders to determine the common binary signature you'd use to identify each of those file types. Now, record those signatures as hexadecimal values. Too short a signature risks false positives; too long a signature risks missing legitimate variants. Show your answers below:

File Type	Binary Signature	Hex Signature
DOC		
GIF		
PDF		
WAV		
XLS		
TXT		

Discussion questions (NOT ASSIGNED):

1. Do all files have unique binary signatures?
2. Do file signatures always start with the first byte of a file?
3. How do you distinguish between the various MS Office files?
4. Can a file's binary signature be changed?
5. Do files have footers (signatures at the end of files)?
6. How are file signatures used by e-discovery service providers and forensic examiners?
7. Can you find a leetspeak message (Google it) in the *hex* headers of Microsoft Office files? (Hint: Look at the first few hex bytes of older Microsoft Office files with DOC, PPT or XLS extensions).
8.



Exercise 9: Encoding: Unicode

GOALS: The goals of this exercise are for the student to:

1. Gain further familiarity with the concept of encoding character sets and code pages;
2. Understand the significance of single byte and multibyte encoding schemes (e.g., Unicode); and
3. Appreciate the role that encoding schemes play in EDD processing and search.

OUTLINE: Students will examine files of like content written in different languages and character sets and, correspondingly, encoded using different multibyte encoding schemes. You might want to re-read the brief discussion of Unicode at p. 118. **NOTE: You DO NOT need to submit any responses to Exercise 9.**

Step 1: Use the files you extracted from www.craigball.com/filetypes.zip in Exercises 8 (in folder TXT).

Step 2: Identify file header signatures for common file types

Using your web browser, go to the Online HexDump Utility at http://www.fileformat.info/tool/hexdump.htm and click "choose File." Using the selection box that will appear, navigate to the folder called TXT. You should see 24 files. Select the file called eula.1033.txt. Click "Open." Now click the blue "Dump" button on the Online HexDump Utility page. You should see this:

```
file name: eula.1033.txt
mime type:

0000-0010:  ff fe 4d 00-49 00 43 00-52 00 4f 00-53 00 4f 00  ..M.I.C. R.O.S.O.
0000-0020:  46 00 54 00-20 00 53 00-4f 00 46 00-54 00 57 00  F.T...S. O.F.T.W.
0000-0030:  41 00 52 00-45 00 20 00-4c 00 49 00-43 00 45 00  A.R.E... L.I.C.E.
0000-0040:  4e 00 53 00-45 00 20 00-54 00 45 00-52 00 4d 00  N.S.E... T.E.R.M.
0000-0050:  53 00 0d 00-0a 00 4d 00-49 00 43 00-52 00 4f 00  S.....M. I.C.R.O.
0000-0060:  53 00 4f 00-46 00 54 00-20 00 56 00-49 00 53 00  S.O.F.T. ..V.I.S.
0000-0070:  55 00 41 00-4c 00 20 00-53 00 54 00-55 00 44 00  U.A.L... S.T.U.D.
0000-0080:  49 00 4f 00-20 00 54 00-4f 00 4f 00-4c 00 53 00  I.O...T. O.O.L.S.
0000-0090:  20 00 46 00-4f 00 52 00-20 00 54 00-48 00 45 00  ..F.O.R. ..T.H.E.
0000-00a0:  20 00 4d 00-49 00 43 00-52 00 4f 00-53 00 4f 00  ..M.I.C. R.O.S.O.
0000-00b0:  46 00 54 00-20 00 4f 00-46 00 46 00-49 00 43 00  F.T...O. F.F.I.C.
0000-00c0:  45 00 20 00-53 00 59 00-53 00 54 00-45 00 4d 00  E...S.Y. S.T.E.M.
0000-00d0:  20 00 28 00-56 00 45 00-52 00 53 00-49 00 4f 00  ..(.V.E. R.S.I.O.
0000-00e0:  4e 00 20 00-33 00 2e 00-30 00 20 00-52 00 55 00  N...3... 0...R.U.
0000-00f0:  4e 00 54 00-49 00 4d 00-45 00 29 00-0d 00 0a 00  N.T.I.M. E.).....
0000-0100:  54 00 68 00-65 00 73 00-65 00 20 00-6c 00 69 00  T.h.e.s. e...l.i.
0000-0110:  63 00 65 00-6e 00 73 00-65 00 20 00-74 00 65 00  c.e.n.s. e...t.e.
```

Note the "dots" that appear between the letters in the document. This is how Unicode text appears when viewed using a tool that treats it like ASCII. Looking at the same content in hex, you can see the second byte used to encode each letter is hex 00. Because the second byte isn't needed for the Latin alphabet, it's 'zeroed out' and appears as a dot separating each letter when treated as ASCII.

Step 3: Open in Default Text Viewer

Now, double click on the file **eula.1033.txt** to open it in your default text viewer application (likely to be Notepad, Wordpad or Word on a Windows machine; TextEdit on a Mac). You may also use the free online application at **http://www.rapidtables.com/tools/notepad.htm.** Chances are, when **eula.1033.txt** opens in the text viewer, it will look "normal;" that is, you won't see any dots or spaces between the letters of each word. That's because your operating system (or the online text editor) is applying a code page that correctly interprets the Unicode data (likely UTF-8 or UTF-16) in the view presented to you.

Point to Ponder: What difference might Unicode encoding make in framing searches for e-discovery?

Step 4: Foreign Language Encodings

Double click on the file **eula.1037.txt** to open it in your default text viewer application. When it opens, it should be in Hebrew with some scattered English text. If you see the Hebrew, it's because your system is applying the correct Unicode character encoding to the data and not attempting to display it to you as ASCII text.

To see what it looks like when the wrong (ASCII) encoding is applied, return to the Online HexDump Utility at http://www.fileformat.info/tool/hexdump.htm and load eula.1037.txt. All you will be able to see in the right column will be the scattered English text. The Hebrew text will be replaced by dots. Like so:

```
file name: eula.1037.txt
mime type:

0000-0010:  ff fe ea 05-e0 05 d0 05-d9 05 20 00-e8 05 e9 05  ........ ........
0000-0020:  d9 05 d5 05-df 05 20 00-e2 05 d1 05-d5 05 e8 05  ........ ........
0000-0030:  20 00 ea 05-d5 05 db 05-e0 05 ea 05-20 00 4d 00  ........ ......M.
0000-0040:  49 00 43 00-52 00 4f 00-53 00 4f 00-46 00 54 00  I.C.R.O. S.O.F.T.
0000-0050:  0d 00 0a 00-4d 00 49 00-43 00 52 00-4f 00 53 00  ....M.I. C.R.O.S.
0000-0060:  4f 00 46 00-54 00 20 00-56 00 49 00-53 00 55 00  O.F.T... V.I.S.U.
0000-0070:  41 00 4c 00-20 00 53 00-54 00 55 00-44 00 49 00  A.L...S. T.U.D.I.
0000-0080:  4f 00 20 00-54 00 4f 00-4f 00 4c 00-53 00 20 00  O...T.O. O.L.S...
0000-0090:  46 00 4f 00-52 00 20 00-54 00 48 00-45 00 20 00  F.O.R... T.H.E...
0000-00a0:  4d 00 49 00-43 00 52 00-4f 00 53 00-4f 00 46 00  M.I.C.R. O.S.O.F.
0000-00b0:  54 00 20 00-4f 00 46 00-46 00 49 00-43 00 45 00  T...O.F. F.I.C.E.
0000-00c0:  20 00 53 00-59 00 53 00-54 00 45 00-4d 00 20 00  ..S.Y.S. T.E.M...
0000-00d0:  28 00 56 00-45 00 52 00-53 00 49 00-4f 00 4e 00  (.V.E.R. S.I.O.N.
0000-00e0:  20 00 33 00-2e 00 30 00-20 00 52 00-55 00 4e 00  ..3...0. ..R.U.N.
0000-00f0:  54 00 49 00-4d 00 45 00-29 00 0d 00-0a 00 ea 05  T.I.M.E. ).......
0000-0100:  e0 05 d0 05-d9 05 20 00-e8 05 e9 05-d9 05 d5 05  ........ ........
0000-0110:  df 05 20 00-d0 05 dc 05-d4 05 20 00-de 05 d4 05  ........ ........
0000-0120:  d5 05 d5 05-d9 05 dd 05-20 00 d4 05-e1 05 db 05  ........ ........
0000-0130:  dd 05 20 00-d1 05 d9 05-df 05 20 00-4d 00 49 00  ........ ....M.I.
```

Why? Because to maximize compatibility with single-byte ASCII text, Unicode also supports ASCII encoding; so, the ASCII viewer in the HexDump tool can see and correctly interpret the ASCII characters. However, the ASCII viewer can't make sense of double-byte encodings (i.e., the Hebrew text) and displays a dot instead.

Data Extraction and Document Filters

If ESI were like paper, a lawyer could open each item in its native application, read it, and decide whether it is relevant or privileged. ESI is different in ways that make this approach impractical and risky:

- ESI collections are exponentially larger than paper collections
- ESI is stored digitally and is unintelligible absent electronic processing
- ESI carries metadata that is always useful and may itself be probative evidence
- ESI is electronically searchable; paper is not
- ESI can be culled, filtered, deduplicated, stored, and transmitted at scale and low cost
- Opening ESI in native applications can alter both content and metadata

These differences make piecemeal review in native applications slow, inconsistent, and unsafe. Search would be unreliable, deduplication impossible, and the benefits of mass tagging, filtering, and production forfeited. Lawyers sometimes conflate *native production*—which may be appropriate or even superior in some circumstances—with *native review*. The two are not the same. Native applications are not e-discovery tools and should not be used for review. Purpose-built e-discovery platforms are essential.

To exploit the advantages of ESI in search, culling, and review, processing tools ingest files and extract text, embedded objects, and metadata. The extracted material is normalized, tokenized, stored in a database, and indexed for efficient search. These steps greatly enhance efficiency but impose tradeoffs in precision and accuracy. Processing is therefore not a neutral act; it reflects design choices that balance speed and scale against fidelity. Understanding those tradeoffs is critical.

Returning to Programmer Ann and her TaggedyAnn file format, Ann could freely define the file's internal structure because she was also writing the software that would read it. Files must be designed to grow and shrink without breaking retrieval, so rigid, fixed-offset addressing is rarely sufficient. Instead, files rely on internal directories, tables, or tags that act as signposts. All content—text, images, formatting, configuration instructions, and metadata—exists as hexadecimal data, but meaning depends entirely on knowing *where* the data resides and *how* it is encoded.

Extraction therefore requires more than locating the start and end of data. The processor must know whether the data represents ASCII text, Unicode, an image, a date-time value, or a bit flag whose numeric value merely signals a condition or setting.

There are two broad approaches to extracting content. One is to use the Application Programming Interface (API) of the application that created the file. The other is to rely on published file specifications or reverse engineering to determine where data resides and how it is encoded.

An API allows a “client” application to request information or services from a “server” application without access to the underlying mechanics. The restaurant analogy fits: the client orders from the menu but never enters the kitchen. APIs may also support undocumented calls intended for a limited audience.

For cloud-hosted data in platforms like Office 365, Dropbox, or OneDrive, APIs are unavoidable. For local files, however, APIs are typically a last resort. They tend to be slow, constrained, and incomplete. Not all applications expose open APIs, and those that do may withhold data critical to e-discovery. For years, one leading processing tool required a licensed copy of IBM/Lotus Notes because Notes files resisted extraction by any means other than the native API.

The alternative is the use of document filters. A document filter is an extraction template that describes where content resides within a file type and how it should be interpreted. Filters may be based on published specifications or on reverse engineering—a labor-intensive process with uncertain accuracy and consistency. Because building and maintaining filters for hundreds of file types is costly, most e-discovery vendors license filters rather than develop their own.

The commercial and open-source ecosystem for document filters and text extraction remains narrow in 2026.

Oracle Outside In continues to be the dominant commercial filter library. It provides SDKs for identifying, extracting, normalizing, and rendering hundreds of unstructured file formats and still sits quietly beneath many e-discovery review platforms, notwithstanding vendors’ reluctance to acknowledge this shared dependency.

Aspose Pty. Ltd. licenses commercial APIs that enable software to read from—and more unusually, write to—formats such as Word, Excel, PowerPoint, PDF, and multiple e-mail containers. Read/write capability remains its distinguishing feature, and writing to complex formats continues to be substantially more difficult than extraction alone.

Hyland Software Document Filters support identification and extraction across roughly 500+ formats, with integrated OCR, redaction, and image rendering. Hyland’s filter stack continues to power third-party e-discovery products, including offerings from **Catalyst** and **Reveal**.

dtSearch remains a widely embedded component, combining content extraction with indexing and lexical search. Its technology continues to underpin tools such as **Relativity**, **OpenText LAW**, **Nuix Discover (formerly Ringtail)**, and **Exterro FTK**.

On the open-source side, **Apache Software Foundation Tika** continues to extract text and metadata from well over a thousand file types. Tika remains closely tied to **Apache Lucene**, the indexing engine that still underlies many commercial search and review platforms.

Beyond these few toolsets, the extractor ecosystem drops off sharply. As a result, most modern e-discovery platforms remain proprietary integrations of the same small pool of shared components for file identification, extraction, OCR, normalization, indexing, and viewing. Interface design and workflow engineering vary widely, but tools built on the same extractors necessarily share the same technical limitations, failure modes, and blind spots.

Recursion and Embedded Object Extraction

Correctly identifying file types and decoding schemas is necessary but not sufficient. A processing tool must extract all components of a file that may contain responsive information.

Modern productivity files—such as Microsoft Office documents—are layered containers historically described as ***compound files*****.** They may embed or link content of other formats, such as an Excel spreadsheet displayed within a Word document. A processing tool must recognize embedded and linked objects and extract and enumerate each one.

E-mail is similarly compound. MIME messages[65] transmit multipart content, including encoded attachments that may themselves be containers. An attachment might be a ZIP archive holding PST files that contain messages with attachments that are themselves compound documents. The process by which a tool identifies, unpacks, and extracts all nested content—repeating until nothing remains—is called **recursion**. Reliable recursion is essential to defensible processing.

Family Tracking and Unitization:

As embedded components are unpacked, the processing tool must preserve relationships among the parts, a concept known as ***unitization***. In e-mail, this appears as ***family tracking***: the message is the parent; attachments are children.

Metadata applying to the parent (such as system metadata and custodian) must be preserved, while metadata unique to each child must remain associated with that child. After processing, every item must be traceable back to its originating file as ingested, including its parent-child relationships and source custodian.

Exceptions Reporting:

No substantial data set processes without error. Encrypted files, corruption, unsupported formats, foreign languages, and image-only documents are routine. Many documents—such as faxes, scans, TIFFs, and PDFs without text layers—are not searchable without OCR.

---

[65] **MIME e-mail** refers to an e-mail message structured in accordance with the *Multipurpose Internet Mail Extensions* (MIME) standard, which defines how e-mail can carry multiple, discrete parts beyond simple ASCII text.
Under MIME, an e-mail message may consist of a series of separately encoded "body parts," including plain text, HTML, inline images, and file attachments. Each part is labeled with headers identifying its content type, character encoding, and disposition (for example, whether the content is an attachment or intended for inline display). We study email anatomy in a future chapter and exercise.

A processing tool must identify these failures and generate ***exceptions reports*** so counsel can secure passwords, repair files, or apply OCR. Exceptions signify missing evidence. Exceptions resolution is key to a defensible e-discovery process. Exceptions must be resolved or disclosed and defended.

A particular danger arises when a tool strips partial text from a misidentified file without flagging an exception. The presence of *some* text does not mean that *all* discoverable content was extracted.

Lexical Preprocessing of Extracted Text

Computers do not read or understand language. They apply rules. Before extracted text can be searched or analyzed, it must be **normalized, tokenized, and segmented**—steps collectively known as ***lexical preprocessing***.

Normalization

ESI is numbers; numbers are precise. Variations in those numbers—however subtle to humans—hinder a computer's ability to equate information as humans do. Before indexing and search, extracted text must be **normalized** to increase recall—that is, to ensure retrieval of responsive documents that do not exactly match a query. How we normalize data and even the order in which steps occur materially affect search outcomes.

Character Normalization

Three character features demand normalization: **Unicode equivalency, diacriticals, and case**.

Unicode normalization: In our discussion of ASCII encoding, we established that each ASCII character has a single, corresponding numeric value (for example, a capital "E" is 0100 0101 in binary, 69 in decimal, and 0x45 in hexadecimal). By contrast, linguistically identical characters encoded in Unicode may be represented by different numeric values because accented letters can be expressed in either *precomposed* or *composite* form.

That is, an accented character may be encoded as a single, dedicated code point (a precomposed character), or it may be constructed by pairing the encoding for a base letter with the encoding for a diacritical mark. For example, the Latin capital "E" with an acute accent (É) may be encoded either as U+00C9 (a precomposed Latin capital letter E with acute accent) or as U+0045 (Latin capital letter E) combined with U+0301 (combining acute accent). Both render identically as É.

Easy enough to grasp—but there are many accented variants of Latin letters, for example:

Ĕ Ę̆ Ê Ê Ê̄ Ê̌ Ề Ê̋ Ể Ễ Ệ Ệ Ẹ̀ Ẹ̆ Ẹ̀ Ẹ̇ Ẹ́ Ẹ̃ Ẹ̣ Ë Ë Ẹ̈ Ë̀ Ë́ Ë̃ Ẹ̈ Ë̀ Ë̀ Ẹ̈ Ë̀ Ë́ Ẹ̈ Ë̄ Ë̀ Ë̋ Ë̃ Ẹ̈ Ę̃ Ę̃ Ę̃ Ę̣ Ę̣ ę̣

Plainly, we do not want to account for every possible Unicode encoding variation when constructing a search query in e-discovery. To eliminate that burden, the Unicode Consortium promulgates

normalization algorithms that reduce equivalent characters to a consistent, canonical encoding. One normalization form converts characters to their precomposed versions; another converts them to decomposed (composite) form.

In e-discovery, normalization often goes further, with accents stripped altogether to improve recall—ensuring that a search retrieves linguistically identical characters regardless of how they were originally encoded.

Diacritical normalization goes further by equating accented characters with their unaccented ASCII counterparts so that searches for "resume" retrieve "résumé," and "cafe" retrieves "café." This often includes normalization of ligatures such as the German Eszett (ß) seen in the word "straße," or "street."

Most processing tools offer filters that fold non-Basic Latin Unicode characters into reasonable ASCII equivalents where possible.

Case normalization addresses capitalization. The Latin alphabet is bicameral, meaning it employs upper- and lower-case letters to enhance meaning.[66] By contrast, languages such as Chinese, Arabic and Hebrew are unicameral and use no capitalization.

Because capitalization is inconsistent—especially in e-mail—indexes are typically case-insensitive, so DOE, Doe, and doe return the same results. In some contexts, however, case-sensitive search is desirable, such as distinguishing DOE (Department of Energy) from doe. Case normalization generally requires specifying a default language because capitalization conventions vary by culture.

Impact of Normalization on Discovery Outcomes

Although processing tools draw on a relatively small set of filters and algorithms for normalization, they do not implement them in the same sequence or with identical default settings. As a result, it is routine to see different tools produce different outcomes in culling and search based solely on differences in character normalization. Whether those differences are material depends on the data and the inquiry, but any service provider or case manager should understand how their chosen tool normalizes text.

Across a multi-million-document collection, normalization effects may appear trivial. In reality, character normalization touches every document and plays an outsized role in what is filtered and found. A working knowledge of processing helps e-discovery professionals "normalize"

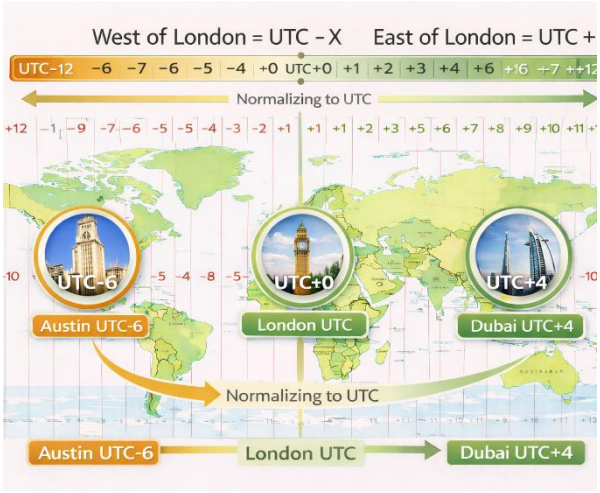
[66] *The terms upper- and lower-case derive from the customary juxtaposition of the shallow drawers or "cases" that held movable type for early printing presses.*

expectations—particularly expectations about what data will be visible and searchable going forward. The most advanced analytics and artificial intelligence can do no better than the data handed to them by processing. If processing is off, it is fancy joinery applied to rotten wood.

Lawyers must fight for quality before review. Review is the stage most lawyers see and understand, so it naturally attracts attention. It is also the costliest phase and the one with the most impressive tools. But here is the bottom line: the most sophisticated MRI scanner cannot save patients who do not survive the trip to the hospital. Triage matters more than equipment. Collection and processing are the EMTs of e-discovery. If quality, completeness, and defensibility are lost before review, review will not save the case.

Time Zone Normalization

You need not be an Einstein of e-discovery to appreciate that time is relative. When reconstructing message threads, it is common to see e-mails sent from Europe to the United States prompt replies that, according to embedded metadata, appear to precede the messages they answer. Time zones and daylight-saving time complicate efforts to place documents and communications in correct chronological order.



Accordingly, a common processing task is to normalize date-and-time values to a single temporal baseline, most often ***Coordinated Universal Time (UTC),*** essentially Greenwich Mean Time, or to another agreed-upon time zone. The ***UTC Offset*** between the source time and the normalized UTC baseline is then expressed as a positive or negative number of hours (for example, UTC-0600 to denote six hours earlier than UTC).

Parsing and Tokenization

Thus far, we have focused on identifying a file's format and extracting its content and metadata—critical tasks, because if you fail to extract and properly decode a file's contents, you have almost nothing to work with. But extraction is only the beginning. Next, the extracted text must be distilled into the linguistic components that convey the file's informational payload. In practical terms, we must isolate the words in the documents and build an index of those words so we can quickly include or exclude files based on their lexical content.

There's a saying that anything a human can do after age five is easy for a computer, while many skills humans master earlier are remarkably hard. Calculate pi to trillions of digits? No problem. Read a *Dr. Seuss* book? Not so much.

Humans instinctively recognize words and sentences, but computers must identify lexical units—called *tokens*—within extracted and normalized character strings. This process is known as ***tokenization***. When search tools look for keywords, they do not scan the extracted text directly. Instead, they consult an index built from the tokens generated during processing. Machines cannot read; they recognize "words" only because patterns of characters satisfy predefined **tokenization rules**.

Those rules vary by system and software. Many indices do not index short words, such as two-letter terms, acronyms, or initialisms. None index single letters or single digits.

Tokenization rules also govern how systems handle punctuation and formatting: compound words like "wind-driven," capitalization and case (will "roof" also find "Roof"?), diacritical marks (will "Rene" find "René"?), and the treatment of numbers and single letters (will a search for "Clause 4.3" work? What about "Plan B?"). Most users assume these searches will work. In many e-discovery tools, they won't.

So how does a computer decide what constitutes a word or a sentence?

Languages based on Latin, Cyrillic, or Greek scripts—such as English and most European languages—are *segmented*: words are typically delimited by whitespace and punctuation. Tokenizers for segmented languages therefore rely heavily on spacing and punctuation to identify word boundaries. What appears simple quickly becomes complicated by hyphenation, contractions, dates, phone numbers, abbreviations, and technical nomenclature.

How does a tokenizer distinguish a line-break hyphen from a true lexical hyphen? How does it distinguish the periods in "Mr." or "G.D.P.R." from periods signaling the end of a sentence? In medicine and pharmacology, many terms include numbers, dashes, and parentheses as integral components. How could one defend a search for *ibuprofen* that failed to retrieve instances of *(RS)-2-(4-(2-methylpropyl)phenyl)propanoic acid*?

Tokenization rules are neither uniform nor comprehensive. Some tools simply cannot index or search certain characters at all, and these exclusions can materially affect discovery. Several years ago, after hard-fought motion practice, a court ordered the parties to search for the term "20%." Only then did everyone learn that their e-discovery tools could not search for the percentage sign.

In Relativity, this limitation exists because the platform uses dtSearch as its indexing engine, and dtSearch reserves the "%" character for another function. No configuration change can make the percentage sign searchable. When you run such a query, you receive no warning that the search is impossible, you simply get no hits.

Using dtSearch (and Relativity) as a further example, there are four mutually exclusive options for how hyphens are indexed. The selected option is baked into the index at creation and cannot be changed without re-indexing the collection. The default behavior treats hyphens as spaces, but dtSearch offers four alternative options for the treatment of hyphens when indexing documents: treat them as spaces, searchable text, ignored, or "all three."

The four options are:

- **Hyphens treated as spaces (default)**
  Hyphens are converted to spaces during indexing and searching.
  "first-class mail" is indexed as *first*, *class*, *mail*.
  Searches for *first class mail*, *first-class mail*, or *first-class-mail* will all match.
  This is the *recommended setting for most e-discovery applications.*
- **Hyphens treated as searchable characters**
  Hyphenated terms are indexed as a single token.
  "first-class" is indexed only as *first-class*.
  Searches for *first class* will **not** match.
- **Hyphens ignored**
  Hyphens are removed entirely.
  "first-class" becomes *firstclass*.
  This setting is rarely useful and often counterintuitive.
- **"All three"**
  Each hyphenated term is indexed in **all forms**:
  first-class, *first class*, and *firstclass*.
  This permits a search for *firstclass* to retrieve *first-class*.

Why selecting "all three" is usually a bad idea:
Although it marginally improves recall for unhyphenated queries, it has serious drawbacks:

- It **bloats the index**, generating up to six indexed terms per hyphenated pair.
- It can produce **unexpected and inconsistent results** in longer phrases or multi-hyphen terms.
- It complicates word positioning, undermining phrase and proximity searches.

Bottom line:

Hyphen handling is a foundational indexing decision, not a cosmetic search preference. In most e-discovery matters, treating hyphens as spaces delivers the most predictable and defensible results. Changing hyphen behavior later requires re-indexing the entire collection, so the choice should be made deliberately and documented.

By default, dtSearch and Relativity treat the following characters as spaces:
!"#$&'()*+,./:;<=>?@[\]^{|}~`

Although some of these characters can be made searchable by changing settings and re-indexing, the following characters **cannot** be indexed or searched in dtSearch and Relativity under any configuration:
( ) * ? % @ ~ & : =

TEST TIP: Remember the difference between **normalization** and **tokenization**:

Normalization is the process of reformatting data to a standardized form, such as setting the date and time stamp of files to a uniform time zone or converting all content to the same character encoding. Normalization facilitates search and data organization.

Tokenization is a method of document parsing that identifies words ("tokens") to be used in a full-text index. Because computers cannot read as humans do but only see sequences of bytes, computers employ programmed tokenization rules to identify character sequences that constitute words and punctuation.

Stop Words

Most search engines exclude common terms—called *stop words* or *noise words*—from the index at the time it is built. E-discovery tools typically exclude **dozens to hundreds of stop words**, meaning those words are never indexed and therefore can never be searched.

By default, dtSearch and Relativity exclude 123 English stop words, along with all punctuation, single letters, and single numbers (see table below). Nuix Discover uses a similar list, but the two tools differ in notable ways. For example, Nuix indexes words such as *between, does, else, from, his, make, no, so, to, use,* and *want* that Relativity excludes, while Relativity indexes words such as *an, even, further, however, not, or,* and *she* that Nuix excludes. Both tools exclude *he* and *her*, both include *hers*, but only Relativity excludes *his*.

Other platforms diverge further. Tools built on the open-source Natural Language Toolkit exclude 179 English stop words, while some commercial systems exclude 500–700 common words. In one notorious instance, the two words that made up the company's own name were both stop words in their e-discovery system, rendering its name—and others like it—unsearchable in its own system.

The practical consequence is stark: **if a word is not indexed, it cannot be searched**. Quotation marks do not help. Search tools do not warn users that a query contains stop words or that the search is incapable of returning hits. A result of "no hits" does not mean "no documents."

Accordingly, those running searches bear responsibility for understanding and disclosing search limitations. If counsel knows a system cannot perform a requested search, that limitation should be affirmatively disclosed. If counsel is uncertain, they must *test and verify* system behavior rather than assume it works as expected.

No e-discovery system is perfect, and perfection is not the legal standard. *But competence requires knowing what a system cannot do.* Only then can counsel work around limitations, adjust expectations, or craft defensible search strategies grounded in reality rather than magical thinking.

Begins with	Noise words excluded by default in *dtSearch* and Relativity
A	a, about, after, all, also, an, and, another, any, are, as, at
B	be, because, been, before, being, between, both, but, by
C	came, can, come, could
D	did, do
E	each, even
F	for, from, further, furthermore
G	get, got
H	had, has, have, he, her, here, hi, him, himself, his, how, however
I	i, if, in, indeed, into, is, it, its
J	just
L	Like
M	made, many, me, might, more, moreover, most, much, must, my
N	never, not, now
O	of, on, only, or, other, our, out, over
S	said, same, see, she, should, since, some, still, such
T	take, than, that, the, their, them, then, there, therefore, these, they, this, those, through, thus, to, too
U	under, up
V	very
W	was, way, we, well, were, what, when, where, which, while, who, will, with, would
Y	you, your

Source: Relativity website, January 17, 2026

Building a Database and Concordance Index

This chapter focuses on processing tools. Databases and viewers are typically associated with review platforms, but a brief look at their interrelationship is useful because processing feeds—and in many ways defines—what review tools can do.

The Review Database

Throughout processing, information derived from and about each item is continuously written to a database. As files are ingested, the database records file names, sizes, and system metadata. Sources—called *custodians* when they are individuals—are identified and stored as database fields. Hash values calculated during processing are added to the database, as are application metadata values such as document authorship and e-mail headers (including sender, recipients, subject, and dates).

The database also maintains pointers to extracted text and to page-level images (such as TIFF or PDF renderings). Each processed item is enumerated—assigned a unique identifier that distinguishes it within the collection. This identifier is separate from any Bates number later applied for production.

The database sits at the center of every e-discovery review platform. It receives much of the output of processing and organizes it for retrieval. But the database is **not** the text index. The **database**, the c**oncordance index**, and the **document viewer** are distinct components, even though they operate so tightly together that they appear as one.

When a user runs a search, the query is executed against the index. The database then supplies information about the items "hit" by the query, and the viewer displays those items—often with search hits highlighted. What users perceive as the quality of a review tool largely reflects how smoothly these components integrate into a coherent and responsive interface.

As with file identification and extraction, most e-discovery vendors do not build database engines from scratch. Instead, they rely on a small number of established database platforms, such as SQL Server or SQLite. Despite marketing claims of uniqueness, many e-discovery tools share similar database foundations. These commonalities are largely invisible to users because the back end is hidden behind a slick interface.

The Concordance Index

A *concordance* is an alphabetical listing or mapping of significant words in a text. Historically, scholars constructed concordances to works like Shakespeare or the Bible by hand. In e-discovery, software builds concordance indices to enable fast lexical searching.

While it is technically possible to search documents one at a time—so-called *serial search*—that approach is prohibitively slow for modern collections.[67] Instead, processors extract text, tokenize it

[67] Computer forensic examiners still use serial searches when the corpus is modest and when employing Global Regular Expressions (GREP searches) to identify patterns conforming to, *e.g.*, social security or credit card numbers.

to identify words, and build an index that maps each token to the documents and locations in which it appears.

As a result, keyword searches in e-discovery do **not** search the evidence itself. They search a concordance index built from tokenized text. This distinction is critical: *search quality is limited by the completeness and accuracy of the index.* If content is not extracted, normalized, tokenized, and indexed, it is effectively invisible to search.

Indexing is therefore the process by which processed text is transformed into a highly efficient cross-reference structure that enables rapid retrieval. Everything that follows in review depends on getting this step right.

Culling and Selecting the Dataset

Processing is not an end in itself. It is the means by which potentially responsive information is exposed, enumerated, normalized, and prepared for search, review, and production. Although much culling occurs later during search and review, the processing phase offers an important—and defensible—opportunity to reduce data volumes using objective criteria.

Once metadata resides in a database and the collection is made text-searchable through creation of a concordance index, the dataset can be filtered by characteristics such as date ranges, file types, Internet domains, file size, custodian, and other neutral attributes. Processing also enables systematic culling through immaterial item suppression, de-NISTing, and deduplication, each addressed below.

The bluntest and most common culling tool remains keyword and query filtering, that is, lexical search. Its mechanics and shortcomings are addressed in later chapters, but one point should already be clear: *search quality is bounded by processing quality.* No review process can assess the substance of items that were never extracted, were misprocessed, or were excluded from the index.

Immaterial Item Suppression

Processing tools must preserve the ability to trace every extracted object back to its originating source file. Every item—container files and their contents alike—must be catalogued and enumerated for forensic completeness. That said, many such items have no independent evidentiary value once their contents are properly extracted.

Immaterial items are those retained for structural or forensic reasons but carrying little or no intrinsic discovery value. Common examples include folder paths and container or wrapper files such as ZIP or RAR archives, mailbox containers like Outlook PST or MBOX files, and forensic image wrappers such as .E0x or .AFF files. These items are relevant only as vessels for their contents.

Accordingly, it's handy to be able to suppress immaterial items once we extract and enumerate their contents. It's pointless to produce a ZIP container if its contents are produced, and it's perilous to do so if some contents are non-responsive or privileged.

De-NISTing

De-NISTing is a culling technique used in e-discovery and computer forensics to reduce review volumes by excluding standard operating system and commercial software files—materials that are ubiquitous, non-user-generated, and almost never relevant. Examples include Windows system files, screen savers, sound files, document templates, clip art, and executable components of applications like Microsoft Word or Excel.

These files are identified by comparing their cryptographic hash values—digital fingerprints unique to file contents—to reference hash sets published by the **National Software Reference Library (NSRL)**, maintained by the National Institute of Standards and Technology (NIST). Files whose hashes match entries in the NSRL are presumed to be standard software components and can be excluded from review. The NSRL hash sets are publicly available and are widely used by e-discovery and forensic tools.

The effectiveness of de-NISTing depends on how the data was collected. It is most valuable when collections are broad or indiscriminate, such as full forensic images of hard drives that include operating systems and installed applications. De-NISTing has far less impact when collections are narrowly targeted to user-created content, such as documents, e-mail, and cloud data.

As a general rule, the more precise the collection effort, the less data de-NISTing will remove. Even so, de-NISTing remains a defensible and well-established method for eliminating digital noise and focusing review on materials that actually matter.

Cryptographic Hashing (Yes, Again—Because It Matters)

We spend a good deal of time in this course hammering home a simple but unintuitive truth: **all ESI is numbers**. We slog through Base2 (binary), Base10 (decimal), Base16 (hexadecimal), Base64, and the difference between single-byte encodings like ASCII and multibyte encodings like Unicode. It can feel like an unnecessarily wonky detour. It isn't. Once you grasp that digital evidence is numerically encoded, you gain Proustian new eyes for ESI and how mathematics can be used to identify, cull, filter, and cluster data at scale.

Because ESI is numbers, we can apply mathematical algorithms to distill and compare it. Chief among these tools are **cryptographic hash functions**, which generate digital fingerprints of data.

Hashing underlies file identification, authentication, exclusion, and deduplication throughout e-discovery.

A hash function is a published, standardized mathematical algorithm into which we input digital data of any size (the *message*). The algorithm outputs a fixed-length numeric value—the *hash value*. That value corresponds almost uniquely to the input data. While it is theoretically possible for two different inputs to produce the same hash value (a *hash collision*), the odds are vanishingly small for legitimate e-discovery use. Accordingly, hash values are not truly "unique," but they are reliable enough to function as digital fingerprints.

Hashing is a **one-way operation**. You cannot reverse a hash value to reconstruct the original data, just as you cannot derive a person's traits from a fingerprint. Hashing identifies data without revealing its contents. A second defining characteristic is the *avalanche effect*: even a trivial change in the input produces a dramatically different hash value. Similar hash values do **not** imply similar data.

Different hash algorithms produce different hash values for the same data. For example, the phrase *"Mary had a little lamb"* yields:

- **MD5:** e946adb45d4299def2071880d30136d4
- **SHA-1:** bac9388d0498fb378e528d35abd05792291af182
- **SHA-256:** efe473564cb63a7bf025dd691ef0ae0ac906c03ab408375b9094e326c2ad9a76

The input is identical; the algorithms are not. Conversely, identical data hashed with the same algorithm will always produce identical hash values. Because hash algorithms are publicly available, matching hash values generated independently are strong evidence the underlying data is identical.

MD5 in E-Discovery

A widely used hash function in e-discovery is **MD5**, published in 1992 by MIT professor Ron Rivest as Request for Comments 1321. Requests for Comments or RFCs are a way the technology community circulates proposed standards and innovations generally relating to the Internet. MD5 produces a **128-bit** hash value, commonly expressed as **32 hexadecimal characters**. Those 32 characters represent a number drawn from a universe of approximately **3.4×10^{38} possible values**—a vast number space: 340 *trillion, trillion, trillion* different possible values (2^{128} or 16^{32}) or in decimal terms, 340,282,366,920,938,463,463,374,607,431,768,211,455.

MD5 is no longer considered collision-resistant for security applications because it is possible to engineer different inputs that produce the same MD5 hash. As a result, MD5 has been abandoned for uses like digital signatures and cryptographic authentication.

That weakness, however, has little practical impact in e-discovery. Tasks such as deduplication and de-NISTing are not adversarial cryptographic contests. They do not depend on resisting deliberately engineered collisions. MD5 remains fast, efficient, and more than adequate for identifying identical files in discovery collections. For these purposes, MD5 "ain't broke," and fixing it would mostly add cost and computational burden without improving outcomes.

Some tools supplement or replace MD5 with stronger algorithms such as SHA-256, and there is nothing wrong with doing so. But the continued use of MD5 in e-discovery reflects pragmatism, not negligence.

Litigation Sidebar: Cryptographic Hashes and Authenticity Under the Rules of Evidence

Cryptographic hash values play a quiet but powerful role in satisfying **authentication requirements** under the Rules of Evidence, particularly **Federal Rule of Evidence 901(a)**, which requires the proponent of evidence to produce sufficient proof that the item is what it is claimed to be. A matching hash shows **data integrity**—that a file has not changed since a prior point in time. When the same file yields the same hash at collection, processing, production, and trial, it is strong technical evidence the contents are identical.

Hashing fits comfortably within **FRE Rule 901(b)(9)**, which permits authentication by describing a reliable process or system. Courts routinely accept testimony that a standard hash algorithm was applied consistently and produced matching values as sufficient to authenticate digital evidence.

Hashes are also useful in **chain-of-custody** disputes. While the Rules do not require an unbroken chain, hash values help rebut claims of tampering or substitution by showing that the digital evidence at trial is bit-for-bit identical to the evidence collected earlier. This is especially persuasive where data passes through multiple hands, systems, or vendors.

It is important to understand what hashes do **not** prove. A matching hash does not establish who created a file, when it was authored, or whether its contents are truthful. Hashes authenticate **identity and integrity**, not authorship or accuracy. Nor do hashes make evidence self-authenticating under Rule 902; they still require testimony or stipulation explaining their significance.

In short, cryptographic hashes are not magic talismans, but when properly generated, documented, and explained, they are among the most effective technical tools lawyers have for demonstrating that digital evidence is what it purports to be—nothing more or less.

Deduplication

Calculating hash values during processing enables several capabilities, perhaps none more valuable than **deduplication**—the identification and suppression of identical information items to reduce review volume.

Because modern storage media hold vast quantities of data, it is far faster and more efficient to compare **hash values**—those compact digital fingerprints—than to repeatedly compare the underlying data itself. Accordingly, as each file or extracted item is ingested, its hash value is compared to those already seen. If a match is found, it's a duplicate. The first encountered instance is retained, while subsequent duplicates are suppressed, with their existence and associated metadata recorded in a deduplication or *occurrence* log.

When deduplication works cleanly, it works because the digital *contents* of the files are identical. That distinction matters.

Hashing typically operates on a file's contents, not on system metadata maintained outside the file. For example, a file name is usually stored in the file system, not inside the file itself such that changing a file's name will not change its hash value. Conversely, any change to the file's contents—however slight—will alter its hash.

This becomes critical when the *same informational content* exists in different digital formats.

A Word .DOCX document is constructed of a mix of text and rich media encoded in Extensible Markup Language (XML), then compressed using the ubiquitous ZIP compression algorithm. It's a file designed to be read by Microsoft Word. When you print the "same" Word document to an Adobe PDF format, it's reconstructed in a *page description language* specifically designed to work with Adobe Acrobat. It's structured, encoded and compressed in an entirely different way than the Word file. Print that PDF and scan it to Tagged Image File Format (TIFF), and you now have a photographic image encoded in yet another format. To humans, these are all the "same" document. To a computer, they are radically different.

To illustrate this, I created a Word document of the text of Lincoln's Gettysburg Address. First, I saved it in the latest .DOCX Word format. Then, I saved a copy in the older .DOC format. Next, I saved the Word document to a .PDF format, using both the Save as PDF and Print to PDF methods. Finally, I printed and scanned the document to TIFF and PDF. Without shifting the document on the scanner, I scanned it several times at matching and differing resolutions.

I then hashed all the iterations of the "same" document. As the table below demonstrates, none of them matched hash-wise, not even the successive scans of the paper document:

FILENAME	MD5 HASH	FILE SIZE
GBA.docx	5074fbb210ed4e9e498e4908a946a871	21Kb
GBA.doc	1aacf60b523eb8cf2829208ffee58005	26Kb
GBA-Save as.pdf	c8d68e84ea573772d14dc536fbe8594e	83Kb
GBA-Word generated.pdf	2be09d776682fee46c79be8ecac03ec5	27Kb
GBA-scan1.tiff	0f5fdbbcbc96abc05b43f356c4e24818	967Kb
GBA-scan2.tiff	04c93ac7eb6716bc96bc3a396fed882a	967Kb
GBA-scan3_600BW.tiff	93e726efa56fe7f25956da6664a32957	1,060Kb
GBA-scan4_600BW.tiff	8d97df97c28414d4b61bb8b88b1db343	1,060Kb
GBA_scan5_300GS.pdf	b558eccee1bdcc5f26de53763f89aef4	2,950Kb
GBA_scan6_300GS.pdf	520be78a7ec81ebebece5a19e9c6e425	2,930Kb

As a result, **hash-based deduplication will not match different electronic forms of the same document**, nor will it match successive optical scans of the same paper document. Each scan introduces small variations in the data sufficient to produce different hash values. Testing confirms this: even multiple scans of the same page, made moments apart, will not hash-match.

By contrast, true copies of the same electronic file—for example, duplicated attachments or forwarded files—will hash-match and deduplicate cleanly, because the underlying data is identical, although the system metadata values will change. This distinction between *copies* and *conversions* is central to understanding deduplication.

Deduplication of E-Mail

Deduplicating e-mail is more complex than deduplicating loose files. Mechanized e-mail deduplication generally follows one or more of these approaches:

1. **Whole-message hashing**
   Each message is hashed as a complete data object. If the hash values match, the messages are identical. This approach often fails with exported e-mail because messages that appear identical to users frequently differ at the byte level due to routing headers, timestamps, or message-ID values.
2. **Segment-level hashing**
   Hashes are calculated for selected portions of the message—such as subject, sender, recipients, body, and attachments—and corresponding segments are compared. This approach requires excluding inherently variable header data and normalizing segments so they are presented consistently to the hash algorithm.
3. **Segment-level textual comparison**

Instead of hashing segments, their text is compared directly to determine whether messages are sufficiently alike to be treated as duplicates. This method resembles segment-level hashing but without cryptographic fingerprints.

In practice, tools may use hybrids of these approaches.

As noted, all methods fail when parties work from different forms of the same data. An e-mail rendered as a TIFF image bears no digital resemblance to the original message, just as a photograph of a rose neither smells nor feels like the rose itself.

The Bottom Line

Changing the form of ESI changes the data, and changing the data changes the hash. **Hash-based deduplication requires identical source data and consistent processing workflows**. When those conditions are met, deduplication is simple, fast, and defensible.

When parties cannot deduplicate e-mail, the causes usually trace to one or more of the following:

1. They are working from different forms of the ESI
2. They fail to exclude inherently variable data (such as message IDs or routing headers)
3. They fail to normalize message content consistently
4. They use different hash algorithms
5. They fail to preserve hash values throughout the workflow
6. They alter the data during processing

Deduplication is not magic. It is math applied to data. When the data changes—or is mishandled—the math cannot save you.

Litigation Sidebar: Deduplication Disputes and Meet-and-Confer Failures

Deduplication often sparks discovery disputes because parties assume they mean the same thing when they say "we deduplicated." That phrase can describe global deduplication across custodians, custodian-level deduplication, attachment-only deduplication, or hybrids. Failure to define the method invites confusion and motion practice.

Disputes also arise when parties work from **different forms of the same data**. Hash-based deduplication cannot operate across natives, TIFFs, and PDFs. Without clear disclosure, this mismatch can be miscast as underproduction or spoliation.

Meet-and-confer breakdowns typically occur when counsel cannot explain:

- the scope of deduplication (global vs. custodian-specific),
- whether messages, attachments, or both were deduplicated,
- what metadata was preserved for suppressed duplicates, and
- which hash algorithm and normalization steps were used.

Courts generally permit deduplication but expect **transparency**. Silent deduplication—especially without occurrence tracking—can look like evidence loss. Parties should disclose their approach early and document it.

Finally, deduplication is not a substitute for judgment. Suppressing duplicates that carry unique custodial or path metadata may undermine proof of notice, distribution, or intent. Done thoughtfully, deduplication reduces volume; done carelessly, it distorts the evidentiary record.

Most deduplication disputes are not technical failures but failures to communicate.

Other Processing Tasks

This chapter addresses the core functions of ESI processing in e-discovery, but there are many other processing tasks that make up today's powerful processing tools. Some examples include:

Foreign Language Detection

Tools may identify the language of content to select appropriate extraction filters, character sets, and diacritical handling, and to route documents to reviewers with suitable language skills.

Entropy Testing

Statistical analysis used to flag likely encrypted files so they can be handled as processing exceptions.

Decryption
Some tools can automatically decrypt password-protected files using known or supplied credential lists.

Bad Extension Flagging
Files whose binary signatures do not match their extensions can be flagged to detect corruption, mislabeling, or data hiding.

Color Detection
Where color conveys meaning, tools can identify color-dependent content and steer such items away from grayscale imaging.

Hidden Content Flagging
Processing can identify embedded or non-visible content—such as Word comments, tracked changes, or PowerPoint speaker notes—that may not appear in standard productions but remain discoverable.

N-Gram and Shingle Generation
Text sampling techniques used to support advanced analytics, including clustering, similarity analysis, and predictive coding.

Optical Character Recognition (OCR)
OCR converts image-based documents (such as TIFFs, scanned PDFs, and JPGs) into searchable text. Some tools perform OCR natively; others require a separate OCR workflow followed by re-ingestion.

Virus Scanning
Collected data may contain malware. Processing tools may scan and quarantine infected files to protect systems and workflows.

Production Processing
Processing does not end with review. Items selected for production are often converted to production formats (TIFF or PDF), redactions are applied, and metadata and data describing the organization of the production, such as pointers to TIFF images and text extractions, is exported via

load files. Production processing also includes Bates numbering and application of confidentiality legends (*i.e.*, "Produced Subject to Protective Order").

These auxiliary tasks round out the processing pipeline, reinforcing the central point of this chapter: **processing shapes what can be searched, reviewed, and safely produced.**

Illustration 1:
E-Discovery Processing Model

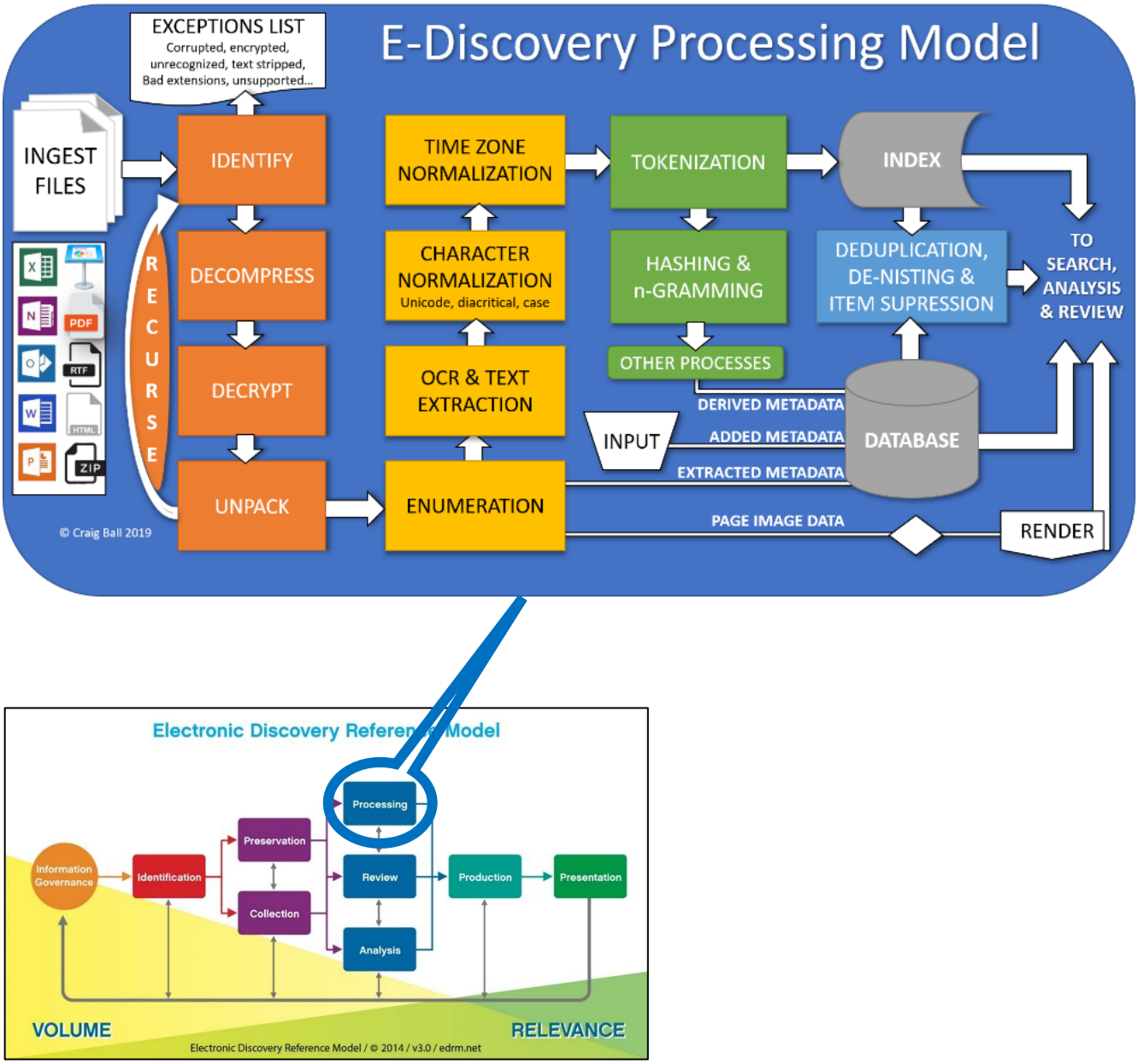
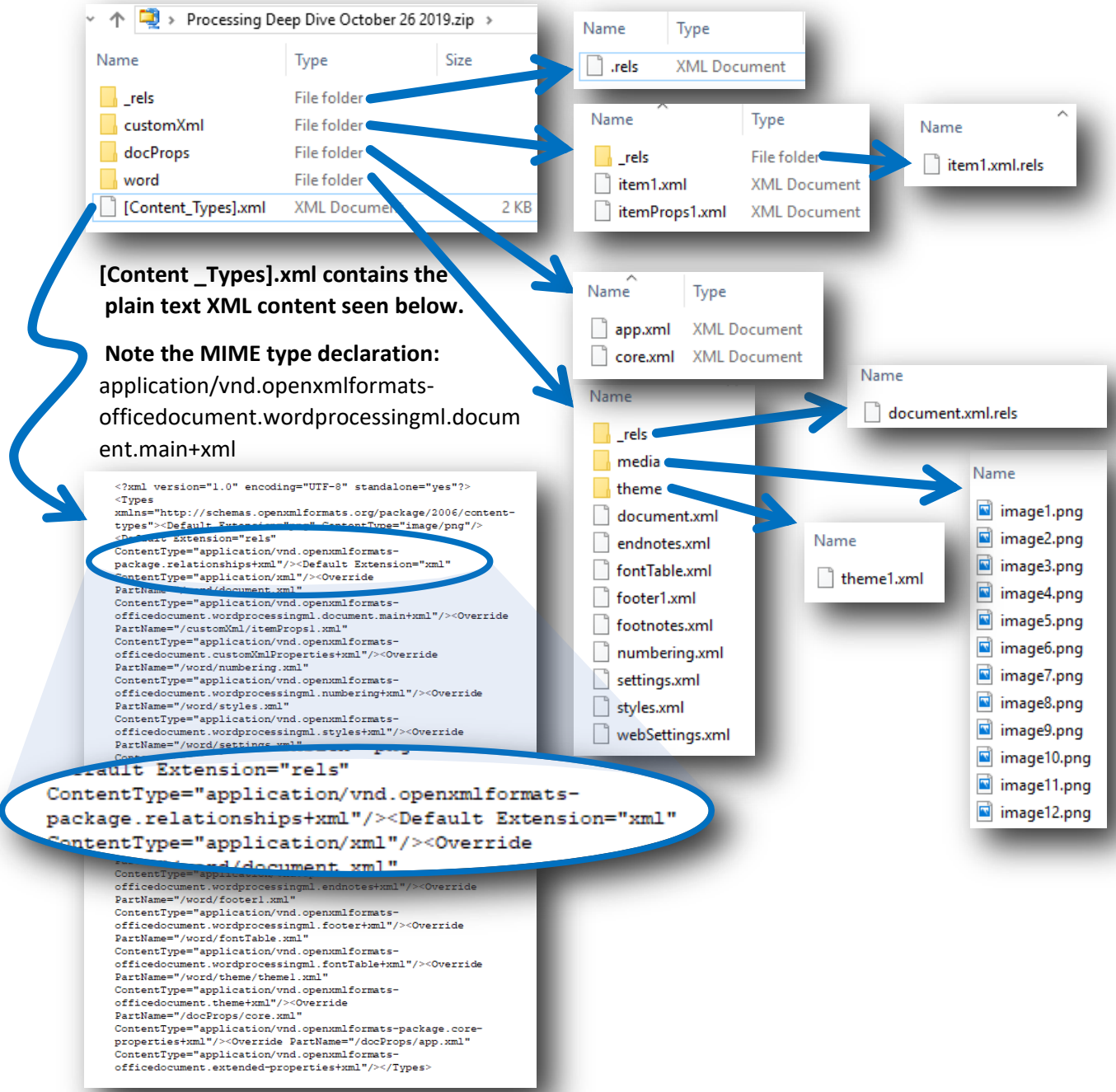


Illustration 2: Anatomy of a Word DOCX File

By changing a Word document's extension from .DOCX to .ZIP, you can "trick" Windows into decompressing the file and sneak a peek into its internal structure and contents. Those contents little resemble the document you'd see in Word; but, as you peruse the various folders and explore their contents, you'll find the text, embedded images, formatting instructions and other components Word assembles to compose the document.

The root level of the decompressed file (below left) contains four folders and one XML file.



Processing Glossary	
Term	**Definition**
ASCII	American Standard Code for Information Interchange. A single-byte character encoding standard mapping 7- or 8-bit integers to 128 or 256 characters. Seven-bit ASCII includes 95 printable characters and 33 control codes. So-called "Extended ASCII" refers to various incompatible 8-bit extensions.
Bates Numbers	Sequential identifiers applied to documents or files in discovery, typically including a prefix identifying the producing party or matter (e.g., DEF_000000001).
Binary File Signature	A short sequence of bytes, typically at the beginning of a file, that identifies the file's format. For example, ZIP files begin with hex 50 4B; most JPEGs begin with FF D8 FF.
Case Normalization	Indexing technique that enables case-insensitive searching by equating uppercase and lowercase forms of the same letters (e.g., Rice, RICE, rice).
Character Normalization	Processes that reduce variation in characters that humans readily overlook but computers treat as distinct, including case, diacritical, and Unicode normalization.
Chain of Custody	Procedures used to document and preserve the integrity of evidence from acquisition through use, demonstrating it has not been altered or corrupted.
Compression	Reduction of data size for storage or transmission. *Lossless* compression removes redundancy without changing content (e.g., ZIP, RAR). *Lossy* compression discards nonessential data (e.g., JPEG).
Container File	A file that holds other files, such as compressed archives (ZIP, RAR), email containers (PST, MBOX), or forensic image wrappers. Containers are unpacked during processing and may then be suppressed as immaterial.
Corruption	Damage to a file that impairs its ability to be decoded or processed, caused by factors such as transmission errors, software faults, or physical media damage.

Custodian	A person or entity that possesses or controls potentially responsive information.
DAT File	A delimited text load file used to transmit metadata and structural information in productions. Fields are separated by defined delimiters and correspond to column headers.
Deduplication	Identification and suppression of identical items within a dataset, typically based on matching hash values or normalized message components.
DeNIST	Use of hash comparisons against the National Software Reference Library (NSRL) to identify and suppress known operating system and commercial software files.
Diacritical Normalization	Indexing technique that equates accented and unaccented characters to improve recall (e.g., résumé ↔ resume).
DTSearch	A commercial content extraction, indexing, and search engine embedded in many e-discovery and forensic tools, including Relativity and Nuix Discover.
Encoding	Mapping of characters and symbols to numeric values for electronic storage and transmission. ASCII and Unicode are character encoding schemes.
Encryption	Transformation of data into unreadable ciphertext to prevent access without a decryption key.
ESI	Electronically Stored Information, as defined in Federal Rule of Civil Procedure 34(a)(1)(A).
Exception Reporting	Identification and tracking of items that fail processing, including encrypted, corrupt, unsupported, or image-only files requiring OCR.
Family Group	In email, the relationship between a parent message and its child attachments.
File Header Signature	Also known as a "binary header signature," "binary file signature" or "magic number." Typically, the first few hex bytes of data in a file identifies the format of the data within the file. For example, ZIP compressed files begin with Hex 504B (or the initials PK in ASCII). Most JPG image files begin with Hex FF D8 FF E0.

Filtering`	Culling data based on objective criteria such as file type, date, size, custodian, hash value, or lexical content.
Forensic Image	A verified, bit-for-bit copy of electronic media that preserves active data, slack space, and unallocated space.
Hash	A fixed-length digital fingerprint generated by a one-way cryptographic algorithm (e.g., MD5, SHA-256). Hashes support identification, deduplication, and exclusion, though collisions are theoretically possible.
Identification	Determination of a file's format and encoding, typically based on its binary signature and extension.
IM	Real-time text communications conducted over the Internet, often in conversational or group formats.
Image Format	A page-oriented representation of content (e.g., TIFF, PDF, JPEG) used to emulate printed documents. Rendering is the process of converting native files to images.
Index	A data structure that enables fast retrieval by mapping tokens to documents and locations. E-discovery relies on full-text indices.
Ingestion	Loading data into a processing system.
Keyword	A term used to query an index or database.
Language Detection	Identification of the language of content to support proper text extraction, normalization, and reviewer assignment.
Load File	A companion file that transmits metadata, text, Bates numbers, and structural information alongside produced files.
MD5	A 128-bit cryptographic hash algorithm widely used in e-discovery for identification, deduplication, and de-NISTing despite known collision vulnerabilities.
Metadata	Data describing other data. Includes system metadata (stored outside a file), application metadata (stored within a file), and review metadata (human judgments and coding).
MIME (Media Type)	A standardized classification system for file types used on the Internet (e.g., application/pdf, image/tiff), maintained by IANA.

Native Format	The original file format created and used by an application (e.g., DOCX for Microsoft Word).
NSRL	The National Software Reference Library maintained by NIST, providing hash sets used for de-NISTing.
Noise Words/Stop Words	Common terms excluded from an index to improve performance and reduce storage.
Normalization	Standardization of data formats, such as time zones or character encodings, to facilitate search and analysis.
OCR or Optical Character Recognition	Conversion of image-based text into machine-readable text for searching.
Processing	The extraction, normalization, indexing, and culling of ESI to prepare it for search, review, and production.
Recursion	Repeated exploration and extraction of embedded content within files until no additional content remains.
Request for Comment (RFC)	A published technical document proposing or defining Internet standards and protocols.
SHA	A family of cryptographic hash algorithms (e.g., SHA-1, SHA-256) used for identification and deduplication.
SMS	Short Message Service, a protocol enabling mobile text messaging.
System Files	Non-user-created `operating system and application files, typically excluded through de-NISTing.
Targeted Collection	A collection strategy that limits acquisition to likely responsive data rather than imaging entire devices.
Threading	Organization of messages into conversational sequences based on reply relationships and timestamps.
Tika	An open-source toolkit for extracting text and metadata from many file types, derived from Apache Lucene.
Time Zone Normalization	Conversion of timestamps to a common temporal baseline, often UTC or a designated local time zone.

Tokenization	Identification of word units (“tokens”) from text for indexing. Tokenization rules vary by language and system.
Unicode	A multibyte international character encoding standard supporting text, symbols, and emoji across languages.
Unicode Normalization	Reduction of canonically equivalent Unicode character sequences to a standard form for consistent indexing.
UTF-8	The most widely used Unicode encoding, backward-compatible with ASCII and variable in byte length.



Exercise 10: Metadata: File Table Data

GOALS: The goals of this exercise are for the student to:

1. Distinguish between system metadata and application metadata;
2. Explore the Windows Master File Table; and
3. Understand that, because system metadata is not stored within the file, you don't preserve system metadata by simply copying the contents of the file; you must grab its system metadata from the file table, too.

OUTLINE: Students will create a file in Notepad and search within and without the file for metadata.

This exercise is belated return to metadata, but purposefully, because we needed some topics covered in the processing chapters, like ASCII and code pages. **Note there is nothing to submit for this exercise.**

Background

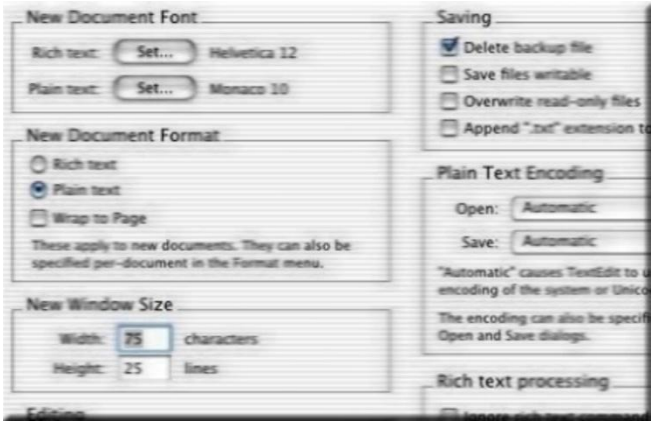
Once again we will use the HexDump utility at http://www.fileformat.info/tool/hexdump.htm, to view every byte in a file as ASCII text and hex values. So, when you know a file has attendant metadata that you can't see within the four corners of the file, it must be *somewhere*. Where does that metadata come from?

Step 1: Create a simple text file

In Windows: On your Desktop, right click on an open area select New>Text Document. Name the file "**me.txt**," then open the file you just created and type your full name. Save the file, then close it. Double click the file to re-open it and confirm that your name appears in the document named me.txt on your Desktop.

In MacOS: You will use the Mac default text editor called TextEdit to create a plain text (ASCII) file. But, since the Text Edit program creates Rich Text (RTF) formats by default, you must first modify some program settings:

a. Open TextEdit.
b. Choose Preferences from the TextEdit application menu.
c. Click the Plain Text radio button for New Document Format.
d. Be sure the checkbox for "Wrap to Page" is deselected.
e. Close the Preferences box.

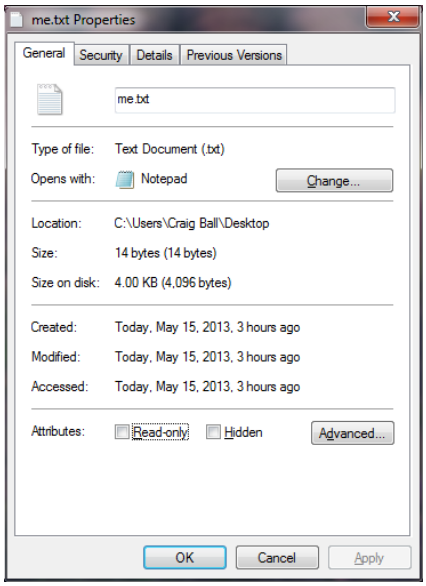


Create a new file using TextEdit and type just your full name in the file. Save the file as "me.txt" to your Desktop and close it. Re-open the me.txt file and confirm that your name appears in the document.

Step 2: Check the Properties

In Windows: Right click on your me.txt file and select "Properties."

In MacOS: Right click and select Get Info



Note the file's size and its size on disk. The first reflects the actual byte count for the data needed to store your name (including spaces). The size on disk is the total size of the cluster(s) allocated to storing the file. On my Windows machine, the drive is logically divided into 4 kilobyte clusters, so every file occupies at least 4KB on disk as seen in the figure at right (but see the discussion of resident MFT data below).

Step 3: Dump the Hex

Using your web browser, go to the Online HexDump Utility at http://www.fileformat.info/tool/hexdump.htm and click "choose File." Using the selection box that will appear, navigate to the file you just created called "me.txt." Click "Open." Now click the blue "Dump" button on the Online HexDump Utility page. You should see something like this (but with *your* name, of course):

```
file name: me.txt
mime type:

0000-000e:  59 6f 75 72-20 66 75 6c-6c 20 4e 61-6d 65        Your.ful l.Name
```

Step 3: Carefully examine the complete contents of the file

Look at the hex. Look at the text. Do you see any data within the file other than your name? Do you see any file path (location) data? Do you see any date or time data? *Do you even see the file's name within the file?* Every file has system metadata, *so where's the metadata if it's not in the file? It's in the MFT!*

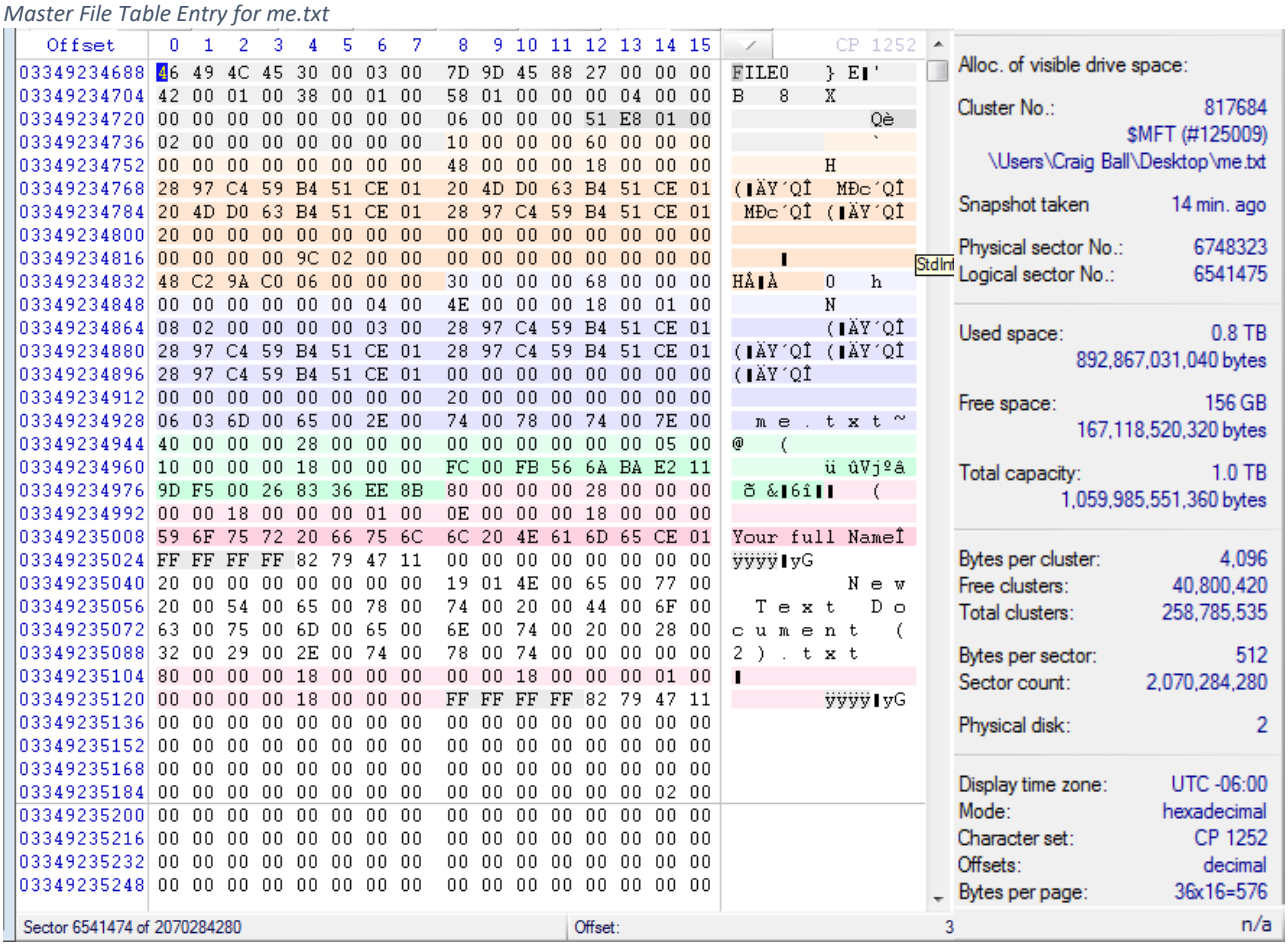
Plumbing the Windows MFT

Recall that **MFT** stands for **Master File Table.** On a Windows system, the MFT is like a library card catalog, storing information about the location of the "book" (file) and describing some of its

characteristics (system metadata). The MFT is where most system metadata reside, in contrast to *application* metadata, which resides within the file it describes and moves with the file when copied.

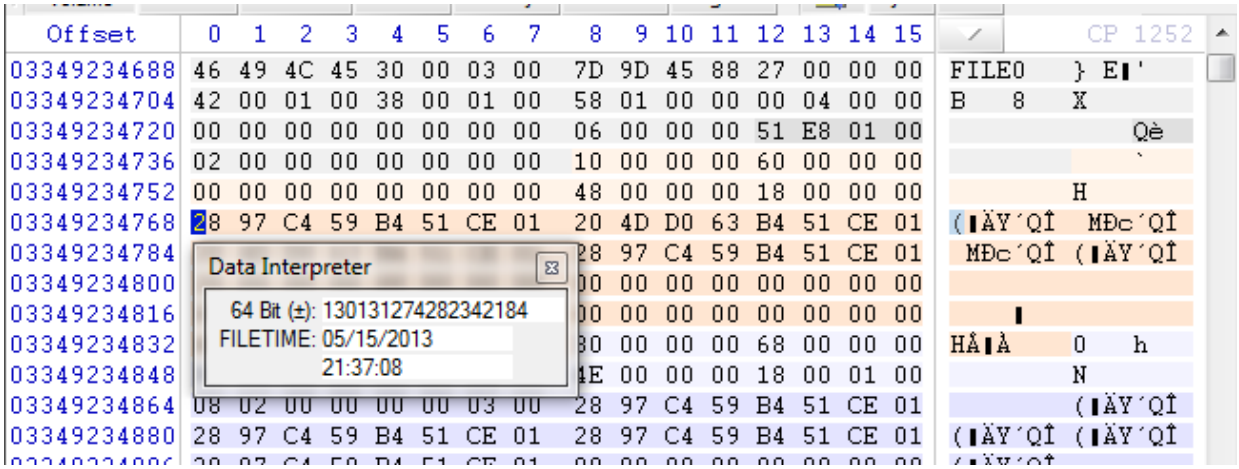
The MFT is made up of numerous 1,024-byte entries, each describing a file or folder stored on the media. The image below is a screenshot of the MFT entry for the **me.txt** file on my Windows Desktop. Like all MFT entries, it begins with FILE0, and after some weirdly encoded stuff, you'll see the name stored in code page 1252 (which if you remember our processing discussion at p. 117 is Microsoft's version of Latin 1). Note the spaces between the letters of the file name, which tells us it is double byte encoded.

Master File Table Entry for me.txt



An interesting aspect of the MFT is that if the contents of a file are sufficiently small (less than about 750 bytes), the operating system doesn't *really* create a file at all. Instead, it stores the contents *right in the MFT* and just *pretends* to create a discrete file. Because the me.txt file holds so little content, we can see that content stored right in the MFT entry (beginning FILE0).

The MFT also stores date and time values reflecting when the file was Modified, Accessed and Created. You can't see them because they are encoded in an extraordinary way. Windows file times are stored as (and I absolutely LOVE this) **a value equal to the number of 100 nanosecond intervals since January 1, 1601**. Thus, if you look at the hex content from the MFT entry, the sixth line down begins with these eight bytes: 0x2897C459B451CE01. This is a 64-bit numeric value equivalent to the decimal 130,131,274,282,342,184. It also happens to *equal the number of 100 nanosecond intervals between January 1, 1601 and May 15, 2013 @21:37:08 UTC, when I created the file.*



Discussion Points to Ponder (NOT homework):

1. If the contents of the file me.txt reside in the MFT, why does Properties state that the file is taking up 4KB of space on disk?
2. When we copy a file from one media to another (as might occur when collecting ESI for processing), what MFT metadata routinely follows the file to its destination? Why these fields? What metadata is lost unless overt steps are taken to collect it? Does the destination medium have its own Master File Table? What data is lost when the file tables of the source and target media are incompatible?

Introduction to Digital Forensics and Examination Protocols

Digital forensics sits at the intersection of rapidly evolving technology and enduring human behavior. The forensic examiner finds the human drama manifesting as digital traces—sometimes buried in vast data haystacks, sometimes distilled to a single byte marking an act like opening, deleting, tagging, or altering a file. It is exacting work. Competence in the discipline requires moving at the speed of technological change while maintaining the deliberate precision demanded by law.

What is Computer Forensics?

Operating systems (such as Windows, macOS, and Linux) and the applications that run on them generate and retain far more information than most users ever see. Only some of this information is **active data**—content readily accessible to users—though even active data may require skilled interpretation to illuminate behavior. Metadata tracked by operating systems and applications is a familiar example. Windows Explorer, for instance, can display geolocation data embedded in photographs, yet few users ever enable such views.

Other active data exist in obscure locations or **encoded**[68] forms that are neither obvious nor intelligible to end users but are highly probative when interpreted and correlated by forensic experts. Log files, system files, and structured or non-text records often reveal how a system was used. In a data theft investigation, for example, decisive evidence may reside in the Windows Registry, where time-stamped entries record when USB storage devices were connected.

Beyond active data lie vast regions of storage that operating systems no longer use but do not erase. These areas—**unallocated space** and **file slack**—are digital boneyards holding **forensic artifacts** discarded over the life of a device. Accessing and interpreting this fragmented detritus requires specialized tools, methods, and experience.

Computer forensics is the expert acquisition, interpretation, and presentation of data drawn from all three domains—active, encoded, and forensic—evaluated in context and often correlated with other evidence such as financial records, access logs, phone data, surveillance video, cloud accounts, social media, documents, and messages.

In litigation, computer forensics is not confined to personal computers and servers. Phones, tablets, external drives, memory cards, and removable media are routine subjects of examination. Increasingly, relevant evidence resides in cloud repositories, Internet-of-Things devices, vehicles, and embedded systems. Like a physical crime scene, the scope of forensic inquiry expands to reflect the available evidence and the questions at issue.

[68] that is, stored in structured or binary formats not meant for direct human reading

How Computer Forensics Differs from Electronic Discovery

Electronic discovery concerns the ESI accessible to litigants; computer forensics concerns the ESI accessible to experts. The boundary is porous. E-discovery may require attention to data deemed "not reasonably accessible" because of burden or cost, while forensic analysis often turns on information that is plainly accessible, such as file timestamps.

The principal distinctions are **expertise**, **purpose**, and **proportionality**. Computer forensics demands specialized technical skills and is unnecessary in many cases—but indispensable in others. It frequently raises contentious issues of expense, delay, and intrusion. E-discovery treats evidence as discrete information items—documents, messages, databases—while computer forensics takes a systemic view, examining how information relates, persists, and reveals what a user did or attempted to do. Finally, e-discovery focuses largely on what exists; computer forensics often focuses on what is missing, how it disappeared, and whether it can be reconstructed.

When to Turn to Computer Forensics

Most cases do not require forensic-level examination of computers or devices. Accordingly, parties and courts should scrutinize requests for access to an opponent's systems to ensure they are grounded in genuine need rather than speculation or a fishing expedition. When the need is debatable, courts can balance burden and proportionality by appointing a neutral examiner, imposing a protective protocol, and initially allocating costs to the requesting party pending evidentiary justification for reallocation.

Some disputes plainly warrant forensic analysis, and in those cases the parties and the court should act promptly to preserve relevant evidence. Claims of data theft are a common example. When a key employee departs to join or form a competitor, forensic examination of business systems, portable storage, cloud repositories and home devices may be necessary to determine whether data was taken, how it was removed, and whether it has been used, shared, or disseminated.

Forensic analysis is also justified where there are credible allegations of destruction, alteration, or fabrication of ESI, as well as claims of system intrusion or misuse. Employment disputes involving harassment or discrimination frequently turn on electronic communications, and electronic devices play a central role in criminal and domestic relations matters. Fraud, infidelity, and concealment rarely occur without leaving a trail of digital breadcrumbs in email, messaging platforms, financial records, mobile devices, and cloud accounts.

What Computer Forensics Can Do

The scope and reliability of forensic findings vary, but a competent examination may reveal:

1. The manner and extent of theft or misappropriation of proprietary data
2. The timing and scope of file deletion or antiforensic activity (*e.g.*, data wiping)
3. Whether, when, and how external storage devices were connected

4. Forgery or alteration of electronic documents
5. Recoverable deleted data, file structures, and associated metadata
6. Internet usage, online activity, cloud access, and e-commerce transactions
7. Breaches, intrusions, and unauthorized system access
8. Social networking activity
9. Clock or calendar manipulation
10. Image manipulation
11. Detailed system usage over time

What Computer Forensics Cannot Do

Despite popular mythology, computer forensics has real limits. An examiner generally cannot:

1. Recover data that has been completely overwritten, even once
2. Reliably identify the individual at the keyboard where credentials are shared or misused
3. Conduct a proper examination without access to the evidence or a forensically sound image
4. Recover data from media suffering catastrophic physical damage
5. Guarantee that a device will not fail during acquisition
6. Instantly hack into a computer at a scene and uncover decisive evidence on demand

That last scenario may make for good television, but it's bad forensics.

NTFS, the Master File Table, and Why It Matters

As a refresher, the **NTFS file system** underlies Windows NT through current versions of Windows (Windows 7–11). NTFS relies on a powerful database called the **Master File Table (MFT)** to track every file and directory on a volume. The MFT records a file's name, size, permissions, timestamps, and storage location. All NTFS file system metadata resides in the MFT, making it one of the most important artifacts examined in Windows forensics.

A distinctive feature of NTFS is its treatment of very small files. When a file is sufficiently small (generally under about 1,500 bytes), NTFS may store the file's contents directly within the MFT record itself rather than allocating separate disk clusters. These are known as **resident files**. Storing both metadata and content together improves performance by eliminating additional disk seeks.



Small but critical artifacts—password lists, account numbers, short emails, notes, or "smoking gun" text—often become resident files. Consequently, forensic examination of the MFT can reveal substantive evidence, not just file metadata. If NTFS were an old library card catalog, it would not merely supply a book's location—it would store books small enough to fit inside the card drawer.

Formatting and Partitioning

Confusion about **formatting** and **partitioning** persists, much of it traceable to obsolete practices. In the early days of personal computing, users sometimes performed so-called "low-level formatting," physically laying out tracks and sectors on a new hard drive. That era is long past. Modern storage devices are low-level formatted at the factory, and users have no ability to alter that physical structure.

What users *do* control is **partitioning** and **high-level formatting**. Partitioning divides a storage device into one or more logical volumes, typically presented to users as drive letters (e.g., C:, D:). High-level formatting then establishes the file system structures within a partition, enabling the operating system to store and retrieve files.

Most computers ship with a single partition that is already high-level formatted and ready for use, though a drive may be subdivided into multiple volumes for organizational, administrative, or forensic reasons.

Clusters, Slack Space, and Their Diminishing Forensic Role

Digital storage media are physically subdivided into fixed-size units called **sectors**. Historically, sectors on electromagnetic hard drives were 512 bytes; today, they are often larger (commonly 4,096 bytes). While sectors are the smallest *physical* unit of storage, they are not the smallest unit the operating system can address. That distinction belongs to the **cluster**—a logical grouping of sectors and the smallest unit a computer can read from or write to.

You might be wondering, "what about bits, bytes and sectors; aren't they smaller?" Sure, but those smaller units aren't *user-addressable*. Computers must "pick up" and "set down" information in cluster-sized chunks.

On Windows systems, clusters are typically 4 KB in size for drives up to 16 terabytes. Files are stored by allocating enough *whole clusters* to contain their contents. Because clusters cannot be subdivided, a file almost never fills its final cluster exactly. The unused space between the logical end of the file and the physical end of the last allocated cluster is called **file slack** (or simply, *slack space*).



A simple illustration makes this concrete. Create a text file containing the word "hello." The file is five bytes long, but it occupies an entire cluster on disk—often 4 KB or more.[69] Now, open the file and change "hello" to "hello there," then save the file. Now, when you look at the file's properties, it has more than doubled in size to 11 bytes (the space between the words requires a byte, too), but the storage space occupied on disk is unchanged because you haven't gone beyond the size of a single cluster Expanding the file does not change its disk footprint until the cluster's capacity is exceeded. The unused portion of that cluster is slack space, and it may contain data unrelated to the current file.

From the perspective of the file system, slack space is wasted capacity. From a *forensic* standpoint, it may contain remnants of data previously stored in that cluster. Importantly, **slack space is not the same as unallocated space** (sometimes called free space). Unallocated clusters are not assigned to any file and may contain the complete contents of deleted files. Slack space, by contrast, can never contain more than one cluster minus one byte of residual data, and it almost never preserves intact file structures.

This distinction matters. Data recovered from unallocated space may span thousands of clusters and often retains a **binary header signature** identifying the file type and decoding method. Slack space rarely does. The identifying header bytes are typically overwritten first, leaving behind fragments whose origin and encoding are uncertain.

Why Slack Space Once Mattered

In the early days of computer forensics—particularly during the MS-DOS and early Windows eras—slack space loomed large. Many applications stored content in **plain text (ASCII)**, making intelligible snippets recoverable even when fragmented. Slack space persisted for long periods on magnetic hard drives because it was overwritten only when the file occupying the cluster was deleted.

Training exercises and early case law reinforced the perception that slack space was fertile ground for "smoking gun" evidence. In theory, that potential remains. In practice, it has sharply diminished.

Why Slack Space Rarely Delivers Today

Three developments have dramatically reduced the evidentiary yield of slack space searches.

First, modern encoding and compression. Since 2007, Microsoft Office and many other applications store content as **compressed, structured data** (for example, ZIP-compressed XML). Viewed at the media level, such content resembles encrypted noise. In slack space, the compression headers and

[69] If you see that the size on disk is zero, then Windows is correctly reporting that the small file is being stored within the Master File Table. You may not see Windows move the data out of the MFT until you reach about 550 characters in the file.

directory structures needed to decode the data are almost always destroyed. Even when relevant content once resided in slack, it is frequently invisible to search tools.

Second, overlapping artifacts. Slack space often contains fragments of multiple deleted files, overwritten piecemeal over time. What remains is a palimpsest: partial remnants without context, order, or reliable attribution. Searching such material is speculative and prone to false positives.

Third, modern storage hardware. Full-disk encryption is now routine. More significantly, **solid state drives (SSDs)** have upended the traditional forensic model. SSDs employ **wear leveling**, continuously relocating data to distribute write cycles evenly across memory cells. Unallocated space is constantly reused, and legacy data is rapidly erased. Operating systems also issue **TRIM commands**, proactively clearing storage pages no longer in use to improve performance.

The combined effect is decisive: on SSD media, neither unallocated space nor slack space reliably retains deleted data. Forensic imaging remains essential for preservation and analysis, but expectations about what can be recovered must be recalibrated.

Contemporary Forensic Implications

Operating systems do not expose slack space. Examination requires a **forensically sound image** and specialized tools. Even then, the results are often thin. Modern slack searches tend to succeed only when:

- The media is a traditional electromagnetic hard drive;
- Encryption is absent or defeated;
- The search is narrowly targeted using unique terms; and
- The effort can be conducted programmatically at low marginal cost.

As a result, slack space today belongs in the examiner's toolkit as a **secondary or incidental artifact**, not as a primary justification for forensic access to a system. Requests demanding production of "slack space contents" misunderstand both the technology and proportionality principles governing discovery.

That does not mean slack space is irrelevant. It can still surface corroborative fragments, contextual clues, or evidence of prior activity. But it is no longer the digital amber it once was.

Proportionality and Preservation

Because active systems constantly overwrite data, potential forensic artifacts—however attenuated—can be lost quickly. The prudent response to credible forensic issues remains the same: secure **forensically sound images** early to preserve the status quo. Imaging protects active data, residual artifacts, and system metadata while allowing courts and parties to resolve discoverability, privilege, and scope without risking spoliation.

But courts and counsel should be realistic. Searching slack space is rarely outcome-determinative in modern cases. It should be undertaken when the cost is modest and the protocol otherwise justified—not as a talismanic rite or a fishing expedition.

Balancing Need, Privilege, and Privacy

A computer forensic examiner sees everything. The erosion of boundaries between personal and professional life means workplace systems routinely contain personal, confidential, and privileged material—sometimes intimate or highly sensitive—while personal devices often hold business data. A hard drive is not a file drawer; it is closer to a private office. It may chronicle the full arc of a user's daily activity, including trade secrets, customer data, compensation information, personal communications, Internet searches, financial records, medical information, and passwords.

The central challenge is how the justice system affords access to discoverable information without exposing everything else. Forensic examination carries an inherent risk of overreach, so courts and parties must proceed with restraint. Access to an opponent's digital media should turn on demonstrated need and relevance, balanced against burden, cost, and potential harm. Compelling forensic access is properly treated as an extraordinary remedy—more a response to discovery failure or misconduct than a routine discovery tool.

Direct access to an adversary's storage media is disfavored and should be ordered only when circumstances justify it—for example, where a party has shown itself untrustworthy, incapable of preserving or producing responsive ESI, or where the requesting party has a proprietary interest in the data or device. Allegations of spoliation, data theft, harassment, or system misuse often supply the necessary predicate, particularly when the computer itself is central to the claims or defenses.

In Texas, these principles are articulated in *In re Weekley Homes, L.P.*, 295 S.W.3d 309 (Tex. 2009), which remains a useful framework for analyzing requests for forensic access. The court emphasized that:

1. Parties seeking deleted ESI must specifically request it and identify the desired form of production.
2. Responding parties must produce reasonably available ESI in the requested form or timely object.
3. Discovery disputes should be resolved by agreement where possible; failing that, the responding party bears the burden to show undue burden or cost.
4. Even if information is not reasonably accessible, production may be ordered if recovery is feasible and the likely benefit outweighs the burden.
5. Direct access to an opponent's devices is discouraged; if ordered, it must be conducted by a qualified expert pursuant to a narrowly tailored protocol protecting privacy and privilege.
6. The requesting party ordinarily bears the cost of extraordinary forensic measures.

Weekley Homes also recognized a critical, forward-looking obligation: early disclosure of relevant information about electronic systems and storage practices to enable informed negotiation of protocols and to equip courts to craft appropriate orders. That expectation is routine in federal practice and increasingly influential in state courts.

What *Weekley Homes* did not supply—and what modern practice must—are practical guideposts for feasibility and protection. A workable forensic protocol balances what counsel seek, what forensic tools can reliably deliver, and what must remain shielded from disclosure. Vague or overbroad protocols invite cost overruns, privilege disputes, and judicial frustration.

Today, courts increasingly rely on neutral forensic examiners—often appointed as FRCP Rule 53 special masters—to conduct examinations *in camera*.[70] Common safeguards include staged review, keyword and time-bound targeting, segregation of clearly privileged material, and initial disclosure to producing counsel for privilege review and logging before any production. Whether conducted by a party expert or a neutral, the order authorizing forensic examination should define objectives with precision and specify how confidential and privileged information will be identified, isolated, and protected.

A recurring mistake is committing to a forensic examination protocol without consulting a qualified examiner about feasibility, cost, and likely yield. Thoughtful sequencing and expert input at the outset almost always reduce expense and intrusion. In most cases, a neutral examiner operating under a clear, proportional protocol offers the most efficient and defensible path—protecting privacy, preserving privilege, and delivering only what the case genuinely requires.

Who Performs Computer Forensics?

Historically, many competent forensic examiners emerged from law enforcement. Today, the pipeline is far broader: private forensic laboratories, incident-response firms, corporate investigation teams, and university programs all produce practitioners. Yet the field remains unevenly regulated. A small number of certifications are widely respected, but many credentials measure classroom exposure more than demonstrated competence. In some states, “computer forensics” is folded into private-investigator licensing, which may regulate business practices without meaningfully screening technical capability.

A “computer expert” is not necessarily a forensic examiner. General IT experience, software development skills, or cybersecurity work may be useful, but they do not substitute for training in evidence preservation, repeatable methods, artifact interpretation, documentation, and courtroom testimony. Digital forensics is as much about disciplined process and defensibility as it is about technical know-how.

[70] **In camera** means conducted privately for the court’s review, outside the presence of the public and opposing parties, typically to protect confidentiality, privilege, or sensitive information.

Drafting Digital Examination Protocols

A computer or phone under forensic examination is less a "file drawer" than a densely populated city. A single device can yield hundreds of thousands (or millions) of discrete artifacts—some obvious to users, most not. If you do not define what you are looking for, where you expect to find it, and what boundaries must be respected, the examiner can burn weeks "searching everywhere," and the parties can spend months fighting about overbreadth, privilege, and cost.

In e-discovery and digital forensics, an examination protocol is a court order or party agreement that defines the scope, method, and safeguards for acquisition, analysis, and reporting of electronically stored information. Protocols exist to (1) target the work to the issues, (2) protect privilege and privacy, (3) reduce disruption and delay, and (4) create a defensible record of what was done and why.

Why protocols are not "standard"

There is no universal digital examination protocol. What is "standard" changes as devices, storage, and security evolve—full-disk encryption, cloud sync, multi-factor authentication, secure enclaves, solid-state media behavior, and explosive data volumes all affect what is feasible and what is proportional. Without a protocol, an examiner can chase an ambiguous mandate indefinitely. With a bad protocol, the examiner is forced into pointless tasks, misses the real issues, or creates avoidable privilege fights.

A helpful analogy: if you imagine "deposing" a device, a protocol is the outline that prevents fishing expeditions while still allowing competent follow-up. The best protocols balance structure with limited discretion—enough to follow the evidence, not enough to roam.

Common elements of a defensible protocol

Most workable protocols do the following:

1. **Identify the examiner and role**
   Make clear whether the examiner is party-retained or neutral, and what duties attach (e.g., "officer of the court," confidentiality obligations, *ex parte*[71] limits). If the examiner is neutral, specify how the neutral communicates (typically jointly to counsel, with disputes routed to the court).

   Exemplar language:
   The parties shall confer in good faith to select a qualified digital forensic examiner ("Examiner"). If no agreement is reached by [DATE], each side shall submit up to two candidates with CV and

[71] **Ex parte** means a communication or proceeding presented to the court or a neutral by one party without notice to or participation by the opposing party.

fee schedule. The Court will appoint an Examiner, who will serve as a neutral and be bound by this Protocol and the Protective Order.

2. **Identify devices and media with precision**
   List serial numbers, asset tags, phone identifiers, user accounts, and relevant cloud repositories. Define what happens if additional devices are identified during the work (this is where many protocols fail and accusations of concealment begin).

Add a practical safeguard:
If examination reveals indications of additional potentially relevant devices or accounts within the subject's custody or control, the Examiner shall notify counsel, describe the basis for the indication, and no additional device/account will be examined absent agreement or further order.

3. **Define scope: topic, objectives, and methods**
   Start with goals, not tools. "Search for X" is weaker than "determine whether X occurred, when, how, and what data left the system."

   A useful structure:
   (a) Issues to be tested (data theft, deletion, fabrication, intrusion, harassment communications, clock tampering, etc.)
   (b) Data sources likely to answer them (user folders, email stores, cloud sync logs, USB artifacts, browser artifacts, messaging databases, etc.)
   (c) Allowed methods (artifact review, targeted searches, hash matching, timeline reconstruction, recovery attempts, etc.)
   (d) Exclusions and guardrails (privilege, personal content, irrelevant categories)

4. **Evidence integrity checks (do this explicitly)**
   This is routinely omitted and should not be. Require baseline integrity assessment: confirm device identity, signs of reimaging, wiping, unusual account creation, suspicious "privacy" tools, time anomalies, and encryption status.

   Exemplar language:
   At the outset, Examiner shall assess whether the evidence appears consistent with its stated origin and use, including indicators of reimaging, wiping, time manipulation, or antiforensic activity. Examiner shall promptly report material irregularities to counsel. The Examiner may corroborate conclusions by reference to multiple independent artifact sources (e.g., MFT/USN Journal, Registry artifacts, link/jumplist artifacts, logs, and snapshots) and shall note material inconsistencies or limitations affecting reliability.

5. **Temporal scope, expressed realistically**
   Date limits are attractive and often illusory. Many artifacts cannot be reliably pre-filtered by date before they are found, and file timestamps may not track authoring or user activity. A better approach is "search broadly, produce narrowly," with a relevance interval used primarily as a production filter, not a hard gate on examination.

   Preferred approach:
   The Relevant Interval guides production and reporting. Examine may review artifacts outside the interval as necessary to understand or validate activity, but should not disclose out-of-interval content unless it bears on integrity, spoliation, or the issues specified.

6. **Privilege and privacy protections**
   This is where protocols often become "artless." Say exactly how privileged material is screened and who sees what first.

   Common workable options:
   Party-retained examiner *under strict non-disclosure and privilege filtering by producing counsel (risk: trust deficit).*
   Neutral examiner *conducts in camera review and provides only responsive, non-privileged output (often best when hostility is high).*
   Hybrid: *neutral does first-pass triage; producing counsel logs privilege; disputes to the court.*
   If the protocol uses keyword filtering, define the privilege keyword process, who supplies terms, how disputes are handled, and how overinclusive hits are controlled.

7. **Cooperation and access credentials**
   Hardened security makes "we'll just bypass it" unreliable. Protocols should require prompt provision of passwords, decryption keys, MFA cooperation, and device unlock sessions, while limiting use of credentials to the examination.

   Exemplar language:
   The producing party shall reasonably cooperate to enable access to encrypted or locked data, including providing passwords, recovery keys, and necessary authentication steps. Credentials shall be used solely for purposes of this examination and not to access other systems or accounts.

8. **Define the work product and reporting**
   Specify what the examiner will deliver: narrative report, timelines, file and hash inventories, exception logs, lists of relevant artifacts, and productions (if any). Set what is "discoverable output" versus "work notes," and how tool-generated reports are handled.

9. **Cost allocation**
   State who pays, how invoices are handled, and whether cost-shifting may occur depending on findings (*e.g.,* if spoliation is shown).

A caution on "template orders"
Borrowing language from another case is risky because forensic tasks are not interchangeable. A protocol focused on hash and filename searches can be entirely appropriate for locating a known set of documents, but it is not a meaningful "forensic examination" when the issues involve deletion, exfiltration, wiping, or fabrication—matters that live in artifacts, logs, and system state.

What lawyers get wrong most often

1. Demanding the impossible (e.g., "identify every deleted file," "prove who was at the keyboard," "recover overwritten data").
2. Treating forensic work like document review (keyword lists as the primary method, with little attention to artifacts and context).
3. Overpromising temporal limits ("search only between these dates") that cannot be reliably enforced at acquisition and artifact levels.
4. Hand-waving privilege protections instead of specifying who sees what, when, and how disputes are resolved.
5. Failing to require integrity checks up front, then litigating about "sterilized" evidence after money is spent.

A practical closing thought on Examination Protocols
A good protocol does at least three things: it tells the examiner what questions to answer, it protects what should not be exposed, and it defines "done." If it does not do all three, it is not a protocol—it is an argument waiting to happen.

How Windows "Deletes" a File (and why it may still be recoverable)
Windows can be downright obstinate in its retention of data you don't want to hang around. Neither emptying the Recycle Bin nor performing a quick format of a disk will obliterate all its secrets. How is that deleting a file doesn't, well, *delete* it? The answer lies in how Windows stores and catalogs files.

Remember that the Windows files system deposits files at various locations on a disc drive and then keeps track of where it has tucked those files away in its Master File Table. So, in Windows, "Delete" is more bookkeeping than obliteration. NTFS records a file's existence and location in the Master File Table (MFT). When a user deletes a file, Windows typically removes the directory entry and marks the file's clusters as available for reuse. The file's content may remain on the drive until overwritten by new data.

Returning to our library card catalog analogy, pulling a card from the card catalog doesn't remove the book from the shelf, although consulting the card catalog, it seems to be gone. Deleting a computer file only removes its "card" (the file table record). The book (the file's content) hangs around until the librarian needs the shelf space for new titles (overwriting the clusters).

That persistence is most reliable on traditional magnetic hard drives (HDDs). On solid state drives (SSDs), the combination of full-disk encryption, wear leveling, and TRIM can cause deleted content in unallocated space to be discarded or rendered unrecoverable quickly. In other words: "recoverable deleted files" is now a fact-dependent question, not a default assumption.



"True, I can't take it with me, but I can take the access codes to it."

© Cartoonbank.com

Practical takeaway for litigation: "I deleted it" often means "Windows stopped pointing to it." Whether it can be recovered depends on media type (HDD vs. SSD), encryption, TRIM, time, continued use, and whether backups/snapshots exist.

System-Created Artifacts: "Happy Accidents" that become evidence

A forensic exam is rarely about the user's documents alone. Windows and applications generate artifacts to improve performance, convenience, and reliability. Their evidentiary value is collateral—happy accidents—often powerful, but always contextual.

Common Windows artifacts that help reconstruct activity (with typical uses):

- **LNK ("shortcut") files** record evidence of file access—paths, volumes, and timestamps—even after the target file is moved or deleted. LNK files are system-generated pointers (not user-created) stored in the user's *Recent* locations to speed access to recently used items. Each LNK preserves metadata about its target—such as timestamps, size, location, and the originating storage device—that can persist after deletion. Windows keeps these artifacts for performance, not for evidence; their forensic value is a byproduct of that design.
- **Jump Lists** are Windows-maintained records of recently and frequently accessed files, organized by application and stored separately from LNK files. They capture file paths, access order, and timestamps, and often persist even after the underlying files are moved or deleted. Because Jump Lists are application-centric (e.g., Word, Excel, Acrobat), they can

corroborate which program opened a file and when, making them especially valuable for reconstructing user activity and cross-validating LNK evidence.

- **Prefetch:** Windows maintains Prefetch files to speed application startup by recording which executables are run and loading needed code in advance. These records typically reflect the *most recent executions* (historically up to ~128 entries), with timestamps and limited run-count data. Forensically, Prefetch is useful to show that a particular application *did execute* and *roughly when*—for example, corroborating use of wiping utilities or data-exfiltration tools.
- **Registry artifacts (multiple hives):** The Windows Registry is the operating system's central configuration database, assembled at boot from multiple files called *hives*. Most hives persist on disk; one—the *Hardware* hive—is rebuilt each boot as Windows inventories connected hardware. Though never designed as an evidentiary log, the Registry yields a wide range of forensic artifacts by happy accident, including device history, program installation and execution traces, MRU (most-recently-used) lists, file associations, user profiles, serial numbers, and configuration data that can corroborate user activity—even where users have attempted to delete visible traces.
- **USB connection artifacts:** When a USB mass-storage device is connected, Windows uses Plug and Play to identify the device, load drivers, and record the pairing. In doing so, Windows writes artifacts to Registry locations such as *ENUM/USBSTOR* and *DeviceClasses*, often capturing device identifiers and evidence of first and last attachment. These records can support conclusions about *what device was attached* and, sometimes, *when*. Because Registry timestamps and fields can be incomplete or misleading, interpretations should be validated against corroborating sources (e.g., event logs, link files, shellbags, or volume metadata). Windows retains this information to speed future device recognition—not to track user misconduct—but, properly interpreted, it carries substantial probative value in investigations of data movement and theft.
- **Shellbags**: Windows Explorer folder view settings stored in the Registry that can indicate a folder path was enumerated or opened in Explorer, even if the folder was later moved or deleted. Useful where the *existence, location, or navigation of a folder* is the issue (e.g., contraband, staging, or exfiltration directories). Shellbags persist beyond deletion and can reflect local and removable media paths, but they do **not** establish that specific files within the folder were opened or viewed. Temporal fields, where present, require corroboration and should be treated as contextual rather than definitive proof of access.
- **Pagefile/hiberfil and related memory-backed artifacts**: can contain fragments of content that lived in RAM—documents, web content, messages, and sometimes credentials—difficult to parse but occasionally decisive.
- **NTFS change and transaction artifacts (e.g., $LogFile, USN Journal):** can corroborate file creation/deletion/renames and timing patterns, subject to retention limits.

Volume Shadow Copies (VSS)

Windows includes a snapshot feature called the **Volume Shadow Copy Service (VSS)** that can preserve earlier versions of files by recording changes to disk blocks before they are modified.

Where enabled, VSS may retain prior versions of user files—not just system data—and can reveal when files were created, changed, or deleted.
VSS is not a backup system and should not be assumed to exist or persist. On modern Windows 10 and Windows 11 systems, shadow copying is often disabled by default, especially on laptops and solid-state drives. Even when enabled, retention is limited by available disk space and system activity. Shadow copies may be pruned automatically during updates, under disk pressure, or by security features designed to reduce ransomware risk. There is no fixed number of snapshots or guaranteed retention period.

When present and accessible, shadow copies may allow recovery of prior file versions through ordinary user interfaces (e.g., “Restore previous versions”), sometimes without specialized forensic tools. In such cases, courts may regard the data as **reasonably accessible ESI**. In many environments, however, users lack the permissions or configuration necessary to view or restore shadow data without administrative access.

Why this matters in protocols
These are examples, and we must recognize that artifacts are different for different operating systems (Windows versus MacOS) and even for different releases of the same operating system. Artifacts are radically different on phones versus computers. It’s complicated, and it changes, literally, every day.

A protocol should define what questions must be answered and what artifact classes are in-scope, then permit reasonable corroboration across multiple sources rather than demanding a single, brittle “one-artifact proves all” approach.

Mobile Hypothetical: The Leaver
A senior sales executive resigns and joins a competitor. Before leaving, she returns her company laptop, which is imaged and reviewed. Nothing remarkable appears—no mass copying, no suspicious deletions, no obvious misconduct. The employer moves to compel a forensic examination of her personal devices.

At first glance, the request looks excessive. But the factual context matters.

The executive conducted nearly all customer communications on her phone. Pricing discussions moved from corporate email to text. Draft proposals were photographed, not saved. Shared folders were accessed through mobile apps. Two weeks before resignation, she disabled corporate email on the phone.

A neutral examiner acquires the device promptly, using a logical and file-system–level extraction. No deleted files are recovered. There is no “smoking gun” document.

What emerges instead is a pattern:

- Message metadata shows sustained late-night exchanges with a contact later identified as a competitor.
- Media artifacts reveal photos of customer whiteboards taken inside secure facilities.
- Application logs show repeated access to cloud storage apps not used elsewhere.
- Cloud sync indicators show outbound transfers shortly before departure.

No single artifact proves theft. Together, they tell a coherent story of timing, opportunity, and conduct—one that laptop forensics alone could never supply.

Mobile Device Forensics and Analysis

Everything discussed thus far has focused on traditional computers—desktops and laptops running general-purpose operating systems, storing data on media that, while encrypted with increasing frequency, still exhibit forensic persistence through redundancy and accident. Mobile devices are different. Profoundly so. What follows is not Windows forensics at smaller scale, but a departure from it—one driven by encryption-by-default, aggressive security models, constant network dependence, and hardware designed to frustrate post-hoc reconstruction.

Mobile devices are no longer peripheral to digital investigations; they are often the center of gravity. Phones and tablets have become the primary computing platform for many users, eclipsing laptops for communications, photography, navigation, authentication, and access to cloud services. As a result, mobile device forensics is now a core competency for litigators and courts—not a specialty niche reserved for criminal cases or exotic disputes.

Yet mobile forensics differs in consequential ways from traditional computer forensics, and those differences matter profoundly to preservation decisions, scope, cost, expectations, and protocol design. Lawyers who approach phones as “small computers” will misjudge both what can be recovered and what it will take to do so.

Why Mobile Forensics Is Different

Unlike general-purpose computers, modern mobile devices are engineered as tightly integrated systems combining specialized hardware, secure firmware, operating system controls, application sandboxes, and constant network connectivity. Full-disk or file-based encryption is universal. Encryption keys are bound to hardware enclaves. Application data is compartmentalized. Operating systems update aggressively and often without meaningful user intervention. Cloud synchronization is the default, not the exception.

From a forensic standpoint, this architecture produces three practical consequences.

First, **deleted-data recovery from physical storage is sharply constrained**. The slack-space and unallocated-cluster artifacts familiar from desktop forensics are largely absent or inaccessible on modern phones. Deletion is more often real and irreversible.

Second, **probative evidence tends to reside in application databases, system logs, and synchronization metadata**, not in conventional file systems. The "documents" model of discovery fits mobile data poorly.

Third, **access—technical, legal, and cooperative—is the gating issue** in nearly every mobile examination. Without access, there is no analysis. Without realistic access expectations, there is no defensible protocol.

Preservation Comes First—and Early

Everything said earlier in this chapter about preservation applies with greater urgency to mobile devices.

Continued use of a phone alters evidence rapidly. Messages are overwritten, databases compacted, logs rolled, caches purged, and synchronization events propagated to cloud services. Automatic OS updates, background application activity, and security housekeeping can permanently destroy artifacts in hours or days, not weeks or months.

Accordingly, the first forensic question is not *what should we examine?* but *how do we preserve what exists right now?*

The gold standard remains **prompt, forensically sound acquisition by a qualified examiner**, authenticated and documented. In practice, that may mean different things depending on platform and circumstances:

- **iOS devices** typically permit only logical or file-system–level acquisitions, often contingent on the device being unlocked and compatible with available tooling.
- **Android devices** vary widely by manufacturer, OS version, and security patch level; acquisition may range from limited logical extraction to broader file-system access in narrow circumstances.

Where immediate acquisition is not feasible, a protocol should at least require **device isolation** (airplane mode, Faraday enclosure), suspension of updates and backups, and preservation of associated credentials and accounts pending further order. Delay is the enemy of mobile forensics.

Acquisition Methods—and Their Limits

Mobile forensic acquisition is best understood as a **spectrum,** not a binary choice.

- A **logical acquisition** captures user-accessible data through operating-system interfaces: contacts, call logs, messages, media, and some application data. It is fast, minimally intrusive, and often sufficient for discovery-oriented questions.
- A **file-system acquisition** accesses a broader range of application databases, logs, and metadata. It yields richer artifacts but usually requires device unlocking and user cooperation.
- A **physical acquisition**, in the classic sense of a bit-for-bit image of raw storage, is now rare on modern devices due to encryption and secure hardware. Where vendors claim "physical" access, courts should demand clarity about what that term actually means.

A defensible protocol does not promise a particular acquisition type in advance. Instead, it authorizes the examiner to employ **the most complete method feasible** without defeating security controls or resorting to invasive exploits absent explicit court approval.

What Mobile Forensics Can Reliably Show

Despite these constraints, modern mobile forensics—properly scoped—is extraordinarily powerful. A competent examination may reveal:

- The timing, content, and metadata of SMS, MMS, and application-based messages (e.g., iMessage, WhatsApp, Signal), subject to encryption and retention limits.
- Call history, voicemail metadata, and contact interactions.
- Photographs and videos with embedded metadata, thumbnails, and cached derivatives.
- Application usage patterns, including installation history and foreground/background activity.
- Location history derived from GPS, Wi-Fi, Bluetooth, and application records—often probabilistic rather than exact.
- Evidence of cloud synchronization, account usage, and external storage interaction.
- Indicators of deletion, wiping, factory reset, or tampering.

As with computer forensics, conclusions are strongest when **corroborated across independent artifact classes**. No single mobile artifact should be treated as dispositive in isolation.

What Mobile Forensics Cannot Do

Equally important are the limits. A mobile examiner generally cannot:

- Reliably recover overwritten data on modern encrypted devices.
- Bypass strong passcodes or biometric protections at will.
- Attribute device activity to a specific individual with certainty where devices are shared or unattended.
- Guarantee completeness of cloud-synchronized content without direct access to the cloud source.

Protocols that ignore these limits invite cost overruns, false expectations, and judicial skepticism.

The Central Role of Cloud Data

Mobile devices are inseparable from cloud ecosystems. Messages, photos, notes, and application data often exist simultaneously on the device, in backups, and on provider servers. The phone frequently serves as the **index** to these remote stores.

A forensic examination of a device may therefore illuminate what cloud data exists, when it was synchronized, and under which account—but not necessarily capture the cloud-resident content itself. Accessing cloud accounts typically requires separate legal process or express consent and must be addressed explicitly in any protocol. Using device-derived credentials to access live accounts without authorization is unethical and improper.

Tools, Workflows, and Examiner Judgment

Mobile forensic tools are mature and widely deployed. Platforms such as **Cellebrite UFED**, **MSAB XRY**, and **Oxygen Forensic Detective** automate much of acquisition and artifact parsing across iOS and Android ecosystems. They are powerful, indispensable—and limited.

Tool output is not evidence. Interpretation is.

A defensible mobile-forensics workflow typically includes:

- Device identification and condition assessment.
- Documentation of OS version, security state, and lock status.
- Selection of the least intrusive acquisition method capable of answering the questions posed.
- Hash authentication of acquired data where applicable.
- Parsing and normalization of application databases and system artifacts.
- Timeline reconstruction and cross-artifact corroboration.
- Reporting that explains not just findings, but limits and confidence.

Courts should be wary of protocols that reduce mobile forensics to “run the tool and produce the report.” That is data dumping, not analysis.

Privacy, Proportionality, and Protocol Design

Mobile devices are intensely personal. They blend business communications with family life, medical information, finances, and intimate relationships. A forensic image of a phone is closer to a diary than a filing cabinet.

Accordingly, mobile forensic examination should be treated as an extraordinary remedy. Preservation is often appropriate; examination should be targeted, staged, and supervised. Neutral examiners, producing-party privilege screening, and in camera review are not luxuries—they are safeguards.

A sound mobile-forensics protocol should:

- Define the questions to be answered, not just the data to be collected.

- Limit scope by time, application, and artifact class where feasible.
- Provide explicit non-waiver protections for privilege.
- Specify reporting format and review sequence.
- Acknowledge technical limits and allow examiner discretion within guardrails.

Mobile forensics is not "computer forensics, but smaller." It is a distinct discipline shaped by encryption, mobility, and cloud dependence. When done well, it can reconstruct a remarkably detailed picture of conduct and intent. When ordered casually or framed unrealistically, it becomes invasive, expensive, and unreliable.

As with all digital forensics, the goal is not to extract everything. It is to extract what matters, in a way the court can trust.

More Protocol Points

Establish Who Pays
Forensic preservation of a desktop or laptop—meaning a forensically sound image—often costs no more than a short deposition. Forensic examination, by contrast, can range widely in cost depending on the nature of the media, the questions posed, and the technical obstacles encountered. Phones, servers, cloud accounts, strong authentication, full-disk encryption, and sheer data volume all drive cost. But nothing influences expense more than scope.

Forensic examiners usually charge by the hour, with rates commonly ranging from roughly $250 to $750 depending on training, experience, reputation, and locale. Poorly targeted or open-ended examinations can quickly run into five- or six-figure territory. Focused examinations—communicated through clearly expressed protocols—tend to cost less, finish sooner, and yield more reliable results. Searches should be carefully evaluated for over- or under-inclusiveness, progress should be monitored, and protocols should be adjusted when reality diverges from expectation. It is prudent to require periodic status reporting when either hourly or cost benchmarks are reached. In all events, the protocol should make clear how, when, and by whom the Examiner is compensated for professional time and reimbursed for expenses.

Exemplar language:
Charges for Examiner's professional time and time in transit shall be timely paid by Plaintiffs at the Examiner's customary rates, together with reasonable and customary expenses pursuant to the rate sheet submitted prior to appointment. In the event Examiner's charges equal or exceed $_____, Examiner shall report progress to the parties and estimate the additional charges reasonably expected to complete the examination.

Address Onsite Acquisition and Supervision

A party may insist on being present during acquisition. That can be reasonable for the collection itself. It is rarely reasonable to "supervise" analysis, which is dominated by machine time (processing, indexing, carving, decryption attempts, password recovery workflows, parsing logs) punctuated by examiner judgment calls. Requiring continuous onsite attendance during machine time converts idle waiting into billable hours and inflates cost without improving reliability.

When possible, acquisition should be done efficiently (often onsite for critical systems; otherwise by shipment to a lab or secure remote collection), with clear chain-of-custody documentation and hashes to verify integrity. Analysis is typically best performed in the examiner's lab environment where tooling, secure storage, and controlled workflows exist, and where the examiner can stop billing during machine time.

If onsite work is necessary, the protocol should define it narrowly: where it will occur, who may attend, whether network access will be required, what credentials will be supplied, and how to handle interruptions caused by business needs.

Exemplar language:
Attendance by a party representative may be required for acquisition. Continuous observation during processing/analysis is not required and shall not be demanded absent good cause. Examiner may conduct analysis offsite in a secure environment and will preserve audit logs sufficient to explain methods and results.

Recovering Deleted Data
Deleted-data recovery is a common reason parties seek forensics. But protocols that demand "recover all deleted files and fragments" are a recipe for cost, delay, and disappointment. Courts occasionally use language like that; it reads as decisive, but it is technologically naïve—especially on modern systems.
Why "Undelete Everything" Fails
Deletion usually removes pointers, not the underlying data. On magnetic drives, remnants can persist until overwritten; on solid-state drives, TRIM and wear-leveling can erase or relocate data quickly, and full-disk encryption can make remnant recovery effectively impossible without keys. Even when remnants exist, they are frequently fragmented, partially overwritten, and stripped of the file header information needed to identify type and structure. The result is a lot of corrupt output and a lot of human review to determine what is corrupt.

What examiners can do—when it's feasible—tends to fall into three buckets:

1. **Signature-based carving.** Scan unallocated space for file headers ("magic numbers") and attempt reconstruction. It is fast on some media, slow on others, and often produces partial or corrupted files—especially when headers are overwritten or files were fragmented.

2. **Directory-structure recovery.** Use residual file-system metadata (when present) to identify deleted entries and attempt to reassemble content. This can be effective on some file systems and scenarios, but it is not universal and can be undermined by reuse and fragmentation.
3. **Targeted keyword searching of unallocated space and slack.** This is labor-intensive, produces many false positives ("noise hits"), and is increasingly ineffective against compressed, encoded, or encrypted content; it is best reserved for distinctive search terms and narrow objectives.

Better practice than "Undelete" is "Try to Find"

A protocol should target the recovery effort to the dispute: relevant file types, relevant locations, distinctive terms, and the specific question to be answered (exfiltration, wiping, fabrication, timing). It should also require the examiner to report feasibility early, because feasibility is often the whole ballgame.

Exemplar language

Examiner shall use reasonable, forensically sound methods to identify and, where feasible, recover deleted user-generated files reasonably likely to bear on the issues, prioritizing [file types] and [locations] and using the following search terms: ___. Examiner is not required to 'recover all deleted files' and shall instead report (a) what recovery methods were attempted, (b) what was feasible on the media at issue (including the impact of SSD/TRIM/encryption), and (c) the volume and reliability limitations of any recovered remnants.

Reporting and Deadlines

In digital forensics, "reporting" can mean very different things to different people. As a lawyer-examiner, I prepare narrative reports that explain—in plain language—what the evidence shows and what my training and experience suggest it signifies. Many forensic examiners, by contrast, treat reporting as a software function. A forensic "report" may consist of dozens or hundreds of pages of machine-generated output: hashes, offsets, tables, and logs that are impenetrable to anyone not steeped in the tool that produced them.

That disconnect should be addressed explicitly in the protocol. If the parties or the court need specific questions answered in narrative form, say so. If counsel expects the examiner to explain conclusions, correlations, limitations, or confidence levels, say so. Conversely, if a tool-generated export will suffice, say that instead.

Deadlines deserve similar candor. If the analysis must be completed by a date certain, the protocol should establish staged reporting—preliminary findings followed by a final report—and should acknowledge whether those deadlines are feasible. At the outset, the examiner has seen no

evidence and usually has more questions than answers. Unrealistic deadlines invite either superficial work or missed dates, neither of which serves the court.

Forensic Acquisition versus Forensic Examination
Parties and courts should distinguish—and apply very different standards to—requests for **forensically sound acquisition** versus requests for **forensic examination**.

Forensically sound acquisition preserves the status quo. It guards against spoliation caused by continued system use or intentional deletion and preserves the ability to conduct a forensic examination later, if warranted. Properly done, acquisition is minimally intrusive, relatively inexpensive, and can often be accomplished without exposing privileged or confidential information.

Forensic examination, by contrast, is intrusive and costly. It entails searching, interpreting, and sometimes reconstructing an opponent's digital life. Direct access to an adversary's computers or storage devices should be an extraordinary remedy, typically reserved for cases involving demonstrated discovery abuse, data destruction, concealment, or misconduct in which the device itself was instrumental. I caution courts against ordering such access reflexively.

Accordingly, courts should be far more willing to order forensic preservation upon a showing of good cause than to order forensic examination. Preservation keeps options open; examination commits resources and risks unnecessary intrusion.

During a forensically sound acquisition, three principles generally apply:

1. Nothing on the evidence media is intentionally altered by the acquisition;
2. Everything on the evidence media is faithfully acquired, subject to disclosed limitations; and
3. The processes employed are validated to confirm successful capture.

These standards cannot be met in every circumstance—notably in the logical acquisition of a live server, cloud system, or mobile device—but any deviation from a "change nothing" standard should be disclosed and justified. Transparency about what could not be captured, why, and with what consequence is essential to defensibility.

Exemplar <u>Acquisition</u> Protocol Order
Adapted from *Xpel Techs. Corp. v. Am. Filter Film Distribs.*, 2008 WL 744837 (W.D. Tex. Mar. 17, 2008), and updated to reflect contemporary storage, encryption, and forensic practice.

ORDER
The motion is GRANTED. Expedited forensic acquisition shall proceed as follows:

A. Examiner

Forensic acquisition shall be performed by _____ ("Examiner"), a qualified digital forensic examiner.

B. Costs

The reasonable costs of acquisition shall be borne by Plaintiff.

C. Examiner Obligations

Prior to commencing work, Examiner shall agree in writing to be bound by the terms of this Order and to serve as an officer of the Court with respect to the acquisition.

D. Acquisition Procedures

Within two (2) business days of this Order, or at such other time as the parties may agree, Defendants shall make available the specified computer systems and electronic storage media for forensic acquisition.

1. **Methodology**
    - Acquisition shall be conducted using forensically sound hardware and software capable of creating a verified, bit-for-bit image of the storage media where feasible.
    - The image shall capture accessible allocated space and, where technically possible and meaningful, unallocated space and file slack.
    - Examiner is not required to defeat encryption, security controls, or hardware-level protections absent further court order or party agreement.
2. **Documentation**

    Examiner shall document, to the extent observable without altering data:
    - Device make, model, serial/service tag;
    - Storage media type (HDD, SSD, NVMe, removable);
    - Operating system and version (if ascertainable);
    - Physical condition and configuration of the device;
    - Date, time, and location of acquisition.
3. **Authentication**
    - All forensic images shall be authenticated using cryptographic hash values.
    - Hash values shall be recorded and preserved as part of the chain of custody.
4. **Clock Assessment**
    - Without altering data, Examiner shall assess and document system clock settings and any observable deviations, recognizing that precise historical offsets may not always be determinable.
5. **Retention and Confidentiality**
    - Examiner shall retain the forensic images in secure storage and strict confidence.
    - Images shall not be examined or disclosed absent further order or written agreement.
    - Images shall be destroyed or returned only upon joint request or court order.

E. Minimizing Disruption

Examiner shall use reasonable efforts to minimize disruption to Defendants' business operations during acquisition.

F. Preservation Obligations

Defendants and their agents shall refrain from deleting, overwriting, encrypting, reimaging, or otherwise altering the subject devices pending completion of acquisition.

Exemplar Forensic Examination Protocol (Windows Laptop)

The following examination protocol was accepted by the Court in a case where the parties sought to determine what a user was doing on a laptop machine on a single day. As the machine was in Illinois, it was determined that the forensic image would be acquired by a local examiner and the image shipped.

I. PURPOSE AND GOALS

This Protocol governs the forensic examination of a Windows laptop computer ("Machine") for the purpose of identifying and interpreting active and latent artifacts bearing on:

1. The nature, timing, and extent of system and user activity during the Relevant Interval ([Date 1] through [Date 2]); and
2. The integrity of the Machine and its data, including indications of deletion, alteration, wiping, or other spoliation-related activity.

II. EVIDENCE PROVIDED

The Examiner shall receive:

- A forensically sound, hash-authenticated image ("Image") of the Machine's storage media;
- Chain-of-custody documentation; and
- Available baseline information regarding system clock accuracy at acquisition.

Unless otherwise agreed, only a qualified forensic examiner shall perform acquisition.

III. IMAGE HANDLING

The Examiner shall create authenticated working and archival copies of the Image. The original Image shall not be altered or used for analysis.

IV. COOPERATION AND ACCESS

The Parties shall cooperate by timely supplying non-privileged credentials reasonably required to access encrypted data contained within the Image.

Credentials shall be used solely for analysis of the Image and not for access to live systems or external accounts.

V. AUTHORIZATION AND SCOPE OF EXAMINATION

The Examiner is authorized to employ reasonable and accepted forensic methods to answer the questions posed, including the following, as technically feasible and proportionate:

1. Load authenticated working copies into appropriate forensic analysis platforms.
2. Assess evidence integrity, including indicators of drive substitution, reimaging, antiforensic activity, or clock manipulation.
3. Evaluate temporal consistency across system folders, operating system artifacts, and event records.
4. Identify metadata-altering, wiping, privacy, or obfuscation tools.
5. Identify recently installed applications relevant to the inquiry.
6. Identify hidden partitions, encrypted volumes, host-protected areas, and container files.
7. Examine Volume Shadow Copies, where present and accessible, recognizing storage-type and configuration limits.
8. Extract and analyze Registry hives for user, device, execution, and configuration artifacts.
9. Analyze relevant file system artifacts including but not limited to:
    - LNK files and Jump Lists;
    - Shellbags;
    - MRU lists;
    - Prefetch artifacts (where present);
    - Thumbnails and cache artifacts;
    - TEMP and orphaned files;
    - Registry entries and logs; and
    - Volume Shadow Services.
10. Generate comprehensive file listings with path, size, hash, and relevant MAC metadata, noting known limitations of timestamp reliability.
11. Where justified, conduct targeted keyword searches of allocated space, unallocated space, and file slack, recognizing diminishing utility on encrypted or solid-state media.
12. Where appropriate and proportionate, attempt recovery of specified categories of deleted files or fragments using reasonable methods, but not a directive to “undelete everything.”
13. Assess network, cloud, synchronization, and removable-media artifacts where supported by available evidence.
14. Correlate findings across artifact classes and assess consistency or contradiction.
15. Summarize findings and limitations and recommend additional inquiry only if likely to yield probative information.

VI. REPORTING

The Examiner shall report findings in a form sufficient to answer the questions posed. Narrative explanation shall be provided where interpretation, limitation, or context is necessary. Machine-generated outputs may be appended but shall not substitute for explanation where conclusions are drawn.

VII. COSTS

Plaintiff shall pay Examiner's reasonable fees and expenses at customary rates pursuant to the Examiner's engagement terms. Examiner shall notify the parties if projected charges materially exceed agreed benchmarks.

Hashing

The order above authorizes the examiner to "hash all files." As discussed earlier, hashing uses mathematical algorithms to compute a fixed-length digital fingerprint—commonly called a *hash value*—for a file or data set. Hashing is foundational to both computer forensics and electronic discovery, and courts increasingly rely on it to validate preservation, scope analysis, and production integrity.

Hashing allows an examiner to demonstrate that a forensic image is a faithful duplicate of its source by showing that the image and original media share identical hash values. It also enables file identification without opening or reviewing content. If two files have the same hash value, they are, for practical purposes, identical.

That matching capability makes hashing indispensable for:

- **Deduplication**, reducing review cost and avoiding inconsistent privilege or responsiveness decisions across identical files;
- **Identification of known files**, including proprietary data alleged to have been exfiltrated; and
- **Exclusion of known noise**, such as operating system and standard application files, through de-NISTing.

Courts may order hash analysis to:

1. Confirm proper preservation by matching hash values between originals and forensic images;
2. Identify files matching the hash values of known confidential or proprietary materials;
3. Exclude known irrelevant system and application files (de-NISTing);
4. Identify ESI produced in native form using hash values in lieu of Bates numbers; and
5. Reliably identify known contraband images, particularly in child exploitation prosecutions.

Hashing is powerful, but it is not magic. Because even a one-bit change alters a file's hash value, hashing is poorly suited to finding modified or partially altered content. Renaming a file does not change its hash value, but even trivial changes to content do. Thus, hashing excels at identifying *known, unchanged files*, not variants.

Moreover, while a changed hash value conclusively shows that a file is different, it reveals nothing about *how*, *when*, or *where* the file was altered. Those questions require artifact analysis, metadata review, and corroboration across other forensic sources.

Other Protocol Provisions to Ponder

Privilege and Confidentiality

The preceding protocol addressed a matter where privilege and confidentiality were not in play. That is atypical. Most forensic examination protocols must anticipate the presence of attorney-client communications, work product, trade secrets, personal data, and other protected material. A sound protocol should expressly provide for non-waiver of privilege, define who may communicate with the examiner, and establish a review sequence that permits privilege assertions before disclosure to adversaries.

Protocols should also address **ex parte communications**—communications with the examiner outside the presence of opposing counsel. Absent court approval or party agreement, such communications should be limited to logistics and administration.

Exemplar Language

To the extent the Examiner has direct or indirect access to information protected by the attorney-client privilege or work-product doctrine, such access shall not constitute a waiver of any privilege or protection. Except for purely ministerial communications necessary to accomplish the tasks set out in this Protocol, there shall be no communication between the Examiner and any party or counsel unless counsel for all parties are included.

All data, artifacts, and analyses governed by this Protocol are deemed Protected Material. Possession of Protected Material is limited to the Examiner, counsel of record in the captioned matter, and their retained experts. Protected Material shall not be shared with any other person, including the parties themselves, absent agreement or court order.

Any data or reporting generated by the Examiner shall first be provided to counsel for the device or media owner for privilege review. No data shall be disclosed to opposing counsel until released by producing counsel. The Examiner shall include an index identifying produced materials in a reasonably practicable manner, including file name (if applicable), file size, and last-modified date, where available. Producing counsel shall identify withheld items and state the basis for withholding. Disputes regarding withheld material shall be resolved cooperatively or, failing agreement, by motion to the Court.

Forms of Production

Forensic examinations rarely yield "documents" in the sense lawyers expect. Many probative artifacts are fragments, logs, records, or encoded values—not files that can be opened, printed, or Bates-stamped. Others exist only within system structures such as unallocated clusters, registries, or logs. Treating forensic outputs as if they were ordinary ESI invites confusion, error, and spoliation. It is often neither feasible nor appropriate to pre-define production formats at the outset. Loading forensic artifacts into document-review platforms may be impractical or misleading, and opening native files in native applications alters metadata and risks changing the evidence itself.

A practical protocol affords the examiner discretion to deliver results in forms reasonably calculated to aid counsel's review and, where necessary, the Court's **in camera** consideration—whether through reports, screenshots, tables, exports, or other defensible representations.

Ethical Boundaries

Unless expressly authorized by the Court or agreed by the parties, a forensic examiner must not use devices, credentials, or information derived from the examination to access data beyond what resided on the tendered media at the time of acquisition. This includes remote accounts and networked resources. Most examiners understand and honor this boundary, but a protocol should state it plainly to avoid doubt and prevent overreach.

Exemplar Language:

The Examiner shall not use the devices or storage media tendered for examination, nor any credentials or information derived therefrom, to access any electronic information not present on the devices or media at the time they were tendered. This prohibition includes accessing cloud repositories, email accounts, social-media platforms, financial accounts, or other remote systems.

Other Matters Commonly Addressed

Depending on the case, a protocol may also need to address disposition and destruction of forensic images, data-retention obligations, responsibility for storage costs, the examiner's amenability to deposition or subpoena, and the applicability of protective orders.

It is often useful to authorize the examiner to recommend additional lines of inquiry or sources of evidence. While courts should be alert to make-work, experienced examiners are frequently best positioned to identify corroborating data sources that materially advance the truth-finding process.

Conclusion

A defensible forensic examination protocol is not a form to be copied and pasted. It is a roadmap—one that reflects what the parties need to know, what the technology can reliably show, and what must be protected along the way. The inevitable tension between constraint and discretion is resolved by giving the examiner just enough latitude to follow the evidence, while anchoring the process in clear objectives, ethical limits, and proportionality.

Frequently Asked Questions

How do I preserve the status quo without requiring a party to stop using its systems?

Continued use of a computer inevitably alters evidence. Files are opened, logs roll, caches refresh, and unallocated space is reused. Where credible allegations suggest forensic examination may be

warranted, the most reliable way to preserve the status quo is prompt creation of a **forensically sound image** of the device or media by a qualified examiner, authenticated by cryptographic hashing.

An acceptable alternative—often faster and less disruptive—is to replace the storage device (or logically isolate it) and sequester the original evidence drive. This allows the user to continue working while preserving the evidence in a fixed state. The key is speed: delay increases both spoliation risk and dispute.

A party wants to have its technicians make clone images of the drives. Are those forensically sound images?
No. Tools commonly used by IT staff to clone drives for deployment or backup—such as Norton Ghost and similar utilities—are not designed for forensic acquisition. They typically do not capture all addressable space, may omit unallocated regions, and often alter the source media during acquisition.

Forensically sound imaging requires **write protection**, full capture of all addressable storage (including slack and unallocated space where relevant), and verification by hashing. Even appropriate forensic tools can yield unreliable results if used by untrained operators. That said, non-forensic imaging may be sufficient to satisfy preservation duties where forensic analysis is not at issue.

Do servers need to be preserved by forensically sound imaging?
Generally, no. Servers—especially virtualized or high-availability systems—are poor candidates for traditional forensic imaging. Their constant activity, storage architectures, and business-critical roles make shutdown impractical and unallocated-space analysis largely unproductive.
For preservation in e-discovery, **live or logical acquisition** of active data is usually sufficient. Forensic imaging of servers may be appropriate in rare circumstances (e.g., intrusion investigations), but courts should weigh business disruption heavily before ordering it.

What devices and media should be considered for examination?
Computers are rarely the whole story. Depending on the facts, candidates may include laptops, desktops, external drives, thumb drives, phones, tablets, removable media, cameras, IoT devices, vehicle data recorders, GPS units, backup systems, and cloud repositories.
Equally important are *where* and *by whom* devices are used. Home systems, legacy machines in storage, and devices used by assistants, family members, or other proxies may all hold relevant artifacts. Forensics follows usage, not job titles.

How intrusive is a computer forensic examination?

Acquisition itself is typically brief and non-destructive. Remote acquisition may be feasible; otherwise, devices are usually surrendered for hours—not days. Properly conducted acquisition installs no software, alters no data, and leaves the system functioning as before.

The intrusion lies in **what examination can reveal**. A forensic image exposes the full history of use: business and personal activity, deleted content, credentials, communications, and sensitive personal information. Once deleted data is implicated, irrelevant material cannot always be neatly segregated. That reality makes **scope limits, privilege protections, and often a neutral examiner** essential.

What does it cost?

Forensic preservation of a single computer often costs no more than a short deposition. Examination costs vary widely. Hourly rates in 2026 commonly range from **$250 to $750**, depending on experience, specialization, and geography. Poorly scoped examinations can escalate quickly into six figures.

Nothing influences cost more than scope. Clear protocols, targeted objectives, and periodic progress reporting keep costs in check. Courts should expect—and require—cost checkpoints and informed course correction.

Additional Questions That Belong in a 2026 FAQ

Does full-disk encryption defeat forensic examination?

No—but it changes the playbook. Encryption often renders unallocated-space analysis unproductive, but it does not eliminate active artifacts, logs, registry data, application metadata, cloud sync traces, or volume snapshots (where present). Cooperation with credentials may be essential, and protocols should address this explicitly.

Can forensic examiners tell who was "at the keyboard"?

Rarely, and never conclusively. Forensics can often show *what* happened and *when*, but attribution to a specific human actor requires corroboration from access controls, usage patterns, and non-digital evidence. Protocols should avoid demanding certainty where none is defensible.

Are cloud and SaaS data part of computer forensics?

Increasingly, yes—but access is usually indirect. Devices often retain sync data, caches, logs, and credentials that illuminate cloud activity. Direct access to remote accounts requires court authorization or party consent and should be addressed expressly in the protocol.

Is forensic preservation different from forensic examination?

Very much so. Preservation fixes evidence in time and is relatively inexpensive and non-intrusive. Examination is invasive, costly, and should be ordered only upon a showing of genuine need. Courts should distinguish between the two and not conflate preservation with permission to rummage.

Should a neutral examiner be used?
Often. Neutral examiners reduce disputes, protect privacy and privilege, and enhance credibility. They are particularly appropriate where personal data, trade secrets, or allegations of spoliation are central. Neutrality costs less than prolonged motion practice.

How current are forensic artifacts—do they change with OS updates?
Constantly. Artifacts differ by operating system, version, configuration, and usage. What persists in one Windows release may disappear or relocate in the next. Protocols should focus on **questions to be answered**, not brittle lists of artifacts, and allow corroboration across sources.

The Big Six: Getting Your Arms Around the ESI Elephant

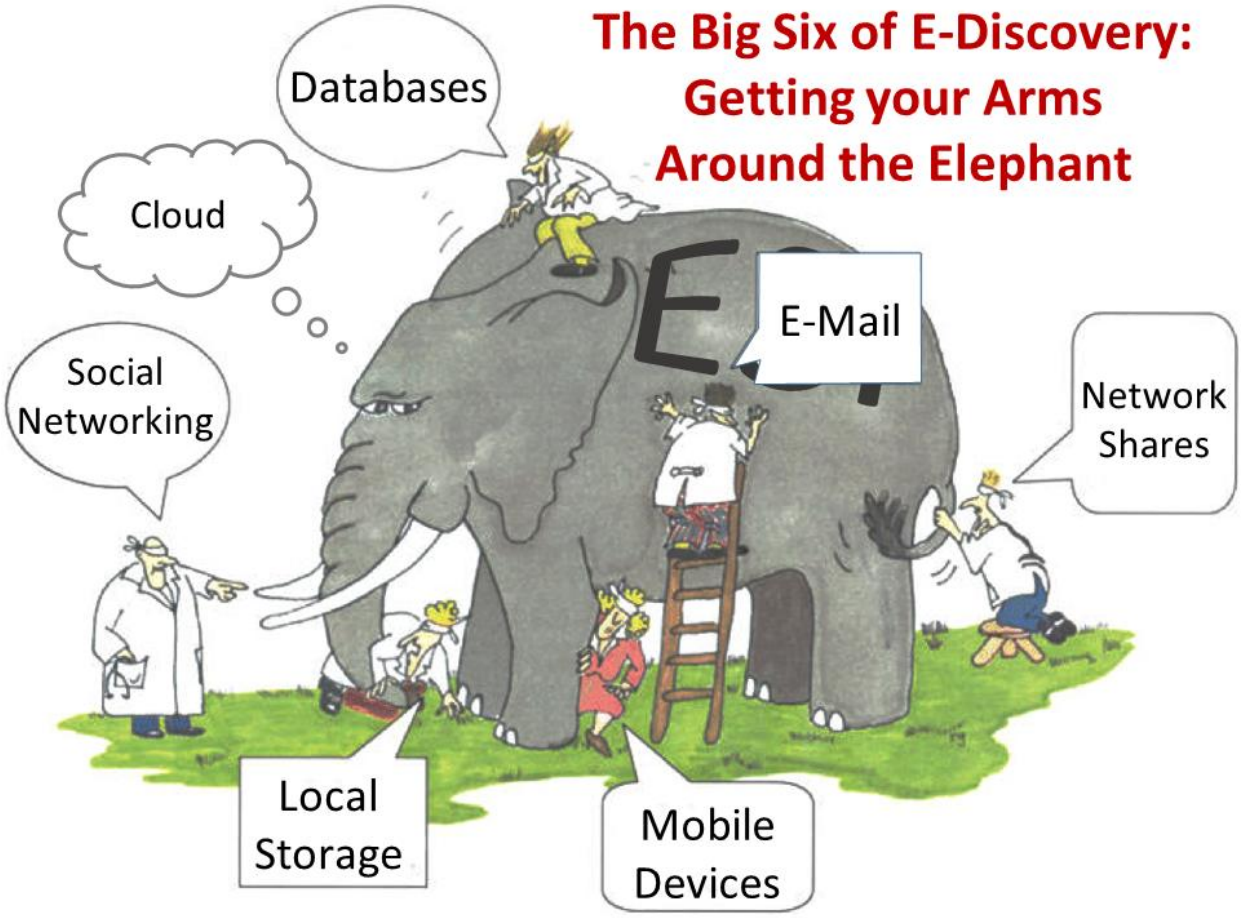
Many cultures share the parable of the six blind men describing an elephant. One grasps the tail and insists the elephant is a snake. Another seizes the trunk and says it's a tree branch. A third hugs a leg and declares it a tree trunk. The man touching the ear says it's a large leaf; the one at the tusk swears it's a spear. None understands the elephant because none considers more than the piece within reach.

Electronic discovery suffers from the same myopia. We cannot grasp the true nature of potentially responsive ESI until we touch many parts of the digital elephant.

There are no forms or checklists that substitute for understanding electronic evidence any more than a Polish phrasebook equips you to try a case in Gdańsk. Still, a few rules of thumb—applied thoughtfully—can help you get your arms around the ESI elephant. We start with the Big Six and work through some necessary geek speak along the way.

VITAL VOCABULARY

Cloud
Container Files
Journaling Server
Key Custodian
Microsoft 365
Microsoft Exchange
Media Rotation
Network Share
OneDrive
Schema
SharePoint
Slack
Structured Data
Teams



The Big Six

Without knowing anything about a party's IT systems, you can safely assume there are **at least six principal sources** of potentially responsive ESI in most matters.

1. Messaging (Email and Texts)

Corporate users still generate a substantial volume of communications through e-mail systems—whether hosted on-premises or via Software-as-a-Service platforms such as Microsoft 365 or Google Workspace. Traditional enterprise platforms like Exchange remain, but increasingly they coexist with—or are overshadowed by—collaboration systems such as Microsoft Teams, Slack, and similar tools that blend messaging, file sharing, and workflow.

Users typically maintain overlapping bodies of communications across servers, local machines, mobile devices, and cloud repositories. On desktops and laptops, e-mail may occasionally reside in container files such as PSTs and OSTs; but in cloud-first environments, it may never touch local storage at all. Archival sources—journaling systems, retention vaults, backups, and legacy mail stores—often persist long after users assume messages are "gone."

Your first task is to pin down **where communications live** and **how long they persist**, paying special attention to automated deletion, retention policies, and backup rotation. The critical questions remain:

- What types of messages exist (sent, received, drafts, chats, channels)?
- What temporal range is available, and where are the gaps?
- What is the scale—message counts versus total data volume?
- How are attachments handled? Are they embedded attachments or linked files?

Until you answer those questions, you don't yet know the shape of the elephant.

2. Shared Repositories

Apart from communications, custodians generate substantial work product as documents—spreadsheets, presentations, reports, designs, and data exports. Historically these lived on local hard drives or mapped network shares. Today, they increasingly reside in shared, cloud-hosted workspaces: SharePoint, OneDrive, Google Drive, Box, deal rooms, and virtual data rooms.

These repositories blur traditional distinctions between "my files" and "our files." Group shares and collaborative folders are easy to overlook because no single custodian feels ownership, yet they often contain consequential work product. Versioning, co-authoring, and permissions metadata can be as probative as document content.

As before, shared spaces may exist on company infrastructure or in leased cloud environments of uncertain geography. Either way, they must be identified, preserved, and assessed with the same care as individual custodians' data.

3. Mobile Devices

Mobile devices are no longer peripheral to discovery; they are the center of gravity. Phones and tablets are the primary platform for messaging, photography, navigation, authentication, and access to cloud services. The assumption that "whatever is on the phone must exist somewhere else" is not merely risky—it is flat wrong.

Modern mobile devices generate dense, time-stamped records of daily life, much of it unique: texts, app messages, photos, videos, location artifacts, and application usage metadata. Many of these artifacts never exist in enterprise systems and may not be preserved elsewhere in recoverable form. If you are not considering phones and tablets, you are almost certainly missing relevant, unique, and often highly probative evidence. Any claim that mobile data is duplicative should be supported by metrics, not optimism.

4. Local Storage

Despite decades of policy and technology designed to corral work product into centralized repositories, users persist in storing data locally. Laptops, desktops, external drives, thumb drives, camera media, and ad hoc storage devices remain fertile ground for unique ESI.

Local storage is often where drafts live, where downloads linger, where users stash files they do not wish to share, and where artifacts of prior activity endure long after users forget them. Assuming local storage is irrelevant is rarely defensible absent documented proof that policy and reality align.

5. Social Media

Social networking and online platforms continue to be rich sources of candid, unguarded information. Posts, messages, photos, comments, reactions, and metadata may reveal intent, knowledge, state of mind, relationships, and chronology. Professionals blog, post, and comment. Businesses receive praise and complaints through social channels. Personal relationships play out in public and semi-public forums.

If you are not asking about social, online and collaborative content—broadly conceived—you are ignoring a substantial portion of the elephant.

6. Databases

Databases underpin nearly every modern organization, from small desktop databases to massive enterprise systems and cloud-hosted platforms such as CRM, ERP, HR, finance, and analytics

services. Whether housed locally or accessed as a subscription service, databases hold operational truth.

The initial task is not to extract everything but to identify what databases exist, what they do, who uses them, and what time ranges and fields they contain. Standard reports may suffice; if not, counsel must understand enough about schema and query capability to request targeted, defensible extractions.

Plus: The Cloud

The Big Six increasingly intersect—and sometimes collapse—into the cloud. Cloud services host communications, documents, databases, analytics, backups, and machine logs. Much data exists *only* in cloud systems; some is replicated across devices and services; some is ephemeral by design. The cloud is no longer merely where data is stored—it is where work happens. Treating it as an afterthought is a recipe for missed evidence and credibility problems.

Pitfalls and Sinkholes

Few organizations preserve all legacy data, but most retain far more than they realize in backups, archives, and dormant systems. Courts may excuse searching such data when burden and cost are proven—but not when parties cannot even say what exists.

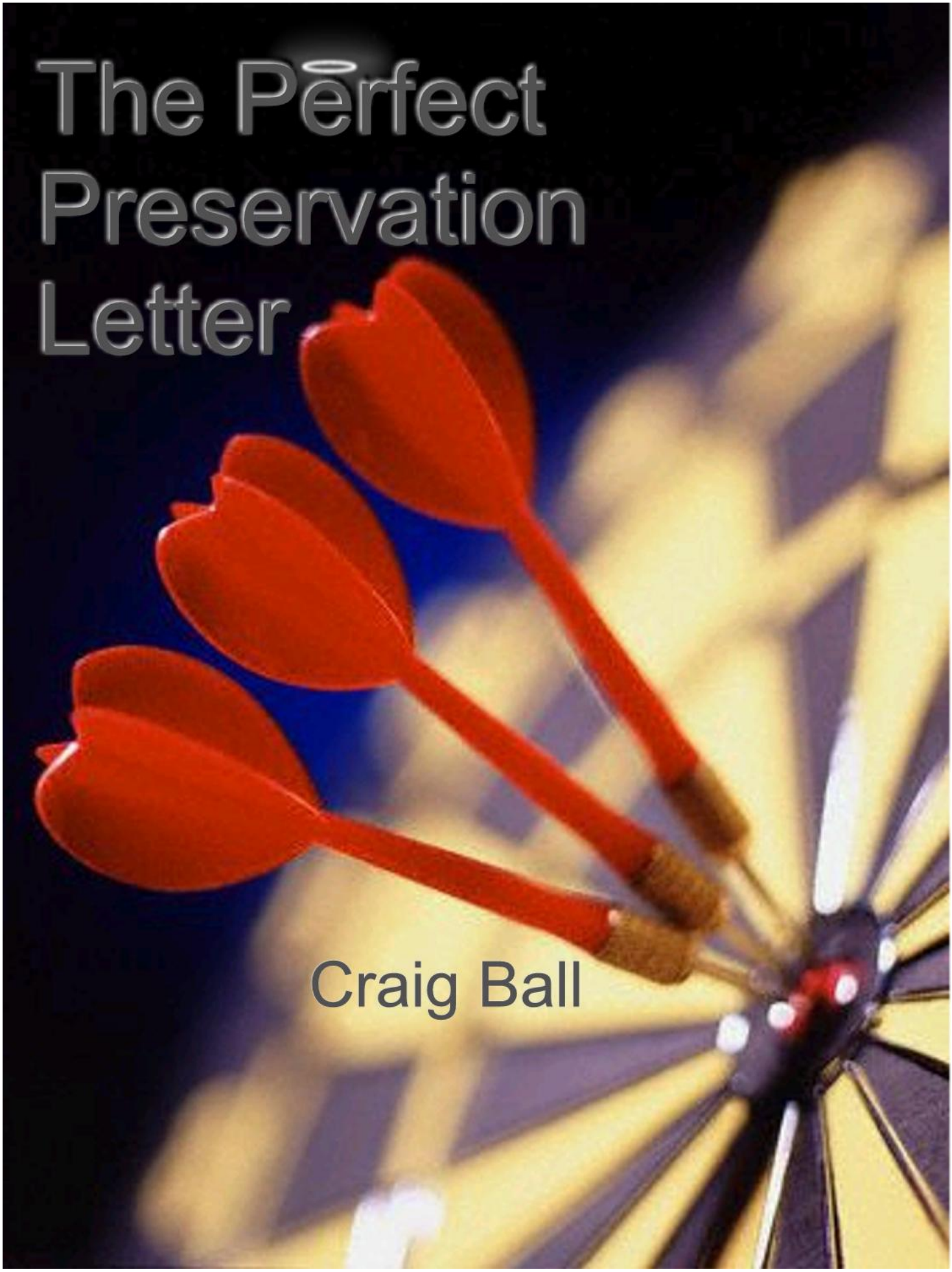
A related pitfall is accepting "it's gone" too readily. With minimal persistence and informed questioning, supposedly vanished data often resurfaces. Custodians and administrators assume; so lawyers must probe and verify.

Lather, Rinse, Repeat

Preservation does not require immediate collection of everything. In high-volume matters, divide and conquer. Start with the most relevant and accessible ESI from the über-key custodians. Use what you learn to refine subsequent discovery. Iterate. Ask again. Trust but verify.

Bottom line: The Big Six (plus the cloud) will not reveal every piece of ESI, but they will reveal most of what matters in most cases—so long as you approach them with curiosity, skepticism, and a willingness to touch more than one part of the elephant.

The Perfect Preservation Letter



Craig Ball

The Perfect Preservation Letter

Craig Ball

©2020

Well, I was drunk the day my Mom got outta prison, and I went to pick her up in the rain; But before I could get to the station in my pickup truck, she got runned over by a damned old train.

From *"You Never Even Called Me by My Name" (a/k/a "The Perfect Country and Western Song")*
By Steve Goodman, performed by David Allan Coe

Outlaw musician David Allan Coe sings of how no country and western song can be "perfect" unless it talks of Mama, trains, trucks, prison and getting drunk. Likewise, no digital evidence preservation letter can be "perfect" unless it clearly identifies the materials requiring protection, educates your opponent about preservation options and lays out the consequences of failing to preserve the evidence. *You won't find the perfect preservation letter in any formbook.* You must custom craft it from a judicious mix of clear, technically astute terminology and *fact-specific* direction. It compels broad retention while asking for no more than the essentials. It rings with reasonableness. Its demands are *proportionate to the needs of the case,* and it keeps the focus of e-discovery where it belongs: *on relevance*. This chapter discusses features of an effective, efficient preservation letter and offers suggestions as to how it can be drafted and deployed.

VITAL VOCABULARY
Preservation Letter
Proportionate
Anticipation of Litigation
Sanctions
Intent to Deprive
System Metadata
Application Metadata
Backup Tape
Computer Forensics
Forensic Image
Bitstream

THE ROLE OF THE PRESERVATION LETTER

You can read the Federal Rules of Civil Procedure from cover to cover and you'll find no mention of preservation letters. So why invest effort creating the perfect preservation letter? Doesn't every lawyer know the law and rules prohibiting destruction of evidence apply to electronically stored information just like any other evidence? Don't all litigators ensure clients take reasonable steps to preserve information in anticipation of litigation and discovery? Fifteen years after amendment of the Federal Rules on these points and countless published decisions post-*Zubulake v. UBS Warburg*, 220 F.R.D. 212 (S.D.N.Y. 2003), the answer remains a sad, resounding "NO." *You cannot rely upon the competence and training of opposing counsel when it comes to electronic evidence.* Too many litigators and in-house counsel remain clueless and careless about information systems. The reality of electronic discovery is it starts off as the responsibility of those who don't understand the technology and ends up as the responsibility of those who don't understand the law. A well-drafted preservation letter helps bridge this knowledge gap.

The reality of electronic discovery is it starts off as the responsibility of those who don't understand the technology and ends up as the responsibility of those who don't understand the law.

At bottom, the preservation letter reminds parties to preserve evidence, *to act,* so evidence doesn't disappear. But the preservation letter also serves as the linchpin of claims for spoliation, helping establish the requisite intent to deprive and conscious disregard for the duty to preserve. The more plainly and practically you convey what evidence must be retained, the greater your client's access to justice when an opponent loses or destroys it.

The more plainly and practically you convey what evidence must be retained, the greater your client's access to justice when an opponent loses or destroys it.

THE RULES OF CIVIL PROCEDURE

Though serving a preservation letter isn't a formal component of civil discovery procedures, it's a wise precursor to the obligations imposed by the federal, state and local rules of procedure imposing discovery "meet and confer" obligations. Rule 26 of the Federal Rules of Civil Procedure requires litigants "discuss any issues about preserving discoverable information, *Fed. R. Civ. P. Rule 26*, and "any issues about disclosure, discovery, or preservation of electronically stored information, including the form or forms in which it should be produced." *Fed. R. Civ. P. Rule 26(f)(3)*. By compelling early consideration of the nature and scope of potentially relevant evidence, often before litigation has begun, the preservation letter serves to frame the agenda for conferences to follow.

The preservation letter plays a key role in a court's consideration of whether a party acted in bad faith in connection with the irreparable loss of data that should have been preserved. Rule 37(e) of the Federal Rules of Civil Procedure states:

> **Failure to Preserve Electronically Stored Information.**
> If electronically stored information that should have been preserved in the anticipation or conduct of litigation is lost because a party failed to take reasonable steps to preserve it, and it cannot be restored or replaced through additional discovery, the court:
> (1) upon finding prejudice to another party from loss of the information, may order measures no greater than necessary to cure the prejudice; or
> (2) only upon finding that the party acted with the intent to deprive another party of the information's use in the litigation may:
> (A) presume that the lost information was unfavorable to the party;
> (B) instruct the jury that it may or must presume the information was unfavorable to the party; or
> (C) dismiss the action or enter a default judgment.

Assessment of intent turns on the subjective awareness of the party failing to preserve evidence. The preservation letter helps establish such awareness, proving a party destroying evidence knew of its discoverability and purposefully disregarded it. A clear and instructive preservation letter that serves to educate your opponent isn't just a professional courtesy; it compels recognition of the duty to intervene to prevent data loss and makes it harder to assert ignorance as a defense.

A clear and instructive preservation letter that serves to educate your opponent isn't just a professional courtesy; it compels recognition of the duty to intervene to prevent data loss and makes it harder to assert ignorance as a defense.

What is Electronic Evidence Preservation?

When evidence was on paper, preserving it was simple: We set the original or a copy aside, confident that it would come out of storage as it went in. Absent destructive forces or tampering, paper stays the same. But despite lawyers' archaic ardor for paper, modern information is born *digitally* and stored *digitally*. Little of it is ever printed save for short-term convenience and then discarded.

Preserving electronically stored information (ESI) poses unique challenges because:

- Touching ESI changes it
- Digital evidence is ill-suited to printing
- ESI must be interpreted to be used
- Storage media are fragile and dynamic, changing all the time
- Digital storage media are disposable and recyclable

TOUCHING ESI CHANGES IT

Route a document through a dozen hands and, aside from a little finger grime or the odd coffee stain, the document won't be changed by moving, copying or reading it. But, open the same document in Microsoft Word, or copy it to a thumb drive, and you've irretrievably changed the document's *system metadata*, the data-about-data metrics, like a document's creation date, that may be evidence in its own right. Open the document in its native application (e.g., Microsoft Word) and embedded *application metadata* values are irreparably altered.

Even the medium employed to copy or transmit data may play a role in altering its metadata. Back when it was common to use recordable optical disks to transfer or produce ESI, few appreciated that merely copying a file from a Windows computer to a recordable CD-R stripped the file of time values. Hard drives, floppy disks, thumb drives and optical media all use different file system architecture such that the CD-R doesn't offer a structure capable of storing all Windows time metadata. Where the Windows NTFS file system offers three "slots" for storing file dates (i.e., Modified, Accessed and Created), the CD-R's Joliet file structure supplies just one. With nowhere to go, temporal metadata is jettisoned in the CD recording process, and the missing metadata misreported on the destination system. Similar incongruities may impact the ability to store long filenames as well as the precision of time values. When ESI is evidence, such differences matter.

DIGITAL EVIDENCE IS ILL-SUITED TO PRINTING

Much modern evidence doesn't lend itself to paper. For example, a spreadsheet displays values derived from embedded formulae, but you can't embed those formulae in paper and see the calculated values. In large databases, information occupies expansive grids that wouldn't fit on a printed page. Sound and video evidence can't make the leap to paper and allocating a full sheet of paper to every text message is insanely wasteful and cumbersome. So, preserving on paper has ceased to be a practical option.

ESI MUST BE INTERPRETED TO BE USED

If legible and in a familiar language, a paper document conveys information directly to the reader. A literate person can interpret an alphabet, aided by blank spaces and a few punctuation marks. It's a part of our grade school "programming." All digital data are just streaming information denoted as ones and zeroes. For these streams of data to convey anything intelligible to humans, the data must be interpreted by a computer using specialized programming called "interfaces" and "applications." Without the right interface and application—sometimes even without the correct *version* of an interface or application—data is wholly inaccessible or may be inaccurately presented. Successfully preserving data may entail preserving legacy applications capable of correctly interpreting the data as well as legacy computing environments—hardware and software—capable of running the applications. Operator's manuals and the schema laying out a database's architecture may be needed as well.

STORAGE MEDIA ARE FRAGILE AND DYNAMIC

If your great grandfather put a letter in a folder a century ago, chances are good that apart from signs of age, you could pull it out today and read it. But changes in storage technology and instant obsolescence have already rendered fifteen-year-old digital media largely inaccessible absent considerable effort and expense. How many of us still have a computer capable of reading an optical disk, let alone a floppy disk? Data stored on back up tapes and other magnetic and optical media fades and disappears over time like the contents of once-common thermal fax paper. Disks expected to last a century are turning up illegible in a few years. Back up tapes stretch a bit each time they are used and are sensitive to poor storage conditions. Long-term data preservation will entail either the emergence of re durable media or a relentless effort to migrate and re-migrate legacy data to new media.

DIGITAL STORAGE MEDIA ARE DISPOSABLE AND RECYCLABLE

By and large, paper is not recycled for information storage; at least not in a way we'd confuse its prior use as someone's Last Will & Testament with its reincarnation as a cardboard carton. By contrast, a hard drive is constantly changing and recycling its contents. A later version of a document may overwrite—and by so doing, destroy—an earlier draft, and storage space released by the deletion of one file may well be re-used for storage of another. This is in sharp contrast to paper preservation, where you can save a revised printout of a document without affecting—and certainly not obliterating-- a prior printed version.

Clearly, successful preservation of digital data isn't as simple as copying something and sticking it in a folder; but your opponent may not appreciate the planning and effort digital preservation requires. When that's the case, the requesting party is at a crossroads: Do you seek to educate the producing party or its counsel about how and why to properly preserve digital evidence, or do you keep mum in hopes that an advantage will flow from your opponent's ineptitude? That is, do you want the evidence or the sanction?

Setting an opponent up for a spoliation sanction is a fool's errand; most of the time, you'll want the evidence.

THE DUTY TO PRESERVE

At what point does the duty to preserve evidence arise? When the lawsuit is filed? Upon receipt of a preservation letter? When served with a request for production?

The duty to preserve evidence may arise before—and certainly arises without—a preservation letter. In fact, the duty can arise long before. A party's obligation to preserve evidence is generally held to arise when the party knows or has reason to know that evidence may be relevant to future litigation. This "reasonable anticipation of litigation" standard means that any person or company who should see a claim or lawsuit on the horizon must act, *even before a preservation letter or lawsuit has materialized*, to cease activities likely to destroy electronic or tangible evidence and must take affirmative steps to preserve such evidence.

The duty to preserve evidence may arise before—and certainly arises without—a preservation letter.

Thus, the preservation letter is but one of several events sufficient to trigger the duty to preserve evidence, *but the preservation letter is an explicit, decisive trigger.* Often, the preservation letter's arrival marks the moment parties awaken to their duty to determine what evidence exists and what must be retained.

Often, the preservation letter's arrival marks the moment parties awaken to their duty to determine what evidence exists and what must be retained.

BALANCE, REASONABLENESS AND PROPORTIONALITY

I've seen producing parties sneer in contempt at preservation letters when they should consider them a gift. A well-crafted preservation demand is well-nigh a checklist of sources and forms of potentially relevant ESI. Does it too-often overreach? Certainly, because most are drafted by lawyers knowing little-or-nothing about an opponent's information systems. Apprehension and ignorance foster everything-but-the-kitchen-sink requests; the perfect preservation letter esteems the "how" and "how much" issues faced by the other side.

A preservation letter seeking everything and a pony or serving to paralyze an opponent's operations won't see compliance or enforcement. Absent evidence of misconduct (e.g., overt destruction of evidence), a court won't sanction a party for failing to comply with a preservation letter so onerous that no one dare turn on their computer for fear of spoliation! For a preservation letter to work, it must be reasonable on its face.

A preservation letter seeking everything and a pony or serving to paralyze an opponent's operations won't see compliance or enforcement.

Take Note: If your goal is to keep the other side from destroying relevant evidence, any judge will support you in that effort *if your demands aren't cryptic, overbroad or unduly burdensome*. In a word: *proportionate.*

If it could be accomplished with paper evidence, judges expect a corollary accomplishment with electronic evidence. Still, digital is different, and some of the ways we approach paper discovery just won't fly for electronic evidence. For example, using the term "any and all" in a request for digital evidence is a red flag for potential over breadth. Demanding that an opponent retain "any and all electronic communications" is nonsense. After all, phone conversations are electronic communications, and it's unlikely that, outside a regulated environment like a retail brokerage, a court would require a litigant to record all calls, though a judge shouldn't hesitate to compel *retention* of recordings (think Zoom meetings) *when conferences are already recorded and relevant.* If what you want preserved is e-mail, or text messaging or social networking content, *spell it out.* Your opponent may squawk, but at least the battle lines will be drawn on specific evidentiary items your opponent may destroy instead of fighting about vague language" The risk to this approach is that your opponent may fail to preserve what you haven't specified. Fear not! To the extent the evidence destroyed was relevant and material, an omnibus request to retain information items bearing on the claims made the basis of the claim will catch it.

Remember: the preservation letter neither creates the duty to preserve *nor constrains it*. Parties must still think for themselves. If the evidence was relevant and discoverable, its intentional destruction is spoliation, even if you didn't cite it in your preservation demand.

Remember: the preservation letter neither creates the duty to preserve nor constrains it. Parties must still think for themselves. If the evidence was relevant and discoverable, its intentional destruction is spoliation, even if you didn't cite it in your preservation demand.

PRESERVATION ESSENTIALS

First and foremost, a perfect preservation letter must seek to halt routine business practices geared to the destruction of potential evidence. It might call for an end to automatic purging of messages, repurposing of drives, overwriting of logs, scheduled destruction of back up media, sale, gift or destruction of computer systems and, (especially if you know computer forensics may come into play) running "privacy" software.. A lot of digital evidence disappears because of a lack of communication ("legal forgot to tell IT") or of individual initiative ("this is MY e-mail and I can delete it if I want to"). So, be sure to highlight the need to effectively communicate retention obligations

to those with hands-on access to systems and suggest steps to forestall personal delete-o-thons. **Remember:** When you insist that communications about preservation obligations ***reach*** every custodian of discoverable data and that such communications stress the importance of the duty to preserve, you are demanding no more than the law requires. See, *e.g., Zubulake, supra.*

Next, get fact specific! Focus on items specifically bearing on the claim or suit, like relevant business units, activities, practices, procedures, time intervals, jurisdictions, facilities and key players (a/k/a "custodians"). Here, follow the "who, what, when, where and how" credo of good journalism. Preservation letters are more than a boilerplate form into which you pack every synonym for document and computer. If your preservation letter boils down to "save everything about anything by everyone, everywhere at any time," it's time to re-draft it because not only will no trial court enforce it, many will see it as discovery abuse.

The preservation letter's leading role is to educate your opponent about the many forms of relevant electronic evidence and the importance of taking prompt, affirmative steps to be sure that evidence remains accessible. Educating the other side isn't a noble undertaking—it's sound strategy. Spoliation is frequently defended on the basis of ignorance; *e.g.,* "Your honor, we had no idea that we needed to do that," and your goal is to slam the door on the "it was an oversight" excuse. Doing so entails more than just reciting a litany of storage media to be preserved--you've got to educate, clearly and concisely.

Don't be so focused on electronic evidence that you fail to direct your opponent to retain the old-fashioned paper variety. Finally, remember that *turnabout is fair play*. Don't expect to hold your opponent to a standard of preservation your client won't meet. Your opponent may face a greater *burden* to preserve a larger volume or variety of relevant information, but *their duty to preserve is no greater than yours*.

Don't expect to hold your opponent to a standard of preservation your client won't meet.

THE NATURE OF THE CASE

As documentary discovery typically follows service of a complaint, parties know what a dispute is about by the time the first request arrives. But a pre-suit preservation letter may be your opponent's first inkling they face litigation. Don't assume those receiving your preservation letter know what the dispute is about: *spell it out for them*. Supply sufficient information about the claim to allow a reasonable person reading the preservation letter to understand what evidence may be relevant. Names of key players, dates, business units, office locations, causes of action and events will all be weighed in deciding what's relevant and must be retained. The more you elucidate, the less likely you are to hear, "*If you wanted Madison's text messages, why didn't you mention Madison in the preservation letter?*"

WHEN TO SEND A PRESERVATION LETTER

The conventional wisdom is that preservation letters should go out as soon as you can identify potential defendants. But there may be compelling reasons to delay sending a preservation letter.

For example, when you face opponents who won't hesitate to destroy evidence, a preservation letter is just the starting gun and blueprint for a delete-o-thon. Instead, consider seeking a temporary restraining order or appointment of a discovery master (but recognize that the Comments to the proposed Rules amendments strongly discourage entry of *ex parte* preservation orders). Deferring the letter may be wise when your investigation is ongoing, and the service of a preservation letter will cause the other side to hire a lawyer or trigger work product privileges running from the anticipation of litigation. There may even be circumstances where you ***want*** your opponent's routine, good faith destruction of information to continue, such as where information unfavorable to your position will be lost in the usual course of business.

WHO GETS THE LETTER?

If counsel hasn't appeared for your opponent, to whom should you direct your perfect preservation letter? Here, the best advice is erring on the side of as many appropriate persons as possible. Certainly, if an individual will be the target of the action, he or she should receive the preservation letter. However, if you know of others who may hold potential evidence (such as a spouse, accountant, employer, banker, customers and business associates), it's smart to serve a *tailored* preservation demand on them, making clear that you are seeking preservation of physical and electronic records in their possession pertaining to the matters made the basis of the contemplated action. Some litigants use the preservation letter to put pressure customers lost to or solicited by a competitor-defendant. ***Beware such tactics!*** The preservation letter isn't a discovery mechanism expressly countenanced by the rules of procedure, so its misuse as an instrument of intimidation may not be privileged and could provoke a counterclaim based of libel or tortuous interference.

If the other side is a corporation, a directive to the wrong person may be ignored or be late in reaching those capable of putting a litigation holds in place. Consequently, if no counsel has appeared, it's wise to direct preservation letters to several within the organization, including, *inter alia*, the Chief Executive Officer, General Counsel, Director of Information Technologies and perhaps even the Head of Corporate Security and registered agent for service of process. You may want to copy other departments, facilities or business units.

Consider who is most likely to *unwittingly* destroy evidence and be certain that person receives a preservation letter. Sending a preservation letter to a person likely to destroy evidence *intentionally* is a different story. The letter may operate as the triggering event to spoliation, so you may need to balance the desire to give notice against the potential for irretrievable destruction.

Of course, preservation letters, like any important notice, should be dispatched in a way enabling you to prove receipt, even if that means via certified mail, return receipt requested.

HOW MANY PRESERVATION LETTERS?

Turning to the obligatory litigation-as-war metaphor, is a preservation letter best delivered as a single giant salvo across the opponent's bow, or might it instead be more effectively launched as several targeted blows? It's common to dispatch a single, comprehensive request, but might it instead be wiser to present your demands in a *series* of focused requests, broken out by, *e.g.*, type

of digital medium, issues, business units, or the roles of key players? Your preservation letter may be destined to be an exhibit to a motion, so when the time comes to seek sanctions for a failure to preserve evidence, wouldn't it be more compelling to direct the court to a lean, specific preservation notice than a bloated beast? Consider supplementing a "master" preservation notice with specific notices directed at key players as the matter proceeds. It's difficult to claim, "*We didn't realize you wanted Elizabeth's Facebook content*" when Elizabeth got her very own, custom-tailored preservation letter.

SPECIFYING FORM OF PRESERVATION

The Federal Rules of Civil Procedure permit a requesting party to specify the form or forms in which the requesting party wants electronic evidence produced. Often, there's no additional trouble or expense for the producing party to generate one format over another and there may be occasions where a non-native production format is preferred, such as when evidence must be redacted to remove privileged content. *But should the preservation letter specify the form in which the data should be preserved?* Generally, not. Your preservation letter should not demand preservation in forms other than those used in the ordinary course of business. However, when your specification operates to *ease* the cost or burden to the producing party or otherwise *help* the producing party fulfill its preservation obligation, an alternate format might be *suggested*.

Your preservation letter should not demand preservation in forms other than those used in the ordinary course of business.

SPECIAL CASES: BACK UP TAPES, COMPUTER FORENSICS AND METADATA

The e-discovery wars rage in the mountains of e-mail and flatlands of Excel spreadsheets, but nowhere is the battle so pitched as at the front lines and flanks called *back up tapes, computer forensics and metadata*. These account for much of the bloodshed and so deserve special consideration in a preservation letter.

BACK UP TAPES

In the "capture the flag" e-discovery conflicts waged years ago, waged, the primary objective was often your opponent's server backup tapes or, more particularly, forcing their retention and restoration. Backup systems have but one legitimate purpose, being the retention of data required to get a business information system "back up" on its feet in the event of disaster. To this end, a business need retain disaster recovery data for a brief interval since there are few instances where a business would wish to re-populate its information systems with stale data. Because only the latest data has much utility in a well-designed backup system, the tapes containing the oldest backed-up information are typically recycled. This practice is "tape rotation," and the interval between use and reuse of a tape or set of tapes is the "rotation cycle" or "rotation interval."

Ideally, the contents of a backup system would be entirely cumulative of the active "online" data on the servers, workstations, laptops and other devices that make up a network. But, because businesses entrust the power to destroy data to every computer user--including those motivated to make evidence disappear--backup tapes are often the only evidence containers beyond the reach of those with the incentive to destroy or fabricate evidence. Going way back to Col. Oliver

North's deletion of e-mail subject to subpoena in the 1980's Iran-Contra affair, it's long been the backup systems that ride to truth's rescue with "smoking gun" evidence.

Another reason backup tape lay at the epicenter of early e-discovery disputes was that many organizations used to retain back up tapes long after they lost their usefulness for disaster recovery. When data has been deleted from the active systems, the stale backup tapes are a means by which the missing pieces of the evidentiary puzzle can be restored.

In organizations with many servers, backup systems are complex, hydra-headed colossi. There may be no simple one-to-one correspondence between a server and a user, and most tape backup systems structure stored data differently from active data on the server, complicating restoration and exploration. Volume, complexity and the greater time it takes to access tape compared to disk all contribute to the potentially high cost of targeting backup tapes in discovery. Compelling a large organization to interrupt its tape rotation, set aside back up tapes and purchase a fresh set can carry a princely price tag, but if the tapes aren't preserved, deleted data may be gone forever. That's been the Hobson's choice[72] of e-discovery.

A preservation letter should target just the backup media likely to contain deleted data relevant to the issues in the case—a feat easier said than done. Whether by Internet research, contact with former employees or consultation with other lawyers who've plowed the same ground, seek to learn all you can about the architecture of the active and backup systems. The insight gleaned from such an effort may allow for a more narrowly tailored preservation request or justify a much broader one.

The responding party need not preserve evidence that is merely cumulative, so once established that data has not been deleted and all relevant information still exists on the servers, the backup tapes should be released to rotation. Again, this is harder than it sounds because it requires three elements often absent from the adversarial process: ***communication, cooperation and trust*.** Hopefully, the adoption of compulsory meet-and-confer sessions in state courts will force litigants to focus on e-discovery issues sufficiently early to stem unnecessary costs by narrowing the breadth of preservation efforts to just those actions or items most likely to yield discoverable data.

DRIVE IMAGING

Data deleted from a personal computer isn't gone. On electromagnetic ("spinning") hard drives, the operating system simply releases the space the deleted data occupies for reuse and treats the space as available for reuse. Deletion rarely erases data. In fact, there are three and *only* three ways that information's destroyed on personal computer:

There are three and *only* three ways that information's destroyed on a personal computer

[72] Thomas Hobson was a British stable keeper in the mid-1600s whose policy was that you either took the horse nearest the stable door or he wouldn't rent you a horse. "Hobson's choice" has come to mean an illusory alternative. Back up tapes are problematic, but the unacceptable alternative is letting evidence disappear.

1. Completely overwriting the deleted data on magnetic media (*e.g.,* floppy disks, tapes or conventional hard drives) with new information.
2. Strongly encrypting the data and then "losing" the encryption key; or,
3. Physically damaging the media to such an extent that it cannot be read.

Computer forensics is the science that, *inter alia,* resurrects deleted data. Because operating systems turn a blind eye to deleted data (or at least that which has gone beyond the realm of the Recycle Bin), a copy of a drive made by ordinary processes won't retrieve the deleted data. Computer forensic scientists use specialized tools and techniques to copy every sector on a drive, including those holding deleted information. When the stream of data containing each bit on the media (the so-called "bitstream") is duplicated to a sequence of files, it's called a "drive image" or "forensic image." Computer forensic tools analyze and extract data from images.

In routine computer operation, deleted data is overwritten by random re-use of the space it occupies or by system maintenance activities; consequently, the ability to resurrect deleted data with computer forensics erodes over time. *When the potential for discovery from deleted files on personal computers is an issue*, a preservation letter may specify that the computers on which the deleted data reside should be removed from service and shut down or imaged in a forensically sound manner. Such a directive might read:

> **Act to Prevent Spoliation**
> You should take affirmative steps to prevent anyone with access to your data, systems, accounts and archives from seeking to modify, destroy or hide potentially relevant ESI wherever it resides (such as by deleting or overwriting files, using data shredding and erasure applications, re-imaging, damaging or replacing media, encryption, compression, steganography or the like).
>
> **System Sequestration or Forensically Sound Imaging [*When Implicated*]**
> As an appropriate and cost-effective means of preservation, you should remove from service and securely sequester the systems, media, and devices housing potentially relevant ESI of the following persons:
>
> **[NAME KEY PLAYERS MOST DIRECTLY INVOLVED IN CAUSE]**
>
> In the event you deem it impractical to sequester systems, media and devices, we believe that the breadth of preservation required, coupled with the modest number of systems implicated, dictates that forensically sound imaging of the systems, media and devices of those named above is expedient and cost effective. As we anticipate the need for forensic examination of one or more of the systems and the presence of relevant evidence in forensically significant areas of the media, we demand that you employ forensically sound ESI preservation methods. Failure to use such methods poses a significant threat of spoliation and data loss.
>
> "Forensically sound ESI preservation" means duplication of all data stored on the

evidence media while employing a proper chain of custody and using tools and methods that make no changes to the evidence and support authentication of the duplicate as a true and complete bit- for-bit image of the original. The products of forensically sound duplication are called, *inter alia,* "bitstream images" of the evidence media. A forensically sound preservation method guards against changes to metadata evidence and preserves all parts of the electronic evidence, including deleted evidence within "unallocated clusters" and "slack space."

Be advised that a conventional copy or backup of a hard drive does not produce a forensically sound image because it captures only active data files and fails to preserve forensically significant data existing in, e.g., unallocated clusters and slack space.

Further Preservation by Imaging

With respect to the hard drive, thumb drives, phones, tablets and storage devices of each of the persons named below and of each person acting in the capacity or holding the job title named below, demand is made that you immediately obtain, authenticate and preserve forensically sound images of the storage media in any computer system (including portable and personal computers, phones and tablets) used by that person during the period from _____ 20___ to _____, 20___, as well as recording and preserving the system time and date of each such computer.

[NAMES, JOB DESCRIPTIONS OR JOB TITLES]

Once obtained, each such forensically sound image should be labeled to identify the date of acquisition, the person or entity acquiring the image and the system and medium from which it was obtained. Each such image should be preserved without alteration and authenticated by hash value.

METADATA

Metadata, the "data about data" created by computer operating systems and applications, may be critical evidence in your case, and its preservation requires prompt and decisive action. Information stored and transmitted electronically is tracked by the system where it resides and by the applications that create and use it.

For example, a Microsoft Word document is comprised of information you can see (*e.g.,* the text of the document and the data revealed when you look at the document's "Properties" in the File menu), as well as information you don't always see like tracked changes, collaborative comments, revision histories and other data the program only displays on request). This *application* metadata is stored within the document file and moves with the file when it is copied or transmitted. Likewise, the computer system on which the document resides keeps a record of when the file was created, accessed and modified, as well as the size, name and location of the file. This *system* metadata is *not* stored within the document. So, when a file is copied or transmitted—as when it's uploaded or copied to thumb drive for production—potentially relevant and discoverable system

metadata is lost or changed. Absent proper steps to protect metadata, it's constantly at peril of loss or alteration.

Metadata is not a crucial evidence in all matters, but it's always enormously important to culling and managing electronic evidence, and to assessing integrity and authenticity. Metadata proves when a document or record was created, altered, copied or deleted. If you reasonably anticipate that metadata will be important—and that's so often the case—you should specifically direct the other side to preserve relevant metadata evidence and warn of the risks threatening its loss and corruption. Because most lawyers have a spotty appreciation of the variety and utility of system and application metadata, the perfect preservation letter defines metadata and informs your opponent where to find it, the actions that damage it and, if possible, the mechanisms by which it should be preserved. *It pays to be specific.* Although specificity is challenging when we know nothing about an opponent's ESI usage, for most of the information deployed in discovery (e.g., e-mail, texts, documents, spreadsheets and presentations), we *CAN* anticipate the metadata of the most common forms and applications. For example, if you know you will need, say, the *Message ID* and *In-Reply-To* metadata fields to thread e-mail, demand that those fields be preserved.

For further information about metadata, see "*Beyond Data about Data: the Litigators Guide to Metadata,*" *infra* and at http://www.craigball.com/metadata.pdf.

DOES IT REALLY MAKE A DIFFERENCE?

Are you prepared to let relevant evidence disappear without a fight? ***No!***
Can the perfect preservation letter really make *that* much difference? ***Yes!***

The preservation letter demands your best effort for a host of reasons. It's the basis of your opponent's first impression of you and your case. A well-drafted preservation letter speaks volumes about your savvy, focus and preparation. A poorly drafted, scattergun missive suggests a lazy formbook attorney who's given little thought to where the case is going or what evidence is required. A letter that demonstrates close attention to detail and preemptively slams the door on cost-shifting and "innocent" spoliation bespeaks a force to be reckoned with. The artful preservation letter serves as a blueprint for meet and confer sessions and a touchstone for efforts to remedy destruction of evidence.

Strategically, the preservation letter forces your opponent to weigh potential costs and business disruption early, often before a lawsuit. If it triggers a litigation hold, everyone from the board room to the mail room may learn of the claim and be obliged to take immediate action. It may serve as the starting gun for a reckless rush to destroy evidence or trigger a move toward amicable resolution. But done right, ***the one thing it won't be is ignored*.**

APPENDIX: EXEMPLAR PRESERVATION DEMAND TO OPPONENT

What follows *isn't* the perfect preservation letter for *your unique* case, so ***don't deploy it as a form***. Instead, use it as a ***drafting aid*** to flag issues unique to relevant electronic evidence, and tailor your preservation demand proportionately, *scaled to the unique issues, parties, and systems in your case.*

(Download as a Word document **here**)

Demand for Preservation of Electronically Stored Information and Other Evidence

I write as counsel for [Plaintiff(s)] [Defendant(s)] to advise you of [a claim for damages and other relief against you] growing out of the following matters (hereinafter this "cause"):

[DESCRIPTION OF MATTER, INCLUDING ACTORS, EVENTS, DATES, LOCATIONS, CLAIMS/DEFENSES]

We demand that you preserve documents, tangible things, and electronically stored information potentially relevant to the issues and defenses in this cause. As used in this document, "you" and "your" refers to **[NAME OF OPPONENT]**, and its predecessors, successors, parents, subsidiaries, divisions and affiliates, officers, directors, agents, attorneys, accountants, employees, partners Assigns and other persons occupying similar positions or performing similar functions.

You must anticipate that information responsive to discovery resides on your current and former computer systems, phones and tablets, in online repositories and on other storage media and sources (including voice- and video recording systems, Cloud services and social networking accounts).

Electronically stored information (hereinafter "ESI") should be afforded the broadest possible meaning and includes (*by way of example and not as an exclusive list*) potentially relevant information electronically, magnetically, optically, or otherwise stored as and on:

- **Digital communications** (*e.g.*, e-mail, voice mail, text messaging, WhatsApp, SIM cards)
- **E-Mail Servers** (*e.g.,* Microsoft 365, Gmail, and Microsoft Exchange databases)
- **Word processed documents** (*e.g.,* Microsoft Word, Apple Pages or Google Docs files/drafts)
- **Spreadsheets and tables** (*e.g.,* Microsoft Excel, Google Sheets, Apple Numbers)
- **Presentations** (*e.g.,* Microsoft PowerPoint, Apple Keynote, Prezi)
- **Social Networking Sites** (*e.g.,* Facebook, Twitter, Instagram, LinkedIn, Reddit, Slack, TikTok)
- **Online ("Cloud") Repositories** (*e.g.,* Drive, OneDrive, Box, Dropbox, AWS, Azure)
- **Databases** (*e.g.,* Access, Oracle, SQL Server data, SAP)
- **Backup and Archival Files** (*e.g.,* Veritas, Zip, Acronis, Carbonite)
- **Contact and Customer Relationship Management Data** (*e.g.,* Salesforce, Outlook, MS Dynamics)
- **Online Banking, Credit Card, Retail and other Relevant Account Records**
- **Accounting Application Data** (*e.g.,* QuickBooks, NetSuite, Sage)
- **Image and Facsimile Files** (*e.g.*, .PDF, .TIFF, .PNG, .JPG, .GIF., HEIC images)
- **Sound Recordings** (*e.g.,* .WAV and .MP3 files)
- **Video and Animation** (*e.g.,* Security camera footage, .AVI, .MOV, .MP4 files)
- **Calendar, Journaling and Diary Application Data** (*e.g.,* Outlook PST, Google Calendar, blog posts)

- **Project Management Application Data**
- **Internet of Things (IoT) Devices and Apps** *(e.g.*, Amazon Echo/Alexa, Google Home, Fitbit)
- **Computer Aided Design/Drawing Files**
- **Online Access Data** (e.g., Temporary Internet Files, Web cache, Google History, Cookies)
- **Network Access and Server Activity Logs**

ESI resides not only in areas of electronic, magnetic, and optical storage media reasonably accessible to you, but also in areas you may deem *not* reasonably accessible. You are obliged to *preserve* potentially relevant evidence from *both* sources of ESI, even if you do not anticipate *producing* such ESI or intend to claim it is confidential or privileged from disclosure.

The demand that you preserve both accessible and inaccessible ESI is reasonable and necessary. Pursuant to the rules of civil procedure, you must identify all sources of ESI you decline to produce and demonstrate to the court why such sources are not reasonably accessible. For good cause shown, the court may order production of the ESI, even if it is not reasonably accessible. Accordingly, you must preserve ESI that you deem inaccessible so as not to preempt the court's authority.

Preservation Requires Immediate Intervention

You must act immediately to preserve potentially relevant ESI, including, without limitation, information with the *earlier* of a Created or Last Modified date on or after [**DATE**] through the date of this demand and continuing thereafter, concerning:

1. The events and causes of action described [**above**] **[in the Complaint] [in the Answer]**
2. ESI you may use to support claims or defenses in this case
3. ….

Adequate preservation of ESI requires more than simply refraining from efforts to delete, destroy or dispose of such evidence. You must intervene to prevent loss due to routine operations or active deletion by employing proper techniques and protocols to preserve ESI. *Many routine activities serve to irretrievably alter evidence and constitute unlawful spoliation of evidence.*

Preservation requires action.

Nothing in this demand for preservation of ESI should be read to limit or diminish your concurrent common law and statutory obligations to preserve documents, tangible things and other potentially relevant evidence.

Suspension of Routine Destruction

You are directed to immediately initiate a litigation hold for potentially relevant ESI, documents and tangible things and to act diligently and in good faith to secure and audit compliance with such litigation hold. You are further directed to immediately identify and modify or suspend features of your information systems and devices that, in routine operation, operate to cause the loss of potentially relevant ESI. Examples of such features and operations may include:

- Purging the contents of e-mail and messaging repositories by age, quota, or other criteria
- Using data or media wiping, disposal, erasure or encryption utilities or devices
- Overwriting, erasing, destroying, or discarding backup media
- Re-assigning, re-imaging or disposing of systems, servers, devices or media
- Running "cleaner" or other programs effecting wholesale metadata alteration
- Releasing or purging online storage repositories or non-renewal of online accounts
- Using metadata stripper utilities
- Disabling server, packet, or local instant messaging logging
- Executing drive or file defragmentation, encryption, or compression programs

Guard Against Deletion
You should anticipate the potential that your officers, employees, or others may seek to hide, destroy or alter ESI. You must act to prevent and guard against such actions. Especially where company machines were used for Internet access or personal communications, you should anticipate that users may seek to delete or destroy information they regard as personal, confidential, incriminating or embarrassing, and in so doing, they may also delete or destroy potentially relevant ESI. This concern is not unique to you. It's simply conduct that occurs with such regularity that any custodian of ESI and their counsel must anticipate and guard against its occurrence.

Preservation of Backup Media
You are directed to preserve complete backup media sets (including differentials and incremental backups) that may contain unique communications and ESI of the following custodians for all dates during the below-listed intervals:

[CUSTODIAN] [INTERVAL*, e.g.,* 1/1/20 through 7/15/20]

Act to Prevent Spoliation
You should take affirmative steps to prevent anyone with access to your data, systems, accounts and archives from seeking to modify, destroy or hide potentially relevant ESI wherever it resides (such as by deleting or overwriting files, using data shredding and erasure applications, re-imaging, damaging or replacing media, encryption, compression, steganography or the like).

System Sequestration or Forensically Sound Imaging [When Implicated]
As an appropriate and cost-effective means of preservation, you should remove from service and securely sequester the systems, media, and devices housing potentially relevant ESI of the following persons:

[NAME KEY PLAYERS MOST DIRECTLY INVOLVED IN CAUSE]

In the event you deem it impractical to sequester systems, media and devices, we believe that the breadth of preservation required, coupled with the modest number of systems implicated, dictates that forensically sound imaging of the systems, media and devices of those named

above is expedient and cost effective. As we anticipate the need for forensic examination of one or more of the systems and the presence of relevant evidence in forensically significant areas of the media, we demand that you employ forensically sound ESI preservation methods. Failure to use such methods poses a significant threat of spoliation and data loss.

"Forensically sound ESI preservation" means duplication of all data stored on the evidence media while employing a proper chain of custody and using tools and methods that make no changes to the evidence and support authentication of the duplicate as a true and complete bit-for-bit image of the original. The products of forensically sound duplication are called, *inter alia,* "bitstream images" of the evidence media. A forensically sound preservation method guards against changes to metadata evidence and preserves all parts of the electronic evidence, including deleted evidence within "unallocated clusters" and "slack space."

Be advised that a conventional copy or backup of a hard drive does not produce a forensically sound image because it captures only active data files and fails to preserve forensically significant data existing in, e.g., unallocated clusters and slack space.

Further Preservation by Imaging

With respect to the hard drive, thumb drives, phones, tablets and storage devices of each of the persons named below and of each person acting in the capacity or holding the job title named below, demand is made that you immediately obtain, authenticate and preserve forensically sound images of the storage media in any computer system (including portable and personal computers, phones and tablets) used by that person during the period from _____ 20___ to _____, 20___, as well as recording and preserving the system time and date of each such computer.

[NAMES, JOB DESCRIPTIONS OR JOB TITLES]

Once obtained, each such forensically sound image should be labeled to identify the date of acquisition, the person or entity acquiring the image and the system and medium from which it was obtained. Each such image should be preserved without alteration and authenticated by hash value.

Preservation in Native Forms

You should anticipate that ESI, including but not limited to e-mail, documents, spreadsheets, presentations, and databases, will be sought in the form or forms in which it is ordinarily maintained (*i.e.,* native form). Accordingly, you should preserve ESI in such native forms, and you should not employ methods to preserve ESI that remove or degrade the ability to search the ESI by electronic means or that make it difficult or burdensome to access or use the information.

You should additionally refrain from actions that shift ESI from reasonably accessible media and forms to less accessible media and forms if the effect of such actions is to make such ESI not reasonably accessible.

Metadata

You should anticipate the need to disclose and produce system and application metadata and act to preserve it. System metadata is information describing the history and characteristics of other ESI. This information is typically associated with tracking or managing an electronic file and often includes data reflecting a file's name, size, custodian, location and dates of creation and last modification. Application metadata is information automatically included or embedded in electronic files, but which may not be apparent to a user, including deleted content, draft language, commentary, tracked changes, speaker notes, collaboration and distribution data and dates of creation and printing. For electronic mail, metadata includes all header routing data and Base 64 encoded attachment data, in addition to the To, From, Subject, Received Date, CC and BCC header fields.

Metadata may be overwritten or corrupted by careless handling or improper preservation, including by carelessly copying, forwarding, or opening files.

Servers

With respect to servers used to manage e-mail (*e.g.,* Microsoft 365, Microsoft Exchange, Lotus Domino) and network storage (often called a "network share"), the complete contents of each relevant custodian's network share and e-mail account should be preserved. There are several cost-effective ways to preserve the contents of a server without disrupting operations. If you are uncertain whether the preservation method you plan to employ is one that we will deem sufficient, please contact the undersigned.

Home Systems, Laptops, Phones, Tablets, Online Accounts, Messaging Accounts and Other ESI Sources

Though we expect that you will act swiftly to preserve data on office workstations and servers, you should also determine if any home or portable systems or devices may contain potentially relevant data. To the extent that you have sent or received potentially relevant e-mails or created or reviewed potentially relevant documents away from the office, you must preserve the contents of systems, devices and media used for these purposes (including not only potentially relevant data from portable and home computers, but also from external storage drives, thumb drives, CD-R/DVD-R disks and the user's phone, tablet, voice mailbox or other forms of ESI storage.). Similarly, if you used online or browser-based e-mail and messaging accounts or services (such as Gmail, Yahoo Mail, Microsoft 365, Apple Messaging, WhatsApp or the like) to send or receive potentially relevant messages and attachments, the contents of these account mailboxes and messages should be preserved.

Ancillary Preservation

You must preserve documents and other tangible items that may be required to access, interpret or search potentially relevant ESI, including manuals, schema, logs, control sheets, specifications, indices, naming protocols, file lists, network diagrams, flow charts, instruction sheets, data entry forms, abbreviation keys, user ID and password rosters and the like.

You must preserve passwords, keys and other authenticators required to access encrypted files or

run applications, along with the installation disks, user manuals and license keys for applications required to access the ESI.

If needed to access or interpret media on which ESI is stored, you must also preserve cabling, drivers, and hardware. This includes tape drives, readers, DBMS other legacy or proprietary devices and mechanisms.

Paper Preservation of ESI is Inadequate
As hard copies do not preserve electronic searchability or metadata, they are not an adequate substitute for, or cumulative of, electronically stored versions. If information exists in both electronic and paper forms, you should preserve both forms.

Agents, Attorneys and Third Parties
Your preservation obligation extends beyond ESI in your care, possession or custody and includes ESI in the custody of others that is subject to your direction or control. Accordingly, you must notify any current or former agent, attorney, employee, custodian and contractor in possession of potentially relevant ESI to preserve such ESI to the full extent of your obligation to do so, and you must take reasonable steps to secure their compliance.

Preservation Protocols
We are desirous of working with you to agree upon an acceptable protocol for forensically sound preservation and can supply a suitable protocol if you will furnish an inventory and description of the systems and media to be preserved. Alternatively, if you promptly disclose the preservation protocol you intend to employ, we can identify any points of disagreement and resolve them. A successful and compliant ESI preservation effort requires expertise. If you do not currently have such expertise at your disposal, we urge you to engage the services of an expert in electronic evidence and computer forensics so that our experts may work cooperatively to secure a balance between evidence preservation and burden that's fair to both sides and acceptable to the court.

Do Not Delay Preservation
I'm available to discuss reasonable preservation steps; however, *you should not defer preservation steps pending such discussions if ESI may be lost or corrupted because of delay.* Should your failure to preserve potentially relevant evidence result in the corruption, loss, or delay in production of evidence to which we are entitled, such failure would constitute spoliation of evidence, and we will not hesitate to seek sanctions.

Confirmation of Compliance
Please confirm by [**DATE**], that you have taken the steps outlined in this letter to preserve ESI and tangible documents potentially relevant to this action. If you have not undertaken the steps outlined above, or have taken other actions, please describe what you have done to preserve potentially relevant evidence and what you will not do. Else we will rely upon you to complete the preservation sought herein.



Exercise 11: Compiling a Checklist of Sources

GOALS: The goals of this exercise are for the student to:

1. Develop a checklist of potential data sources that are candidates for legal preservation;
2. Explore published ESI checklists; and
3. Identify sources missing from published checklists, improving organization and utility.

In a forthcoming exercise, you will compile a data map of your own digital footprint. The key takeaways from that data mapping exercise are a deeper appreciation of the variety of sources documenting our digital lives and how many of these sources we overlook. Most students are surprised by how little they know about the native forms and data volumes of their own information, particularly that reposed in third parties. Also, student data maps are idiosyncratic, rife with different units, source names and descriptors. In an enterprise setting, data maps for e-discovery must be consistent and complete for all custodians.

So, what might help you prepare to do a more complete, consistent job in identifying sources of evidence? ***How about a well-ordered and complete checklist of sources?***

There are loads of e-discovery checklists extant, geared to assist lawyers and litigation support personnel in the identification and preservation of discoverable information. Some are comprehensive; most are spotty. Older checklists tend to omit important new sources like Slack or Instagram. For example, former President Trump's lawyers "borrowed" the verbatim text of a preservation letter from an article I published more than a dozen years ago without updating it to include later-developed sources. The ironic upshot is that the Twitter President failed to include tweets in his preservation demand. (*See*, https://craigball.net/2018/01/05/the-sincerest-form-of-flattery/).

Assignment: You are to develop the best checklist of potential data sources that are candidates for legal preservation in any engagement—good enough that you'll want to keep it to use in your own work. It does NOT have to be entirely original, and I encourage you to start with several of the best existing checklist(s) you can find and then make updates and improvements. Certainly, you should seek to add missing sources, but you may also see a better way to organize and present the content. Remember that *this checklist is a list of sources and varieties not methods* and, crucially, *it should serve to enhance the completeness of the personal data map you will create (though it should NOT be limited to the scope of your personal digital footprint alone)*.

You will (1) start from several published ESI checklists, (2) improve them for modern reality, and (3) organize the result so it is easy to use in a later data-mapping assignment.

Introduction: This assignment has three deliverables in one Word file. You will assemble a modern, well-organized checklist of ESI sources, document how you improved on published checklists, and show the sources you consulted.

What to submit (one Word file)

- File name: Ex11_ESI_Checklist_LastName.docx (example: Ex11_ESI_Checklist_Smith.docx)
- Include, in order:
    1. **My Checklist**
    2. **What I Improved**
    3. **Sources and Tools Use**d (with "Research Screenshots" at the end)

Time expectation: 60–90 minutes.

1. My Checklist

Purpose: Create a practical checklist of potential ESI sources for individuals and businesses. Focus on sources, not methods.

Minimum coverage:

- At least 30 items total across all categories below.

Use **no fewer** than the following category headings:

A. Devices and local storage
B. Email and messaging
C. Cloud storage and collaboration
D. Social media and web accounts
E. Work/business systems (keep it simple—I know most of you have no personal exposure)
F. Backups and archives

Other categories you may wish to include: IoT/Smart Home; Financial/Transactional; Vehicles/Telematics.

Required structure

Use categories and subcategories (simple outline is fine). Avoid vagaries like "online sources." Each item should be a source, not a method.

Item format (one line per item)-- (do better if you have a better idea)

- Source | Examples | Why it matters

Copy/paste table template for each category:

Source	Examples	Why it matters
[Write item here]	[App/service or "N/A"]	[Plain English, one short clause]

Example item (format and level of detail to emulate):
Table

Source	Examples	Why it matters
Text messages on phones	iMessage; Android Messages	Often the most candid, time-stamped communications.

Formatting tips:

- Keep each "Why it matters" to one line in plain English.
- Use product/app names sparingly—just enough to jog memory.
- Do not list collection methods or tools here.

2. What I Improved (brief change log)

Write 8–10 bullets comparing your checklist to the one(s) you started from. For each bullet, briefly state:

- What you added, removed, renamed, or reorganized; and
- Why (e.g., outdated, missing, redundant, clearer grouping).

Example bullets:

- Added "Workplace chat (Slack/Teams)" because older lists omitted it and it's now a primary communications channel.
- Split "Cloud storage" into "Personal" and "Business" to reflect different permissions and version histories.

3. Sources and Tools Used

A. Sources (required)

- List the checklists/references you relied upon (title + source). If you borrowed language closely, say so in a sentence.

B. Research Screenshots (required)

- Number of screenshots: At least three (3).
- Distinct sources: Each from a different non-AI source.
- Each screenshot must show: (1) the source/page title and visible URL or document title; and (2) some relevant content (not just a homepage).
- Placement: Insert under a sub-heading titled "Research Screenshots" at the end of your Word file.
- Integrity rule: *Do not submit AI chat screenshots as sources.*

C. AI/LLM use (allowed but limited; disclosure required) Answer these three questions briefly:

1. What did you ask the LLM to do (1–3 sentences)?
2. What parts of your submission are meaningfully yours (1–3 sentences)?
3. Highlight two items you personally added or changed after reviewing LLM output.

Limits on LLM use to keep this your work

• You may not submit an LLM-generated checklist without meaningful editing.
• You must make at least 8 change-log entries that reflect your own judgment.
• You must include at least 10 items that are **not copied verbatim** from **any single source** (human or AI). That means your final checklist must contain at least 10 source entries that are written in *your own words*, not pasted word-for-word from anywhere else—whether that "anywhere else" is a website checklist, a form book or an AI output.

Practically:

- If you start from a published checklist and copy 25 items exactly as written, that's allowed only if you cite it—but you still must add or rewrite at least 10 entries so the checklist reflects your own judgment and phrasing.
- "Not copied verbatim" can be satisfied by (a) adding genuinely new items the source didn't include, or (b) rewriting items in your own words (with the same idea) rather than copying the exact wording.
- "Any single source" is there to prevent "I copied everything from Source A." You can borrow from multiple sources with citations, but you must still contribute meaningful original work.
- If your "Sources and Tools Used" section suggests you used an LLM to generate the whole deliverable with minimal changes, you should expect a low score.

How it will be graded

- **Completeness**: Covers common, often-forgotten sources.
- **Organization**: Logical, easy to skim; no repetition across categories.
- **Clarity**: Specific, plain-English items; concise "Why it matters."
- **Improvements:** Clear, justified changes from your starting checklist(s).
- **Integrity**: Proper credit; screenshots and AI disclosures included.

Final submission checklist

- One Word file named Ex11_ESI_Checklist_LastName.docx
- Section order: My Checklist → What I Improved → Sources and Tools Used
- My Checklist has at least 30 items across required categories A–F
- Every item uses the three-column format (Source | Examples | Why it matters)
- What I Improved: 8–10 bullets explaining changes and why

- Sources listed; borrowed language acknowledged
- At least 3 screenshots from 3 different non-AI sources, with visible titles/URLs and one-sentence captions, placed under “Research Screenshots”
- AI/LLM disclosure: all three questions answered
- Plain English; concise; no methods listed as items

Reminders

• This is a checklist of sources, not a preservation plan.
• Borrowing is permitted; plagiarism is not. Credit what you use.
• Technology changes quickly. Your job is to notice what older checklists miss and fix it.

Computer Backup Systems

Backup is the spare tire in the trunk of the e-discovery world. Nobody talks about it, nobody maintains it…and everyone suddenly becomes a believer at 11:30 p.m. on the shoulder of I-35. Backup is pretty much ignored until headquarters burns to the ground and it turns out the old tapes in the basement hold the only copy of the all-important TPS reports the judge ordered be produced.

Would you be surprised to learn that backup has quietly reinvented itself while most lawyers were arguing about search terms? Consider the:

Backup is no longer an IT backwater. It has quietly become one of the fastest-evolving areas of information technology, driven by

- Migration of data to the "cloud" *(Minsk! Why's our data in Minsk?);*
- Explosive growth in hard drive capacities *(Ten terabytes! On a desktop!?!);*
- Ascendency of virtual machines *(Not another Terminator movie?);* and
- Increased reliance on replication *(D2D2T? That's the cute Star Wars droid, right?).*

If you don't understand how backup systems work, you can't reliably assess whether discoverable data exists—or what it will cost in sweat and coin to find it.

Vital Vocabulary

Look for these key terms:

- *disaster recovery*
- *full backup*
- *differential backup*
- *incremental backup*
- *tape restoration*
- *tape rotation*
- *legacy tapes*
- *replication*
- *drive imaging*
- *bitstream*
- *backup set*
- *backup catalog*
- *tape log*
- *linear serpentine*
- *virtual tape library*
- *D2D2T*
- *RAID*
- *striping*
- *parity*
- *hash value*
- *single-instance storage*
- *non-native restoration*
- *Cloud backup*

The Good and Bad of Backups

Ideally, the contents of a backup system would be entirely cumulative of the active "online" data on the servers, workstations, and laptops that make up a network. But businesses entrust the power to alter and destroy data to every computer user—including those motivated to make evidence disappear—and configure systems to purge ESI pursuant to records-retention policies. As a result, backups may be the only surviving evidence beyond the reach of neglect, spoliation, or convenience.

History offers no shortage of cases where backups supplied the "smoking gun" after active data vanished. Backup media can also invite abuse. They are a tempting target for fishing expeditions launched without regard to cost or burden, or demanded prematurely before readily accessible

sources have been exhausted. That tension—between evidentiary value and disproportionality—defines modern backup discovery disputes.

Grappling with Backup Tapes

Backup tapes were designed for **disaster recovery**: rebuilding damaged or corrupted systems. They are often described as "snapshots," and like photographs, they capture only what is in focus at a particular moment. To conserve time and space, backups typically exclude commercial software that can be reinstalled, concentrating instead on user-created data. **Full backups** capture everything within scope; **incremental** and **differential backups** capture only what has changed. Together, they reassemble Humpty-Dumpty in a process called **restoration**.

Tape remains cheap, durable, and portable—qualities that made it indispensable for decades and still relevant for long-term archival storage. But tape is also slow and mechanically cumbersome, limitations long tolerated because restoration was expected to be rare.

Because backup systems exist to get a business "back up" on its feet after disaster, they only need to retain data for a short recovery window. No enterprise wants to restore its systems as they existed months—or even weeks—before a failure. In theory, older tapes are recycled through overwriting, a process known as **rotation**.

In practice, organizations often retain tapes far longer than planned, sometimes beyond their ability to read them. These legacy tapes are business records—sometimes the last surviving copy of historical data—but they are rarely managed as such. Old media is boxed, shelved, and forgotten in offsite storage or a corner of the data center. Like the DeLorean in *Back to the Future*, legacy tapes allow time travel. It just costs money instead of plutonium.

Duplication, Replication and Backup

We protect data from loss through three related but distinct mechanisms: duplication, replication, and backup.

Duplication is the simplest: making a copy of a file. Copies on the same medium protect against user error; copies on different media protect against hardware failure; copies stored offsite protect against catastrophe.

You might reasonably ask whether backup is just duplication by another name. In part, it is. But enterprise backup adds features that distinguish it from casual copying, most importantly that backups occur automatically and systematically, not at the discretion of individual users.

Replication is duplication without discretion. Data is mirrored periodically or continuously from one system to another. Replication may be as simple as RAID mirroring or as complex as maintaining a geographically remote failover environment.

Unlike duplication and replication, **backup** deliberately transforms data. Backup systems compress, encrypt, segment, and catalog information. They track versions, manage schedules, and record where every piece of backed-up data resides. That metadata—the catalog—is often more important for discovery than the media itself.

Major Elements of Backup Systems

Understanding backups requires grasping three elements: **source data**, **backup sets**, and **catalogs**.

1. **Source Data (Logical or Physical)**

Users think in files and folders, but backups may operate at broader logical levels—volumes, partitions, or entire systems. Drive imaging, used in forensics and some enterprise backups, captures a bit-for-bit representation of storage media. These images are usually segmented into manageable chunks to facilitate storage and handling.

2. **Backup Sets (Full and Changed-File)**

A "backup set" may refer to the physical media holding the data or to the logical collection of files and metadata comprising a backup event. Sets may be full or incremental. Full backups capture everything within scope; changed-file backups capture only additions or modifications.

The changed-file components further subdivide into **incremental backups**, **differential backups** and **delta block-level backups**. The first two identify changed files based on either the status of a file's archive bit or a file's created and modified date values. The essential difference is that every differential backup duplicates files added or changed since the last *full* backup, where incremental backups duplicate files added or changed since the last *incremental* backup. Block-level methods store only the portions of files that have changed, producing smaller backups and faster operations at the cost of greater system sophistication.

3. **Backup Catalog vs. Tape Log**

Unlike duplication and replication—where little or no record is kept—the catalog is the beating heart of enterprise backup. The catalog records what was backed up, when, from where, and where each component resides within the backup set. Without it, restoration is guesswork.

The catalog tracks file identity, metadata, source system, backup date, and physical or logical location. It enables targeted restoration and makes selective extraction possible without rebuilding entire systems.

Just as importantly, catalogs enable **single-instance storage**. Identical files—common across users and systems—are stored once, with pointers used to reconstruct them at restore time. This dramatically reduces storage requirements and restoration volume.

Lose the catalog, and Humpty Dumpty stays broken.

It's critical to distinguish the catalog from a **tape log**. The latter is usually a skeletal record listing dates, systems, and media identifiers. A tape log may tell you *that* a backup occurred; only the catalog tells you *what's inside*.



Backup Media: Tape and Disk-to-Disk

Tape Backup

Though backup tape may seem antiquated, it has evolved alongside modern computing. Early enterprise tape systems measured capacity in megabytes and speed in seconds per megabyte—figures that now seem almost comical.



Today's **LTO-9** cartridges—about four inches square and less than an inch thick—hold **18 terabytes natively (45 TB compressed)** and sustain transfer rates approaching **400 MB/sec**. That's orders of magnitude more capacity and speed in a fraction of the physical footprint.

Many modern tape systems employ **linear serpentine recording**, a back-and-forth scheme in which data is written in parallel tracks running the length of the tape. Multiple tracks are read or written simultaneously as the tape shuttles past the heads, requiring hundreds of passes to read or write a full cartridge. The result is impressive density, but at the cost of mechanical complexity and latency.

Alternate recording schemes—such as the **helical scan** used in SAIT systems—once offered competitive density by writing diagonal tracks across the tape, much like an old VCR. Development of SAIT ceased more than a decade ago, and such formats are now encountered chiefly in legacy archives.

Why Tape Is Still Slow

Tape's limitations are not technological so much as physical. All that serpentine shuttling back and forth over miles of tape is a mechanical process. It occurs at a glacial pace relative to the speed with which computer circuits move data. That is, electrons move faster than atoms, and even the brightest legal minds have yet to find loopholes in those pesky laws of physics.

Tape restoration remains constrained by the need to physically move media past read/write heads, often repeatedly, and across multiple cartridges. Restoration is frequently cumulative, requiring data from full and incremental backups to be combined before anything useful emerges.

As storage volumes continue to grow, gains in tape speed are largely offset by the sheer quantity of data that must be read end-to-end.

How Long Does Restoration Take?

Years ago, the defunct tape restoration firm **eMag Solutions, LLC** compared theoretical tape restore times to real-world experience, identifying variables that materially affect performance, including:

- tape format and drive generation;
- device interface and firmware;
- compression and block size;
- file count and data structure; and
- tape condition and retry rates.

Their findings were blunt: **real-world restoration typically takes about twice as long** as raw capacity and transfer rates would suggest. Even generating a catalog requires reading the entire tape. Accordingly, delivering thousands of tapes to a vendor does not translate into rapid results, regardless of budget.

While the specific drive models tested are now obsolete, the proportionality lesson remains valid. A practical rule of thumb for estimating real-world restoration time is:

Restore Time (hours) ≈ Native Capacity (GB) ÷ (1.8 × Native Transfer Rate in MB/sec)
Applied to a modern **LTO-9** cartridge (18 TB at ~400 MB/sec), a single tape requires on the order of **7 hours** of sustained read time—before accounting for tape changes, verification, cataloging, error handling, or human labor.
At scale, these variables compound. Multiple drives can shorten elapsed time, but operational realities—media handling, staffing, quality control, balky tapes, weekends, and holidays—remain stubbornly resistant to Moore's Law.

Common Tape Formats
The LTO tape format is the clear winner of the tape-format wars, having eclipsed all contenders apart from disk and cloud storage, which now dominate active backup workflows.

As of 2026, **LTO-9** is the current enterprise standard, holding **18 TB natively (45 TB compressed)** with sustained transfer rates approaching **400 MB/sec**. Roadmaps for **LTO-10 and beyond** promise further gains, but tape's trajectory is no longer exponential in practical e-discovery terms.

Tape use is down, but not out—not for some time—because it remains **cheap, durable, portable, and entrenched** in enterprise archives.

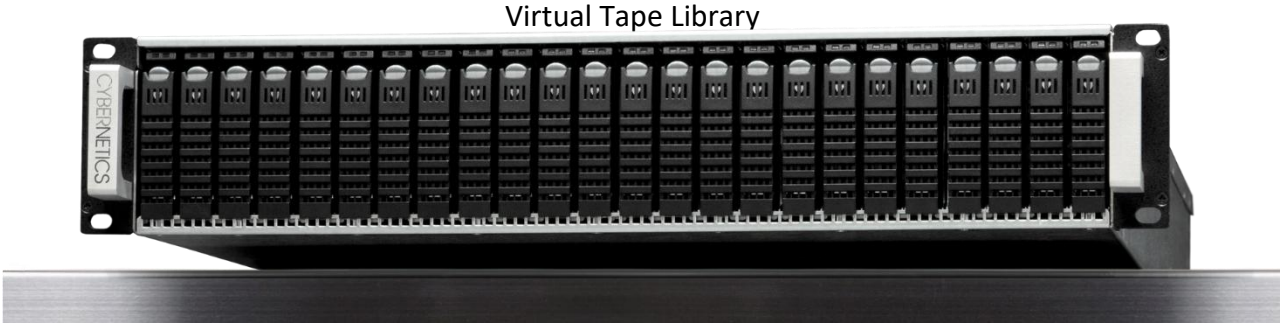
That entrenchment matters. The dusty catacombs beneath Iron Mountain still brim with legacy tape formats that continue to surface in e-discovery disputes, especially where litigation reaches back a decade or more. Counsel need not master every extinct format, but they must recognize that **old tapes persist long after the systems that created them are gone**.

Disk, Tape, and the Vanishing Meaning of "Backup" (Consolidated)
Tapes are stable, cheap, and portable—historically making them a natural medium for moving large data volumes without consuming bandwidth or disrupting operations. But advances in deduplication, compression, and inexpensive high-capacity disks have eroded many of tape's former advantages. When data is deduplicated to unique content and further compressed, vastly more information fits into far less space. With today's drive capacities, disk-based backup intervals often rival traditional tape rotation cycles, making disk-to-disk backup routine rather than exotic.

Hard-disk arrays can now retain weeks or months of disaster-recovery data at costs competitive with tape. As a result, tape has been displaced from primary recovery and relegated largely to **l**ong-term archival storage. In practice, most enterprises rely on disk or cloud-based systems for rapid recovery, reserving tape for deep history.

The death of tape has been predicted for decades. For the first time, however, the demand curve has unmistakably bent downward. Disk-to-disk backup initially arrived wearing tape's sheep's clothing: early disk arrays emulated tape drives so existing software and workflows need not change. These **virtual tape libraries (VTLs)** eased the transition but obscured the real shift—backup was becoming faster, more searchable, and less mechanical.

Virtual Tape Library



As disk supplants tape for recovery, the need remains for **a cheap, offline, and portable medium** for long-term retention—the digital attic of the enterprise. Tape continues to fill that role, giving rise to **D2D2T (Disk-to-Disk-to-Tape)** architectures, in which disk handles recovery and tape handles history. By design, tape now holds archives, not operational lifelines, ensuring its continued—but narrowing—relevance to e-discovery.

What lawyers still call "backup," however, increasingly no longer looks like tape—or even like backup at all. Modern systems rely heavily on **snapshots, replication, and object storage**, blurring the once-bright line between active data and backup. A snapshot is a near-instantaneous, point-in-time view preserved through copy-on-write technology. Snapshots are fast, cheap, and frequent, and because they often persist for weeks or months, they are frequently more accessible sources of historical ESI than legacy tape.

At the same time, enterprises increasingly store data in **object-based repositories** rather than traditional file systems. In object storage (such as Amazon S3, Azure Blob, or Google Cloud Storage), data is addressed by unique identifiers and rich metadata, not file paths. These systems are inherently redundant, geographically dispersed, and version-aware. From an e-discovery standpoint, they behave less like backups and more like continuously available archives.

Software-as-a-Service platforms push this evolution further. In Microsoft 365, Google Workspace, Slack, Salesforce, and similar ecosystems, backup is no longer something the customer schedules. It happens continuously, invisibly, and under the provider's control. What survives, for how long, and in what form is governed by **retention policies, licensing tiers, and provider architecture—not tape rotation charts**.

The practical consequence for discovery is straightforward but often overlooked:
• "Backup" data may be as accessible as active data
• Historical versions may exist without any restoration step
• The most probative evidence may reside in **snapshots, version histories, or SaaS retention stores**, not on tape

When lawyers reflexively argue that "backups are not reasonably accessible," they usually mean tape. Increasingly, that assumption is wrong.

Compression
The design of backup systems is driven by considerations of speed and cost. Perhaps surprisingly, the speed and expense with which a system can be brought back online after failure is often less critical than the speed and cost of each backup. The reason is simple: failure is (one hopes) rare, while backup is routine. No one seriously contends that restoring a failed system from piles of magnetic tape is the fastest or cheapest way to recover. Tape's advantage has always been its low cost per gigabyte to store data, not to restore it.

Again, electrons move far faster than atoms; that is, all those atoms that make up machines. The slowest parts of any backup system are mechanical: spinning reels, moving heads, and human beings swapping media. One way to wring efficiency from such systems is to increase **areal data density**, storing more information per inch of tape or disk so fewer volumes must be written, handled, and read.

Because computing operations occur orders of magnitude faster than mechanical motion, backup systems devote processing power to **r**estructuring data into more compact representations before it is written to storage. This is the logic behind compression: re-expressing information so it consumes less space and moves faster, at the cost of modest computational effort.

Recalling our discussion in the processing chapters, compression is an essential, ubiquitous technology. Without it, there would be no streaming video, digital photography, cloud storage, or economical backup. And without compression, backing up modern enterprise systems would require far more time, media, and money than anyone would tolerate.

While file-level compression relies on relatively standardized formats (e.g., ZIP), **backup compression is often proprietary**, tied to particular software or hardware versions. As a result, restoring legacy backup tapes may require not only compatible drives and catalogs, but specific software capable of reversing the original compression—a complication that regularly surprises litigants who assume data can be read once media is mounted.

Deduplication

Enterprises that retain backup data for months or years accumulate enormous redundancy. Each full backup captures a snapshot of largely unchanged systems, so identical files recur again and again. An email that sits in a user's mailbox for a year may appear in dozens of backups. Restoring those backups without deduplication would multiply identical messages, inflate review volumes, and invite inconsistent treatment of the same evidence.

The overlap between successive backups routinely exceeds **80–90%**. Add to that the repetition created by enterprise-wide emails, shared documents, and replicated attachments, and the scale of duplication becomes staggering.

Deduplication is the process of identifying and eliminating identical content before review. Deduplicating within a single custodian's data is known as **vertical deduplication** and is straightforward. But backup media typically contain data for many custodians, producing duplication both within accounts and across users. Eliminating identical content across custodians is called **horizontal deduplication**.

Horizontal deduplication dramatically reduces review cost and minimizes inconsistent privilege or relevance determinations. But it comes with a tradeoff: once duplicates are collapsed, it may be harder to reconstruct a complete picture of what resided in any single custodian's collection at a given moment in time.

Accordingly, deduplication serves two critical functions in e-discovery involving backup data. First, it eliminates massive repetition between backup iterations. Second, it reduces review burden by collapsing identical material within and across custodians—*so long as metadata or reporting preserves custodial attribution.*

Modern backup systems typically deduplicate before storage, using single-instance storage. Each file or data block is assigned a unique hash value as it is ingested. If an identical block already exists, the system stores only a reference rather than another copy. This may occu**r in-line,** as data is written, or **post-process**, after storage. Either way, what looks like a mountain of backup media often resolves into far less unique information than its raw size suggests.

Data Restoration



Data preserved in backup systems is a bit like furniture from Ikea: it has been taken apart, compacted, and packed for storage, not left ready for immediate use. When backup data is needed for e-discovery, it must be reconstituted and reassembled, often expanding dramatically in size. That restored data has to go somewhere—traditionally, into a computing environment resembling the one from which it came.

Historically, restoration required recreating the native hardware, operating systems, applications, and configurations used when the backup was made. For small and mid-size organizations, this meant acquiring idle servers, storage, and software licenses solely to perform restoration. Even large enterprises hesitated to repurpose disaster-recovery infrastructure for discovery, fearing it might compromise readiness in the event of a real emergency.

These burdens helped cement the notion—embraced by courts and counsel alike—that backup media were **"not reasonably accessible."** But that presumption has not aged well.

Non-Native Restoration

A major inflection point in backup accessibility has been the rise of **non-native restoration**. Modern tools allow vendors (and sometimes in-house teams) to restore data **without recreating the original computing environment** or locating obsolete backup software. Instead, data can be extracted directly from backup sets and delivered into contemporary review platforms.

Equally important is the ability to **index backup contents and selectively extract data**—by custodian, date range, or file type—without restoring entire systems. In many cases, it is now feasible to recover only targeted mailboxes or document repositories rather than terabytes of irrelevant material.

As a result, backups are no longer uniformly burdensome. In some matters, they are no harder—and occasionally easier—to mine for evidence than active systems.

The Fading Presumption of Inaccessibility

Lawyers still reflexively assert that "backup data is not reasonably accessible," often meaning tape. Increasingly, that shorthand is inaccurate. The accessibility of backup data now depends less on media type and more on architecture, tooling, and scope.

Where backup data can be indexed, deduplicated, selectively restored, and loaded directly into review systems, the traditional cost and burden arguments may carry little weight. Conversely,

where legacy media, missing catalogs, or proprietary compression remain obstacles, inaccessibility arguments may still be well-founded.

The point is not that backups are always accessible—but that ***they are no longer categorically inaccessible***.

When Backups Are the Smart First Stop

Consider a dispute turning on a narrow window of email communications, where custodians have departed or local mailboxes are incomplete. A targeted restoration of a server-level backup from the relevant period may be **faster, cheaper, and more reliable** than forensic collection of endpoints or enterprise-wide searches of active systems.

In the right case, backups are not a last resort. They are the most efficient path to the truth.

Sampling

Sampling backup data involves restoring **d**iscrete snapshots in time, not arbitrary numbers of tapes. Because full restoration may require multiple media (full plus incremental or differential backups), sensible sampling depends on access to—and understanding of—the backup catalog.

Sampling is like drilling for oil. You identify promising intervals, test them, and proceed only if results justify further effort. Courts increasingly expect parties to **sample before they speculate**, particularly where backup data may be outcome-determinative.

Backup and the Cloud

Nowhere is the adage that *"the Cloud changes everything"* more apt than when applied to backup. Microsoft, Amazon, Google, and a host of others have made it practical—and cheap—to abandon local backups in favor of storing data securely in leased cloud repositories. The cost per gigabyte is now measured in **pennies** and continues to fall as scale explodes.

Adoption of cloud computing among enterprises is no longer aspirational; it is dominant. Barring catastrophic missteps, that migration will continue inexorably as bandwidth improves and security concerns fade into the background noise of routine risk management.

But the cloud's most consequential impact is not that it offers *another place* to put backups. It is that, for many systems, it **eliminates the very notion of user-managed backup**. As data and infrastructure move into SaaS and cloud-native platforms, backup ceases to be something customers schedule or even see. It becomes a **perpetual, invisible obligation of the provider**, implemented through redundancy, replication, snapshots, and geographically distributed data centers. The customer no longer "backs up" data; the provider ensures persistence.

How a cloud provider fulfills that obligation—whether through disk, tape, or other media—is largely irrelevant to the customer so long as availability, integrity, and retention expectations are met. From the user's perspective, backup fades from view.

Welcome to the Future

For much of the digital era, *backup* meant distance. Data was written to tape, boxed up, and sent away. That physical and technical remoteness shaped how lawyers and courts came to think about backup data in discovery: expensive to restore, difficult to search, and often "not reasonably accessible."

In 2009, Harvard Law professor Lawrence Lessig observed, *"We are not going back to the twentieth century. In a decade, most Americans will not even remember what that century was like."* That decade has passed. Yet many of the assumptions lawyers still carry about backup systems come straight from that earlier century—when recovery meant forklifts, offsite vaults, and weeks of restoration. Modern backup rarely looks like that.

Today, what lawyers call "backup" may take the form of snapshots, replicated disks, version histories, or cloud retention layers. These systems often preserve prior versions automatically and continuously, sometimes without any formal "backup event" at all. In many environments, historical data is no farther away than an administrative console or an upgrade in licensing.

This shift matters because **Rule 26 does not care what data is called**. The proportionality analysis under Rule 26(b)(1) asks whether the burden or expense of obtaining the information outweighs its likely benefit. It does not ask whether the data resides on tape, disk, or in the cloud. Labels do not decide accessibility—architecture does.

As a result, the familiar refrain that "backups are not reasonably accessible" must now be treated as a hypothesis, not a conclusion. Sometimes it will still be true—particularly with legacy tape. Increasingly, it will not. In many cases, the most accessible historical evidence will come not from active systems, but from the quiet memory layers beneath them.

Backup has not vanished from discovery. It has changed roles. Tape now preserves history rather than continuity. Disk and cloud systems quietly remember what users forget. The challenge for lawyers is no longer restoring data from oblivion, but recognizing when it was never truly gone.

The practical lesson is simple but essential: **effective discovery begins with understanding how information is actually stored, retained, and versioned—not with assumptions inherited from obsolete technology**. That is as true for backup systems as it is for every other source of electronically stored information.

TEN PRACTICE TIPS FOR BACKUPS IN CIVIL DISCOVERY (2026)

1. **Backup ≠ Inaccessible.**
   Do not assume backup data falls outside discovery. Rule 26(b)(2)(B) requires a party claiming inaccessibility to *identify* the sources and *prove* undue burden or cost. Assertions are not evidence. Be prepared to support your position with metrics, sampling results, and competent testimony.
2. **Understand how your client actually uses backup—not how the policy says it's used.**
   Determine whether your client:
   • routinely restores backup data (e.g., user error, ransomware response, audits);
   • treats backups as *de facto* archives or historical repositories;
   • has in-house capability (or vendor support) to restore or extract data; and
   • can search or index backup content in a meaningful way.
   If backups are routinely accessed, they are unlikely to be deemed "not reasonably accessible."
3. **Preserve context, not just media.**
   Backup data does not exist in isolation. Preserve catalogs, indices, manifests, retention schedules, and configuration data along with any media. Without them, you may be forced to restore data merely to learn what it contains—a costly and unnecessary exercise.
4. **Sampling beats speculation.**
   Courts increasingly expect parties to test assumptions. Metrics gleaned from a reasonable sampling protocol is more persuasive than blanket claims of burden—and cheaper than wholesale restoration.
5. **Test for uniqueness and materiality early.**
   Backups frequently duplicate data available elsewhere. Sampling that shows minimal unique or probative content is powerful evidence against full restoration. Like drilling for oil, a few dry holes can justify moving on.
6. **Show that more accessible sources exist—and were pursued.**
   Judges are reluctant to force restoration when responsive ESI is available from active systems, snapshots, SaaS version histories, or user devices. Sampling results and expert testimony help here.
7. **Know what backup tools can't do.**
   Most backup platforms were not designed for e-discovery. Ask whether they can:
   • search inside common file types;
   • handle compressed or container formats;
   • preserve metadata accurately; and
   • support defensible exports.
   If not, restoration may still be required—but that limitation must be explained, not assumed.
8. **Expect scrutiny of backup hygiene.**
   If a court examined your client's backup environment—physical or virtual—would it reflect intentional governance or digital hoarding? Disorder undermines credibility.
9. **When data lives in the cloud, consider reviewing it there.**
   With cloud-based backup and SaaS platforms, it may be faster, cheaper, and more defensible to deploy review tools directly against cloud-resident data than to extract it wholesale and process it.
10. **Align retention with reality.**
    Backup is for recovery, not archaeology. If data is too old to restore systems, ask why it is still retained? Indefinite retention without purpose increases cost, risk, and discovery exposure.

Databases in E-Discovery

Years ago, when I set out to write on databases in e-discovery, I went to the literature to learn prevailing thoughts and ensure I wasn't treading old ground. What I found shocked me.

I found next to no literature on the topic! What little authority existed made brief mention of flat file, relational, and enterprise databases, noted that discovery from databases is challenging, and then fled to other topics.[73] A few commentators mentioned *In re Ford Motor Co,*[74] the brief 2003 decision reversing a trial court's order allowing a plaintiff to root around in Ford's databases with nary a restraint. Although the Eleventh Circuit cancelled that fishing expedition, it left the door open for access to an opponent's databases on different facts—such as where the producing party fails to meet its discovery obligations.

The constant counsel offered by any article touching on databases in e-discovery is "get help." That's good advice, but not always feasible or affordable.

Because databases run the world, we can't avoid them in e-discovery. We must know *enough* about how they work to deal with them when the case budget or time constraints make hiring an expert impossible. We need to know how to *identify and preserve databases*, and we must learn how to *gather sufficient information about them* to frame and respond to discovery about databases.

Databases run the world

You can't surf the 'net, place a phone call, swipe a badge, tap-to-pay, use an ATM, book a flight, refill a prescription, unlock a hotel room, or get admitted to an emergency room without a database making it happen.

Databases touch our lives all day, every day. Our operating systems and e-mail applications are databases. Your spell checker is a database. Search engines are databases. Legal research platforms are databases. Social media, retail sites, streaming services, ride-shares, and delivery apps are all big honkin' databases—often spread across many machines, many locations, and more than one vendor.

Yet, when it comes to e-discovery, we tend to fix our attention on documents, without appreciating that most electronic evidence exists only as a flash mob of information assembled and organized on the fly from a dozen or thousand or million discrete places. In our zeal to lay hands on documents

---

[73] Happily, after I first published, others waded in and produced practical scholarship. Here are links to two:
Requests for Production of Databases: Documents v. Data, by Christine Webber and Jeff Kerr
The Sedona Conference Database Principles Addressing the Preservation & Production of Databases & Database Information in Civil Litigation

[74] 345 F.3d 1315 (11th Cir. 2003)

instead of data, we make discovery harder, slower, and costlier. Understanding databases—and acquiring the skills to peruse and use their contents—gets us to the evidence better, faster, and cheaper.

Databases are also changing the way we think about discovery. Historically, parties weren't obliged to create documents for production; you produced what you had on file. Today, many "documents" don't exist until you generate them. Tickets, statements, dashboards, call-detail records, website pages, inventories, price lists, register receipts, and system logs are often just ad hoc reports generated by databases. They don't take tangible form until you render them—on screen, as a PDF, or as an export—and only a few thousandths to a few hundredths of one percent of the world's newly created digital information will emerge as ink on paper. Litigants must be adept at crafting queries to elicit responsive data, and mastering ways to interpret and use the data stream that emerges—because that stream is often the evidence.

Introduction to Databases

Most of us use databases with no clue how they work. Take e-mail, for example. Whether you know it or not, each e-mail message you view in Outlook, Gmail, or through a web browser is a report generated by a database query—an assemblage of fields culled from a complex dataset—presented in a user-friendly arrangement determined by the client's capabilities and your settings.

That an e-mail message is not a single, discrete "document" is confusing to some. The data segments or "fields" that make up an e-mail are formatted with such consistency from application-to-application (and look so similar when printed) that we mistake e-mail for fixed paper. But each message is better understood as a customizable report from the database that stores your mail—often in the cloud and in multiple tables and indexes you will never see.

When you see a screen or report from a database, you experience something that *feels* like a document. But the data that comes together to create what you see may be drawn from different tables, different indexes, different storage tiers, and sometimes different systems entirely—changing moment to moment. The "same" report run twice can differ because the underlying data moved, was updated, or was filtered differently.

Understanding databases begins with mastering a few simple concepts and a little specialized terminology. Beyond that, the distinction between your e-mail database and a global platform's is mostly differences in scale, optimization, and security.

Constructing a Simple Database

If you needed a way to keep track of the cases on your docket, you'd probably begin with a simple table of columns and rows on a legal pad. You'd list clients by name. Then you might add the other

parties, case number, court, judge, and trial date. If you still had room, you'd add addresses, phone numbers, settlement demands, insurers, policy numbers, opposing counsel, and so on.

In database parlance, you've constructed a "**table**." Each separate information item you entered (e.g., name, address, court) is a "**field**." The group of items you assembled for each client (organized in columns and arranged in a row) is a "**record.**" Because the client's name is the field that governs the contents of each record, it would be termed the "**key field**."

Pretty soon, your table would be unwieldy and push beyond the confines of a sheet of paper. If you added a new matter and wanted to keep the list alphabetized, you'd constantly be rewriting the list.

So, you might turn to index cards. Now each card is a "record" listing the "fields" pertinent to each client. It's easy to add cards and re-order them. Then sometimes you'd want to order matters by trial date or court. To do that, you'd either extract bits from each card to compile a report, re-sort the cards, or maintain multiple sets of cards—one by name, one by trial date, another by court.

Those decks of cards comprise a database of tables. They are still deemed "tables" even though you used a card per record instead of a row per record. Each deck—organized by a chosen key—is a "**flat file database,**" distinguished by the characteristic that all the fields and records live in a single file, with no relationships or links between records except the table structure (the order of the deck and the order of fields on the cards).

Of course, you'd need to keep all cards up-to-date as phone numbers and addresses change. When a client has more than one matter, you'd write the same client data on multiple cards and update each card, one-by-one, trying not to overlook any card. What a pain.

So, you'd automate, turning first to something like a spreadsheet. Now you're not limited by paper. Add a new case anywhere and re-sort by name, court, or trial date. Search electronically. Filter. Report. Still, your spreadsheet is just a table in a flat file database. If the same address appears in ten rows, you must update it in ten places (though "find and replace" makes this less awful), and when you want to extract or transform information you generally work with the entire table.

What you need is a system that allows a change to one item to update every place that information appears—across all tables in the database. You need a system that identifies the *relationship* between common fields of data, updates them when needed, and uses that common relationship to pull related information together. Think of it as adding rudimentary intelligence to a database, allowing it to "recognize" that records sharing common identifiers relate to common entities.

Databases that do this are called "**relational databases**," and they account for most of the databases used in business, ranging from simple tools like Microsoft Access or QuickBooks to enormous enterprise applications marketed by Oracle and SAP.[75]

To be precise, only the tables of data are the "database," and the software used to create, maintain, secure, and interrogate those tables is the **Database Management System** or **DBMS**. In practice, the two terms are used interchangeably.

Relational databases
(Your sample tables work fine pedagogically. The only update I'd suggest is to avoid anchoring the trial dates to a bygone decade unless you want the reader constantly reminded that time is cruel. You can keep the tech-titan names if you like the flavor; just don't let the dates signal "museum.")
As we look at these tables, note that each has a unique key field called the "primary key" for that table. [Verify footnote: 75F.] The essential characteristic of a primary key is that it cannot repeat within the table for which it serves as primary key, and a properly designed database will prevent duplicate primary keys.
Many databases simply assign a unique primary key to each row—either a number or a non-recurring value. For example, a key might be an automatically incrementing integer, a UUID, or a composite identifier. The primary key supports indexing, making it efficient to search, sort, link, and perform other operations on the data.

Relational Databases
Let's re-imagine your case management system as a relational database. You still have a CLIENTS table organized by client name. A client record includes name, address, and one or more case numbers. Even if a client has multiple matters, there is still one authoritative client record. Then you maintain a CASES table for the matters themselves, and a COURTS table so you can keep current on where cases are tried and the presiding judge.
CLIENTS TABLE

CLT_LAST	CLT_FIRST	ST_ADD	CITY	STATE	ZIP	CASE_NO
Bezos	Jeff	9218 NE 38th Str	Yarrow Pt.	WA	98004	001, 005
Chambers	John	5608 River Way	Buena Park	CA	90621	002
Dell	Michael	3400 Toro Canyon Rd.	Austin	TX	78746	003, 007
Ellison	Lawrence	745 Mountain Home Rd.	Woodside	CA	94062	004
Gates	William	1835 73rd Ave. NE	Medina	WA	98039	001, 005
Zuckerberg	Mark	1462 Edgewood Drive	Palo Alto	CA	94301	006, 009
Palmisano	Samuel	665 Pequot Ave.	Southport	CT	06890	007

[75] One of the most important and widely used database applications, MySQL, is open source; so, while great fortunes have been built on relational database tools, the database world is by no means the exclusive province of commercial software vendors.

It's essential to keep track of cases and upcoming trials, so you create another table called CASES:

CASES

CASE_NO	TRL_DATE	MATTER	TYPE	COURT
001	2026-02-14	U.S. v. Amazon	Antitrust	FDDC-1
002	2027-01-09	EON v Cisco	Patent	FEDTX-2
003	2026-02-15	In re: Dell	Regulatory	FWDTX-4
004	2026-05-16	SAP v. Oracle	Conspiracy	FNDCA-8
005	2027-01-09	Bezos v. Washington D. of Labor	Injunction	FWDWA-6
006	2026-12-06	FTC v Meta	Antitrust	FNDCA-8
007	2026-10-31	Dell v. Travis County	Tax	TX250
008	Null	Hawkins v. McGee	Med Mal	FUSSC
009	2026-12-05	Zuckerberg v. City of Palo Alto	Zoning	CASMD09

You also want to stay current on where your cases will be tried and the presiding judge, so you maintain a COURTS table for all the matters on your docket:

COURTS

COURT	JUDGE	FED_ST	JURISDICTION
FNDCA-8	Chen	FED	Northern District of California (SF)
FDDC-1	Boasberg	FED	USDC District of Columbia
FWDTX-4	Pitman	FED	Western District of Texas
TX250	Liu	STATE	250th JDC, Travis County, TX
CASMD09	Pegg	STATE	Santa Clara County Superior Court, CA
FEDTX-2	Crone	FED	Eastern District of Texas
FWDWA-6	Jones	FED	Western District of Washington
FUSSC	Hand	FED	United States Supreme Court

As we look at these tables, note that each has a unique key field serving as the "**primary key**" for that table.[76] The essential characteristic of a primary key is that it cannot repeat within the table for which it serves as primary key, and a properly designed database will prevent duplicate primary keys.

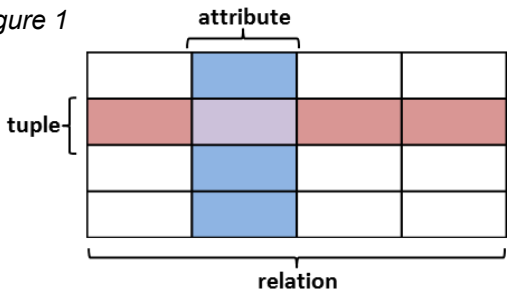
[76] Tables can have more than one primary key.

Many databases simply assign a unique primary key to each row—either a number or a non-recurring value. For example, a key might be an automatically incrementing integer, a UUID,[77] or a composite identifier built from, *e.g.* elements like the first four letters of a name, first three numbers in the address, first four letters in the street name and the Zip code. For example, an assigned key for Jeff Bezos derived from data in the CLIENTS table might be BEZO921NE3898004. The primary key supports indexing, making it efficient to search, sort, link, and perform other operations on the data.

Tuples and Attributes

Now we need a bit of database lingo. The contents of each row in a table is called a "**tuple**," defined as an ordered list of elements.[78] In the COURTS table above, there are seven tuples, each consisting of four elements. The columns (the elements, in their defined order) are called "**attributes.**" And what we've called "tables" in the flat-file world are termed "**relations**" in the relational world. Put another way, a relation is a set of tuples that share the same attributes (See Figure 1).

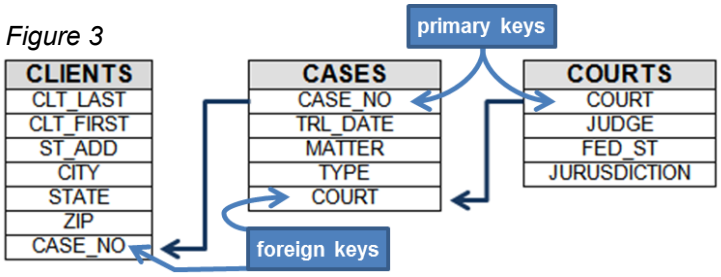
Figure 1





The magic happens in a relational database when tables are "**joined**" by referencing one table from another, much like the cube at left.[79] This is done by incorporating the primary key from the referenced table as a "**foreign key**" in the referencing table. The referenced table is the parent; the referencing table is the child. In Figure 3, COURTS is the parent table to CASES with respect to the COURT field. In the CASES table, the foreign key for COURT points back to COURTS, assuring that current court data populates the field. In turn, CLIENTS can reference CASES via CASE_NO.

Figure 3



What you are seeking is consistency—so you don't build a database with conflicting addresses, mismatched courts, or case numbers that mean one thing on Monday and another on Tuesday.

---

[77] A Universally Unique Identifier (UUID) is a 128-bit label used to uniquely identify information in computer systems.

[78] Per Wikipedia, the term "tuple" originated as an abstraction of the sequence: single, double, triple, quadruple, quintuple, sextuple, septuple, octuple...n-tuple. The unique 0-tuple is called the null tuple. A 1-tuple is called a "singleton," a 2-tuple is a "pair" and a 3-tuple is a "triple" or "triplet." The *n* can be any positive integer. For example, a complex number can be represented as a 2-tuple, a quaternion can be represented as a 4-tuple, an octonion can be represented as an octuple (mathematicians use the abbreviation "8-tuple"), and a sedenion can be represented as a 16-tuple. I include this explanation solely to remind law student readers why many of us went to law school instead of studying applied mathematics. You DO NOT have to know this *n*-tuple stuff for this course.

[79] Although unlike the cube, a relational database is not limited to just three dimensions of attachment.

Relational databases avoid conflicts by using the parent primary key as the definitive source. By pointing child tables to that definitive parent via foreign keys, you promote "**referential integrity**." Remember: a primary key must be unique within its table, but it can be referenced as many times as needed as a foreign key. As in life, parents can have multiple children, but a child can have but one set of (biological) parents.

Field Properties and Record Structures

When you were writing case data on index cards, you were unconstrained. You could abbreviate, write dates as words or numbers, and include as much or as little as space allowed. But databases are fussy—field contents are *constrained*—to promote data integrity. You don't want four or ten letters entered into a field reserved for a Zip code. You don't want duplicate primary keys. You don't want to open a case without a client name, and if a field is designed to store a U.S. state, you don't want it to accept "Zambia" or "female."

Accordingly, databases enforce field properties such as:

1. **Field size:** limiting number of characters or permitting variable length entries for memos;
2. **Data type:** text, currency, integers, date/time, e-mail, and masks for phone numbers, SSNs, Zip codes, etc.;
3. **Unique fields:** primary keys must be unique, and other fields may be constrained (*e.g.,* one SSN per person);
4. **Group or member lists:** fields populated only from a defined set (states, departments, account numbers);
5. **Validation rules:** ranges and formats that make sense (no negative age, no "25:00pm," guard against date notation confusion, and so on); and
1. **Required data:** fields made mandatory where absence destroys utility (*i.e.,* a car rental database requiring input of a valid driver's license number).

Fixed-length and variable-length records (a glimpse at the ugly guts)

You'll appreciate why demanding production of the raw tables in a database can be an untenable approach when you consider how databases store information. Tables are often stored as sequences of bytes optimized for speed and integrity—not for human reading—and they may be distributed across files, partitions, and storage tiers. The DBMS exists to make that mess usable.

In simplified terms, data may be stored using **fixed-length fields** (where every record consumes the same amount of space) or **variable-length fields** (where records can differ in size). Fixed-length designs make offset-based retrieval straightforward: the database can jump to record N by calculating N × record_size. Variable-length designs require pointers or offset tables to locate variable content (like memos or long text), so the record can expand without breaking the structure.

Fixed-Length Field Records

Fixed length fields are established when the database is created, and it's important to appreciate that the data is stored as long sequences of data that may, to the untrained eye, simply flow together in one incomprehensible blob. A fixed length field record may begin with information setting out information concerning all of the fields in the record, such as each field's name (e.g., COURT), followed by its data type (e.g., alphanumeric), length (7 characters) and format (e.g., only values matching a specified list of courts).

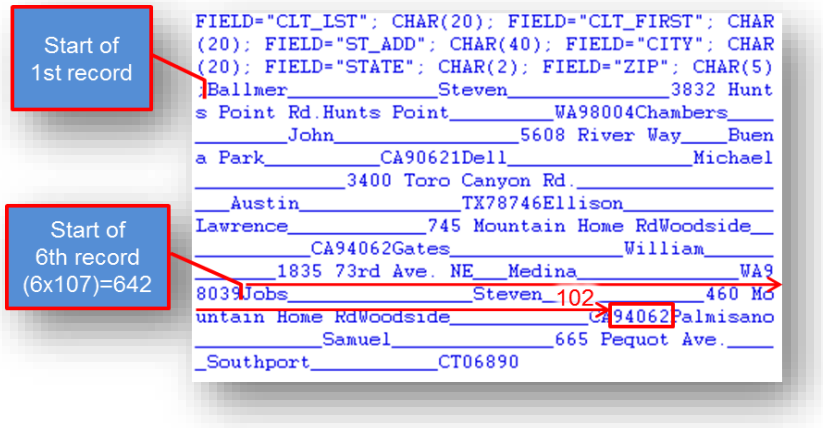
Figure 4

```
FIELD="CLT_LST"; CHAR(20); FIELD="CLT_FIRST"; CHAR
(20); FIELD="ST_ADD"; CHAR(40); FIELD="CITY"; CHAR
(20); FIELD="STATE"; CHAR(2); FIELD="ZIP"; CHAR(5)
;Ballmer_____Steven_____3832 Hunt
s Point Rd.Hunts Point_____WA98004Chambers____
_____John_____5608 River Way____Buen
a Park_____CA90621Dell_____Michael
_____3400 Toro Canyon Rd._____
___Austin_____TX78746Ellison_____
Lawrence_____745 Mountain Home RdWoodside__
_____CA94062Gates_____William_____
_____1835 73rd Ave. NE___Medina_____WA9
8039Jobs_____Steven_____460 Mo
untain Home RdWoodside_____CA94062Palmisano
_____Samuel_____665 Pequot Ave.____
_Southport_____CT06890
```

A fixed length field record for a simplified address table might look like **Figure 4**.

Note how the data is one continuous stream. The name, order and length of data allocated for each field is defined at the beginning of the string in all those "FIELD=" and CHAR(x) statements, such that the total length of each record is 107 characters. To find a given record in a table, the database software simply starts accessing data for that record at a distance (also called an "offset") from the start of the table equal to the number of records times the total length allocated to each record.

Figure 5



So, as shown in Figure 5, fixed-length records let the computer "jump" to any record by counting. If each record is 107 bytes long, the fourth record begins 428 bytes from the start of the table (4 × 107). Within that record, each field also begins at a fixed position from the start of the record.

For example, to extract Steve Jobs' Zip code from the sample table, the DBMS first jumps to the Jobs record (the sixth record), which begins 642 bytes from the start of the table (6 × 107). Then it jumps to the Zip code field within that record. Because the Zip code field is defined to start 102 bytes into each record (after the earlier fixed-length fields), the Zip code begins at byte 744 from the start of the table (642 + 102). Offset-based retrieval like this is miserable work for humans, but it's effortless for computers.

Variable-Length Field Records

Fixed-length records are tidy, but they can be brittle. Sometimes the space allocated to a field proves insufficient in unanticipated ways, or you simply need to allow records to expand on the fly without redesigning the whole table.

To handle that, databases employ variable-length field records whose size can change from one record to the next. Variable-length fields often rely on **pointers**—offset values that redirect retrieval to the place where the overflow or long-text content is stored (commonly in a separate memo or overflow area). The DBMS follows the pointer, reads the variable-length data, and stops when it reaches an end marker or another pointer to the next segment of related content.

Forms, Reports and Query Language

Now that you've glimpsed the guts, you can appreciate why databases employ management software to enter, update, and retrieve data. Though DBMS software serves many purposes—indexing, optimizing, securing, logging—its most familiar role is providing a user interface for forms and reports.

There's little difference between forms and reports except that we tend to call the interface used to input or modify data a "form," and the interface used to extract data a "report." Both are simply user-friendly ways to implement commands in a "**query language**."

Query language is the set of commands used to retrieve information from a database. The best known is **SQL** (**Structured Query Language**). Officially it's "ess-cue-ell," but most everyone calls it "sequel." SQL is a computer language, but unlike Java or C++ (which can be used to construct applications), SQL's purpose is the creation, management, and interrogation of databases.

The phrase "query language" makes it sound like SQL's job is only to get data out, but SQL also handles database creation and data insertion. It includes subsets for data definition, manipulation, and control. SQL syntax is beyond the scope of this chapter, but a short snippet gives the flavor.

```
CREATE TABLE COURTS
        (COURT          varchar(7) PRIMARY KEY,
        JUDGE           varchar(18),
        FED_ST          varchar(5),
        JURISDICTION  varchar(40));
CREATE TABLE CASES
        (CASE_NO        int IDENTITY(1,1) PRIMARY KEY,
        TRL_DATE        date,
        MATTER          varchar(60),
```

```
    TYPE            varchar(40),
    COURT           varchar(7),
    CONSTRAINT FK_CASES_COURTS
FOREIGN KEY (COURT) REFERENCES COURTS(COURT));
```

In these few lines, the COURTS and CASES tables are created, named and ordered into various alphanumeric fields of varying specified lengths. Two primary keys are set and one key, CASE_NO, is implemented so as to begin with the number 1 and increment by 1 each time a new case is added to the CASES table.

Who Owns SQL?

Nobody "owns" SQL, but several giant software companies—notably Oracle and Microsoft—have built significant products around it and produced proprietary dialects. When you hear "SQL Server," you're hearing a Microsoft product name. Microsoft doesn't own SQL; it sells a database platform that uses (its version of) SQL.

SQL has much to commend it, being both simple and powerful; but even the simplest computer language is too much for the average user. So databases employ graphical user interfaces (GUIs) to put a friendly face on SQL. When you enter data into a form or run a search, you're triggering a series of pre-programmed query commands.

In e-discovery, if the standard reports supported by the database are sufficiently encompassing and precise to retrieve what you need, great! You still must arrive at a suitable form of production and wrangle over scope, privilege, and confidentiality; but the path to the data is clear.

However, because most companies design their databases for operations—not litigation—standard reporting won't always retrieve the information required in discovery. In that event, you'll need more than a SQL doctor on your team; you'll also need a good X-ray of the database you're about to plumb.

Schemas, Data Dictionaries, System Catalogs, and ERDs

The famed database administrator, Leo Tolstoy, remarked, "*Great databases are all alike; every ordinary database is ordinary in its own way.*" Although it's tongue-in-cheek that I invoke Tolstoy's line about happy and unhappy families, it's apt here: you can only assume so much about the structure of an unfamiliar database. After that, you need the manual and a map.

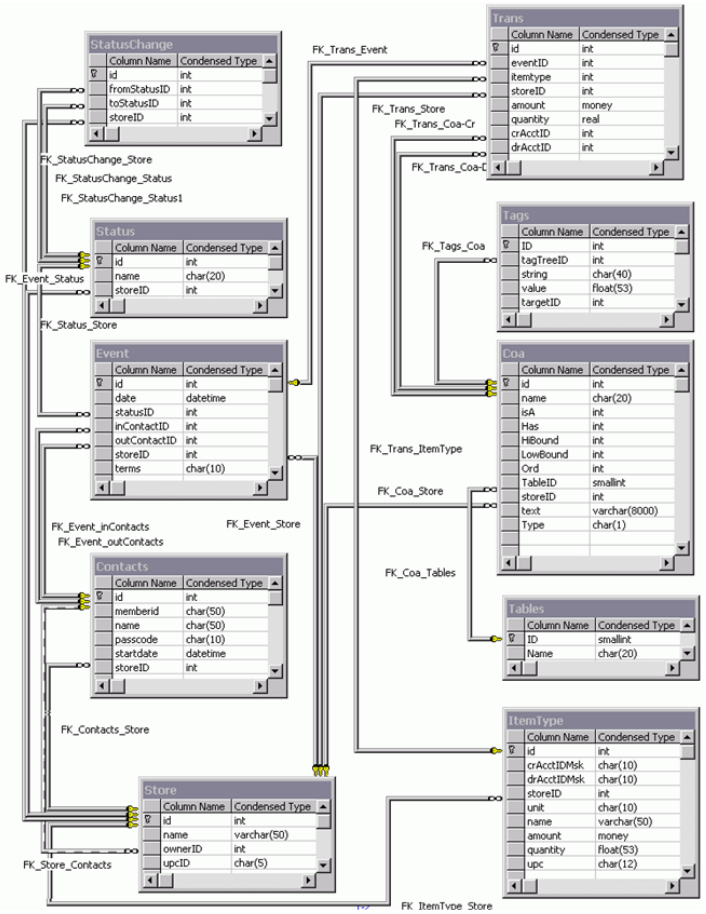
In database land, the "map" is the schema, and it's housed (one way or another) in the system's data dictionary or catalog. The schema may be logical—detailing table structures, attributes (fields), relationships, joins, and views—or physical—setting out the hardware and software

implementation across machines, storage, and networks. As Tolstoy might have said, “A logical schema explains death; but it won’t tell you where the bodies are buried.”

Information in a database is mostly gibberish without the metadata that gives it form and function. In many systems, the compendium of that metadata is called the **system catalog**. In practice, the terms schema, data dictionary, and system catalog are often used interchangeably, and vendors don’t help by using their own labels. Whatever you call it, the most important lesson is this: there is a map—or one can be generated—so get it.

Unlike that elusive Loch Ness monster of e-discovery, the “enterprise data map,” the schemas of databases tend to exist and may be maps in the plain-English sense of depicting the structure. They are often graphical depictions of the database’s structures showing what tables exist, what fields are in each table, and how tables relate to one another.

Figure 6: ERD of Database Schema



In the relational world, the most common way to draw that picture is **Entity-Relationship Modeling (ERM)**, a system and notation used to lay out the conceptual and logical schema of a relational database. The resulting diagrams—akin to flow charts—are called **Entity-Relationship Diagrams** or **ERDs** (Figure 6).

If you’ve never seen an ERD, don’t let the name intimidate you. An ERD is the database’s road map. It shows the “entities” (usually tables), the “attributes” (fields) inside them, and the lines connecting them—the relationships—typically built around primary keys and foreign keys. Those connecting lines are not decoration. They tell you how data in one table links to data in another, and they reveal what the database can say (and what it can’t) without guesswork.

That’s why ERDs and data dictionaries are gold in e-discovery. When someone asks for “all transactions tied to this customer,” or “all users who accessed this record,” or “all changes to this account,” the answer lives in relationships: which tables hold the transactions, which table defines

"customer," what field is the linking key, and where (if anywhere) the system records access and change events. With the ERD and the data dictionary in hand, you can identify the right tables, the right fields, and the right join points. Without them, you're reduced to rooting around and hoping you recognize something—never a sound discovery plan.

So, before anyone starts talking about exporting "the database," demand the map. Ask for (1) the schema, (2) the data dictionary/field definitions, (3) any system catalog information the DBMS can generate, and (4) the ERD or other schema diagram.

Practice Pointer: **Discovery About Databases**

With databases, the first fight is often "discovery about discovery": learning enough about the system to frame workable requests and insist on production in usable forms—before your opponent claims it "can't be done" or dumps out junk and calls it compliance.

The following topics can be used (and tailored) as interrogatories, requests for production, or Rule 30(b)(6) topics:

For each database or system reasonably likely to contain responsive information concerning [IDENTIFY SUBJECT MATTER], identify and describe:

1. System identity and purpose: system name, business function, owner/administrator, hosting (on-prem/cloud/hybrid), and vendor/product/version.
2. Data scope and retention: date ranges, retention/purge practices, archives, and whether older data or logs fall outside default search scope.
3. Structure ("the map"): schema, data dictionary/system catalog, ERD or data model diagram, field definitions/constraints, and code/lookup tables.
4. Reports and exports: standard and custom reports/templates, fields included, available filters, and whether outputs can be produced in searchable electronic form.
5. Structured exports: ability to export to CSV/delimited, JSON, XML, or vendor-native formats, preserving identifiers and relationships (keys, thread IDs, parent/child links).
6. Search/query methods: available tools (GUI reports, query builders, SQL, APIs/admin consoles) and limits (permissions, business-unit scope, indexing, rate limits).
7. Logging and history: existence/location/retention of audit, access, and change logs, and ability to show who did what, when.
8. Persons with knowledge: identities/roles of administrators, analysts, and contractors responsible for the system and for searches/exports.
9. Burden and feasibility: practical steps, cost, and effort to identify, preserve, search, extract, and produce responsive information faithfully.

These are not magic words. Use only what you need and know why you need it.

(For why "usable form" often turns on preserving fielded data—and why productions too often destroy it—see Appendix A.)

Two Lessons from the Database Trenches

The importance of securing the schema, manuals, data dictionary, and ERDs was borne out by my experience serving as Special Master for Electronically Stored Information. In a drug product liability action involving thousands of plaintiffs, I was tasked to expedite discovery from as many as sixty different enterprise databases, each more sprawling and complex than the next. The parties were at loggerheads, and serious sanctions were in the offing.

The plaintiffs insisted that the databases would yield important evidence, and their team included support personnel technically astute enough to get deeply into the weeds with the systems. Plaintiffs were willing to narrow the scope of database discovery—to abandon systems unlikely to be responsive and to sharpen their requests—but to do that, they had to know what the systems were and what they contained.

For each system, we faced the same questions:

1. What does the database do in the business?
2. What is it built on (platform, DBMS, vendor, version)?
3. What information does it hold, and for what date ranges?
4. What content is likely relevant, responsive, or privileged?
5. In what form does the information exist (tables, views, reports, logs, exports)?
6. How can it be searched effectively—using what query language or tools?
7. What reporting and export capabilities exist out-of-the-box?
8. What form or forms of production will be functional, searchable, and cost-effective?

It took a three-step process to turn things around.

Step one: Plaintiffs did the homework, and the defense supplied the curriculum. The defense was required to identify each system and produce a spreadsheet detailing, at a minimum:

• Name of system and business purpose
• Application/vendor/platform (including DBMS, if known)
• Date range covered and retention/purge practices
• Approximate size and growth rate
• User groups and business owners
• Where the system lived (on-prem, hosted, cloud/SaaS) and what environments existed (production, test, archive)
• Available documentation (operator’s manuals, schema/ERD, data dictionaries, and any export specs)

This alone was a revelation. It allowed plaintiffs to prioritize demands toward the systems most likely to matter, instead of chasing every database that ever drew breath. I then directed defendants to furnish operator's manuals, schema information, and data dictionaries for the most relevant systems.

Step two: I ordered narrowly-focused meet-and-confer sessions between technical personnel for both sides. These were conducted by telephone, and each session was limited to one (or a small number) of databases. The defense had to make knowledgeable personnel available. Plaintiffs had to confine their questions to the nuts-and-bolts: what tables existed, what fields meant, how the system logged activity, what was indexed, what could be exported, what was feasible, and what would break things.

When the telephone sessions concluded, plaintiffs served revised requests for production targeted to what the databases could actually yield. In most instances, plaintiffs learned enough about the systems that they could propose SQL queries to be run, or at least specify the fields, joins, date ranges, and filters needed to generate reliable exports.

Step three: because the case was especially contentious, we needed a decisive step to break the database logjam. I convened a face-to-face session. Counsel and technical liaisons, joined by the database specialists, met with me for two days. We worked through each database and arrived at specific agreements concerning scope, searches, sampling where appropriate, quality checks, timing, and form of production. The devil is in the details, and the goal was to nail down every detail. It took two such sessions, but in the end, disputes over databases ceased. Production changed hands smoothly, and the parties could refocus on the merits.

The heroes in this story were the technical personnel who collaborated to share information and forge solutions when the lawyers could see only problems. The lesson: **get the geeks together**, then get out of their way.

Lesson Two: Don't assume the search was complete—or competent

In another case where I served as special master, the Court questioned the adequacy of defendants' search of their databases. Defendants used many systems to run far-flung operations, ranging from legacy mainframe platforms housed in national data centers to homebrew applications cobbled together in Access or Excel. But whether big or small, I found with disturbing regularity that the people tasked to query the systems for responsive data either didn't know how to use them, didn't understand what the data meant, or lacked the rights needed to access the data they were obliged to search.

The takeaway is simple: *never blindly assume that a DBMS query searches all potentially responsive records, and never assume that the operator knows what they are doing.*

Database systems employ a host of techniques to optimize performance and protect confidentiality. For example:
• Older records may be purged from indexes, archived, or moved to cheaper storage that isn't searched by default;
• Users may lack the privileges needed to access all potentially responsive data (or to query across business units, regions, or subsidiaries);
• Queries may be restricted to certain date ranges, regions, record types, or "views" that exclude data sitting in related tables;
• Tables may not be joined in the ways needed to gather the data sought, or joins may be ambiguous without the schema and field definitions;
• Audit logs may exist but be segregated, retained briefly, or accessible only to administrators; and
• In cloud/SaaS systems, the "search" available to ordinary users may be a convenience feature, not a litigation-grade export, and it may omit fields, metadata, or historical states of records.

Any of these can cause responsive data to be missed even when a query appears sensible and the operator seems confident.

Establishing operator competence can be challenging, too. Ask the person running queries whether they have the privileges required for a comprehensive search and you're likely to get a dirty look and an insistence that they do. In truth, they probably don't know their own limitations. What they have are the privileges they need to do their job day-to-day. Those privileges may be nowhere near sufficient to elicit all the responsive information the system can yield.

How do you preserve a database in e-discovery?
Talk to even tech-savvy lawyers about preserving databases and you'll likely hear how databases are gigantic and dynamic, or how incomprehensibly risky and disruptive it is to mess with them. The lawyer who responds, "Don't be ridiculous. We're not preserving our databases for your lawsuit," isn't protecting the client.

On the other side, opposing counsel may say, "Preserve our databases? Sure, no problem. We back up the databases all the time. We'll just set aside some tapes." This agreeable fellow isn't protecting his client either. When it comes time to search that data, Mr. Congeniality may learn that restoration will displace the server currently in use, require scarce expertise, and cost real money. Worse, the "backup" may not even capture what matters—indexes, logs, permissions, code tables, or the configuration needed to reproduce the reports everyone is fighting about.

What both lawyers should say is, "Let me explain what we have and how it works. Better yet, let's get our technical advisors together. Then we'll try to work out a way to preserve what you really need in a way you can use it. If we can't agree, I'll tell you what my client will and won't do, and you can go to the judge right away if you think we haven't done enough."

Granted, this conversation almost never occurs. Counsel may have no idea what the client has or how it works. The duty to preserve may attach before opposing counsel emerges. Or counsel believes cooperation is anathema to zealous advocacy and wants only to scorch the Earth. Still, if you're serious about preservation, you start the same way every time: identify the system, understand how it changes, and decide what must be frozen to keep the evidence reasonably accessible and meaningful.

First, identify what "the database" really is. Is it a classic relational database on-premises? A mainframe system? A homebrew Access file? A cloud service with a web front end? In 2026, many of the systems that matter most are SaaS platforms and cloud services. In those systems, preservation is less about "tapes" and more about retention settings, legal hold features, and defensible exports—sometimes through administrative tools, sometimes through an API. If you don't understand the service's retention and purge rules (including how long audit logs persist), you can lose the past while confidently believing you preserved it.

Second, decide whether the system is accretive or mutative. Databases are dynamic by design, but not all databases change in ways that impair preservation. Many databases—particularly accounting and other ledger-like systems—are accretive: they add new data as time goes on but do not surrender the ability to search historical data thoroughly. For accretive systems, preservation may be as simple as ensuring historical data isn't going anywhere for the life of the case and that retention and purge settings won't silently erase it.

Mutative systems are the troublemakers. If records are routinely updated in place, overwritten, rolled up, or reclassified—and if logs are short-lived—you may need to preserve responsive states of the data before they change. If the issue is temporally sensitive—wholesale drug pricing in 2022, reduction-in-force decisions in 2024, or who had access to what and when—you need to preserve not only the data, but the ability to prove what the system would have shown *at the relevant time.*

Third, preserve what gives the data meaning. Don't preserve data alone, like a bug pinned to cardboard. Preserve the schema, the data dictionary, and the code tables that translate cryptic values into business meaning. Preserve report definitions and the business rules embedded in views, stored procedures, or application logic. Preserve role-based access and permission structures when "who could see what" is at issue. And preserve audit/change logs and their retention settings, because those are often the first things to disappear unless you act quickly—especially in cloud systems.

Fourth, use snapshots and backups as backstops, not the whole plan. Creating snapshots of data stores—or capturing a full backup set for a relevant period—can be sensible insurance against spoliation. If the likelihood of litigation materializing is remote, or there is little chance the

preserved backups will ever be restored, pulling backups may be sufficient and economical. But if litigation is certain and discovery from the database is likely, the better approach is to duplicate and/or segregate the particular dynamic data you'll need, or export it to forms that preserve searchability and utility. The goal is to keep essential data reasonably accessible and to shield it from changes that will impair relevance and probative value.

Done right, database preservation isn't magic. It's discipline: identify the system, learn its structure, decide what must be frozen (data, metadata, logs, and reporting capability), and document what you did so the result is defensible and usable.

Forms of Production

In no other corner of e-discovery are litigants quite so much as the dog that caught the car as when dealing with databases. Data from specialized systems and enterprise platforms often don't play well with off-the-shelf applications—not surprising, considering the horsepower, complexity, and cost of the systems tasked to run big-iron operations. Still, there is always a way.

Sometimes a requesting party demands a copy of an entire database, with too little thought about what success would look like. If the "database" is a modest Access file or other small platform, it may be feasible to duplicate enough of the producing party's environment to use it as intended. But if the dataset is enormous, distributed, or tied to an enterprise DBMS—or, increasingly, to a cloud service—mirroring the environment is usually a non-starter.

Even when you can rent computing muscle by the hour in the cloud, you may not be able to reproduce the application layer, the permissions, the business rules, the indexes, and the integrations that make the data intelligible. And in SaaS systems, "give me the database" may be about as meaningful as "give me the internet."

A more realistic path is to narrow the dataset using filters and queries, then produce the responsive information in forms that are usable outside the source system. That usually means exporting structured data—CSV or other delimited text, JSON, or XML—paired with the schema, field definitions, and code tables needed to interpret what the fields mean. In other instances, it means generating reports (standard or custom) and ensuring that the output is electronically searchable and complete, not a stack of paper or dead-image page renders that can't be sensibly analyzed.

Before negotiating a form of production, investigate what the system can actually do. *Do not* assume the database administrator has ever had occasion to perform litigation-grade exports, even if they know the system cold for day-to-day operations. And don't accept "it can't be done" without a follow-up question: "Can it be exported?" It is the rare system—relational, NoSQL, or cloud—that can't export data in some structured form.

Have a proposed production format in mind. Better still, be prepared to describe what makes the export usable: delimiters, encoding, date formats, time zones, record identifiers, and whether relationships are preserved through keys.

Also remember that resistance may not come from the producing party or the DBMS administrator. More often, the insistence on static page images is driven by lawyers who are more comfortable with paper than with data. In that instance, it helps to establish early—on the record—that a structured export is feasible and that a "report" can be produced in a form suited to search and analysis rather than as a page-by-page photograph of what the system displayed.

As with other forms of e-discovery, be careful not to accept production in formats you don't want. Like it or not, many courts give one bite at the production apple. If you accept TIFF images for the sake of expediency, you may close the door on re-production in more functional forms. If the data matters, fight the format fight early.

Even when the parties agree on an electronic form of production, get a test export first. Evaluate it before anyone pulls the lever on a high-volume export. A test run will reveal the gotchas that wreck usability: truncated fields, missing identifiers, cryptic codes without code tables, orphaned tables with no keys, inconsistent date/time formats, unexplained null values, and exports that silently omit archived tiers or older records. Better to find those problems in a thousand rows than in ten million.

Closing Thoughts

When dealing with databases in e-discovery, requesting parties should avoid the trap of "You have it. I want it." Lawyers who would never be so foolish as to demand the contents of a file room will blithely insist on production of "the database." For most, if they succeeded in that foolish quest, they'd wind up holding an obscure collection of inscrutable information they can't possibly use.

Things aren't much better on the producing party's side, where counsel routinely fail to explore databases on the theory that if a report hasn't already been generated, it doesn't have to be created for litigation. That thinking is increasingly perilous. Modern business records are often nothing more than database outputs, and the "document" exists because the system can generate it, not because someone printed it last Tuesday. Even when counsel acknowledges a duty to search databases, few appreciate how pervasively embedded databases are in their clients' operations, and fewer still possess the skills to translate an amorphous request into precise, effective queries and defensible exports.

Each side is trading on ignorance, and both do their clients a disservice.

But these are increasingly problems of the past. There is cause for cautious optimism in how lawyers and litigants approach databases in discovery. Counsel are starting to ask about databases earlier,

to demand the maps—schemas, data dictionaries, ERDs or their modern equivalents—and to focus on practical, usable productions instead of theatrical ones. Databases are not going away, and neither is the evidence they hold. The sooner lawyers stop treating database discovery as exotic—and start treating it as ordinary—the sooner discovery gets cheaper, faster, and more reliable.

Appendix A: Fielding Matters—Don't Strip the Data Bare

One feature that sets databases apart from many other forms of ESI is the critical importance of fielded data. Preserving the fielded character of data is essential to preserving its utility and searchability. "Fielding" means that information is stored in defined locations dedicated to holding particular pieces of information—sender, recipient, date, amount, status, author, and so on. Fielding lets you search, sort, filter, and cull using just that information. It's a capability we take for granted in databases but one that is routinely crippled—or eradicated—when data is produced in e-discovery. Before you negotiate a production format, think hard about whether the fielding will survive, whether the data is supplied as a report or as a delimited export.

Fielding isn't new. We did it when data lived on paper. Take a typical law firm letter: the letterhead identifies the firm; the date below the letterhead is understood to be the date sent; a Re: line denotes the matter; then come the addressee and salutation. The recipient is understood to be identified at the start and the sender at the end. These conventions about where information belongs are what make correspondence intelligible and sortable.

Similarly, the common productivity file types encountered in e-discovery—Microsoft Office formats, PDF, and e-mail—employ fielding to support utility and functionality. Native "documents" are natively fielded; that is, content and metadata are structured so that particular pieces of information reside in defined locations within the file. Native applications—and tools designed to exploit file architecture—understand that structure and make use of it.

Yet we act inconsistently, inefficiently, and irrationally when we deal with fielded information in e-discovery. In contrast to a few years ago, not even the most Neanderthal counsel challenges the need to produce the native fielding of spreadsheet data. Production of spreadsheets in native form has become routine and largely uncontentious. To get there, workflows were modified, Bates numbering procedures were tweaked, and despite dire predictions, none of it made the sky fall. We can—and should—do the same with other structured sources, including presentations, word processing files, and database exports.

"*What's vice today may be virtue tomorrow*," wrote novelist (and jurist) Henry Fielding.

Take e-mail. All e-mail is natively fielded data, and the architecture of e-mail messages is established by published standards (RFCs)—structural conventions that e-mail systems must follow so messages can traverse servers and be reliably parsed by clients. Those standards define where the sender, recipients, subject, date, message identifiers, routing information, and other components belong.

But when we produce e-mail in discovery, the "accepted" practice has often been to deconstruct each message and produce it in a cruder fielded format that is incompatible with the native structure and unrecognizable to any e-mail tool or system. Too, the production is often incomplete compared to the native content.

This deconstruction is accomplished by **field mapping.** Values from native fields are extracted and inserted into a new matrix of fields that may reuse native names or rename them altogether. The source data is "mapped" to a new name and a new location. At all events, the mapped fields do not mirror the field structure of the source file.

Ever? No, never.

The jumbled fielding doesn't always destroy the ability to search within fields or to cull and sort by fielded content, but it does force lawyers to obtain tools that can reassemble and interpret the restructured data. Worse, information is often omitted—not because it's privileged or sensitive, but because…well, um, er, we just always do it that way.

And the omitted information is not necessarily "useless metadata." Often it materially aids comprehension. For example, message threading and conversational context can depend on fields such as Message-ID and reply/reference headers. Correct ordering across time zones depends on preserving date/time values with offsets. In modern practice, these same issues recur across chat and collaboration systems, where thread IDs, channel IDs, parent/child relationships, and edit history can be the difference between context and confusion.

"Why do producing parties get to recast and omit this useful information," you ask? The industry responds: "These are not the droids you're looking for." "Hey, is that Elvis?" "No Sedona for you!"

The answer is that counsel—especially requesting counsel—are too often asleep at the wheel. Producing parties have been getting away with this for so long, unchallenged, they've come to view it as a birthright. Reform is coming, but at the glacial pace for which we lawyers are justly reviled, I mean revered.

E-discovery practices have evolved to acknowledge that e-mail must be supplied with some fielding preserved; but there is no sound reason to produce e-mail with shuffled or omitted fields when faithful, complete exports are feasible. It doesn't cost more to be faithful to the native or near-native architecture or to be complete in supplying fielded content. Producing parties often pay more to degrade production, and what emerges costs more to review.

Perhaps the hardest thing for lawyers and judges to appreciate is how central fielding is to efficiency:

- It's efficient to cull and sort files by date.
- It's efficient to search only within e-mail recipients.
- It's efficient to distinguish Speaker Notes in PowerPoint or filter by Author in Word.

Preserving the fielded character of data makes that possible. Preserving fielding and native architecture opens the door to a broad array of tools. Restructuring and degrading fielding limits use to a handful of pricey tools that understand peculiar and proprietary production formats.

It’s not enough for producing parties to respond, “But you can reassemble the kit of data we produce to make it work somewhat like the original evidence.” In truth, you often can’t, and you shouldn’t have to try.

It ties back to the Typewriter Generation mentality that keeps us thinking about “documents” and trying to force everything into a document-shaped box. Much of what we seek in discovery today is not a purposeful precursor to something that will be printed. Most modern evidence is data—fielded data. Modern productivity files aren’t blobs of text; they’re ingenious little databases. Powerful. Rich. Databases. Their native content and architecture are key to utility and efficient searchability in discovery. Get the fielding right, and functionality follows.

Seeking discovery from databases is a key capability in modern litigation, and it’s not easy for the technically challenged (although it’s probably a whole lot easier than your opponent claims). Getting proper data in usable forms demands careful thought, tenacity, and more-than-a-little homework. Still, anyone can do it—alone with a modicum of effort, or aided by a little expert assistance.



Exercise 12: Forensic Imaging

GOALS: The goals of this exercise are for the student to:

1. Distinguish forensically sound images from copies, clones and targeted collections; and

2. Create and validate a forensically sound image of evidence.

CAVEAT: In this exercise, we must proceed without write protection of the evidence media; something a forensic examiner would avoid in real-life. Accordingly, the techniques used in this exercise should not be employed against real evidence in litigation absent the use of suitable write-protection techniques.

When users empty deleted files from Windows recycle bins, they aren't gone. The operating system simply ceases to track them, freeing the clusters the deleted data occupies for reallocation to new files. Eventually, these unallocated clusters may be reused and their contents overwritten, but until that happens, Windows turns a blind eye to them. Because Windows only *sees* active data, it only *copies* active data. Forensically sound preservation safeguards the entire drive—data and metadata in allocated clusters PLUS artifacts in slack space and unallocated clusters--including the deleted data they hold.

Accordingly, think of the Three Commandments of forensically sound preservation as:

1. Don't alter the evidence;

2. Accurately and thoroughly replicate the contents; and

3. Be prepared to prove the preceding objectives were met.

These standards aren't observed in every situation—notably, in the logical acquisition of a live server or physical acquisition of a phone or tablet device—but parties deviating from a "change nothing" standard should weigh, disclose and defend any deviation.

Distinguishing "Clones" and "Collections" from "Images"

Even lawyers steeped in electronic data discovery may confuse active file imaging (i.e., "Ghosting" or targeted collection), cloning and forensically sound imaging. You shouldn't. If someone suggests an active data duplicate is forensically sound, set them straight and reserve "forensically sound" to describe only processes preserving all the information on the media (allocated clusters, unallocated clusters and slack), coupled with the ability to demonstrate identicality by hash authentication.

The terms "clone" and "image" are often used interchangeably, along with others like "bit stream" and "mirror copy." So long as the duplicate is made in a forensically sound way and can be reliably verified to be so, the name attached to the duplicate doesn't make much difference.

A "targeted collection" is the copying of some or all *the active data* from an evidence source. Targeted collections omit the contents of unallocated clusters and file slack space and typically leave behind operating system artifacts of importance to forensic analysis. Accordingly, targeted collections are frequently sufficient for the limited purposes of e-discovery but may be wholly insufficient to preserve data for forensic analysis.

The term "drive image" is most closely associated with a method of forensic duplication whereby all of the data and metadata on the source drive are stored in a file or series of files which, though structurally different from the source drive, can be reconstituted ("restored" or "mounted") in such a way as to be a forensically-sound duplicate of the source drive. A drive image may be used with compression algorithms to store the source drive data in a more compact format. Though a drive image is capable of being restored to a wiped physical disk to create a "clone drive," forensic analysis tools are designed to "virtually restore" the drive, reading directly from the image file(s) and "seeing" the forensically sound duplicate drive without the necessity of physical restoration.

How do you make a "forensically-sound" duplicate of a drive?

Although forensic examiners use similar techniques and equipment, there is no single "recognized" or "approved" way to create a forensically-sound duplicate of a drive. There are several hardware and software tools well-suited to the task, each with strengths and weaknesses, but all can create a forensically-sound duplicate of a hard drive when properly deployed. Keep in mind that there are many different types of digital media out there, and a tool well-suited to one may be incapable of duplicating another. Examiners simply need to know what they are doing and match the tool to the task

Don't Alter the Evidence

Forensic examiners are careful to prevent the preservation process from effecting changes to the source evidence in a process called "write protection" or "write blocking." Write blocking is achieved in three principal ways:

1. Using write blocking *hardware* interposed between the evidence drive and any computer. Write blockers intercept "writes" to disk or other alterations of the evidence;
2. Using write blocking *software* tools to achieve the same prophylactic interception; or
3. Using operating systems (*e.g.,* Linux) configured so as not to write to the evidence.

We forego the use of write blocking tools or software in this exercise (relying instead on evidence media with a built-in write blocking switch), but forensic examiners wouldn't proceed in this manner absent compelling justification considering the risk posed to the integrity of the evidence.[80]

Accurately and Thoroughly Replicate the Contents

[80] By way of example, when a drive is attached to a computer, the operating system (either PC or Mac) writes data to the root of the drive. The change is minor, but *any* change is enough to alter the drive' hash value of and writing to evidence media overwrites unallocated clusters that could hold otherwise-recoverable deleted data.

Again, forensic collection methodologies for drives fall into two camps: those which create a drive *image* (a file or collection of files which can be virtually or physically restored to match the source evidence) and those which create a *clone* drive (a fully operational one-to-one copy of the evidence drive). Cloning was dominant decades ago but is now a long obsolete approach that has entirely given way to drive imaging. Done right, either approach works but drive imaging enjoys significant advantages deriving from its encapsulating evidence data as authenticable image files unaffected by the file system for the target drive. A clone, being a fully operational physical device, is prone to corruption by improper evidence handling. Simply attaching a clone to a computer without the use of write blocking may serve to destroy the integrity of the evidence and render it incapable of authentication by hashing.

Contemporary approaches to forensically sound duplication range from generic software utilities capable of producing a bitstream capture (e.g., RAW or Linux dd with hash authentication) to purpose-built software applications for forensic drive duplication (*e.g.*, Nuix Imager, FTK Imager, Encase Imager) to dedicated hardware devices that automate the process.

Prove the Image is Forensically Sound

Considering the size of modern hard drives, one way you *can't* prove the validity of your duplicate is by manually comparing the data. It's just impossible. So, the process of verification must be automated and reliable. To appreciate the solution, take a moment to ponder the problem: how can you examine billions or trillions of sectors on a duplicate drive and be certain that every one of them has precisely the same value and is in the exact same relative location as on the source drive? Not just be *certain* but be more reliably certain than fingerprints or DNA evidence! This is where we once more say "thanks" to the mathematical geniuses who gave up normal human interaction to dedicate their lives to algorithms, arrays and one-way computations. These are the brainiacs who thought up "hash functions" and "message digests."

Recall that a hash function accepts a value of any size as its input, performs a complex calculation on that input and returns a value of fixed length as its output. The output value functions as a unique representation of the input, a digital fingerprint. Put in a complex "message" and out pops a long, fixed string "digest" composed of letters and number bearing no discernable relationship to the "message" but which can only be generated by that one input. Accordingly, the output is called a "message digest" (MD). The truly marvelous feature of this is that the computation only works in one direction ("one way"). It's considered "computationally infeasible" to decode the input from the output. Since the input message can be anything, someone had the very bright idea to use the entire contents of a hard drive or thumb drive as the input and—*voila*!—the output becomes a fingerprint of that drive's contents and layout. Change so much as one single bit somewhere—*anywhere*--on the drive and the message digest changes dramatically. Since the fingerprint is

functionally unique to the input message (here, the data on the drive) only a forensically-sound duplicate of the drive could generate the same message digest.[81]

Most software and hardware tools sold for the purpose of forensically sound drive duplication integrate hashing for immediate authentication. These will typically create a file record or ancillary file memorializing the hash values of the source and target as well as other data identifying the evidence and demonstrating the integrity of the process. We will use one such application in this exercise.

Exercise 12: Creating a Forensic Image of an Evidence Thumb Drive

Two workflows for this exercise follow; one for MacOS users (beginning immediately below) and another for Windows users (beginning after the Mac OS steps). The individually numbered evidence drive supplied to you has a USB-A connector on one end and a USB-C (Thunderbolt) connector on the other for your ease of use. Whatever machine you are using should support one or both USB interfaces. You will need 8GB of storage space on your machine.

Forensic Imaging Exercise (macOS): Imaging a USB Thumb Drive Named "EVIDENCE"

What you are doing

You will create a **forensic image** of a USB thumb drive. A forensic image is a **complete, bit-for-bit copy** of the entire device, including deleted space and data you cannot see in Finder. This is not the same as copying files. You will use built-in macOS tools. You are **not modifying** the USB drive (except for the minor change your machine will make to the thumb drive when attached because you are not using write protection).

Before you begin

- You will be using your own Mac laptop
- Be sure you know your Mac login password
- The USB thumb drive is labeled **EVIDENCE#**
- Do **not** copy files using Finder
- Do **not** rename the USB device or the image file.

Step 1 — Insert the USB drive

Insert the USB thumb drive labeled **EVIDENCE**. Your drive may be numbered as well, *e.g.,* EVIDENCE DRIVE 1

[81] In a world of infinite variations, the number space of a hash value is gargantuan but still finite (e.g., 2^{128} for MD5). Accordingly, hash collisions—where two different inputs produce a matching output—can occur and have been fabricated for common hash algorithms. In practice, the potential for an accidental collision is so remote as to be deemed effectively non-existent.

If a Finder window opens showing files, **close it**. Do not interact with the contents of the drive.

Step 2 — Open Terminal

1. Click the **magnifying glass** (Spotlight) in the top-right corner
2. Type **Terminal**
3. Press **Return**

A window with text appears. This is Terminal.

You will type commands **exactly as shown**. Spacing and capitalization matter.

Step 3 — Identify the USB device (do not guess!)

In Terminal, type:

diskutil list

Press **Return**.

Look for a disk that:

• Says **external, physical**
• Is approximately **4 GB**
• Has a volume named **EVIDENCE**

Example:

/dev/disk4 (external, physical):

Write down the disk number.
In this example, the USB device is:

/dev/disk4

⚠️ IMPORTANT: **If your disk number is different, substitute it everywhere below.**

Step 4 — Unmount the USB drive (required)

Unmounting prevents macOS from writing to the device.

Type:

diskutil unmountDisk /dev/disk4

Press **Return**.

You should see confirmation that the disk was unmounted.
If you see an error, stop and seek assistance.

Step 5 — Create the forensic image (written to your Desktop)

You will now create a forensic image named **EVIDENCE.dd** and save it **on your Desktop**.

Type the following command **exactly**:

sudo dd if=/dev/rdisk4 of=~/Desktop/EVIDENCE.dd bs=4m conv=noerror,sync status=progress

Carefully check your command for typos. There is no space between noerror,sync

Press **Return**.

You will be prompted for your **Mac login password**.
As you type, nothing will appear on the screen. This is normal.

Press **Return** after typing your password.

If you see a pop up asking whether to allow Terminal to access files on a removable volume, click "Allow."

If you encounter a "No such file or directory" error when setting your output file destination (of=) using the suggested target, "of=~/desktop/EVIDENCE.dd", try substituting the path to YOUR desktop, which may be (for example) "of=/users/[*your user name*]/Desktop/EVIDENCE.dd" Note that there is no tilde (~) in the substitute path.

You will see progress information as it images the 4GB contents. This may take 2-3 minutes.
Do not close Terminal. Do not remove the USB drive.

When the command finishes, the prompt will return.

What "~/Desktop" means

- ~ means **your home folder**
- ~/Desktop means **your Desktop**

This command guarantees that the file **EVIDENCE.dd** will appear on your Desktop when finished. You do not need to find or create a folder manually.

Step 6 — Calculate the MD5 hash of the USB device

Type:

sudo md5 /dev/rdisk4

Press **Return**.

Allow 1-3 minutes for the hash to be calculated

IMPORTANT: Record the ***first six characters*** full MD5 value shown.

Step 7 — Calculate the MD5 hash of the image file

Type:

md5 ~/Desktop/EVIDENCE.dd

Press **Return**.

The MD5 value **must exactly match** the value from Step 6.

If the values match, your image is verified. If not, you should repeat the process. If they do not match after two attempts, record the differing hash values and, if you have time, seek my help.

I**MPORTANT:** Again, record the ***first six characters*** of each full MD5 value shown.

Step 8 — Confirm the image size

Type:

ls -lh ~/Desktop/EVIDENCE.dd

Press **Return**.

The file size should be approximately **4 GB**, even if the USB drive appears partially unused.

What to turn in for this exercise:

DO NOT turn in your image files. Instead, turn in the following to me on Canvas:

1. USB device identifier on your machine (example: /dev/disk4)
2. First six characters of the MD5 hash of the USB device
3. First six characters of the MD5 hash of EVIDENCE.dd
4. File size as reported in Step 8
5. Your macOS version

Keep the image files because you will need them for a future exercise.

Troubleshooting

"I can't find my image file."
Look on your **Desktop** for a file named **EVIDENCE.dd**.
If it is not there, type this in Terminal:

ls ~/Desktop

If the file still does not appear, stop and ask for help before proceeding.

"My hashes don't match."
Do not delete anything. Let me know, if you have time to do so. This usually means:
• The wrong disk number was used, or
• The device was not unmounted before imaging.

"The command scares me."
That is normal. This exercise uses the command line because it exposes exactly what happens to evidence. You are not expected to memorize commands—only to follow them carefully.

Why this exercise matters

Courts care less about tools than about **process**.
This workflow demonstrates:

• Full-device acquisition
• Protection against alteration
• Integrity verification using hashing

These same principles apply to laptops, servers, and cloud exports.

Forensic Imaging Exercise (Windows PC): Imaging a USB Thumb Drive Named "EVIDENCE"

Imaging a USB Thumb Drive Named "EVIDENCE" Using FTK Imager (EnCase E01 Format)

What you are doing
You will create a **forensic image** of a USB thumb drive in the **EnCase .e0x evidence file format**.. This format produces a **compressed, segmented image set** and automatically calculates and verifies cryptographic hashes.

A forensic image is a **complete, bit-for-bit copy** of the entire device, including deleted space and data you cannot see in File Explorer. This is not the same as copying files. You will use a commercial forensic imaging software application called **FTK Imager.** You are **not modifying** the USB drive (except for the minor change your machine will make to the thumb drive when attached because you are not using write protection).

Before you begin, be sure:

• You are logged into your Windows PC
• FTK Imager is downloaded and installed on your Windows PC (see below)
• The USB thumb drive is labeled **EVIDENCE**
• Do **not** copy files using File Explorer
• Do **not** rename the USB device

Installing FTK Imager:

In the FILES>Software Installer Files area of Canvas, download the FTK Imager installer file from the FTK Imager folder. Double clicking the .ISO file should mount the container and afford you access to the .EXE file that installs the application (Exterro® FTK Imager 8.2.0.33).

Step 1 — Insert the USB drive

Insert the USB thumb drive labeled **EVIDENCE**.

If File Explorer opens automatically, **close it immediately**.
Do not open or interact with the contents.

Step 2 — Launch FTK Imager (as Administrator)

1. Click the **Start Menu**
2. Type **FTK Imager**
3. Right-click **FTK Imager**
4. Select **Run as administrator**
5. Click **Yes** if prompted

FTK Imager will open.

Step 3 — Begin the imaging process

From the menu bar:

File → Create Disk Image…

A *Select Source* window will appear.

Step 4 — Select the source type

Choose:

Physical Drive

Click **Next**.

Step 5 — Select the correct USB device (*do not guess*)

From the list of physical drives, select the one that:

• Is approximately **4 GB**
• Is removable
• Corresponds to the USB labeled **EVIDENCE**

⚠️ **Do not select your internal system drive.**
If you are uncertain, stop and ask for assistance.

Click **Finish**.

Step 6 — Choose the image type (EnCase)

In the *Create Image* window, click **Add…**

Select:

E01 (EnCase)

Click **Next**.

This format:
• Compresses the image
• Segments it if needed
• Stores metadata
• Supports verification

Step 7 — Set the destination (Desktop)

You will now choose where the image files will be saved.

1. Click **Browse…**
2. Select **Desktop**
3. In the *Image Filename* field, type: **EVIDENCE**

FTK Imager will automatically create four files:

EVIDENCE.E01
EVIDENCE.E02
EVIDENCE.E03 and
EVIDENCE.E01.txt (This is the sole file you submit for credit)

Click **Finish**.

Step 8 — Image information

You will see a form requesting case information.

Enter the following (if required):

• Case Number: (Exercise)
• Evidence Number: EVIDENCE
• Examiner: your name
• Description: Blue USB thumb drive labeled “EVIDENCE”

This information does **not** change the data being imaged.

Click **Next**.

Step 9 — Compression and verification settings

Before starting, confirm the following:

- **Compression**: Checked (default is acceptable)
- **Verify images after they are created**: Checked
- **Calculate MD5 hash**: Checked
- (SHA-1 may also be checked; leave defaults unchanged)

Do not change segment size.

Click **Start**.

Step 10 — Imaging in progress

FTK Imager will display progress information.

- Imaging may take several minutes
- Do **not** remove the USB drive
- Do **not** close FTK Imager

Wait until FTK Imager reports that imaging and verification are complete.

Step 11 — Review verification results

When finished, FTK Imager will display:

- MD5 hash of the source device
- MD5 hash of the image
- Confirmation that verification **passed**

Step 12 — Confirm the image files

On your Desktop, confirm that you see:

EVIDENCE.E01 (and possibly EVIDENCE.E02, etc.)

You will also find a small (<2kb) text file named **EVIDENCE.E01.txt.** (This is the sole file you submit for credit). Again, DO NOT submit your image files, only the text file. **Keep the image files because you will need them for a future exercise.**

Troubleshooting

“I don’t see my image file.”
Check the Desktop for EVIDENCE.E01.
If it is missing, return to FTK Imager and confirm the destination path.

“Multiple files were created.”
This is normal. EnCase images may be split into numbered segments.

“My hashes don’t match.”
Do not delete anything. Notify the instructor.
This usually indicates a read error or incorrect source selection.

“Why use E01 instead of dd?”
E01 adds:
• Compression
• Built-in metadata
• Automatic segmentation
• Integrated verification

The forensic principles are the same.

Why this matters

This exercise mirrors professional workflows used in civil litigation, criminal investigations, and digital forensics labs:

• Full-device acquisition
• Read-only handling
• Cryptographic integrity verification
• Tool-independent defensibility

Different formats. Same evidentiary obligations.

Keep track of the image you created <u>as you will use it again in a later exercise</u>!

You're done! You can remove the evidence thumb drive. It's yours to keep, but **please hang on to it and do not use it for storage until the end of the semester** in case you should forget to preserve and protect the image set you've just made.



Exercise 13: Data Mapping

VITAL VOCABULARY
Productivity Files
Information Governance
Possession/Custody/Control
Data Mapping
Trust But Verify
SharePoint

GOALS: The goals of this exercise are for the student to:

1. Consider the breadth and complexity of potentially discoverable information;
2. Appreciate the metrics, judgments, and tradeoffs attendant to building a useful (that is, utile) data map; and
3. Develop a defensible data map of your own data footprint.

We inhabit a world teeming with digital data streams Every day, we generate, touch, transmit, revise, store, duplicate, and discard electronically stored information[82]—ESI—often without the slightest awareness of how much of it exists, where it lives, or how fragile it may be. Our business work product—letters, reports, memoranda, spreadsheets, financial statements, presentations, marketing materials—now manifests as discrete productivity files,[83] and when paired with electronic communications such as email, texts, chats, and collaboration platforms, it accounts for the lion's share of the information preserved, pursued, and produced in discovery.

When business records lived on paper, we relied on physical order to impose informational order: folders, drawers, cabinets, and file rooms. The migration from paper to personal computers, networked systems, and mobile devices dissolved those familiar constraints. Information that once existed as a single, tangible artifact now exists in many forms, in many iterations, and as fragments scattered across devices, systems, accounts, and media.

[82] As amended in 2006, Rules 26 and 34 of the Federal Rules of Civil Procedure (FRCP) employ the phrase "electronically stored information" but wisely do not define same in recognition of technology's ability to outpace law. The phrase supplants the prior use of "data compilations" and per the Rules' Comments "includes any type of information that is stored electronically." The *Guidelines For State Trial Courts Regarding Discovery of Electronically-Stored Information* defines electronically-stored information as "any information created, stored, or best utilized with computer technology of any type. It includes but is not limited to data; word-processing documents; spreadsheets; presentation documents; graphics; animations; images; e-mail and instant messages (including attachments); audio, video, and audiovisual recordings; voicemail stored on databases; networks; computers and computer systems; servers; archives; back-up or disaster recovery systems; discs, CD's, diskettes, drives, tapes, cartridges and other storage media; printers; the Internet; personal digital assistants; handheld wireless devices; cellular telephones; pagers; fax machines; and voicemail systems."

[83] The term "productivity files" refers to the common Microsoft Office application files most often seen in business settings (.doc or .docx Word documents, .xls or .xlsx Excel spreadsheets, .ppt or .pptx PowerPoint presentations; the "dotted" three- or four-letter references being the file extension) and Adobe .pdf Portable Document Files.

The consequence of this sea change in **information governance**[84] is predictable and well documented: individuals and organizations ("custodians"[85]) typically have a poor understanding of the nature, quantity, location, and form of the ESI in their **possession, custody, or control.**[86] A recurring theme in cases imposing sanctions for e-discovery failures is not bad faith or sabotage, but ignorance—failing to know what data existed, who had it, where it resided, how it was stored, or what forces threatened its alteration or loss.

Before you can preserve, review, collect, or produce ESI, you must first know what you have, where you have it, the forms it takes, and how much of it you've got. The disciplined effort to answer those questions is called **data mapping**.

Before you can preserve, review or produce ESI, you must first know what you have, where you have it, the forms it takes and how much of it you've got. The process by which parties and counsel build inventories of potentially relevant ESI is called **data mapping**.

Introduction to Data Mapping

Data mapping encompasses the methods used to identify, inventory, and memorialize sources of ESI. It is an essential prerequisite to everything in the eDiscovery Reference Model that lies east of Information Governance.

The term "data mapping" is an unfortunate one. It suggests a visual diagram—often prompting lawyers to ask for Visio-style schematics resembling IT network maps showing servers, switches, and IP addresses. Those diagrams may be useful to IT, but they are rarely sufficient—and often irrelevant—to the needs of e-discovery.

Unless created expressly for litigation readiness, few organizations possess anything that qualifies as a defensible data map. Network diagrams from IT and retention schedules from Records and Information Management are not data maps, though each may provide valuable clues about where ESI resides and how it is handled.

A data "map" is seldom a map at all. It is more often a list, table, spreadsheet, or database. I favor spreadsheets because they make it easier to sort, filter, and total volumes, but the form matters far less than the substance. Some organizations employ specialized software to manage data-mapping

[84] The Gartner consulting firm defines Information Governance as "the specification of decision rights and an accountability framework to encourage desirable behavior in the valuation, creation, storage, use, archival and deletion of information. It includes the processes, roles, standards and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals."

[85] Though the term "records custodian" is customarily defined as the person responsible for, or the person with administrative control over, granting access to an organization's documents or electronic files while protecting the data as defined by the organization's security policy or its standard IT practices, the term tends to be accorded a less precise definition in e-discovery and is best thought of as anyone with possession, custody or control of ESI, including a legal right or *practical ability* to access same. *See*, *e.g.*, *In re NTL, Inc. Securities Litigation*, 244 F.R.D. 179, 195 (S.D.N.Y. 2007).

[86] FRCP 26(a)(1)(A), and sometimes articulated as *care*, custody or control

and preservation efforts. Others rely on narratives or hybrids of narrative and tabular data. Whatever the form, clients almost never have a data map just sitting around. It must be built—often from scratch.

What matters is not how attractive the data map looks, but whether it does the work. If the form you choose allows you to implement defensible preservation, project burden, guide collection, and accurately answer questions at the meet-and-confer table or in court, then it is the right form—even if it's ugly.

Scope

The duty to identify ESI is the most expansive obligation in e-discovery. You cannot preserve sources you have not identified. You cannot collect, review, or produce what you do not know exists. The Federal Rules of Civil Procedure go further still, requiring parties to identify sources of potentially responsive information they deem "not reasonably accessible," even when they do not intend to search them.

For that reason, a "data map" might more honestly be called an *information inventory*. Like a merchant's inventory, it seeks to answer four basic questions: what is it, how much is there, where is it, and what is it worth (or, in litigation terms, how burdensome or risky is it).

Creating a competent data map is also an exercise in reconstructing institutional memory. It requires understanding:

- People and their roles (cases are still mostly about people);
- Information systems and how they have evolved over time; and
- Projects, facilities, vendors, and tools.

A data map spans both logical and physical sources of information. *Bob's email* is a logical source that may span multiple physical devices and repositories. *Bob's laptop* is a physical device that may contain many distinct logical data sources. Logical and physical collections overlap, but they are rarely the same.

Depending on the case, a data map may capture, among other things:

- Custodian or source;
- Physical and logical location;
- Device or medium;
- Currency of contents;
- Volume (bytes);
- Numerosity (messages, records, attachments);
- Time span and gaps;
- Purpose and function;
- Usage patterns;

• Native form(s);
• Fragility and risk of loss.

This list is illustrative, not exhaustive. You do not map email the same way you map databases, backup media, or SaaS platforms. A data map is not an exercise in clerical excess. The level of detail should track the likely relevance, materiality, and risk associated with the source. Precision matters—but so does judgment.

Tips for Better Data Mapping

Custodial interviews (*i.e*., questioning people who hold data) are indispensable—and unreliable. Custodians know things about their data that no system log will ever reveal. They also misunderstand, misremember, minimize, and omit. Some of that is innocent. Some of it is not. Much of it reflects a simple desire to get you out of their office and out of their business. ***Trust, but verify.***

Do not become so devoted to checklists that you stop listening. When a custodian claims never to have used a thumb drive, a personal email account, or a home computer for work, do not reflexively move on. *Probe*. Ask how they managed exceptions, emergencies, travel, or after-hours work. Common sense is your most valuable forensic tool.

Avoid leading questions that invite denial. "*You didn't take confidential data home, did you?*" is designed to elicit exactly one answer. Open-ended questions, paired with incentives to disclose, are far more effective. "*It would really help us if you still had your 2020 email archive*" is more likely to yield usable information.

Legacy hardware and media have a way of fading into the background—even when they are right under your nose. Old CPUs become furniture. Boxes of backup tapes become scenery. You must cultivate what Proust called "new eyes." Ask what's in the closet. Ask what's under the desk. Ask to see it. Do not let the blind lead the blind.

> **The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.**
>
> **Marcel Proust**

Accounting records are an often-overlooked resource. Companies amortize and depreciate expensive systems, devices, and software. Those records can point you to servers, laptops, mobile devices, and enterprise platforms that otherwise escape notice.

Three Pressing Points to Ponder

• **Accountability matters.** If someone says "*that's gone,*" record who said it and test the claim. Data mapping is not a group shrug. Someone must own the result. The buck must stop *somewhere*—with *someone* who cares, who has access, and who is willing to get their hands dirty.

• **Where you start matters less than when and with whom.** Start early. Start with the key custodians. Look at offices, closets, servers, and C: drives. Go where the evidence leads!

• **Imperfect does not mean useless.** No data map is ever complete or current. That reality does not excuse sloppiness. Effective data mapping is the bedrock of any sound e-discovery effort.

Quest for E-Discovery: Creating a Corporate Data Map

The following framework offers a practical, business-oriented approach to enterprise data mapping. Treat it as guidance, not gospel.

1. **Identify all systems.** Expect surprises. IT's list is a starting point, not the finish line. Talk to business users, contractors, and vendors. Hidden systems, utilities, and home-grown applications are the rule, not the exception.
2. **Document system details.** Capture ownership, function, data types, formats, dependencies, access controls, retention, purge practices, audit trails, hosting, backups, and archival processes. The goal is not encyclopedic detail but operational understanding.
3. **Inventory business processes.** Map processes to systems to ensure no category of ESI is overlooked. Business processes explain how data is created, used, and transformed.
4. **Identify custodians and roles.** Understand who could do what, when, and why. Privileges matter.
5. **Trace information flows.** Determine where data originates, how it moves, and which systems and people touch it along the way.
6. **Focus on email and messaging.** Email and messaging remain ground zero in most discovery disputes. Understand retention, local archives, personal accounts, and policy compliance before opposing counsel does.
7. **Include collaboration tools.** Slack, Teams, SharePoint, and their cousins are prolific generators of ESI. Unmanaged growth here is a recurring discovery nightmare.
8. **Do not forget offsite and Cloud storage.** Backup tapes, archived media, Cloud repositories and forgotten boxes can outlive their supposed destruction dates. Inventory what exists—even if access will be burdensome. Recognize that the Cloud has grown to be the richest source of ESI in enterprise environments.

What a Data Map Should Look Like

There is no universal template. Industry, size, regulation, geography, and business practices all matter. Still, good data maps tend to share common traits.

- **They are usable.** Organization and clarity matter more than polish.

- **They support change.** Choose formats that allow easy updates. Start simple.
- **They emphasize content quality over column count.** Accuracy beats ornamentation.
- **They are accessible but controlled.** Most users need read-only access.
- **They are maintained.** Ownership, handoffs, and update cycles must be defined.
- **They are used.** A data map that lives in a drawer is a failure. Its value extends beyond litigation to governance, compliance, and operational insight.

A data map is not busywork. It is institutional self-knowledge. In e-discovery, ignorance is expensive. Data mapping is how you stop guessing.

Exercise 13: Mapping your own Data

This is an important exercise. It's the first chance to gauge your level of thought and diligence in mapping an informational footprint. If you cannot fathom your own discoverable corpus of data, how could you assist clients to meet their obligation to identify and preserve their data? Students who have performed poorly on this exercise did so by overgeneralization and lack of exploration. In e-discovery, it's absurd to say the form of data is "electronic." Responding "unknown" respecting form or volume is insufficient if a little digging would prompt a more enlightened response.

Submission at a Glance

• Excel spreadsheet: Complete the 2026 Data Map Template (do not rename columns or change their order).

• Word document (one file, in this order):

1. **Discovery Activity Log** (6–10 entries)
2. **Uncertainty/Assumptions** (three items, with what you tried and what remains unknown)
3. **AI/Tool Disclosure** (if used; 2–5 sentences)

Time guidance: Aim for 2–4 hours. If it takes more than four hours, you are likely overdoing it. If under two hours, you likely missed sources or metrics.

Scenario:
You are the client in this scenario, and you've been sued in federal court. Your lawyer tells you the court has ordered all parties to preserve information, whether on paper or stored electronically. She instructs you to create a list of every source of ESI *"in your custody or possession or subject to your control*" where you've stored information, or others have stored information for you, in the last four years (and any older sources still accessible to you), including every medium you've employed to regularly communicate in writing over the same time period. She adds, "*Don't forget phones and those thumb drive thingies; and be sure to include online stuff and mail, financial data and social networking and work-related data because the other side might subpoena that stuff from third-parties, like your bank and mobile phone company who may have it even if you don't have access anymore. I'm sure the other side will try to prove you missed something, so be very thorough*."

Your lawyer explains that "reasonably accessible" information includes any information that you have in your custody as well as that which you routinely access or use, or that you could access and use. To make your job easier, your lawyer supplies a spreadsheet for your use in helping construct a data map. You protest that you should only have to deal with what's relevant and that's not clear from the claims; but your lawyer is adamant that you should NOT be selective in identifying the full reach of your digital footprint. "*Do your best,*" she adds, "*but remember, this judge is pretty serious*

about e-discovery, and we don't want to lose the case because we failed to list something the other side might find out about later. Don't worry about deciding what's relevant or privileged, that's my job; but I need complete information on the spreadsheet."

Assignment: Complete the spreadsheet (data map) as your lawyer directed.

In addition, **submit:**

A. A short **Discovery Activity Log** (6–10 entries) describing what you checked (device settings, account dashboards, export options, storage usage pages, etc.) and what you learned.
B. A short **Uncertainty/Assumptions** section listing three items you are not fully sure about (form, volume, retention, accessibility), what you did to try to learn more, and what remains unknown.
C. **AI/Tool Disclosure** (if used; 2–5 sentences)

Please note:

1. *This is the scenario. It's about you*, not someone else or your current or prior employer. It's personal. You're not the lawyer here. You can't fire your lawyer or persuade her that, by your reading of the law, her request to you is overbroad and unduly burdensome. Neither can you respond that, without knowing what the case is about, you can't comply. The uncertainty respecting scope and relevance is not an oversight here—it's not a bug, it's a feature. The nature of the request closely parallels the paucity of guidance and lack of restraint commonly seen in practice when lawyers frame legal hold instructions.

2. Once more, you are the client here. You will be bearing the cost of preservation and are spending your own hard-earned cash; so, consider how you should balance the obligation to preserve against the cost of your contemplated method. Unless you're wealthy and wasteful, don't assume that some expert will image everything for you. Also, remember that evidence can be both inculpatory and exculpatory. You may be barred from introducing information that helps you, if you fail to identify it in a timely way.

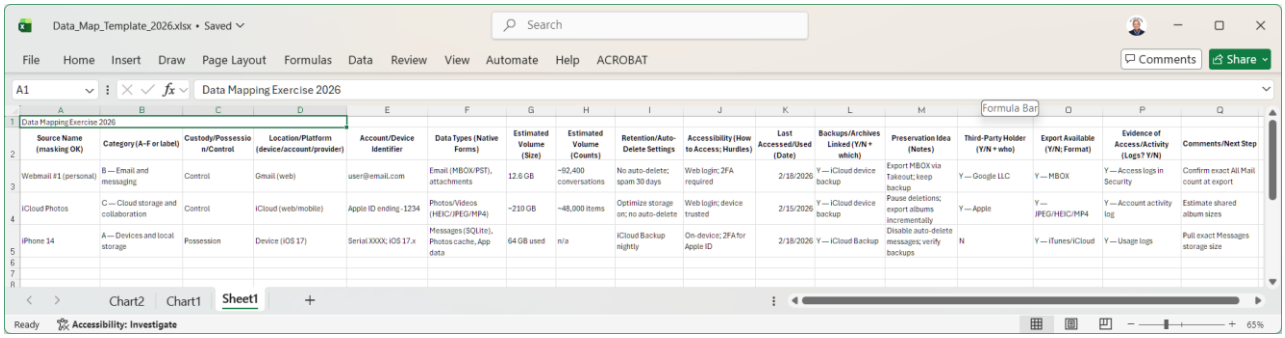
3. The duty to preserve encompasses more than what is in one's custody. It can include material that is subject to your control. So, consider whether you have access to data in another's custody (like your bank or Cloud service provider) or in the care or custody of a person or entity with whom you are in contractual privity or over which you have some practical control (like an attorney, CPA, doctor, family member or close friend).

4. The goal of the exercise is not to invade your privacy. You are mapping your own data because it's easiest for you. Mapping your own data doesn't require you to reach out to others as a lawyer must do. If identifying a genuine source seems too intrusive, feel free to change the name. That is, if you don't want to list that you have a Gmail account, you can call it something like "web mail account #1." If you don't want anyone to know you once

used Facebook, you can call it Social Networking Site 1. X-Box Live can be "Online Gaming Community." Again, the purpose is not to intrude upon private matters but to promote your learning to map data sources accurately, thoroughly and in cogent ways that facilitate meeting obligations to identify, preserve, search and produce evidence in discovery.

5. Different sources demand different solutions; so, don't imagine that all sources can be defensibly preserved by pat solutions like, "I won't delete it" or "I'll have it forensically imaged." Some will. Some won't. Ponder options, consequences, risks and cost. Also, "electronic" or "digital" isn't what we mean by "form" in e-discovery. *What's the native form the data occupies?* Do you even know? Can you find out with a little exploration? If you state "Unknown" for form or volume, you must also state what you did to *try* to determine it and why it remains unknown.

6. **Template:** Use the provided 2026 Excel template spreadsheet downloadable from HERE (web) or HERE (Canvas). ***Please delete the examples supplied in the first three rows before submission.*** Do not rename columns or change their order. Populate every column for each source to the extent reasonably determinable.



7. The time required to complete the assignment will vary depending upon the number and variety of sources and your ability to ascertain the required metrics. If it takes more than four hours, you're overdoing it. If it takes under two hours, chances are you haven't considered all sources or collected enough metrics. You need not consider paper document sources but instead concentrate your efforts on electronically stored information.

8. For this exercise, you are free to seek information from any source so long as you do not delegate the core effort to anyone else. You need not contact other persons (*e.g.,* employers, schools, family) for specific metrics. If you use an AI tool, disclose how (*e.g.,* organizing, proofreading, brainstorming categories). Delegating the actual inventory, estimates, or judgment is not permitted (and is likely infeasible anyway).

9. If you have questions, e-mail them to me at craig@ball.net. Text me if you don't get a response in an hour or two.

10. As you work on this project, please reflect on the pervasiveness and variety of digital information; then, consider what might be required to data map a government agency or a

corporation facing a class action or regulatory inquiry. Observe how little you may know about the nature and extent of data others hold for and about you. Further, be sensitive to any reluctance you feel about disclosing information and the thought and time required to marshal the data. How might such feelings in your clients and their employees impede a thorough and accurate data mapping effort? What strategies might attorneys employ to elicit information and prevent clients from falsely checking "none" on a questionnaire or furnishing incomplete data?

Key Objectives:

1. **Identify All Potential Data Sources:** List every device, online account, and digital storage medium you've used or that contains your data within your custody, possession, or control in the last four years (and older sources *still accessible*). Think beyond the obvious; include less apparent sources like wearable technology, voice-activated home devices, and online platforms used sporadically.

2. **Detail the Types and Volumes of Data:** For each source listed, describe the native forms and types of data stored (*e.g.*, PST/MBOX for email; HEIC/JPEG/MP4 for photos/video; SQLite/JSON for app data) and estimate the volume of data (counts and/or storage size). Where possible, provide specifics (e.g., 5,000 emails; 200 GB of photos).

3. **Examine Online and Third-Party Sources:** Pay particular attention to data held by third parties (*e.g.*, social media platforms, cloud storage services, financial institutions). Indicate where you have practical access or control.

4. **Consider Data Accessibility and Retention**: Record how you access each source (device path or account portal), any retention/auto-delete settings, and hurdles such as fees, delays, or export limits.

5. **Propose Practical Preservation Ideas (in Notes):** Suggest a proportionate way you might preserve each source if needed (*e.g.*, export MBOX, disable auto-delete for 60 days, keep iCloud device backup). This is not a full preservation plan—capture a brief, realistic option.

Tips for Success:

• Thoroughness is Key: Aim to leave no stone unturned in identifying potential data sources. Separate similar sources (e.g., multiple webmail accounts) as distinct entries.

• Detail Matters: Replace vague entries ("photos") with concrete descriptions ("iCloud Photos — HEIC/JPEG/MP4 — ~48k items, 210 GB").

• Count Where Possible: Use device/account storage and usage screens. Capture what you saw in your Activity Log.

• Feasible, Not Fancy: Preservation ideas should be realistic for a paying client—you. Avoid "image everything" as a default.

Assessment Criteria:

• Coverage: Includes common and often-overlooked sources (devices, cloud accounts, messaging, photos, backups, third-party repositories).

• Accuracy and Specificity: Native forms and reasonable volume estimates; retention/access noted.
• Investigative Effort: Activity Log shows real checking (settings, dashboards, export pages).

• Clarity and Organization: Spreadsheet complete and consistent; Word sections concise and in the required order.

• Integrity: Privacy handled responsibly; AI/tool use (if any) disclosed; no delegation of core work.

Additional Deliverables:

Discovery Activity Log (include in your Word file; only 6–10 entries)

For each entry, list Date | Where you looked | What you checked | What you learned. Examples:

• 2026-02-18 | iPhone → Settings → General → iPhone Storage | Photos, Messages storage; iCloud Backup | Photos ~210 GB (~48k items). Messages 14 GB. iCloud Backup enabled; last backup 2026-02-15. • 2026-02-18 | Gmail (web) → Settings → See all settings → Accounts & Import/Labels | All Mail count; export option | ~92,400 conversations; Takeout export (MBOX) available.

• 2026-02-19 | iCloud Photos (web) → Account storage | Library size; formats | 210 GB; formats HEIC/JPEG/MP4; Shared Albums enabled.

• 2026-02-19 | WhatsApp (mobile) → Settings → Storage and Data | Storage usage | 6.1 GB across 8 chats; per-chat exports available.

Uncertainty/Assumptions (include in your Word file; **list three items)** For each, state What is uncertain | Steps you took | What remains unknown | Next reasonable step.

Examples:

• WhatsApp message counts | Checked app storage and export options | App reports GB but not message counts | Next: Export two largest threads to estimate counts.

• 2020 external HDD | Drive powers intermittently; Windows shows 480 GB used | Directory visible; files fail to open | Next: Attempt read-only mount on another PC; if still inaccessible, mark "not reasonably accessible without specialist."

AI/Tool Disclosure (if used; 2–5 sentences) Briefly state: • What you asked AI/tools to do (e.g., brainstorm categories; wording cleanup; where to find counts). • What decisions, estimates, and entries are yours. • Two examples of items you added or quantified based on your own checking.

Template Spreadsheet (2026 Data Map Template — columns and sample row) Note: Use the provided downloadable template. If the link is unavailable or won't load into Excel for you, you may copy these columns into a new Excel workbook (do not change column names or order).

Column order (left to right):

1. Source Name (masking OK)
2. Category (A–F or label)
3. Custody/Possession/Control
4. Location/Platform (device/account/provider)
5. Account/Device Identifier
6. Data Types (Native Forms)
7. Estimated Volume (Size)
8. Estimated Volume (Counts)
9. Retention/Auto-Delete Settings
10. Accessibility (How to Access; Hurdles)
11. Last Accessed/Used (Date)
12. Backups/Archives Linked (Y/N + which)
13. Preservation Idea (Notes)
14. Third-Party Holder (Y/N + who)
15. Export Available (Y/N; Format)
16. Evidence of Access/Activity (Logs? Y/N)
17. Comments/Next Step

Accepted category labels for "Category" column:

A. Devices and local storage
B. Email and messaging
C. Cloud storage and collaboration
D. Social media and web accounts
E. Work/business systems
F. Backups and archives

Reminder: Do not rename columns or change their order. Populate each field to the extent reasonably determinable. Where "Unknown" is unavoidable, record your investigative steps in the Activity Log and note why it remains unknown.

Mastering E-Mail in Discovery

Introduction

Get the e-mail! It's long been the war cry in e-discovery, a recognition of e-mail's enduring importance and ubiquity even as today's matters also implicate chat, mobile messaging, and collaboration platforms that must be considered alongside mail for a complete record. We pursue e-mail because it still captures much of business life and because users continue to let their guards down, committing plainspoken truths to messages they'd never memorialize in a formal memorandum.

If you're on the producing end, you worry about what the messages say and whether you can find, preserve, and produce all responsive items across cloud tenants and any residual local stores without gaps. Questions like these should keep you up nights:

- *Will clients conceal damning messages?*
- *Will employees delete "their" mail?*
- *Will searches reach the right venues?*
- *Will review unwittingly betray privilege?*

Those same questions persist in 2026, but the venues now include Microsoft 365 and Google Workspace alongside devices and webmail, so your plan must address each with clarity.

Meeting these challenges begins with understanding e-mail technology well enough to craft a sound, defensible strategy and, as a requesting party, to assess your opponent's efforts with specificity. In practice, that means learning how modern tenants store messages, how legal hold and export work, and where local caches or legacy stores still lurk, so you can explain scope, preservation, search, and production in plain English to the court. E-mail remains a cornerstone of discovery, but it now shares the stage with system-generated links, shared locations, and message threads spread across services; mapping those relationships early prevents gaps and surprises later.

Not Enough Eyeballs

"*Any sufficiently advanced technology is indistinguishable from magic*," Arthur C. Clarke wrote; but *"it's magic to me, Your Honor*," won't help when messages pull a disappearing act, because judges expect counsel to understand the systems well enough to make them yield complete and reliable results. Lawyers must do more than tell clients to "produce the e-mail;" they must understand systems and procedures, ask the right people the right questions, and recognize when answers aren't trustworthy, because virtually all business records are born digitally and only a fraction are ever printed. Volumes are enormous, user behavior is imperfect, and retention is sticky—mail persists across servers, devices, webmail, and cloud—so a workable plan aligns legal duties with where mail actually lives and how it's managed today. Most users and many IT professionals still

don't know precisely where messages lodge or how long they linger, which is why discovery planning must translate architecture into practical steps for preservation, search, and collection.

A lawyer managing electronic discovery must understand enough to frame scope and methods credibly with the other side and the court, not delegate core obligations to support personnel or assume legacy habits endure in modern, cloud-first environments. The rules don't demand the impossible; they demand diligence, reasonableness, and good faith—acting swiftly, following a sensible plan, and communicating openly when you need agreement or protection—because courts distinguish error from arrogance. The through line from the paper era to today's cloud tenants is unchanged: know your client's environment, identify the right sources, preserve them effectively, search them competently, review them prudently, and produce them in forms that preserve content and context.

Test Your E.Q.

Suppose opposing counsel serves a preservation demand or the court orders preservation of electronic messaging. Are you confident your client can and will faithfully preserve responsive data across cloud tenants, endpoints, and any residual local stores? Even if it's practicable to preserve the current mailboxes of key custodians, are you really capturing most of the discoverable communications-or are critical items living in chats, shared locations, or personal accounts outside your net? How much is falling through, and how will you assess and explain its significance?

Here are a dozen questions a lawyer *should* be able to answer about a client's mail systems:

1. What messaging environments does the client use for business communications—Microsoft 365 Exchange Online and/or Google Workspace Gmail—and are there any secondary or legacy systems still in play (*e.g.,* on-prem Exchange) that retain historical mail for key custodians?
2. Do all discoverable communications flow through corporate channels, or do employees also use chat (Teams/Slack), SMS/iMessage, WhatsApp, or other apps for business? If so, are policies and controls in place to preserve those sources alongside e-mail?
3. How is retention configured in the tenant? For Microsoft 365, what retention policies/labels and litigation holds apply to mailboxes; for Google Workspace, what Google Vault retention rules and holds are in force for users and organizational units?
4. Can the client disable or override user deletion for specific custodians (*e.g.*, Purview litigation hold or Google Vault hold) and confirm when such holds were applied?
5. Are "linked attachments" used (files shared via OneDrive/SharePoint or Google Drive links instead of being embedded), and how will those items be identified, preserved, collected, and produced alongside their parent messages?
6. Do custodians sync mail locally (e.g., Outlook OST/PST, Apple Mail), and where do those local stores reside on laptops/desktops? Are there controls to prevent client-side purge or tampering once a hold is in place?

7. Are there shared or delegated mailboxes, distribution lists, or service accounts used for business functions, and are they in scope for preservation and search?
8. Do employees access corporate mail from personal devices (BYOD) or personal accounts, and what mobile/device management exists to enforce policy, containerize data, and support defensible collection when needed?
9. What backup/journaling/archiving solutions (if any) exist today, and do they capture complete, policy-compliant copies of messages and metadata, including from departed users or deprovisioned mailboxes?
10. How effective are native search tools for your needs, including limits on query syntax, time ranges, partial indexing, encrypted items, and throttling/quotas? What is the plan to validate search completeness?
11. What export forms will be used (*e.g.,* PST/EML/MBOX), and how will threading, deduplication, time zones, and language/encoding be handled to preserve content and context?
12. What policies govern use of personal webmail or side channels for business, how are violations detected/remediated, and how will potentially responsive items from those sources be addressed?

Despite decades of dealing with e-mail in discovery, most lawyers can't answer many of these questions. But these <u>are</u> lawyer issues when lawyers meet and confer, negotiate scope, explain preservation, or defend processes to the court. So, lawyers must own these questions, get plain-English answers, and align legal duties with where communications actually live and how they are managed today.

Staying Out of Trouble

Fortunately, the rules of discovery don't require the impossible; they require diligence, reasonableness, and good faith, demonstrated by acting swiftly, following a sound plan, and taking steps that reasonable minds would judge adequate to the task. Judges expect open lines of communication with the other side and the court, seeking agreement where it helps and protection where it's needed; as Justice Holmes quipped, "Even a dog knows the difference between being stumbled over and being kicked," and courts distinguish error from arrogance when evaluating sanctions.

Paper discovery was self-limiting: physical files, folders, and cabinets organized by person, topic, transaction, or time, plus the cost and space of paper, pushed businesses to cull or avoid creating records. Electronic communications flipped the script: ephemerality, ease of duplication and distribution, and the low direct cost of storage fueled staggering growth in electronic evidence. At roughly 150 e-mails per day, a 100,000-employee company may store nearly 4.5 billion e-mails in a year—a scale that reshapes preservation and search from the outset.

Volume is only part of the challenge because a single user's e-mail is typically stored in massive data blobs housing tens of thousands of messages and attachments, with loose subject lines, drifting threads, and forwarded detritus that obscure topical boundaries. Responding to a request for production can demand skimming vast message sets or naively relying on keywords; if older items live off current systems, counsel can be forced into rummaging through archival folders, old machines and disks, backup tapes without readers, and unlabeled media—an exercise in frustration and risk.

And that's just for one person. At enterprise scale, responding to "all e-mail" invites a tedious, time-consuming trek across far-flung servers and local drives, unless counsel secures a better way through scope, workflow, and validation. Seasoned practitioners report there's no magic mousetrap: enterprise-wide efforts are gargantuan and uncertain, so narrowing by sampling, analytics, time span, offices, sectors, and key players is essential; trying to capture everything enterprise-wide is trawling with a net full of holes.

The market's response has included enterprise search across remote and local stores and enterprise e-mail archiving that congregates messages in a single, searchable repository with deduplication, journaling, and other features, plus advanced analytics (TAR/predictive coding, and now AI) to cut cost and time; but these remain costly and complex, lack dominant standards or a single leader, and risk turning serial litigants into "keep everything" shops increasingly dependent on fraught electronic search to separate wheat from chaff.

New Tools

The market has responded in recent years with tools that either facilitate search of remote e-mail stores, including locally stored messages, from a central location (*i.e.,* enterprise search) or which agglomerate enterprise-wide collections of e-mail into a single, searchable repository (*i.e.,* e-mail archiving), often reducing the volume of stored data by so-called "single instance deduplication," rules-based journaling and other customizable features.

These tools, especially enterprise archival and advanced analytics termed "TAR" or "Predictive Coding," promise to make it easier, cheaper and faster to search and collect responsive e-mail, but they're costly and complex to implement. Neither established standards nor a leading product has emerged. Further, it remains to be seen whether the practical result of a serial litigant employing an e-mail archival system is that they—for all intents and purposes--end up keeping every message for every employee and becoming increasingly dependent upon fraught electronic search to cull wheat from chaff.

E-Mail Systems and Files

In contemporary enterprises, e-mail predominantly lives in cloud tenants rather than in server rooms. The dominant business platforms are **Microsoft 365** (Exchange Online with Outlook as the primary client) and **Google Workspace** (Gmail with the web client and supported desktop/mobile apps). On-premises Microsoft Exchange and IBM/HCL Domino still exist—often for historical mail, niche workflows, or tightly regulated environments-but they are the exception, not the rule. Truly legacy platforms like Novell GroupWise are exceedingly rare and typically encountered only in long-lived public-sector deployments or as remnants preserved for compliance. What matters for discovery is locating which tenant(s) and any legacy servers hold each custodian's mail and understanding the administrative levers those systems provide for hold, search, and export.

Outside large organizations, most users rely on webmail and ISP-provided accounts, with Gmail and Outlook.com the most common. Even in small offices that once ran local servers, cloud mail is now the norm because it is easier to administer, more accessible from anywhere, and bundled with collaboration tools. Client software remains diverse-Outlook, Apple Mail, mobile mail apps, and web browsers-but these front ends increasingly act as windows onto server-side mailboxes rather than as the systems of record. That shift is critical in discovery: the authoritative copy is almost always in the tenant's mailbox, not the laptop's cache, even though endpoint artifacts may still be relevant for timelines, deleted items, or offline folders.

Mail Protocols

An *application protocol* or API is a bit of computer code that facilitates communication between applications, *i.e.,* your e-mail client and a network like the Internet. When you send a snail mail letter, the U.S. Postal Service's "protocol" dictates that you place the contents of your message in an envelope of certain dimensions, seal it, add a defined complement of address information and affix postage to the upper right-hand corner of the envelope. Only then can you transmit the letter through the Postal Service's network of post offices, delivery vehicles and postal carriers. Omit the address, the envelope or the postage and Grandma gets no birthday card! Likewise, computer networks rely upon protocols to facilitate the transmission of information. You invoke a protocol—*Hyper Text Transfer Protocol*—every time you type *http://* at the start of a web page address.

Common protocols matter today chiefly because they signal where evidence resides and how it was created. At a high level, mail is transmitted using SMTP between servers, while end-user access *historically* used POP3 or IMAP for retrieval. Microsoft's ecosystem layered MAPI (and, for mobile, Exchange ActiveSync) to support rich Outlook features and offline files (PST/OST). In cloud platforms, HTTP/HTTPS APIs now power most client interactions. For e-discovery, you rarely need to recite protocol acronyms, but you need to recognize whether messages are server-resident with optional synchronized local copies, or whether a legacy workflow left the only copy on a user's hard drive.

POP3 (Post Office Protocol 3) deserves only a short word today. It was once common for ISPs and small offices: the client downloaded messages and, unless configured to leave copies, the server deleted them. That model created scattered, local-only collections in desktop mail stores and backups. You still see POP3 artifacts in older PSTs and on retired machines, so you cannot ignore it; but you should treat POP3 as a legacy edge case that signals "look to endpoints and old backups," not a primary enterprise source.

IMAP, by contrast, keeps the authoritative mailbox on the server and mirrors foldering and state to clients. Many non-Microsoft business accounts, universities, and older hosted providers used IMAP for years, and traces persist in historical accounts and archives. Discovery value comes from knowing that the server retained both messages and user folder structure, while laptops may hold synchronized local copies that differ in subtle ways.

MAPI/Exchange and today's cloud APIs follow the same core idea: server-resident mailboxes with optional local synchronization, so your default preservation and collection posture should be server-first via tenant-native tools, with targeted endpoint collections when there is a specific need (*e.g*., suspected local-only stores, offline PSTs/OSTs, or forensics).

Finally, browser-based mail **(HTTP/webmail**) is the default user experience for many custodians, especially in Google Workspace. Although users may still enable POP/IMAP in personal contexts, the authoritative data for business accounts resides in the cloud mailbox governed by enterprise retention and legal hold. That reality simplifies scoping: you start by identifying the tenant, the custodial mailbox, and the applicable retention/hold controls; then determine whether any local stores or legacy servers exist for the period at issue. Backup tapes once loomed large in disputes over cost and accessibility; in cloud tenants, the analogous questions focus on retention policies, litigation holds, recoverable items folders, and whether journals or third-party archives were used during the relevant time. Handle the tenant first; treat endpoint caches and truly legacy protocols as supplements guided by facts, not as default hunting grounds.

Outgoing Mail: SMTP and MTA

Just as the pipes that bring water into your home differ from the ones that carry wastewater away, the system that delivers mail to you is distinct from the one that transmits outgoing messages. Everything above addressed how users *retrieve* e-mail; now we turn to how messages are *sent*.

The plumbing for sending is called **SMTP**, short for **Simple Mail Transfer Protoco**l. "Protocol" just means a set of rules. In plain English: SMTP is the rulebook mail servers follow to hand your message from place to place until it reaches the recipient. Your device (Outlook, Apple Mail, or a web page) gives the message to a sending server called an MTA—**Message Transfer Agent**—which relays the

message to the next server, and so on, until it reaches the recipient's server. In business cloud systems like Microsoft 365 and Google Workspace provide these sending servers behind the scenes. As a message travels, each server it passes through stamps the envelope with a "Received" line (you can see these in the full message headers). Those stamps form a simple timeline of where the message was and when. In discovery, those headers help you answer: Was it sent? When? By whom? Through which system? They also help spot oddities (for example, a time zone mismatch or an unexpected detour in the plumbing).

You may also see terms like SPF, DKIM, and DMARC. These are anti-spoofing checks that help receiving systems tell whether "the sender really is who they claim to be" and whether the content changed in transit. They are helpful signals, not perfect proof; treat them as context you can read, not as lie detectors.

Cloud services add security layers along the way (for example, Microsoft's Exchange Online Protection or Google's security filters). These may scan links or add warning banners. That can make what the recipient sees look slightly different (a rewritten link or a notice at the top), but the underlying message content and the headers are still what matter for legal purposes. When collection accuracy is critical, pull from the authoritative source, *viz,* the sender's and recipient's mailboxes in the cloud tenant.

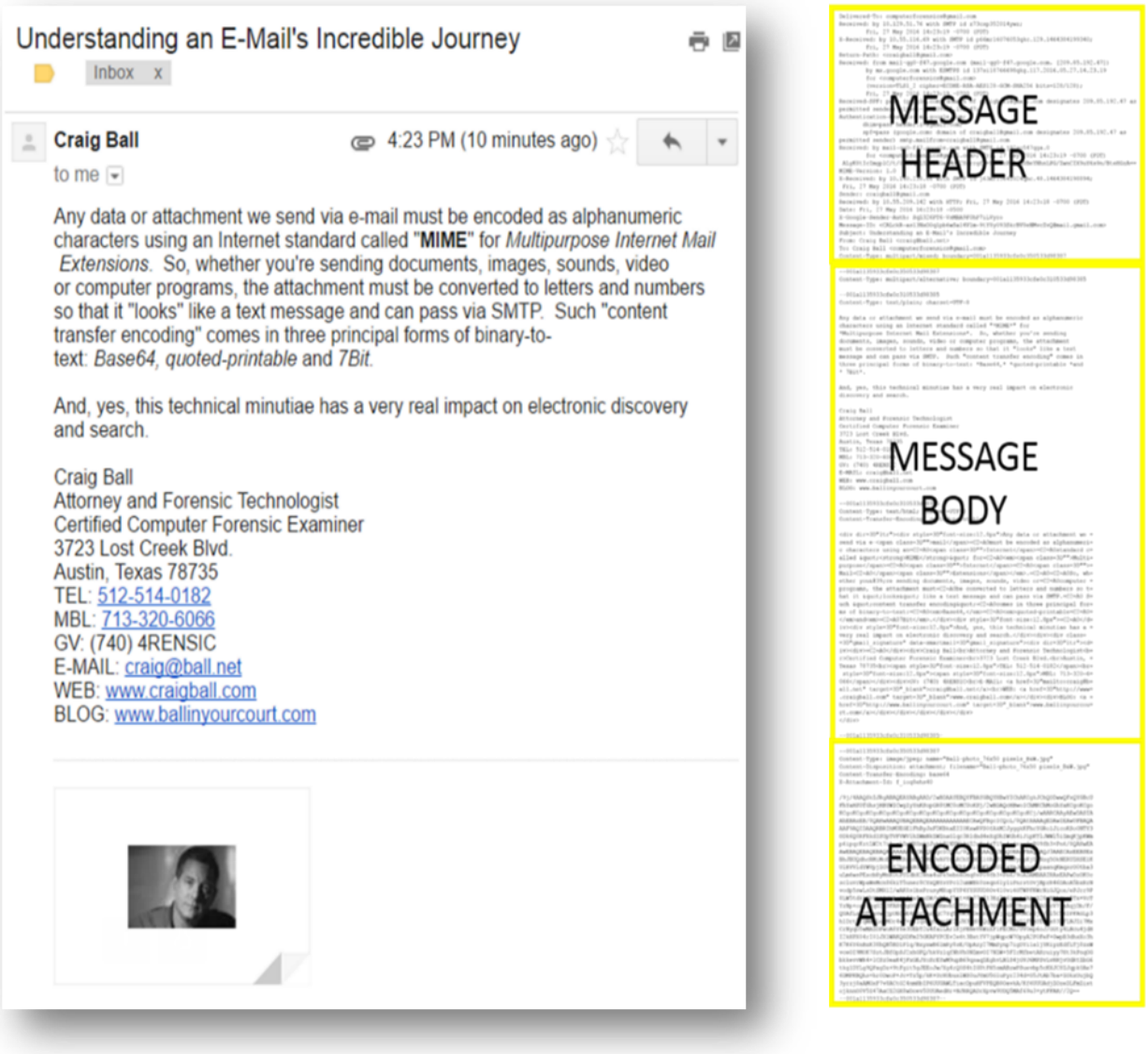
Anatomy of an E-Mail

Now that we've navigated the alphabet soup of protocols, let's look inside a single message. Considering the complex systems around it, an e-mail is surprisingly simple in structure. Internet standards (the RFCs) require rigid formatting, which makes a message easy to dissect once you know where to look. Most of the headaches arrive later-when messages pile up in large mailboxes, are copied to local stores, or are folded into archives and backups.

An e-mail is, at heart, a plain-text document. Even when it carries non-text files—a Word document, a PDF, a photo, a spreadsheet—those files are converted into text for the trip and then converted back on arrival. That's why we can say that **everything in an e-mail is text** during transmission—even the "binary" parts like images and attachments. This matters in discovery because what looks like a separate attachment in your mail app is actually text embedded in the message source (or may be linked as a so-called Modern Attachment).

E-Mail Autopsy: Tracing a Message's Incredible Journey

The leftmost image below is an e-mail I sent to computerforensics@gmail.com from my alias craig@ball.net using my Gmail account craigball@gmail.com. A tiny JPG photograph was attached. A user might see the e-mail below left and mistakenly assume that what they see is all there is. Far from it! The rightmost image in the image below contains the source code of the e-mail.[87] Viewed



in its "true" and complete format, it's too long to legibly appear on one page. So, let's dissect it by looking at its constituent parts: message header, message body, and encoded attachment.

[87] While viewing a Gmail message, you can display the source code for a message by selecting "Show original" from the message options drop-down menu. By default, Outlook makes only some encoded header content readily viewable at message Properties—the complete source code of incoming e-mail is not recorded absent a system Registry edit, which is not a casual operation!

In the e-mail's header (pictured below), each line beginning with "Received" or "X-Received" represents a handoff between two e-mail servers. The sequence is reverse-chronological: lines



closest to the top were added last, and the very top line reflects delivery to the recipient's e-mail service and account-in this instance, computerforensics@gmail.com. As the message passes through intervening hosts, each adds identifying information along with the date and time of transit. Those "Received" lines become a timeline you can read.

The area of the header labeled **(A)** contains the human-facing fields: sender, addressee, date, time, and subject. These are the only header fields most recipients ever see. Note that a 24-hour message time may be recast as a 12-hour format in Gmail's display.

In the line labeled "Date," you'll see both the date and time of transmittal plus a time-zone offset. Here, the time is 16:23:18, and the "-0500" that follows denotes the difference between the sender's local time and Coordinated Universal Time (UTC). On May 27, 2016, "-0500" corresponds to U.S. Central Daylight Time, which suggests the sending machine's time zone. Time and time-zone details can occasionally help test claims of fabrication or align events across systems.

E-mail must adhere to structural conventions. One is the use of a Content-Type declaration and clearly defined content boundaries that separate the header from the body and attachments. The line labeled **(B)** advises that the message is "multipart/mixed," which simply means "this message has multiple parts"-for example, a body plus one or more attachments-and that different parts may be encoded differently. To keep those parts straight, the system generates a unique boundary string. The first boundary here-"001a1135933cfe0c350533d98387"-separates the message header from the sections that follow and also marks the end of the message when it appears with a closing indicator.

The message was created and sent using the Gmail web interface; consequently, the first hop **(C)** shows the message was submitted over HTTP and first received by IP address 10.55.209.142 at 14:23:18 -0700 (PDT). The server marks time in Pacific Daylight Time, suggesting a west-coast data center. The message is immediately handed off to another IP address, 10.140.238.66, using Simple Mail Transfer Protocol (SMTP). Next comes another SMTP handoff to Google's server "mail-qg0-f47.google.com," and so on until delivery to my account, computerforensics@gmail.com.

In the line labeled **(D),** the header declares "MIME-Version: 1.0," indicating the message uses MIME formatting.[88] Ironically, no other MIME version has ever been used; nevertheless, this line appears in essentially every modern message.

[88] MIME, which stands for Multipurpose Internet Mail Extensions, is a seminal Internet standard that supports non-US/ASCII character sets, non-text attachments (e.g., photos, video, sounds and machine code) and message bodies with multiple parts. Virtually all e-mail today is transmitted in MIME format.

Proceeding to the message body, at line **(E)** (right) we encounter our first boundary value ("--001a1135933cfe0c350533d98387"), which marks the transition from header to body. At line **(F),** another Content-Type declaration advises that this segment is "multipart/alternative," meaning the same message appears in two forms-plain text and HTML-so any mail program can display at

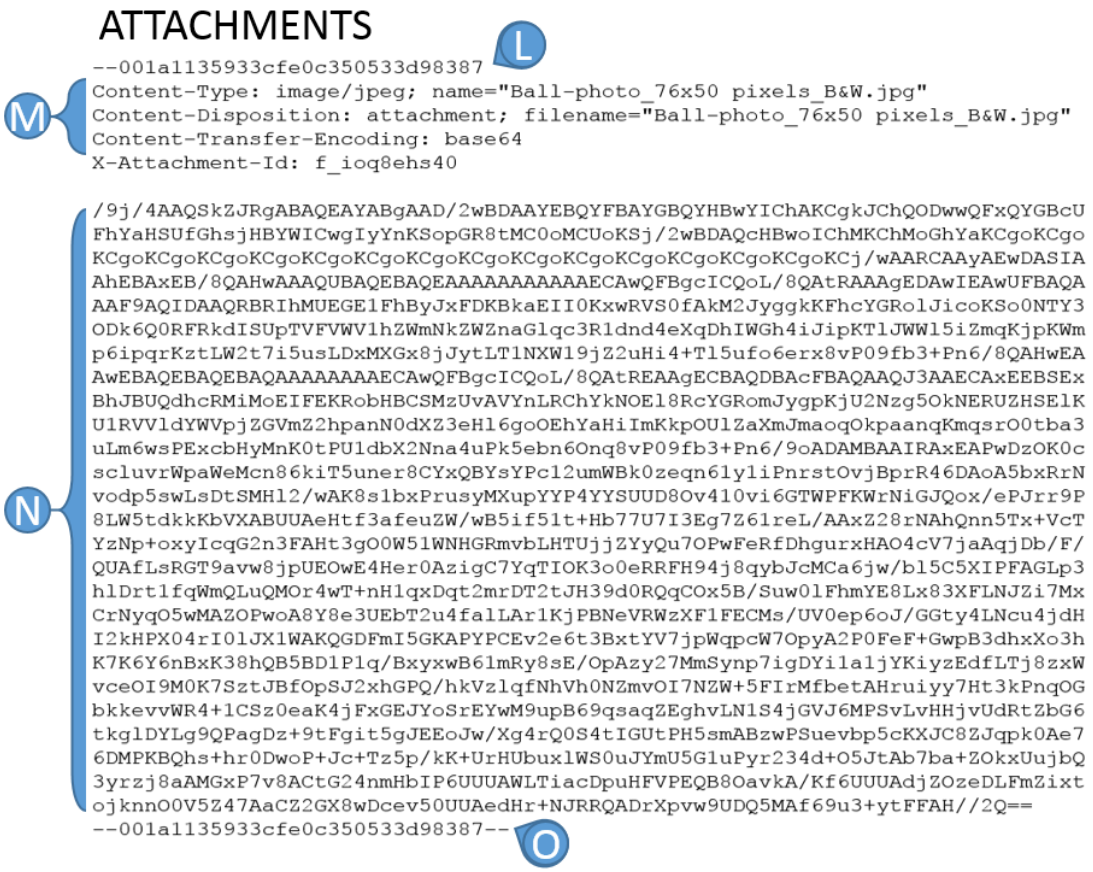
MESSAGE BODY

```
--001a1135933cfe0c350533d98387 (E)
Content-Type: multipart/alternative; boundary=001a1135933cfe0c310533d98385 (F)

--001a1135933cfe0c310533d98385 (G)
Content-Type: text/plain; charset=UTF-8 (H)

Any data or attachment we send via e-mail must be encoded as alphanumeric
characters using an Internet standard called "*MIME*" for
*Multipurpose Internet Mail Extensions*.  So, whether you're sending
documents, images, sounds, video or computer programs, the attachment
must be converted to letters and numbers so that it "looks" like a text
message and can pass via SMTP.  Such "content transfer encoding" comes in
three principal forms of binary-to-text: *Base64,* *quoted-printable *and
* 7Bit*.

And, yes, this technical minutiae has a very real impact on electronic
discovery and search.

Craig Ball
Attorney and Forensic Technologist
Certified Computer Forensic Examiner
3723 Lost Creek Blvd.
Austin, Texas 78735
TEL: 512-514-0182
MBL: 713-320-6066
GV: (740) 4RENSIC
E-MAIL: craig@ball.net
WEB: www.craigball.com
BLOG: www.ballinyourcourt.com

--001a1135933cfe0c310533d98385 (I)
Content-Type: text/html; charset=UTF-8 (J)
Content-Transfer-Encoding: quoted-printable

<div dir=3D"ltr"><div style=3D"font-size:12.8px">Any data or attachment we =
send via e-<span class=3D"">mail</span>=C2=A0must be encoded as alphanumeri=
c characters using an=C2=A0<span class=3D"">Internet</span>=C2=A0standard c=
alled &amp;quot;<strong>MIME</strong>&amp;quot; for=C2=A0<em><span class=3D"">Multi=
purpose</span>=C2=A0<span class=3D"">Internet</span>=C2=A0<span class=3D"">=
Mail=C2=A0</span><span class=3D"">Extensions</span></em>.=C2=A0=C2=A0So, wh=
ether you&amp;#39;re sending documents, images, sounds, video or=C2=A0computer =
programs, the attachment must=C2=A0be converted to letters and numbers so t=
hat it &amp;quot;looks&amp;quot; like a text message and can pass via SMTP.=C2=A0 S=
uch &amp;quot;content transfer encoding&amp;quot;=C2=A0comes in three principal for=
ms of binary-to-text:=C2=A0<em>Base64,</em>=C2=A0<em>quoted-printable=C2=A0=
</em>and<em>=C2=A07Bit</em>.</div><div style=3D"font-size:12.8px">=C2=A0</d=
iv><div style=3D"font-size:12.8px">And, yes, this technical minutiae has a =
very real impact on electronic discovery and search.</div><div><div class=
=3D"gmail_signature" data-smartmail=3D"gmail_signature"><div dir=3D"ltr"><d=
iv><div>=C2=A0</div><div>Craig Ball<br>Attorney and Forensic Technologist<b=
r>Certified Computer Forensic Examiner<br>3723 Lost Creek Blvd.<br>Austin, =
Texas 78735<br><span style=3D"font-size:12.8px">TEL: 512-514-0182</span><br=
 style=3D"font-size:12.8px"><span style=3D"font-size:12.8px">MBL: 713-320-6=
066</span></div><div>GV: (740) 4RENSIC<br>E-MAIL: <a href=3D"mailto:craig@b=
all.net" target=3D"_blank">craig@ball.net</a><br>WEB: <a href=3D"http://www=
.craigball.com" target=3D"_blank">www.craigball.com</a></div><div>BLOG: <a =
href=3D"http://www.ballinyourcourt.com" target=3D"_blank">www.ballinyourcou=
rt.com</a></div></div></div></div></div>
</div>

--001a1135933cfe0c310533d98385-- (K)
```

least one version. A second boundary string-“001a1135933cfe0c350533d98385”-is declared for those alternatives. Note that the first boundary ends in 387 and the second in 385. The second boundary is used at **(G)** to mark the start of the first alternative message body, declared as text/plain at **(H).**

We then see the second boundary used at **(I)** to denote the start of the second alternative message body, with Content-Type declared as text/html at **(J).** The second boundary appears again to signal the conclusion of the multipart/alternative content. You did not draft separate plain-text and HTML versions-Gmail created the pair automatically for compatibility. The HTML version carries formatting-bold, italics, hyperlinks-while the plain-text version carries the same words without decoration.

Turning to the last segment of the message, we see at **(L)** the transition from the body to the attachment section, marked again by our “387” boundary.



At **(M),** another Content-Type declaration appears-now “image/jpeg,” which is common for digital photos. The “name” parameter identifies the encoded item, and “Content-Disposition” designates

how the item should be handled on delivery; here, as an attachment to be assigned the same filename when decoded. But where is the JPG photo?

Recall that to travel as an e-mail attachment, binary content (like photos, sound files, video, or machine code) must be converted into printable text characters. The photograph has been encoded as **base64**, which substitutes 64 printable ASCII characters (A-Z, a-z, 0-9, + and /) for any binary data or for characters outside plain ASCII (such as Cyrillic or Chinese) that can be mapped to this set.[89] Note the declaration in **(M),** "Content-Transfer-Encoding: base64."

Accordingly, the attached JPEG with the filename "Ball-photo_76x50 pixels_B&W.jpg" has been transformed from non-printable binary into 26 lines of seemingly random text-nearly 2,000 plain-text characters-labeled **(N).** That text can traverse mail systems safely and is converted back to the original binary file when the message is received.



Finally, the message concludes with the original boundary string at **(O)** accompanied by a closing marker, indicating there are no further parts.

The lesson is that what you see in your mail client isn't the whole message. It's an arrangement of selected parts of the message-often modified by the client-and frequently supplemented by metadata unique to your software and setup. What appears as a tidy paperclip icon is actually text encoded into the message body. The displayed time is usually adjusted to your device's time zone and daylight-saving settings. Even the sender's name may be altered by your address book. For proof of routing, timing, recipients, and content, consult the message source and headers-the canonical record of what actually traveled.

Hashing and Deduplication

The ability to "fingerprint" data using hash algorithms makes it possible to identify identical files without opening or reading them. If the hash values of two files are identical, the files are

[89] A third common transfer encoding is **quoted-printable**, or **QP encoding**. It facilitates the transfer of non-ASCII 8-bit data over 7-bit channels by representing such bytes as three ASCII characters—an equals sign (=) followed by two hexadecimal characters (0–9, A–F). Quoted-printable is typically employed where the content is predominantly ASCII text with occasional non-ASCII characters. Its principal advantage is that the encoded data remains largely intelligible to human readers, unlike more opaque encodings such as Base64

identical. As discussed earlier, this powers deduplication—culling duplicate items before review to save money and reduce the risk of inconsistent privilege and responsiveness decisions.

Hashing is powerful, but it is also unforgiving. As a prior exercise demonstrated, the tiniest change in a file—even a single character — yields a completely different hash value. That brittleness limits straightforward "whole-message" hashing for e-mail because, as our e-mail autopsy revealed, messages carry unique IDs, timestamps, and routing data that legitimately differ between copies. Look at the message as a whole and you find that multiple recipients of the "same" message hold versions with different hash values.

So, to make deduplication work for e-mail, we hash only the parts that matter for legal review—for example, selected header fields (From, To, Subject, Date), the message body, and the encoded attachments—while omitting volatile routing and system fields. When those selected segments match, the messages are practically identical for our purposes even if the full "wire copy" differs in harmless ways.

Selective hashing also allows "near deduplication," which groups messages that are substantially the same but not byte-for-byte identical—say, a thread that differs only by a footer, an auto-added banner, or trivial format changes. Near deduplication can cut review cost and fatigue; but it is essential to understand what your process is excluding and why, and to tell your opponent you are using it so there is no surprise about what was suppressed as "near-duplicate."

EDRM Message Identification Hash (MIH) Cross-Platform Duplicate Identification
https://edrm.net/edrm-projects/dupeid-2/

What it is. MIH is a standard way to tag the "same" e-mail across different systems and formats—whether you see it as a native message, a PDF/TIFF image, a PST, or an MBOX. If two messages share the same MIH, you can likely treat them as duplicates.

Why it matters. Different software can process the same e-mail slightly differently, which can defeat ordinary duplicate checks. MIH gives you a common, portable duplicate tag that survives those differences. In productions, MIH is carried in the **load file** as a simple metadata value alongside each document—think of a small spreadsheet-like table that lists metadata fields (From, To, Date, etc.) for each item.

How it works. Every e-mail has a standard header value called the Message-ID. MIH takes a normalized version of that value and hashes it in a consistent way. The result is a stable ID: the same message yields the same MIH no matter where it appears. Matching MIH values signal duplicates; non-matches deserve review.

How to use it. In your request for production or protocol, ask parties and vendors to include MIH in the load file for e-mails. With MIH, you can avoid re-reviewing the same message across rolling productions or between parties, compare different production formats without reprocessing, and safely reuse prior coding. MIH complements—not replaces—normal deduplication inside a case.

Hash deduplication of e-mail is tricky in practice. Time values may vary, attachment order may flip, and small formatting differences can defeat duplicate detection. If expected duplicates are not deduplicating, stop and investigate — those misses may signal a broader processing issue, a normalization gap, or mixed sources that need tailored handling.

Local E-Mail Storage Formats and Locations

Suppose you face a discovery request for a client's e-mail and there's little time or budget to secure expert help. Where do you look in 2026—and what forms will you encounter? The old answer—"it's on the company's mail server"—is incomplete. Today, the authoritative copy of business e-mail usually lives in a cloud tenant, and meaningful copies, caches, and strays can exist on endpoints and in older systems. A tenant is simply the organization's cloud account — the fenced-off slice of Microsoft 365 or Google Workspace that holds its users, settings, and data. The true location and extent of mail depends on how that tenant is configured, what the retention policies say, user habits, mobile access, and past migrations.

Start with the tenant—that is almost always the system of record. In Microsoft 365, custodian mailboxes live in **Exchange Online**, which is Microsoft's cloud e-mail server. Each mailbox may also have an **Online Archive** (a server-side "older mail" area) and a **Recoverable Items** store (a protected area that keeps copies of items even after users delete them). **Microsoft Purview eDiscovery** is Microsoft's built-in legal toolkit—it lets authorized personnel place legal holds, run searches, review results, and export data without moving it off the platform.

In Google Workspace, user mailboxes live in Gmail, and **Google Vault** is Google's built-in legal toolkit for retention, legal hold, search, and export. In both ecosystems, "**linked attachments**" are common—instead of embedding a file, the message includes a link to a file stored in OneDrive or SharePoint (Microsoft) or Google Drive (Google). When those links are in scope, treat the e-mail and the linked file as a pair (a "family")—identify the link target, capture the right version of the file, and preserve the relationship to the parent message.

Next, check for secondary or legacy systems that still matter. Some organizations retain on-premises Microsoft Exchange or HCL (*AKA* Lotus *AKA* IBM) Domino for historical mail, legal holds, or niche workflows. Others have third-party archives or **journaling systems** that hold copies from earlier years or for departed employees. Journaling copies every inbound and outbound message to a separate, locked mailbox or archive—creating an independent record that users cannot alter—so confirm whether journaling was enabled during the period at issue and where those journal copies reside today. **Migrations** (shifts of data from old systems to a newer ones) leave footprints—ask whether mail moved between platforms and where migration exports or "frozen" mailboxes now reside.

Endpoints (a fancy word for the user's computer) still count, but differently than before. Outlook typically keeps a synchronized cache called an OST on Windows and a profile database on macOS; Apple Mail and other clients maintain similar caches. These reflect the server mailbox and may include offline folders, user-made archives, or strays from past settings. PST files—portable Outlook data files—persist despite corporate efforts to eliminate them, often created by users to free space or move mail. Laptops and home computers can also harbor exports from personal accounts. Treat endpoint collections as targeted supplements, useful for local-only stores, historical PSTs, or timeline artifacts, while keeping preservation centered on the tenant.

Personal and non-corporate accounts remain a risk. Employees sometimes forward work messages to personal Gmail or iCloud, or conduct business in side channels. Policies and **mobile-device management** (MDM) reduce this, but *you still must ask* and, when justified, *account* for those sources. Likewise, mobile devices usually present a synchronized view of server mail—they rarely hold unique, authoritative copies of modern business e-mail, but they *can* contain cached content, attachments saved to device storage, and evidence of use that matters in specific scenarios.

Backups are not archives. In cloud platforms, disaster-recovery backups are tenant-level and short-lived; discovery typically relies on retention policies, legal holds, and the Recoverable Items (Microsoft) or Vault holds (Google), not on restoring backups. For legacy environments, tape or snapshot backups may exist, but accessibility, cost, and proportionality govern whether you must reach them. Message-tracking logs in mail gateways and tenants can corroborate sending and delivery, but they expire quickly—if delivery is disputed, ask early and preserve before logs roll off.

To recap, a practical search plan in 2026 proceeds in this order: identify the tenant or tenants and place legal holds; scope custodian mailboxes, Online Archives, and Recoverable Items (Microsoft) or Vault holds and search scope (Google); account for linked attachments in OneDrive, SharePoint, or Google Drive; identify any third-party archives or legacy servers; then target endpoints for PSTs, offline folders, and unique local artifacts. Calibrate effort to issues, time periods, and key players—and document what exists even if a source is not reasonably accessible. It is wiser to know precisely what the client has—and why particular sources were scoped out—than to discover it at a sanctions hearing.

Looking for E-Mail 101

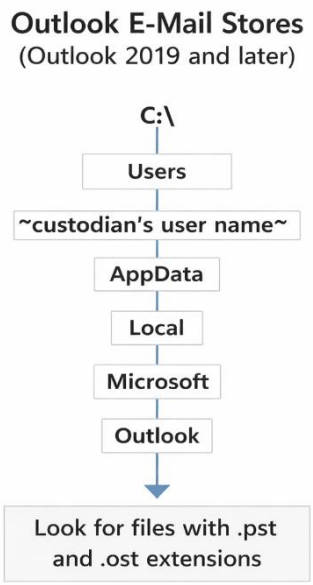
An e-mail is a structured text record, but client software rarely stores every message as a separate text file. Most clients use container databases. Some only display server-resident mail with small local caches; others keep larger local stores. *The only way to know whether e-mail is stored locally is to look for it.*

Policies and settings alone are not reliable—users change settings, create new profiles, install other clients, migrate machines, or export mail without remembering it. *If they do not know it is there, they cannot tell you it is not.* Effective triage means knowing what to look for and where to look — and *then actually looking.*

User storage paths differ because operating systems support multiple users and create separate profile areas. Even a single-user laptop may have multiple profiles over time. Finding mail stores hinges on knowing the user's account name and checking common application paths — then validating by file type, size, and last-modified dates.

Finding Outlook E-Mail

On Windows, Outlook commonly uses two file types. A PST is a portable Outlook data file that can hold mail, calendars, and folders—users often create PSTs to move or archive mail, and they can live on internal drives, external drives, network shares, or cloud-sync folders. An OST is the synchronized cache of a server mailbox—it mirrors the live mailbox and enables offline work; it may briefly contain items no longer visible online, plus offline folders or content from earlier configurations. On macOS, Outlook stores data in a profile database within the user's Library folder; Apple Mail maintains its own mailbox folders and index there as well. Also remember that when users open attachments directly from messages, Outlook places copies in temporary cache locations—these caches can outlast the original message and are worth checking when attachments matter.



In sum: prioritize server-side preservation first—then use endpoints to capture local-only items such as historic PSTs, user-made archives, or unique offline folders. Search drives for .pst and .ost files and validate by size and recency. For macOS, locate the Outlook profile and Apple Mail mailbox directories and check last-modified dates. Document what you found and what you did not—and why.

Microsoft 365 and Exchange — What Matters Now

Most enterprise Outlook users now sit in Microsoft 365. A Microsoft 365 tenant is the organization's Microsoft cloud environment where its users, mailboxes, and settings live. Key discovery controls

exist inside that tenant. Retention policies and labels, along with Litigation Hold, preserve mailbox items; the Recoverable Items area holds server-side copies even if a user deletes content.[90]

Again, Microsoft Purview eDiscovery—Microsoft's built-in legal hold, search, review, and export toolset—lets authorized users preserve and collect data in place. Exports can include mailbox items with system metadata and an export manifest that keeps threading and relationships intact. Because messages often link to OneDrive or SharePoint files, you must identify link targets, capture the correct file versions, and maintain the association to the parent messages.

On-premises Exchange *(i.e.*, the company's mail server) remains in some environments, typically for historic mail or specific use cases. Treat it as a legacy server source—coordinate with IT administrators for legal holds and targeted exports, and account for any companion archives that captured **journaled mail** in prior years.

Google Workspace — What Matters Now

A Google Workspace tenant is the organization's Google cloud environment. Gmail holds user mailboxes, and Google Vault—Google's built-in legal hold and retention tool—governs preservation, search, and export. Vault exports provide message data for review. As in Microsoft 365, linked Drive content must be addressed when the e-mail references shared files—identify the link, capture the correct file version, and keep the association to the parent message. For consumer Gmail or personal use, **Google Takeout** can export mail as **MBOX** when cooperation is obtained—useful if a personal account is in scope.

Webmail and Personal Accounts

Webmail cannot be ignored. If employees used personal accounts for business, preservation may require their cooperation or court intervention. Look for signs of forwarding in corporate mailboxes to identify likely personal sources. When appropriate and proportional, collect via provider tools—for example, Takeout for consumer Gmail—and integrate using standard containers such as MBOX or single message formats such as EML or MSG.

Accessibility and Proportionality

"Accessible" and "less accessible" describe practical effort, not legal destiny. It may be cheaper and more proportionate to examine a narrow set of less accessible sources rich in relevant data than to

[90] In older on-premises Exchange, 'double-deleted' items lived briefly in a hidden store nicknamed the **Dumpster**—a short-term safety net for mail emptied from Deleted Items; in modern Exchange and Microsoft 365, the Dumpster has been replaced by the Recoverable Items subsystem—protected folders (Deletions, Versions, Purges, Discovery Holds) that retain soft- and hard-deleted items, and even prior versions, while a hold or retention policy applies. Practically, this means that even if a user 'emptied' their trash, you can still retrieve the message from Recoverable Items while a legal hold or applicable retention is in place.

pull a broad swath of easily accessible mail with low yield. The burden continuum is not a straight line. Tie your plan to issues, players, and time periods—and memorialize why you included or excluded each location.

Legacy Corner — What You May Still Encounter

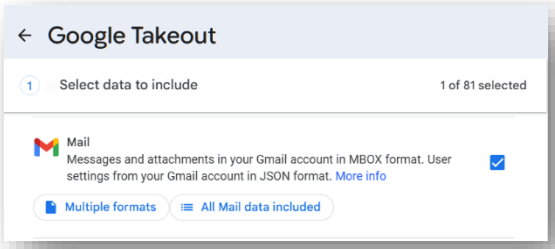
ExMerge and **Recovery Storage Groups** are older Microsoft Exchange collection tools you will still see in historical IT documents and expert reports—modern tenants use Purview eDiscovery and Litigation Hold instead, but legacy exports may exist and should be cataloged if relevant. Lotus Notes and Domino are largely retired, but some organizations retain NSF stores for historic projects or regulatory preservation. If Notes appears, collect the user's ID file and password along with NSF containers, and check for unsynchronized local replicas on laptops from the relevant period.

Practical Checklist for New Lawyers in 2026

Recap: Identify the tenant or tenants and confirm that legal holds or retention protections are in place. Enumerate custodian mailboxes, Online Archives, and Recoverable Items (Microsoft) or the Google Vault holds and search scope. Surface linked attachments—map OneDrive, SharePoint, or Google Drive targets and capture the right versions while preserving the association to parent messages. Ask about legacy servers, third-party archives, and migration exports. Target endpoints for PSTs, offline folders, and viewed-attachment caches; validate by size and dates. Address personal accounts and side channels if facts warrant—and document your methodology. Confirm the availability and retention of message-tracking logs if delivery is disputed. Record scope and source decisions—what you included, what you excluded, and why.

Google Data Tools: Takeout

Google's self-service export tool—is the simplest way for an individual user to export personal Gmail. From the user's Google Account, go to Data and Privacy, select Download your data (Google Takeout), choose Mail, and decide whether to export all mail or only items with specific labels. Exports are delivered as MBOX files—a standard format that preserves full message headers and includes label information—and may arrive as several download archives if the mailbox is large. For business e-mail in Google Workspace, the preferred preservation and export tool is Google Vault; Takeout may be disabled by the organization's administrator and is generally appropriate only with user cooperation in personal-account scenarios or narrow, agreed workflows.



Forms of Production for Email Productions

What users see on screen is a *curated view* of an e-mail—a rendering of header fields, the message body, and any attachments — influenced by the mail app and its settings. Your discovery obligation is to produce a form that preserves the content and the useful context of that source. Getting there depends on choosing suitable forms of production.[91]

"**Native production**" for e-mail rarely means handing over the entire server database. No one produces an Exchange .EDB with everyone's mail. Even when a party produces Outlook mail "natively," they typically export a new PST that contains only the responsive, non-privileged items. In practice, native for e-mail means using forms that *closely* preserve the source's content and functionality—often a container such as PST (for Outlook/Exchange) or MBOX/EML (for Gmail/IMAP)—or near-native single-message formats such as MSG and EML that, with some companion information, deliver comparable usability.

Static image-based productions are common. Producing TIFF or PDF images requires companion "load files"—simple, spreadsheet-like text files that carry searchable text and metadata fields (*e.g.*, From, To, Cc, Bcc, Subject, Date/Time, Bates numbers, attachment links) keyed to each page image—so review platforms can search, sort, and display messages. Either way, images alone are not ideal if you want to preserve optimum searchability, fields, families, and threading.

Native in Transit: Text per a Protocol

What "in transit" means: When e-mail travels between servers, it exists as plain text—ASCII or "7-bit"—so any system can read it. Early e-mail had no attachments or formatting; everything was just text.

RFC Standards: The Internet community standardized e-mail using "**Requests for Comment**"—**RFCs.** Today's core message format is **RFC 5322**, and attachments, non-English characters, and rich content are handled by **MIME**—a companion set of RFCs. Together, they specify how headers, bodies, and attachments are represented as text.

Why care: RFC-compliant messages hold the complete e-mail—headers, body, and attachments encoded as text—in a form that many tools can parse and preserve.

File formats that reflect this: A single message saved as EML mirrors the in-transit form; multiple messages grouped together are often stored as MBOX. These are widely compatible, human-readable text representations that travel well across systems.

Caveat. The in-transit form may not include application-specific fields added later by mail servers or clients. For discovery, EML/MBOX preserve the essentials of what actually traveled—and you can pair them with additional metadata if that context matters.

[91] We will have cover forms of production extensively in future readings and exercises, but it warrants a peek now.

Many producing parties prefer delivering page images to "stamp" Bates numbers or protective-order legends on the images. That works—kinda--but can be more expensive and less functional than native or near-native production. Costs and tradeoffs should be weighed against the case's needs and the parties' tools and skills

There is no single "perfect" form for all e-mail. **Hybrid productions** are sensible— using static images where they are most useful or cost-effective, and producing native or near-native forms where they best preserve content and function. As a rule of thumb, hew close to the source format when you can—PST for Outlook, MBOX/EML for Gmail/IMAP—and ensure the recipient can use what you produce.

When selecting a form, aim to:

- Preserve layout and visual cues — formatting, tables, colors, emphasis, hyperlinks, inline and embedded images, emoji.
- Keep the message searchable and supply header fields as discrete, sortable metadata.
- Maintain families—the parent message and its attachments—and identify any linked attachments so the relationship is clear.
- Convey the original folder path or label context.
- Include metadata fields that meet legitimate needs.
- Support redaction and Bates numbering without corrupting content.
- Normalize dates and times so threads sort correctly across custodians and time zones.

Native and Near-Native — What That Means in 2026

Outlook/Exchange. A PST is the native container for a set of Outlook items; MSG is a near-native for a single Outlook message. Both preserve rich Outlook fields. For Microsoft 365 custodians, exports commonly originate in Purview eDiscovery—Microsoft's built-in legal toolset for hold, search, review, and export in the tenant—and can be packaged as PST with a manifest that preserves threading and relationships.

Gmail/IMAP. EML is the Internet-standard single-message text format; MBOX is a simple container that holds multiple EML-format messages. In Google Workspace, Google Vault—Google's built-in legal hold and export tool—is the preferred tool for preservation and export. For consumer Gmail, Google Takeout can export MBOX with label information when user-cooperative workflows are appropriate.

Practice pointer—Linked attachments. In cloud systems, many messages point to files stored in OneDrive or SharePoint (Microsoft) or Google Drive (Google) instead of embedding a copy. When those linked files are in scope, identify the target, capture the correct version (as feasible), and

maintain the parent-child relationship in your production. Two messages that look alike may point to different file versions — that difference may matter.

Near-Native vs. Images—A Practical Litmus Test

Ask: can the recipient import what you produced into common mail tools or review platforms and still search by fields, see families, and follow threads? If the answer is no, you likely strayed too far from the source in selecting a form of production. Images with complete text and fielded load files *can* be "good enough" for many matters—but they usually cost more to load and host and can be materially less functional than PST/MSG/EML/MBOX.



Exercise 14: E-Mail Anatomy

GOALS: By the end of this exercise, you will be able to identify the essential parts of an e-mail—header, body, and attachments—and locate key header fields such as Message-ID, MIME-Version, boundaries, and time-zone offsets.

Outline. You will send yourself a short e-mail with a tiny image attachment, view the message's "source," and answer a few targeted questions using that source. Plan for 15–20 minutes.

Background

E-mail is still a leading source of probative ESI, and its structure is standardized. The Internet "rulebook" for e-mail is a set of standards called RFCs. RFC 5322 defines the text layout of e-mail headers and bodies. A companion family of standards called MIME explains how to include rich content—formatting, non-English characters, and attachments—while keeping everything represented as text. Because the format is standardized, anyone who can read the source can verify what was sent, when it moved, and to whom.

Review Glossary:

RFC — "Request for Comment," an Internet standard that specifies how something should be formatted so different systems interoperate.

MIME — "Multipurpose Internet Mail Extensions," which defines how to package multiple parts, formats, and attachments inside one e-mail.

Message-ID — a globally unique identifier assigned to a message, typically by the sending system.

Boundary — a unique string that separates parts inside a MIME message (for example, between plain-text, HTML, and attachments).

Step 1 — Draft and send a message with a tiny attachment

Using any account you control, **draft a short, impersonal message to yourself**. Optimally, send it to a different account than the one you're sending from—for example, send from your school or work account to a personal account, or vice versa.[92] **Attach a very small image** (under 5 KB, GIF or JPG). Be sure to use a small image file because you're going to paste the entire message into a text document and you don't want that document to run to pages and pages of base64 encoding.

Send the e-mail.

[92] When you send a message to yourself at the same mail account, the message need not traverse the Internet. Consequently, some of the features we seek to explore in this exercise will be absent. For example, if you send from Gmail, send it to your business or school account. If you have no alternate account, try sending it to a friend's or family member's account and have them forward the message back to you.

Step 2 — View the message source and save it as text

Open the *received* copy and reveal the source:

- Gmail — open the message ▸ More (three dots) ▸ Show original.
- Outlook for Web — open the message ▸ Click the three dots upper right of message ▸ Download ▸ Download as EML ▸ Open the EML in a plain text editor (e.g., Notepad on Windows, TextEdit on macOS in plain-text mode).
- Outlook desktop (Windows) — open the message ▸ File ▸ Properties ▸ Internet headers (copy). For full source, use "Message Options" or export as .eml if available.
- Apple Mail — select the message ▸ View ▸ Message ▸ Raw Source.

Select **all** of the source — it should include the full header, the body, and the base64 block for the image — and **paste it into a plain-text editor**. Avoid using Word if possible. Save as: **Your_Surname_Exercise_14.txt**.

Important Note: In Outlook for Web (e.g., our UT mail), accessing the message source using View ▸ View Message Details will show you the header but will NOT show the encoded attachment, so don't try that. Instead, download it as EML instead (see above) and save as a TXT file.

Step 3 — Dissect your message

Use the saved source to answer the questions below. Copy and paste values exactly as they appear. If your client obfuscates or wraps lines, include the full value even if it spans lines.

Question 1 — MIME version Prompt: What is the MIME-Version? Answer: _____

Boundaries Your message should define at least **two different boundaries** in the header (look for boundary="..."). When used as separators, a boundary line begins with two hyphens, and the final boundary line ends with two hyphens. The information above the first boundary definition is the "**Message Header**." Note that the message header contains the important To, From, Subject and Date information for the message.

Question 2 — First two boundaries:

Copy the *first* boundary string: Answer: _____

Copy the *second* boundary string: Answer: _____

Note: The first boundary separates the major sections (header, body, attachments). The second boundary separates alternative body formats (plain-text and HTML).

Message-ID Every standard e-mail should include a unique Message-ID in the header. It often looks like angle-brackets around a long string, for example: <0100019bfc5631c4-45b2a9f4-90b9-4a95-a302-9198255a85eb-000000@email.amazonses.com>.

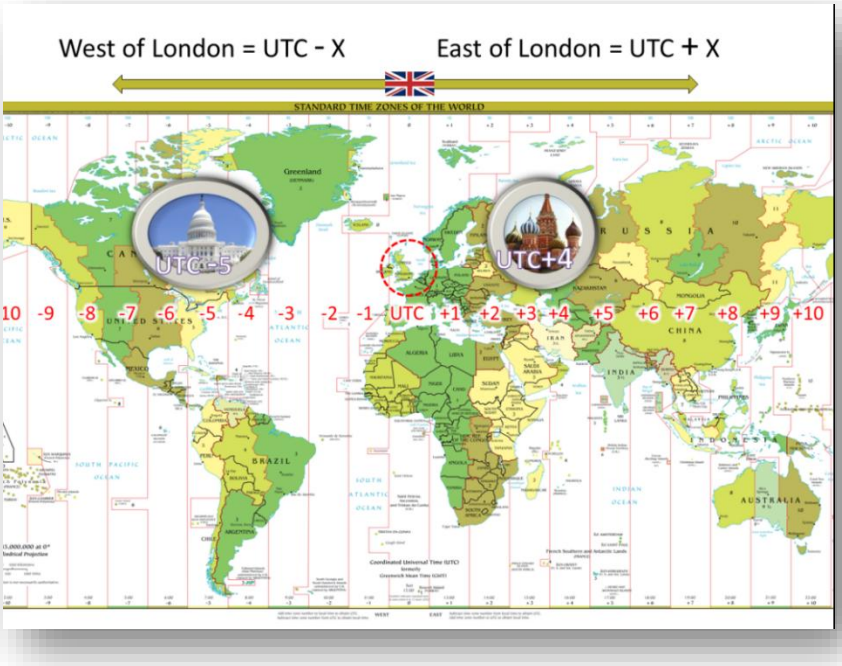
Question 3 — Message-ID: Copy the Message-ID value: Answer: _____

Evidence tip: Email forgers often change body text but forget to alter the Message-ID — a telltale mismatch.

Attachments Scroll to the attachment part — look for Content-Type: image/ and Content-Transfer-Encoding: base64, followed by a long block of encoded text.

Question 4 — System metadata for the image Prompt: In the text that precedes the encoded attachment—*and apart from the filename itself*--do you see system or application metadata for the image (for example, file size, creation or modification dates) in the attachment region?
Answer (yes/no): _____ If yes, copy it here: _____

Coordinated Universal Time (UTC): Time values in e-mail message headers are often expressed with reference to Coordinated Universal Time (*Temps Universel Coordonné* in French), the primary world time standard. UTC is often called Zulu time by the military and aviators and is essentially the same as Greenwich Mean Time (GMT), the time on the official clock located at the Royal Observatory in Greenwich, England. A **UTC offset** value expresses the difference between the stated local time and UTC, allowing messages to be **normalized** on a consistent timeline, notwithstanding differing time zones and daylight savings time settings.



Coordinated Universal Time (UTC) offsets Headers include dates and time-zone offsets (for example, −0500). These offsets allow you to normalize events across time zones.

Question 5 — List all UTC offsets you see in the message source

SENT UTC OFFSET: _____ RECEIVED UTC OFFSET: _____ OTHER UTC OFFSET: _____

These will likely be expressed as negative values (*e.g.*, -0800 (PST)):

Exercise 14 Submission checklist:

- Submit the file *Your_Surname_Exercise_14.txt* (containing the full message source).
- Paste your answers to Questions 1-5 in order under clear labels (Q1–Q5). Use exact copy/paste for values (no paraphrasing). Better yet, you may prefer to use the answer sheet I supply in Canvas>Assignments>Exercise 14.
- If you could not locate a field, briefly note what you tried to do that didn't work.



Exercise 15: Encoded Time Values in Message Boundaries

GOALS. By the end of this exercise, you will:

1. Apply your understanding of encoding, e-mail boundaries, and decimal/hexadecimal conversion to evaluate whether an e-mail is legitimate or forged.
2. Decode boundary-embedded timestamps to assess the integrity of an attached message.

Outline. Students will decode e-mail boundaries to assess whether an attachment's claimed send time is authentic.

"Time heals all wounds." "Time is money." "Time flies."

To these memorable mots, add one more: "*Time is truth*."

A defining feature of electronic evidence is its connection to temporal metadata — timestamps. Electronically stored information is frequently described by when it was created, modified, accessed, transmitted, or received. Clues to time are clues to truth because temporal metadata helps establish — or refute — authenticity, accuracy, and relevance.

Time is tricky in digital evidence. It hides in peculiar places, takes unfamiliar forms, and doesn't always mean what we assume. Because time is truth, it is valuable to know where to find temporal clues and how to interpret them correctly.

Everyone who works with electronic evidence encounters Windows file timestamps often called "MAC times," which have nothing to do with Apple Macs or network MAC addresses. Here, MAC is an initialism for Modified, Accessed, and Created.

That doesn't sound tricky. Modified means changed, Accessed means opened, and Created means authored — right? *Wrong.* A modified time can change due to events invisible to users and not reflecting user edits. Accessed times can change during events — like antivirus scans — few would call "access." Windows also stopped reliably updating access times by default beginning with Windows Vista and turned it back on a few years later. Created *may* match the authoring date, but it often reflects when a file was copied to a new location ("created *in that location*"). Copying a file in Windows produces an object that appears to have been created *after* it was modified.

It's crucial to protect metadata integrity in e-discovery, so changing file creation times by copying is a big no-no. Accordingly, commercial collection and processing tools copy files and then restore MAC times on the copies to match the originals. Done correctly, original metadata values are preserved in the copy and hash values do not change. *Remember:* system metadata values are not stored inside the content of most files, so they are not part of a file's content used to generate the

file's hash value. The upshot is that changing a file's *system* metadata — including filename and MAC times — does not affect the file's hash value.

Conversely — and ironically — simply opening a Microsoft Word document can change its hash because the application may update internal metadata such as the editing clock. Yes, there is even timekeeping inside Office applications.

Time also means nothing without place. When we raise our glasses and quip, "it's five o'clock somewhere," we are acknowledging that time is anchored to a time zone, adjusted for Daylight Saving Time, and often expressed with a UTC offset — the number of hours ahead of or behind Coordinated Universal Time (UTC), essentially Greenwich Mean Time, the time at the Royal Observatory in Greenwich, England atop the Prime or "zero" Meridian, hence "Zulu" or "Z" time, military shorthand for zero meridian time.

File times are typically stored in UTC. When displayed, operating systems shift them to the user's local time zone and DST rules. In e-discovery and forensics, it is essential to know whether a time is a local display or a UTC value. UTC is preferred because it supports normalization — ordering events from different locales on a single consistent timeline.

Time values are especially important when ordering e-mail communications. Many e-mails appear in conversational threads that mix live messages (with full headers and metadata) and embedded quotations from prior messages in the thread. Different time zones can make replies appear to precede prompts, and embedded text can be edited. This is one more reason to seek e-mail in native or near-native forms — not just static images. Headers hold authenticity and integrity data you will never see in an image-only production.

Underscoring that point, this exercise explores time values embedded in message boundaries.

Time Values Embedded in Mail Boundaries

If you know where to look, you will find time values hidden like Easter eggs. When overt timestamps are absent or untrustworthy, examiners turn to encoded time values to build timelines or expose forgeries.

E-mail must follow structural conventions to traverse the Internet. One is the use of a Content-Type declaration and unique content boundaries that separate the header from the body and any attachments. These are the boundaries you flagged in the prior exercise.

The next illustration is a simplified snippet from a forged Gmail message. To view a Gmail message's underlying code, select "Show original" from the message options menu (the three dots).

The line partly outlined in red states that the message will be "multipart/alternative," indicating that there will be multiple versions of the content supplied; commonly a plain text version followed

by an HTML version. To prevent confusion of the boundary designator with message text, a complex sequence of characters is generated to serve as the content boundary. The boundary is declared to be "000000000000063770305a4a90212" and delineates a transition from the header to the plain text version (shown) to the HTML version that follows (not shown).

```
MIME-Version: 1.0
Date: Tue, 24 Dec 2019 10:57:42 -0600
References: <CALckR-bSRHH=M8+sDCiYX0n1GBtQhADigbeG-BOY6xyvb+LE5w@mail.gmail.com>
In-Reply-To: <CALckR-bSRHH=M8+sDCiYX0n1GBtQhADigbeG-BOY6xyvb+LE5w@mail.gmail.com>
Message-ID: <o8ud3ne02mhqhwm7odsv@convertkit-mail.com>
Subject: Re: May 2020 Bootcamp - It's a Deal!
From: Bob Jones <bobjones@gmail.com>
To: Ernest Svenson <ernie@lawfirmautopilot.com>
Content-Type: multipart/alternative boundary="000000000000063770305a4a90212"

--000000000000063770305a4a90212
Content-Type: text/plain; charset="UTF-8"

Then, consider me booked!  I will turn down all other opportunities and
protect the date.

Bob Jones
bobjones@gmail.com
1523 Laurel St.
New Orleans, LA 70115
Tel: 504-555-6066
```

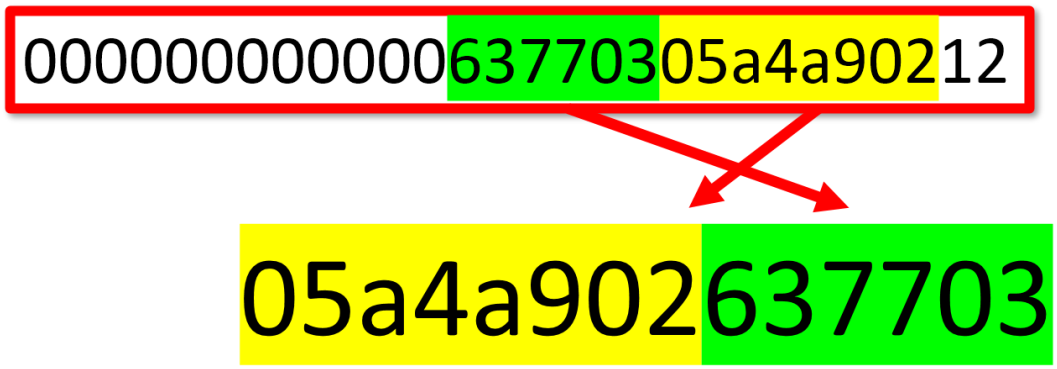
A boundary's sole purpose is to separate parts of an e-mail — but because it must be unique, programmers often integrate time data into the boundary text to reduce collisions with message text and other boundaries.

Here is how we will decode that integrated time data. It's a bit weird, so watch closely.

Here's our boundary, and I've highlighted fourteen hexadecimal characters in red:

000000000000**63770305a4a902**12

Next, I've divided (parsed) the highlighted text into six- and eight-character sequences (strings), *reversed their order* and concatenated ("strung together") them to create a new hexadecimal number:



A decimal number is Base 10. A *hexa*decimal number is Base 16. They are merely different ways of notating numeric values. So, Hex 05a4a902637703 is just a really, really big number. If we convert it to its decimal value, it becomes: 1,588,420,680,054,531. That's 1 quadrillion, 588 trillion, 420 billion, 680 million, 54 thousand, 531. Like I said, a BIG number.

But a big number…of *what*?

Here's where it gets amazing (or borderline insane, depending on your point of view).

It's the number of microseconds that have elapsed since January 1, 1970 (midnight UTC), not counting leap seconds. A microsecond is a millionth of a second, and 1/1/1970 is the "Epoch Date" for the Unix operating system. An Epoch Date is the date from which a computer measures system time.

When you make that curious calculation, the true Sent date extracted from the forged email dated December 24, 20219 proves to be Saturday, May 2, 2020 6:58:00.054 AM UTC-05:00 DST. It's not magic; it's just math.

Had the timestamp been created by the Windows operating system, the number would signify *the number of 100 nanosecond intervals between midnight (UTC) on January 1, 1601 and the precise time the message was sent.*

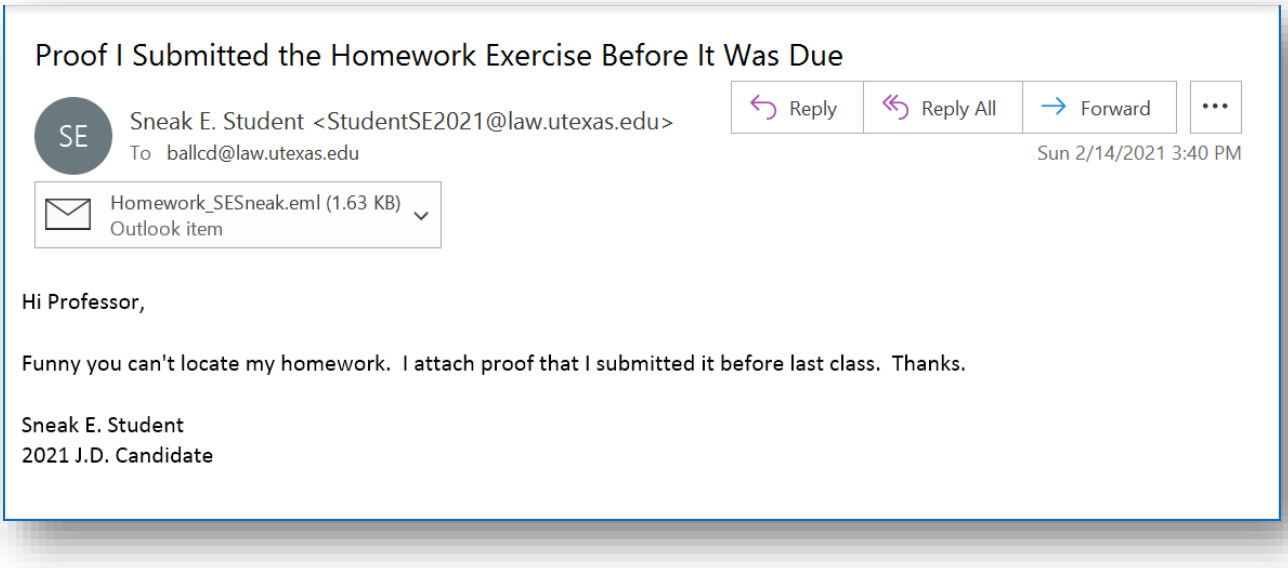
Why January 1, 1601? Because that's the "Epoch Date" for Microsoft Windows. Again, an Epoch Date is the date from which a computer measures system time. Unix and POSIX elected to measure time in seconds from January 1, 1970. Apple used one second intervals since January 1, 1904, and MS-DOS used seconds since January 1, 1980. Windows went with 1/1/1601 because, when the Windows operating system was being designed, we were in the first 400-year cycle of the Gregorian calendar (implemented in 1582 to replace the Julian calendar). Rounding up to the start of the first full century of the 400-year cycle made the math cleaner.

Timestamps are everywhere in e-mail — in boundaries, Message-IDs, DKIM signatures, and SMTP IDs. Each relay hop adds its own timestamp. It is the rare forger who finds and correctly modifies every embedded time.

Exercise 15: Background

It is Sunday, February 14, 2021, and Sneak E. Student failed to turn in the "Find the Secret Word" exercise due before class on February 10. Asked about it, Sneak swears he emailed the answer just before class and promises to forward a copy of the original transmittal.

That afternoon, Sneak e-mails the following.



Here is the message source from the email he sent with an attachment claimed to be timely:

[Server path data intentionally omitted]

```
MIME-Version: 1.0
Date: Sun, 14 Feb 2021 15:40:02 -0600
Message-ID: <CALckR-aWbcreMVY0-
A0E2r4LVkOUy3QU42qn7OgnjLSEBLqi2Q@mail.gmail.com>
Subject: Proof I Submitted the Homework Exercise Before It Was Due
From: Craig Ball <craig@ball.net>
To: ballcd@utexas.edu
Content-Type: multipart/mixed; boundary="000000000000762e5205bb52b5dd"

--000000000000762e5205bb52b5dd
Content-Type: multipart/alternative; boundary="000000000000762e4f05bb52b5db"

--000000000000762e4f05bb52b5db
Content-Type: text/plain; charset="UTF-8"

Hi Professor,

Funny you can't locate my homework.  I attach proof that I submitted it
before last class.  Thanks.
```

```
Sneak E. Student
2021 J.D. Candidate

--000000000000762e4f05bb52b5db
Content-Type: text/html; charset="UTF-8"
Content-Transfer-Encoding: quoted-printable

<div dir=3D"ltr">Hi Professor,<div dir=3D"ltr" class=3D"gmail_signature" da=
ta-smartmail=3D"gmail_signature"><div dir=3D"ltr"><div dir=3D"ltr"><div dir=
=3D"ltr"><div dir=3D"ltr"><div dir=3D"ltr"><div dir=3D"ltr"><div dir=3D"ltr=
"><div dir=3D"ltr"></div></div></div></div></div></div></div></div></div><d=
iv><br></div><div>Funny you can&#39;t locate my homework.=C2=A0 I attach pr=
oof that I submitted it before last class.=C2=A0 Thanks.</div><div><br></di=
v><div>Sneak E. Student</div><div>2021 J.D. Candidate</div></div>

--000000000000762e4f05bb52b5db--
--000000000000762e5205bb52b5dd
Content-Type: message/rfc822; name="Homework_SESneak.eml"
Content-Disposition: attachment; filename="Homework_SESneak.eml"
X-Attachment-Id: f_kl5o9k6v0
Content-ID: <f_kl5o9k6v0>

MIME-Version: 1.0
Date: Wed, 10 Feb 2021 14:39:26 -0600
References: <CAJVHkuuwoSEgdJcq9i0BxVEks0txdqRNDB_OkPe_D-
G35aGzmA@mail.gmail.com>
In-Reply-To: <CAJVHkuuwoSEgdJcq9i0BxVEks0txdqRNDB_OkPe_D-
G35aGzmA@mail.gmail.com>
Message-ID: <CALckR-Y6XoDJZ-2Mo-nL_z-yKTP_ySM7G2pJQGL0MNfv-
vZm9Q@mail.gmail.com>
Subject: Re: Here are my Answers to the Homework
From: Sneak E. Student <studentS2021@law.ut.edu>
To: Craig Ball <craig@ball.net>
Content-Type: multipart/mixed;
      boundary="A_0000000000003376a205bb4c5824"

--A_0000000000003376a205bb4c5824
Content-Type: multipart/alternative;
      boundary="B_0000000000003376a205bb4c5824"


--B_0000000000003376a205bb4c5824
Content-Type: text/plain; charset="UTF-8"
Content-Transfer-Encoding: quoted-printable

Here you go, Professor.  Just under the wire!

Sneak E. Student
J.D Candidate 2021

--B_0000000000003376a205bb4c5824
Content-Type: text/html; charset="iso-8859-1"
Content-Transfer-Encoding: quoted-printable

<html>
<head>
<meta http-equiv=3D"Content-Type" content=3D"text/html; charset=3Diso-8859-=
1">
<style type=3D"text/css" style=3D"display:none;"> P {margin-top:0;margin-bo=
ttom:0;} </style>
</head>
<body dir=3D"ltr">
<div style=3D"font-family: Calibri, Arial, Helvetica, sans-serif; font-size=
```

```
: 12pt; color: rgb(0, 0, 0);">
Here you go, Professor.  Just under the wire!
<div><br>
</div>
<div>Sneak E. Student</div>
J.D  Candidate 2021<br>
</div>
</body>
</html>

--B_0000000000003376a205bb4c5824--

--A_0000000000003376a205bb4c5824
Content-Type: application/msword; name="SStudent_secret word answer.rtf"
Content-Description: SStudent_secret word answer.rtf
Content-Disposition: attachment; filename="SStudent_secret word answer.rtf";
      size=214; creation-date="Wed, 10 Feb 2021 08:39:26 GMT";
      modification-date="Wed, 10 Feb 2021 08:39:26 GMT"
Content-Transfer-Encoding: base64

e1xydGYxXGFuc2lcYW5zaWNwZzEyNTJcZGVmZjBcbm91aWNvbXBhdFxkZWZsYW5nMTAzM3tcZm9u
dHRibHtcZjBcZm5pbFxmY2hhcnNldDAgQ2FsaWJyaTt9fQ0Ke1wqXGdlbmVyYXRvciBSaWNoZWQy
MCA2LjMuOTYwMH1cdmlld2tpbmQ0XHVjMSANClxwYXJkXHNhMjAwXHNsMjc2XHNsbXVsdDFcZjBc
ZnMyMlxsYW5nOSBUaGUgc2NyZXQgd29yZCBpcyB6YXJmXHBhcg0KfQ0KAA==

--A_0000000000003376a205bb4c5824--
--000000000000762e5205bb52b5dd--
```

Exercise 15: Investigate the Integrity of the Digital Evidence

Using timestamps embedded in message boundaries, can you determine if Sneak E. Student submitted his answer to the Find the Secret Word assignment before 2:30pm CST on February 10, 2021?

Steps to Solve This Exercise:

1. Select the *correct* message boundary value to analyze for its embedded timestamp.
2. Extract and compose the hexadecimal timestamp in the boundary.
3. Convert the hexadecimal timestamp to a decimal timestamp value.
4. Determine the date and time the timestamp was created.

Step 1: Select the correct message boundary value.

Putting our knowledge of email anatomy to work, we see from the message source that Sneak sent an e-mail dated **Wed, 10 Feb 2021 14:39:26 -0600** as an attachment to an transmitting e-mail dated **Sun, 14 Feb 2021 15:40:02 -0600.** The earlier e-mail, in turn, holds its own attachment: a rich text format (RTF) text file named **SStudent_secret word answer.rtf.**

Both the attached e-mail message and its own embedded attachment reflect February 10 dates and times which, if accurate, suggest a prior timely submission of the homework assignment. By design, e-mails are made up entirely of plain text so they can traverse even the oldest legacy e-mail systems. Accordingly, the date and time values are plain text and can be easily altered with a text editor. If that's all we had to go on, we would be forced to conclude that the assignment was timely.

Still, we know that e-mails may contain timestamps encoded in message boundaries revealing the genuine date of a forged communication. If we look at the timestamps, we may be able to find evidence corroborating or refuting the plain text dates. However, there are four distinct boundaries declared in the message source.

```
boundary="000000000000762e5205bb52b5dd"
boundary="000000000000762e4f05bb52b5db"
boundary="A_00000000000003376a205bb4c5824"
boundary="B_00000000000003376a205bb4c5824"
```

There's no cause to question the timing of the February 14 transmittal, so we would expect those boundaries to resolve to February 14 timestamps. Instead, we want to look at the boundaries in the e-mail attachment purporting to be created on February 10, 2021. Because the purportedly older e-mail message is an attachment to a subsequent transmittal, we should not be surprised to find that the boundaries of the attachment are flanked by the boundaries of the transmittal.

The boundaries declared in the attachment are:

```
boundary="A_00000000000003376a205bb4c5824"
boundary="B_00000000000003376a205bb4c5824"
```

Step 2: Extract and compose the hexadecimal timestamp in the boundary.

We want to parse the boundary string in the manner demonstrated above (i.e., dropping the two final characters and creating a concatenated string of the reversed eight and six characters in the boundary):

Question 1: What is the **hexadecimal** timestamp of Sneak's purported February 10 transmittal?

Answer: _____

Step 3: Convert the hexadecimal timestamp to a decimal timestamp value.

Now, you must determine the decimal equivalent of the hexadecimal value just found You *could* do this manually, of course, but it's must faster to use a conversion tool. You could download a conversion app for the task, or you can locate an online converter. If you Google *Hex to Decimal Converter*, many suitable options come up. I used the calculator at https://www.rapidtables.com/convert/number/hex-to-decimal.html, but again, any converter will do.

Question 2: What is the **decimal** value of the hex value in your answer to Question 1?

Answer: _____

Step 4: Determine the date and time the timestamp was created.

With luck, the decimal number you've just determined reflects the number of microseconds that elapsed between the Epoch date of January 1, 1970 and the time the timestamp was created. You

could make such a calculation manually knowing there are 86,400,000,000 microseconds in a day, but that's the sort of task where computers excel. So, again, we can turn to the web to find a suitable conversion tool. If you Google, say, "convert decimal to epoch date," you'll find a range of online tools that will complete the conversion from a decimal value to a date value. I used https://www.epochconverter.com/ but any should do.

Question 3: What is the date and time of the decimal value in your answer to Question 2, above? Make sure the date you supply is expressed as U.S. Central Standard Time, not GMT or UTC.

Answer: _____

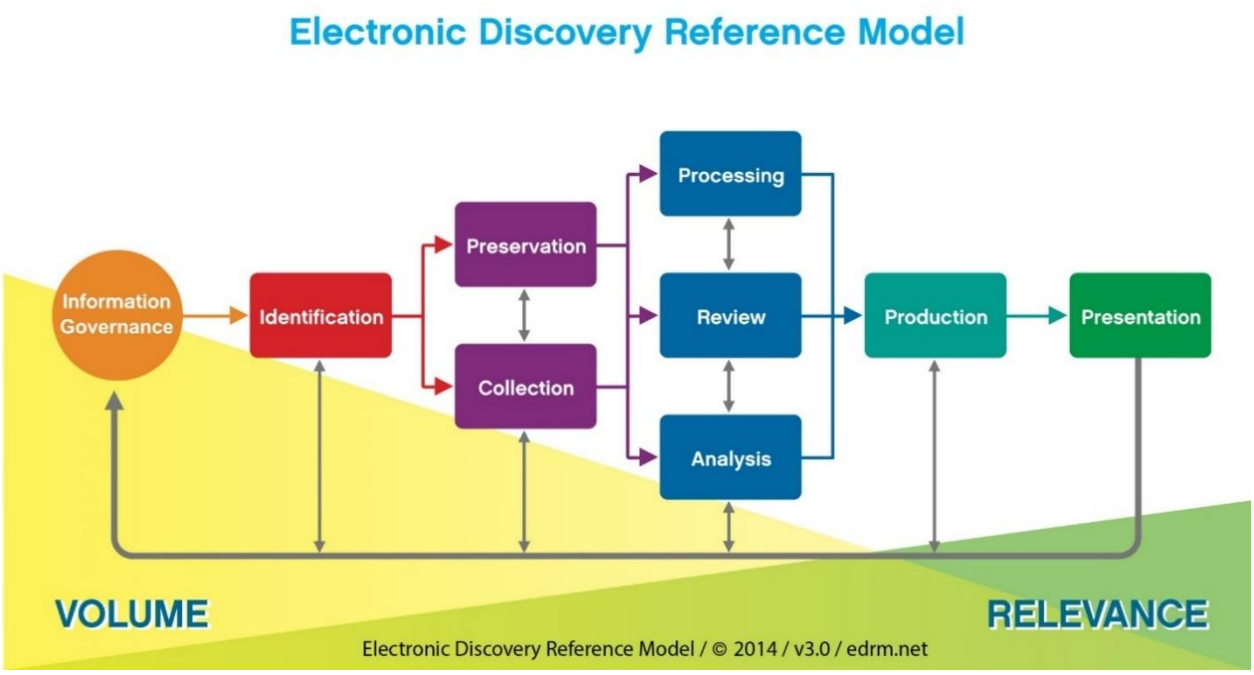
Submission checklist:

I've supplied an answer sheet in Canvas at Assignments>Exercise 15

- Show the exact concatenated hex string you used.
- Show your intermediate decimal value derived from the hex value.
- Provide the *actual* sent date/time of the embedded transmittal in CST.
- If your result does not align with the claimed send date, explain the discrepancy in one sentence.

The Next Arc of Exercises

Young lawyers in large practices first encounter electronic evidence at the review stage — after identification, preservation, collection, and processing have transformed the raw data. Evidence appears as if by magic, yet dozens or hundreds of upstream decisions determine whether relevant material sees the light of day or whether the matter slides into costly discovery disputes. To competently defend those decisions — and to spot when something has gone wrong — you must understand the earlier EDRM stages well enough to explain them clearly. So, the coming workbook exercises entail tasks like processing and producing ESI, that lawyers don't *do* but are often called upon to *explain* and *defend* in court.



Identification, Preservation, and Collection

We have surveyed systems and media that house ESI and the data and metadata they hold. Remember that *ESI is dynamic.* Messages arrive and are deleted by automated processes. Collaboration sites and social platforms update constantly. Logs roll over. Mobile devices push and pull data continuously. Even in cloud mail, retention rules can purge or move content unless a legal hold intercedes. *Data is always in flux and in motion* — so doing nothing, or acting too slowly, invites loss.

You can't preserve what you haven't identified, and you can't collect what you haven't preserved. Collection for discovery and evidence requires more exacting processes than ordinary backup. Integrity turns on preserving content and metadata and documenting chain of custody. No ESI should be produced or offered in evidence absent the ability to trace it to its source and show it

was not altered—or if altered, how and why. That entails reliable, repeatable methods; capturing relevant metadata; validating with hashes; and recording who did what, when, and with what tools.

In a perfect world, you could search, hold, and review data in place without collecting it. Some platforms now support robust preservation-in-place and tenant-side search and export — for example, Microsoft Purview eDiscovery in Microsoft 365 and Google Vault in Google Workspace — but most litigation still requires collection for processing, culling, search, review, and production.

Two collection approaches dominate: **targeted collection** and **forensic imaging**. Targeted collection duplicates potentially relevant ESI by specified criteria — file types, locations, custodians, and date ranges. A collection from Jose Rodriguez' Windows laptop of all Word documents and Excel spreadsheets in the Projects folder having last modified dates between January 1, 2023 and December 31, 2025 is an example of a targeted collection. Forensic imaging duplicates the *complete* contents of the laptop — the active data *and* the forensically significant areas that the operating system does not show, such as unallocated space and slack.

Each approach has champions because each carries trade-offs. Targeted collection reduces volume — which typically reduces processing, hosting, and review costs — but it depends on scoping decisions that may miss relevant material. If you later discover a missed location, file type, or date window, you may need a do-over — assuming the source still exists in a comparable state. Forensic imaging captures everything on the medium and preserves options for later analysis if spoliation is alleged, but it requires specialized tools and skills and still leaves you to separate wheat from chaff. Imaging is often superior for preservation; but it is not the most cost-effective path to a review set.

The Cloud changes this calculus. In public cloud services — multitenant platforms managed by providers like Microsoft and Google — disk-level forensic artifacts such as unallocated clusters are not exposed to customers. Collection is necessarily targeted, usually via tenant-native tools that can place legal holds, search and export. Preservation-in-place may be as simple as enabling a litigation hold or retention policy in the tenant.

Whatever path is taken, you must protect integrity and document processes. Collection tools must preserve or restore original timestamps, avoid altering content, and generate cryptographic hashes to prove authenticity and support deduplication. Chain-of-custody records should identify the source, the scope, the tool and version, dates and times, and the persons responsible. If choices materially limit what can be searched or reviewed, counsel must decide whether to disclose those limitations and seek agreement or court guidance before proceeding.

In your exercises, you have forensically imaged an evidence thumb drive and will perform a targeted collection as a precursor to processing in a commercial e-discovery platform. You will cull and search, then configure and generate a production with appropriate identifiers. In practice, tech specialists may perform these steps for you — but when things go wrong, having done them, even once, equips you to ask the right questions, defend sound process, and remediate defects quickly.

Custodial Hold: Trust but Verify



Many years ago, the late Browning Marean presciently observed that the ability to frame and implement a legal hold would prove an essential lawyer skill. He was right. "Legal hold" is more than a memo or checkbox — it is a multipronged, adaptive *process* tailored to the needs of the case, like a well-fitted tux. Repeatable, defensible workflows matter, but the most successful and cost-effective holds still show the bespoke touch of competent counsel who understands the client's systems and the stakes.

The evolving, case-specific character of legal holds is precisely what some people dislike — that, and the cost. Many want a checklist, a form notice, a one-size-fits-all "best practice." All have value; none obviate the need for someone who understands the ESI environment and will be accountable for getting the hold right. Effective holds balance two aims at once — maximize retention of relevant, material, non-duplicative information and minimize cost, complexity, and business disruption. You can't choose one or the other — ***you need both.***

Both — Subjective and Objective Preservation

"Custodial hold" relies on the human beings who create and hold data to preserve it. That makes sense: custodians often know best where their data lives and what it means. They can relate information to the actions and decisions at issue. But a custodial hold is necessarily subjective *and perilous* — it depends on what people understand and what they choose to do. It's said that three things can happen when you throw a football, and two of them are bad. The same is true for custodial hold. Some will do it well. Some will not do it at all. Some will do it badly because they don't understand what's going on. Others will do it badly because they understand quite well what's afoot.

When you make custodians think about how the information they hold relates to a dispute, you stir them to consider the consequences of others scrutinizing the information they hold. Maybe they start to worry about being blamed for the problem that gave rise to the litigation or maybe they worry about getting in trouble for something that has nothing to do with the litigation, but which looms large as an item they don't want discovered. Either way, it's "their" information, and they aren't going to help it hang around if it might look bad for them, for a co-worker or for the company.

Recent decisions underscore the risks of relying on custodial holds—particularly where auto-deletion, personal accounts, or modern communications tools are in play. In *DR Distributors, LLC v. 21 Century Smoking, Inc.*, the court cataloged years of preservation and collection failures, sharply criticized counsel's reliance on custodians to self-identify sources and self-collect ESI, and imposed

substantial sanctions, underscoring that preservation cannot be delegated and forgotten.[93] *WeRide Corp. v. Kun Huang* likewise imposed severe sanctions where a party failed to suspend automatic email deletion and engaged in destructive account and device practices despite foreseeable litigation and a court-ordered preservation injunction—demonstrating that leaving default retention settings in place is indefensible once the duty to preserve attaches.[94] Together, these cases reflect a consistent judicial expectation: lawyers must actively supervise preservation, understand how client systems actually retain and delete data, and verify that preservation measures are effective. Trust, *but verify*. Or, as Finley Peter Dunne put it, "*Trust everyone, but always cut the cards.*"

So, both. Trust and verify. Backstop custodial holds with objective, system-level measures. Put another way, the limitations of custodial holds don't mean you don't use them–*you must use them in almost every case*. It means you don't use them *alone*.

Design legal hold efforts around a mature recognition of human frailty. Accept that people will fail to preserve or will destroy data and recognize that you can often easily guard against such failure by adding a measure of objective preservation to your hold strategy.

In 2026, objective preservation commonly includes:

- Tenant-side legal holds and retention in cloud suites. In Microsoft 365, apply Litigation Hold or retention policies/labels to custodians' mailboxes and sites via Microsoft Purview eDiscovery — preserving items in Recoverable Items (Exchange's modern "dumpster") and preserving prior versions in SharePoint/OneDrive. In Google Workspace, use Google Vault holds and retention for Gmail and Drive.
- Hold coverage extends beyond e-mail. Include collaboration data — Teams channels, chats and meeting artifacts (Zoom); Slack messages and files; SharePoint/OneDrive and Google Drive files; and any approved text/messaging platforms used for business.
- Gatekeeping for Bring Your Own Device (BYOD) and side channels. Mobile device management and acceptable-use policies help, but you still must ask about personal accounts and devices when facts warrant, then document your decisions.
- Journaling and system safety nets. Where journaling was enabled historically, locate the journal mailbox or archive; it can corroborate that traffic existed even if custodians failed to preserve. Confirm how long audit and message-trace logs persist and preserve them if delivery or access is disputed.

Practical techniques to "cut the cards" include quietly snapshotting key mailboxes and network shares; preserving linked-file targets referenced by e-mail; imaging select endpoints where

[93] DR Distributors, LLC v. 21 Century Smoking, Inc., 513 F. Supp. 3d 839 (N.D. Ill. 2021).
[94] WeRide Corp. v. Kun Huang, 2020 WL 1967209 (N.D. Cal. Apr. 24, 2020).

spoliation risk is acute; and raising server-side preservation thresholds for key custodians so “double-deleted” items remain retrievable during the hold.

Elements of an Effective Legal Hold Notice

Lawyers are trained to distill cases into black-letter propositions and then *do something*: draft a checklist, tweak a form, adjust boilerplate. Modern lawyering is necessarily programmatic, especially under flat-fee arrangements or insurer oversight. Thinking is a liability when billable hours are capped. As a result, the matter-specific instructions essential to an effective litigation hold quickly collapse into boilerplate—language so broad and so generic that it serves little purpose beyond allowing counsel to say, "I told you so," when something inevitably goes missing.

How do we keep the legal hold from becoming just another omnibus notice—so vague it confuses and so expansive it paralyzes?

We cannot dispense with forms. They are institutionalized. But we can recalibrate how we use them so they reflect how data is actually created, stored, deleted, and lost in 2026. What follows is *not* a catalogue of best practices. It is a set of *defensive corrections*—ways to make a hold less brittle and less performative.

Ask Why, Not Why Not

Lawyers do not eliminate risk; we manage it. But in 2026, over-preservation is no longer merely expensive; it is operationally disruptive. Cloud retention extensions, mailbox holds, journaling, and exports impose storage, licensing, and review costs almost immediately. A hold notice goes too far when it compels a client to "preserve everything." That's not caution; it's malfeasance--and the "sanction" is immediate and self-inflicted. A competent hold now requires articulating *why* particular sources merit preservation and *why others do not*, informed by how systems retain data by default.

Get Real

Custodians are no more reliable today than they were a decade ago—but the systems they use are far more complex. Asking individuals to identify, segregate, and preserve responsive material across email, chat, shared workspaces, and personal devices assumes they understand retention settings, sync behavior, and deletion semantics. Most do not. Some will delay. Some will misunderstand. Some will actively delete. A modern hold must assume human failure and compensate for it by pairing instructions to people with objective, system-level preservation wherever feasible.

In many cases, it is cheaper and safer to apply a tenant-level hold, snapshot a mailbox, or preserve a shared workspace than to rely on individual compliance. Hope is not a defensible hold plan.

Focus on the Fragile First

Most enterprise data does not vanish easily. Financial systems, ERP platforms, and core databases remain resilient. The fragility has shifted. Today's preservation failures cluster around:

• short-retention messaging platforms
• collaboration tools with per-channel or per-workspace deletion
• auto-expiring chat, voice, and video
• reissued laptops and mobile devices
• personal accounts used for business communications

Email remains relevant, but it is no longer the sole—or even primary—risk vector. An effective hold prioritizes sources with short default retention and decentralized control, not those already governed by enterprise backup and audit regimes.

Separate Human Holds from System Holds

In 2026, a single "legal hold notice" is often insufficient. Custodians need clear behavioral instructions: stop deleting, stop reusing devices, identify off-system accounts. IT, security, and platform administrators need technical directives: suspend deletion, apply retention locks, preserve audit logs, and prevent device reimaging for named users.

Low-tech measures still matter, too. Physical labeling of devices subject to hold remains inexpensive and effective. But they must be paired with platform-level controls. A hold that relies solely on human restraint, without technological backstops, is indefensible.

Forms Follow Function

Generic notices fail because modern systems are role-segmented. Business users interact with content; IT controls systems; security manages logs; HR controls devices and off-boarding. A single omnibus notice addressed to "all employees" is functionally meaningless. Preservation directives must be tailored to the recipient's authority and competence. Ask people to do only what they are positioned to do—and make clear who is responsible for what they cannot.

Get Personal

Personalization isn't just a courtesy; it is a safeguard. Add one or two custodian-specific sentences that show you understand their role and likely data sources. Stating why a particular custodian matters—and what unique sources they are likely to control—forces counsel to think concretely about evidence and signals to the recipient that the request is neither theoretical nor optional. Personalization prompts diligence, improves comprehension, and plays well on later scrutiny.

Don't Sound Like a Lawyer

An effective legal hold notice prompts action. It tells people what to do, how to do it, and by when. If the duty is ongoing, say so plainly. Get to the point in the first paragraph. Write for a bright twelve-year-old. Supply concrete examples of sources to be searched and items to be preserved.

Verify and Test: ***Trust but cut the cards.***

Track acknowledgements, chase non-responders, and spot-check preservation with system reports. For example: confirm Litigation Hold/Vault holds are applied to the named accounts and working; confirm that the custodian's storage footprint isn't continuing to drop because items are still being deleted or purged; capture screenshots or reports; ticket device change-freezes; and note audit/message-trace log retention and any preservation you execute before logs roll off.

Ten Elements of a "Perfect" Legal Hold Notice

1. **Timely**
2. Delivered through **effective channels**
3. Sent by individuals with **clout**
4. Directed to all **necessary custodians and system owners**
5. Conveys **gravity** and **accountability**
6. Supplies sufficient **factual context** to guide preservation decisions
7. Provides clear, practical **instructions with deadlines**
8. Sensibly scopes **sources and formats**
9. Identifies a clear point of **contact for questions**
10. Includes **acknowledgement, verification, follow-up, and refresh**

The Leery Lawyer's Guide to AI and LLMs in Trial Practice



Updated 2026

Craig Ball

The Leery Lawyer's Guide to AI and LLMs in Trial Practice
By Craig Ball
© 2026

Introduction

This paper is a wholesale update to my 2025 primer, *Practical Uses for AI and LLMs in Trial Practice*. It remains aimed at tech-wary, veteran trial lawyers, but it has been broadened to serve those who have cautiously woven AI tools into their litigation practice. Since large language models like ChatGPT burst onto the scene three years ago, the hype and fearmongering haven't let up. Every day brings fresh alarm: ETHICAL PITFALLS! DEEP FAKES! JOB LOSS! COPYRIGHT INFRINGEMENT!

The noise has only grown louder. In 2025–26, the legal press ran out of synonyms for *transformational*, and vendors now insist that every litigation tool has "AI inside," whether or not that claim would survive a Daubert hearing. Still, beneath the marketing froth lies something real and durable: AI has become a routine—if imperfect—instrument in modern trial practice. Nothing in my forty-odd years at the Bar has had such a swift and profound effect on the way lawyers work.

Some lawyers point to AI's occasional tendency to "hallucinate"—to fabricate information—as an excuse to avoid adoption altogether. Yet we all use AI daily, even if only by reading the summaries atop Google searches or skimming Amazon reviews. What drives the relentless drumbeat of AI, AI, AI isn't simply the integration of smarter agents into our software (so-called *agentic AI*); it is the paradigm-shifting species of generative AI known as *large language models*, or LLMs—an initialism that no longer denotes a Master of Laws, but the most consequential development in technology since the Internet.

LLMs can talk, read, write, and draw pictures. They can speak in a thousand voices—including yours—dishing out sense and nonsense in equal measure. Their uncanny ability to "sound" human and to string words together in ways that convincingly mimic (and sometimes exceed) human language skills is at once exciting and unsettling.

Despite high-profile instances of lawyers getting into trouble by using LLMs carelessly, the case for employing AI in trial practice today is compelling—even if only as a time-saver for document generation and summarization, tasks where AI's capabilities are undeniable. The economics are now clearer. Sensibly implemented AI reduces drafting time, accelerates motion practice, sharpens discovery management, and improves comprehension of large records. Used prudently, it functions as the world's fastest junior associate—minus the hourly rate and parking validation. Used unwisely, it can embarrass you or worse.

The earliest iteration of this guide grew out of the Litigation Section of the State Bar of Texas devoting the Winter 2024 issue of *The Advocate* to artificial intelligence. Every article was wise and

well written—several penned by close friends—but none was *practical* in the sense of helping lawyers use AI in their daily work. That struck me as an unmet need.

As I looked around, I found no articles aimed at guiding trial lawyers who wanted to use LLMs safely, strategically, and without turning their practice upside down. Here, I've tried to fill that gap by highlighting important considerations and illustrating applications you can try now with minimal expertise, anxiety, or expense. LLMs won't replace your professional judgment, but they can frame issues, suggest language, and translate complex doctrines into plain English. In truth, they can do just about anything that mastery of facts and language can achieve—so long as you remain firmly in the driver's seat.

But Know This…

LLMs are unlike any technology tool you've used before. Most of the digital tools in our lives are characterized by consistency: you put the same things in, and the same things come out, in a rigid and replicable fashion. Not so with LLMs. Ask an LLM the same question multiple times, and you are likely to receive somewhat different answers each time. That takes some getting used to.

There is also no single "right" way to interrogate an LLM to be assured of an optimal result. There is no strict programming language or magic set of keywords calculated to achieve your goal. Instead, there are myriad ways to elicit useful information, and—unlike the inflexible and unforgiving tech tools of the past—the most effective approach is often the simplest: interact with your AI tool in a natural, conversational manner, refining your requests as you go.

Throughout this paper, I reference both **ChatGPT** and **Harvey**. ChatGPT now exists in both consumer and enterprise deployments. Harvey is purpose-built for law firms, offering matter-centric workspaces, access controls, retention policies, and auditable logs—features designed to satisfy skeptical general counsel and their insurers. Both platforms are useful. Both are powerful. And both require adult supervision.

Marketplace Models and Versions

Fueled by a trillion+ dollars in corporate investment, the Generative AI market didn't cool in 2025; it matured. LLMs diversified into two main channels:

1. **consumer/general-purpose models** such as OpenAI's GPT series and Anthropic Claude, and
2. **enterprise-governed systems** like Harvey, deployed behind stronger privacy and audit controls.

Here are current market leaders, with no claim that the list is exhaustive or permanent—because it isn't:

- **OpenAI:** Offers the GPT series, including GPT-5.2 and multi-modal models capable of ingesting text, images, spreadsheets, and transcripts. These models are accessible through products like ChatGPT and are integrated into various applications.
- **Google DeepMind:** Developed the Gemini series (successor to Bard), with versions like Gemini 3 Flash and Gemini 3 Deep Think. These models power AI capabilities in Google's services.
- **Anthropic:** Provides the Claude series, including Claude Opus 4.5 designed for conversational AI applications.
- **Meta AI:** Developed the Llama series, with Llama 4 Scout and Maverick being notable versions. These models support multimodal inputs (i.e., text and images).
- **Microsoft:** Introduced the Phi series, such as Phi-4 (that understands text, images and audio) and Phi-4-Mini, marketed as a "small language model" for efficient AI tasks.
- **Harvey AI:** Built expressly for law firms, Harvey operates within firm-controlled environments, supporting matter-level governance, logging, retention limits, and privacy controls appropriate for supervised legal work.

This article will focus on OpenAI's ChatGPT, specifically its ChatGPT-5 product. I don't feature it because it's necessarily the best or brightest; it's the LLM tool most people use—and increasingly its enterprise deployments reduce privacy anxiety. The current $20 per month subscription tier remains surprisingly affordable compared to billable time squandered staring at a blank pleading page.

I've also expanded my focus to the NOT inexpensive Harvey AI, which you might think of as "enterprise ChatGPT with governance." If you're a solo or small-firm lawyer on a budget, think "ChatGPT—used prudently—can still be safe and effective."

Whatever LLM tool you choose, remember this: *There's a huge difference in the capabilities and user experience between the free versions and the subscription offerings.* ***If you're going to integrate an LLM into your practice, even just to kick the tires, opt for the pay-to-play versions.*** The paid tier unlocks document upload, longer context windows, and improved accuracy. The enhanced performance of GPT-5.2 justifies the modest investment, particularly when compared to the hourly cost of professional staff or consultants.

So, yes, doing the things I describe here will cost you twenty bucks, but you can try it for a month or two, then cancel if you don't see value in continuing to subscribe. I *guarantee* you will see value, if only from the ability to upload content to the tool for analysis and editing.

Getting Started

Getting started with ChatGPT is surprisingly simple, even for lawyers who aren't tech-savvy but possess intellectual curiosity and critical thinking skills honed through years of trial practice. At its core, ChatGPT is a conversational artificial intelligence tool designed to assist with tasks such as drafting, research, and document analysis. To begin, visit chat.openai.com and create an account using your email address and date of birth. Upgrade your account to a $20.00 per month **ChatGPT Plus** subscription.

Harvey Users: Access occurs inside your firm's workspace. Treat it like a secure review platform, not a personal app. Matter access is controlled, and work is logged—assume someone responsible is watching, because they should be.

Once logged in, you can enter **prompts**—questions or instructions—to generate **responses**. Think of ChatGPT as a hyper-efficient research assistant who can respond instantly to requests for drafting demand letters, summarizing testimony, or brainstorming trial strategies.

To maximize ChatGPT's usefulness, *clarity and specificity* in prompts are key. The tool thrives on detailed instructions.

For example, instead of saying,

"*Draft a request for production for a baby powder cancer case*,"

You'll get a superior result by uploading your complaint to the tool and prompting, *e.g.,*

> *"You are plaintiff's counsel in a wrongful death case in a Harris County, Texas civil district court. Draft 25 numbered Requests for Production to defendant Johnson & Johnson regarding alleged cervical cancer caused by exposure to talcum powder containing asbestos.*
>
> *Drafting requirements:*
>
> - *Use Texas practice and terminology; track Texas Rules of Civil Procedure (e.g., 192/196) and proportionality.*
> - *Include Texas-style Definitions and Instructions (parties, "document," "ESI," date range, custodians, form of production).*
> - *ESI specifics: identify relevant systems/custodians; request native or near-native production with agreed metadata; de-duplication and email threading; reasonably usable format; index/cover letter.*
> - *Avoid overbreadth; tie each request to issues, custodians, systems, and time period."*

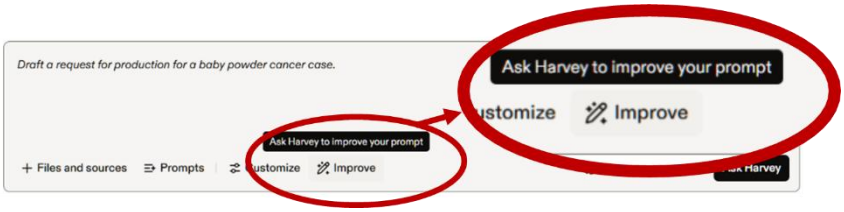
Why the detailed prompt is superior

- **Clarity about role and forum.** By specifying that the LLM is to perform as "plaintiff's counsel" in a Harris County, Texas civil district court, the model defaults to lawyerly

language in a plaintiff-side posture and Texas terminology, instead of mixing federal/state or other jurisdictions.

- **Concrete facts and theory**. Naming Johnson & Johnson, cervical cancer, talc, and asbestos anchors the requests to issues you actually intend to discover (design, warnings, testing, adverse events, marketing, regulatory knowledge), not generic product-liability boilerplate.
- **Scope and deliverable defined.** "Draft 25 numbered RFPs" tells the model the format and volume, so you get a serviceable, complete set—not a handful of examples that require rebuilding.
- **ESI and proportionality baked in.** Calling out "include the language needed to obtain electronically stored information" prompts the model to add definitions, form-of-production, metadata, deduplication/threading, and usability terms that make your set meet-and-confer-ready and defensible under Texas proportionality standards.
- **Hallucination control.** Specific party, product, injury, and venue reduce the model's temptation to invent authorities or wander into non-Texas practice. Constrained prompts produce fewer "creative" but unusable requests.
- **Editing burden drops.** When the model knows the audience, issues, and output format, you spend time doing substantive tailoring instead of structural clean-up.
- **Ethical and practical defensibility.** A prompt that elicits targeted, proportional, ESI-aware requests makes your paper trail stronger when opposing counsel (or the court) asks why your terms or formats are reasonable.

One of the simplest and most effective ways to improve your prompts is by asking an LLM to do it for you. I'll often give ChatGPT a prompt and then ask it to streamline and improve my prompt before I run it. Harvey makes that easier by supplying an "Improve" button dedicated to prompt refinement (right).

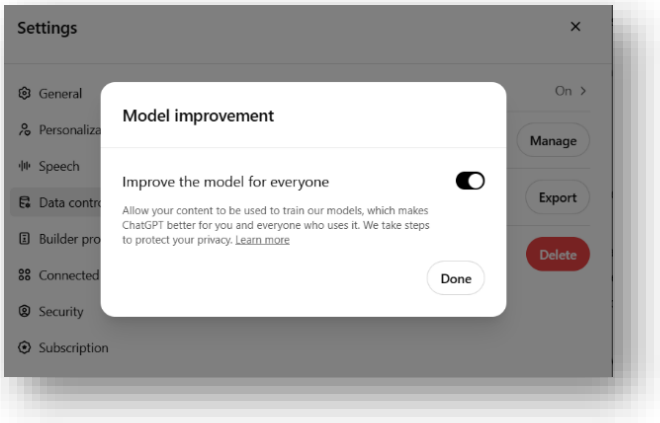


Privacy

As lawyers, we're entrusted with sensitive, privileged, and private information we are obliged to protect. That obligation does not stop at the courthouse door—or at the login screen of an AI platform.

If you're concerned that interactions with ChatGPT may expose confidential information by being used to improve the model, you should adjust the platform's privacy controls before using it for legal work.

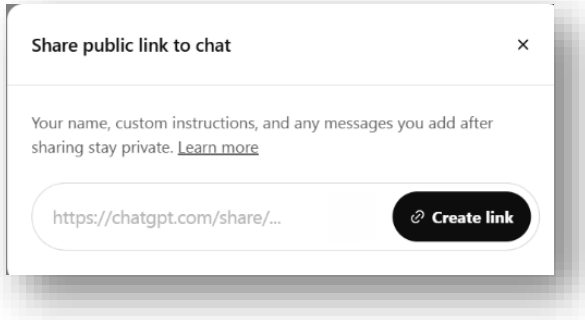
To modify ChatGPT's privacy settings (consumer and Plus plans):

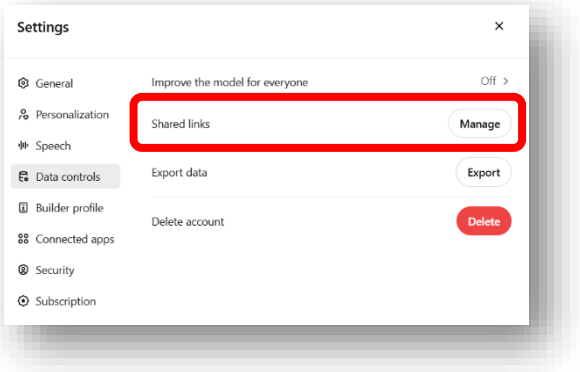


1. Open the ChatGPT app or website.
2. Go to **Settings → Data Controls**.
3. Toggle OFF **"Improve the model for everyone."**
4. Confirm your changes.

Disabling this setting prevents your chats from being used for model training, but it does **not** make the conversation ephemeral or invisible. Chats may still be retained for operational, safety, or legal reasons, and anyone with access to your account can see them.

ChatGPT also allows users to generate **shared links** to conversations. Anyone with the link can view the conversation exactly as it existed when the link was created. These links are convenient—and dangerous. They can be forwarded, indexed,





or accessed outside your control. If you share them at all, share them sparingly, only with authorized recipients, and periodically review or revoke them under **Settings → Data Controls → Shared Links → Manage**. As a rule, do not use shared links for privileged or client-confidential material.

Protecting sensitive data does not end with flipping a toggle. It is a necessary step, but no substitute for a layered approach that includes judgment, restraint, and supervision. When in doubt, abstract facts, anonymize parties, or describe issues at a level that conveys the problem without exposing the client.

Harvey (enterprise law-firm deployments)

Harvey is designed for legal practice and offers a substantially different security posture from consumer ChatGPT. Work occurs inside **matter-centric workspaces** governed by firm controls. Inputs are not used to train public models, access is role-based, and activity is logged—features that matter to general counsel, insurers, and courts.

That said, Harvey does not eliminate risk. Security depends on **configuration and discipline**:

- Limit workspace access to lawyers and staff assigned to the matter.
- Treat uploads as you would a shared litigation repository: only what belongs in the matter goes in.
- Assume all work is auditable and supervised. That is a feature, not a flaw.
- Do not rely on the platform to decide what is privileged; that judgment remains yours.

Whether using ChatGPT or Harvey, the governing principle is the same: AI is a powerful non-lawyer assistant. You may delegate tasks, but not responsibility. Confidentiality, privilege, and professional judgment remain non-delegable duties.

Hallucinations

Working with LLMs is like teaching a teenager to drive. Sure, teens have better vision, hearing, and reflexes; but, at any moment, you may find yourself screaming, *RED LIGHT! RED LIGHT!* ChatGPT is merely a tool, one only as capable and trustworthy as your ability to direct and monitor its use. Like any power tool, when you fail to respect its risks and limitations, it can hurt you; operated skillfully, it will make your work easier.

ChatGPT's tendency to hallucinate—*i.e.*, to fabricate information or confidently fill gaps it does not actually know—requires constant vigilance. **ALWAYS cross-check its output**, particularly when legal accuracy is critical. If the tool makes up a case, statute, or quotation, *you* are the one who will have to answer for it. I cannot emphasize this enough. Courts do not want to hear that you failed to verify your citations because "the AI said so."

In practice, hallucinations most often arise when the model is asked to operate beyond the record supplied or is pressed for conclusions where the source materials are silent or incomplete. You can substantially reduce this risk by **instructing the model how to behave**, for example:

- "Use only the materials I provide; do not rely on outside knowledge."
- "If the record is silent or ambiguous, say so."
- "End with a list of assertions that require human verification."

These instructions do not eliminate hallucinations, but they make them easier to spot—and easier to defend against. See also, "Practice Tip: Verification Checklist for Use in Prompts" on the following page.

Harvey reduces hallucination risk by constraining the model to a defined matter and document set, but it does not eliminate it. Even when analysis is limited to uploaded materials, the model can still

infer incorrectly, over-generalize, or misread context. Whether you use ChatGPT or Harvey, hallucinations remain a supervision problem, not a software bug.

Practice Tip: Verification Checklist for Use in Prompts

To guard against harm from hallucinations, include the following instruction at the end of any prompt involving legal authority, factual assertions, or record-based analysis:

Verification Checklist:
Before finalizing your response:

1. Identify every legal rule, case, statute, or procedural requirement referenced and flag each for independent attorney verification.
2. Identify every factual assertion not directly quoted from the supplied materials and flag it for confirmation.
3. For summaries of testimony or documents, include page and line numbers or document IDs for all quoted or paraphrased material.
4. If the record is silent, incomplete, or ambiguous on any point, state that explicitly rather than inferring an answer.
5. Do not fabricate citations, quotations, or facts. If you cannot confirm something from the provided materials, say so.

The goal is not to eliminate lawyer review, but to make review faster, more reliable, and defensible.

Customizing Settings for Litigation Efficiency

Modern LLM platforms offer customization features that can materially improve efficiency in litigation and trial work, but they operate differently across tools.

ChatGPT

- **Response Style (formerly "Temperature"):**
  While earlier versions exposed a numerical "temperature" setting, most current ChatGPT interfaces control precision through prompt language rather than sliders. For litigation tasks—discovery requests, deposition outlines, motion drafts—explicitly instruct the model to be *formal, precise, and conservative*. Reserve creative latitude for brainstorming themes, demonstratives, or voir dire framing.

- **Custom Instructions:**

ChatGPT allows persistent instructions that shape behavior across sessions. For example: *"Assume I am a Texas trial lawyer. Use plain English. Do not invent authority. Flag any statement requiring verification."*
These instructions reduce rework and reinforce disciplined outputs. They are set at **Settings → Personalization → Custom Instructions**.

- **Memory:**
  ChatGPT's memory feature can retain preferences or context across chats. This can be helpful, but it also poses risks in legal work, including cross-matter contamination and inadvertent retention of sensitive information. Review and manage memory regularly at **Settings → Personalization → Memory → Manage**, and consider disabling it entirely for client-specific work.

Harvey

Harvey does not rely on temperature sliders or long-term conversational memory. Instead, its efficiency controls are structural:

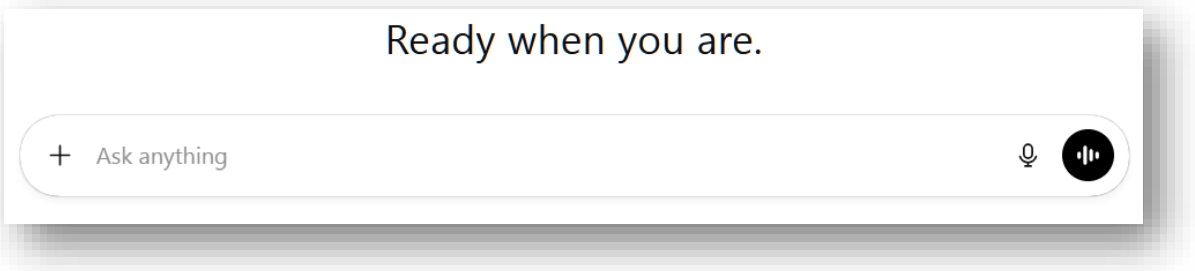
- **Matter-centric scope:** Analysis is confined to the documents and prompts within a specific matter.
- **Document constraints:** Outputs can be limited to the uploaded record, reducing drift.
- **Auditability:** Prompts and outputs are logged, reinforcing disciplined use.

These features make Harvey better suited for sustained case work, but they do not replace the need for careful prompt design and lawyer review. Precision still comes from how you frame the task, not from the platform alone.

The ChatGPT User Interface and its Token, Upload, and Context Limits

When users interact with ChatGPT, they typically do so through the **Chat Window**, a text box with icons in the lower left corner. Users enter prompts and receive responses in the Chat Window. The plus sign initiates a file upload (and more).[95]

[95] That black circle with a waveform on the right of the Chat Window opens Voice Mode, allowing you to converse with ChatGPT as you would with the world's best-informed colleague who is unfailingly polite and helpful. Prepare to be awe-stricken.



If you've experimented with ChatGPT, chances are you've been frustrated by limitations stemming from a failure to understand the tool's **token, upload, and context limits**, even if you didn't appreciate *why* things weren't working.

Tokens are units of text used to measure both input and output. Tokens can be as small as a single character or as large as a short word, depending on structure. Every prompt, response, space, and punctuation mark consumes tokens, and longer or compound words consume multiple tokens.

As a rule of thumb, one page of text equals roughly **375–400 tokens**, and **10,000 tokens ≈ 7,500 words ≈ 25 pages**. That estimate remains useful even as platform limits change.

Think of the **context window** as the size of a workspace where the model can "hold" a finite amount of information at one time. The **upload limit** is more like the size of an entire library. You may upload a very large document, but only a portion of it can be actively considered at once. The model can retrieve and move portions of the document into the workspace as needed, but it cannot reason over the entire library simultaneously.

This distinction explains why large uploads sometimes appear to "stall," truncate, or produce incomplete answers. The document may be present, but the relevant portion may not be in the active workspace when the model responds.

As I write this in early-2026, ChatGPT's paid tiers allow users to upload large files and work with extended context windows. However, the precise limits vary by model and plan and change often. In **ChatGPT-5.2** (single-user monthly subscription) these limitations are substantially expanded but still very real:

Token Limit (per chat/working context):

ChatGPT 5.2 supports a **very large context window—on the order of hundreds of thousands of tokens per chat window**, encompassing both prompts and responses. That translates into *hundreds of pages* of text under ideal conditions—far beyond what was possible a year ago. But this is a working limit, not a guarantee that everything supplied will be considered simultaneously.

Upload Limits:

Users may upload **multiple large documents per chat** (commonly on the order of a few dozen), with individual file size limits typically in the **hundreds of megabytes** for text-based documents and lower limits for spreadsheets and images. Upload capacity is best understood as the size of the *library*, not the workspace. Uploading a very large file does not mean the model will reason over it all at once. Instead, ChatGPT retrieves portions of uploaded files as needed, subject to the context window. In practice, very large or complex files may process slowly, partially, or fail altogether—particularly when multiple large files are uploaded at once.

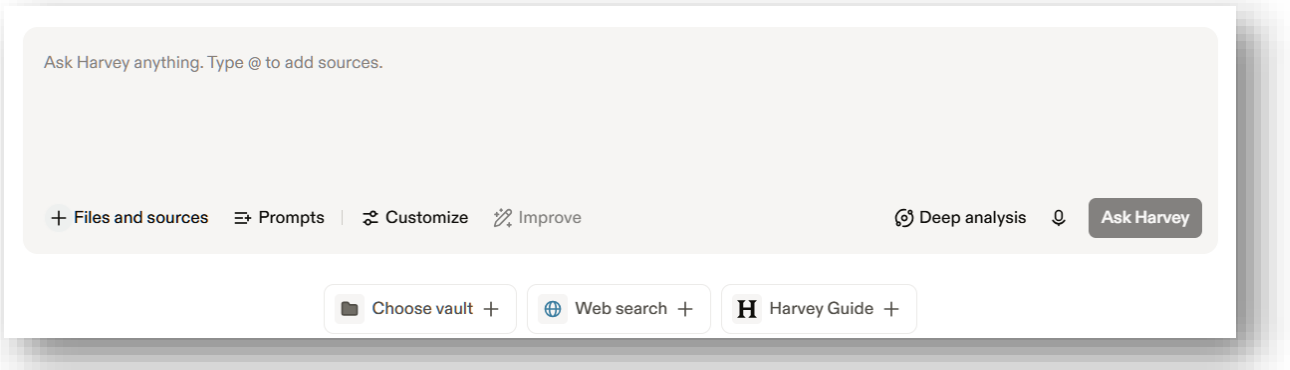
Context Limits:

The context window—the amount of information the model can actively "hold in mind" during a single interaction—is now dramatically larger than in early 2025, but it remains finite. As conversations grow, earlier prompts and responses are displaced by newer material. Even uploaded documents may fall outside the active context unless key portions are summarized or reintroduced.

Is this confusing? You bet! Think of the context window as the size of a workspace where you can spread out a very large—but still limited—set of papers at one time. The file upload limit is more like the size of the entire library. You can upload a large book (or several), but only part of it can be open on the workspace at once. The expanded upload and context limits in ChatGPT 5.2 make incremental and sectional analysis far more practical than before, but they do not eliminate the need to break large records into manageable pieces.

The Harvey User Interface and its Token, Upload, and Context Limits

When users interact with **Harvey**, they do so through a conversational workspace anchored to a **specific vault or matter**, rather than a general-purpose chat window. At the top of the interface is a single prompt field ("Ask Harvey anything"), where users enter questions or instructions. Unlike



consumer chat tools, this field is tightly integrated with sources and controls that govern what Harvey may see and use.

Directly beneath the prompt field are controls that shape the scope and behavior of the interaction. The **"Files and sources"** option allows users to attach documents or select existing materials already stored within the chosen vault. The **"Choose vault"** control determines the matter or repository to which the interaction is confined—a critical distinction, as Harvey's analysis is limited to the documents and sources available within that vault.

Additional interface controls—such as **"Prompts," "Customize,"** and **"Improve"**—allow users to reuse structured prompts, adjust output preferences, or refine prior responses. The **"Web search"** and **"Harvey Guide"** options signal whether the model may consult external reference material or remain confined to internal sources. When enabled, **"Deep analysis"** instructs Harvey to spend more time reasoning across the available materials, though it does not alter the underlying limits on how much information can be considered at once. The interaction is executed by selecting **"Ask Harvey."**

If you've experimented with Harvey, you may still encounter frustrations traceable to a misunderstanding of the platform's **token, upload, and context limits**, even though those limits are mediated through matter-based controls rather than exposed as raw numbers.

As with other large language models, **tokens** remain the unit used to measure both input and output. Every prompt, response, space, and punctuation mark consumes tokens, and longer or compound words consume multiple tokens. The same practical rule of thumb applies: roughly **375–400 tokens per page** of text. This estimate remains useful when deciding how much material to analyze in a single request.

As of early-2026, Harvey supports the ingestion of substantial document sets within a vault and provides extended working contexts compared to consumer chat tools. However, these capabilities do not eliminate fundamental constraints. The precise limits depend on firm configuration and deployment and may change over time. As with ChatGPT, understanding how Harvey *operates* is more valuable than memorizing numerical ceilings.

Think of a Harvey **vault** as a curated library tied to a legal matter. The model has access to the contents of that library, but its **active reasoning context**—the workspace in which it can "hold" and analyze information at one time—is necessarily smaller. Harvey can retrieve documents from the vault as needed, but it cannot reason across the entire contents of the vault simultaneously.

This distinction explains why Harvey may sometimes return incomplete, overly narrow, or seemingly inattentive answers even when the relevant document exists within the vault. The document may be present and available, but it may not have been drawn into the active context at

the moment the response is generated. Effective use of Harvey therefore depends on directing the model's attention—by selecting appropriate sources, constraining prompts to defined subsets of material, and building analysis incrementally rather than attempting to process an entire matter in a single pass.

Why Tokens Matter—and How Uploading Files Actually Works

ChatGPT

These limits mean that **ChatGPT may fail to upload files that are too large** or may truncate longer documents such that ingestion simply peters out when upload or token limits are reached. Even when a document uploads successfully, ChatGPT may process only portions of it at any given time.

As a conversation grows, ChatGPT will eventually **"forget" earlier prompts, responses, and instructions** as the context window fills and older material is displaced by newer exchanges. This is one reason refining answers late in a long session can be frustrating: the background you carefully established may no longer be "in view." Preserving continuity often requires summarizing key points and reintroducing them in later prompts so they remain within the active context.

In short, **LLMs have strict working-memory limits**. Do not expect ChatGPT to digest or reason coherently across an entire case file in a single interaction.

That reality also shapes how file uploads work. In current versions of ChatGPT, documents are added using the **plus sign ("+") control located alongside the prompt field**. Selecting the plus sign allows users to attach files from their local system or, depending on plan and configuration, from connected repositories. Once attached, those files become available for analysis within that conversation.

Uploading documents enables ChatGPT to perform a wide range of useful tasks: summarizing deposition transcripts, analyzing medical records, extracting timelines, reviewing spreadsheets, and drafting work product informed by the uploaded materials. ChatGPT works best with **Word documents, Excel spreadsheets, PDFs containing selectable text, and CSV files** exported from databases. Scanned or image-only PDFs remain less reliable unless they include usable text layers.

Attaching a document, however, does not mean the model has fully absorbed it or will continuously remember it. Large documents may be processed only in part, and details may fall out of context as the conversation progresses. Effective use requires **explicit direction**—for example, telling the model which sections to focus on—and often follow-up prompts that pull additional portions of the same document into the active workspace.

The upshot is that the dream of uploading every deposition, exhibit, and production in a case and letting ChatGPT "figure it out" is not yet practical. But within these limits, ChatGPT can still manage minor miracles—so long as you are willing to **break work into stages, process long documents in sections, and divide large datasets into manageable chunks**.

Harvey

Harvey changes the container, but not the underlying physics. Documents are added through the **"Files and sources"** control beneath the prompt field and are ingested into a selected **vault or matter**, where they persist across sessions. This persistence can make it feel as though Harvey "knows" the entire case.

In reality, Harvey remains subject to the same fundamental constraints as other large language models. Although documents reside within the vault, any single response is generated within a bounded **active analysis context**. Harvey can retrieve documents from the vault as needed, but it cannot reason across the entire contents of a large matter simultaneously.

This explains why Harvey may sometimes return incomplete or narrowly focused answers even when the relevant document exists within the vault. The document may be present and indexed, but it may not have been drawn into the active context at the moment the response was generated. As with ChatGPT, effective use depends on **directing attention**—by identifying specific documents, narrowing the scope of analysis, or building conclusions incrementally across multiple prompts.

Harvey's matter-centric structure reduces organizational friction and improves governance, but it does not eliminate cognitive limits. Depositions are best summarized one at a time. Expert reports benefit from section-by-section analysis. Chronologies are most reliable when built iteratively. In both platforms, the practical lesson is the same: **design your workflow to fit the size of the workspace, not the size of the library.**

Practice Tips (Avoiding the Most Common Missteps re: Context Windows)

Upload ≠ analyze. Attaching a document doesn't mean the model has read or reasoned over all of it.

Expect forgetting. As conversations grow, prompts and instructions fall out of context. Restate key instructions.

Name your sources. Specify which documents or excerpts to use—particularly in Harvey's vault environment.

Work incrementally. If answers seem incomplete, narrow the task and rebuild rather than starting over.

Chunking Large Uploads

"Chunking" is the practice of breaking large uploads into pieces small enough for an AI tool to handle accurately and reliably. Because large language models can only "hold" a limited amount of material at one time, effective use depends less on uploading everything and more on deciding what to present, when, and in what order. Chunking mirrors how trial lawyers already work: we don't absorb an entire case file in one sitting, and we don't try every issue at once. By dividing records, transcripts, and productions into logical segments—and building conclusions incrementally—we improve accuracy, reduce omissions, and retain control over the analysis. The playbook below offers concrete ways to chunk common litigation materials so the AI works within its limits instead of stumbling over them.

Chunking Playbook (With Sample Prompts)

Depositions

Work one witness at a time. Begin with a high-level summary, then process longer transcripts in logical segments (liability, damages, experts). Synthesize only after the parts are complete.

Sample prompt:
You are a Texas trial lawyer. Using only the attached deposition transcript, prepare a concise narrative summary of the witness's testimony on liability. Cite page and line numbers. If the transcript is silent or ambiguous on any point, say so. End with a Verification Checklist.

Expert Reports

Divide expert reports into discrete sections—background, methodology, assumptions, opinions, and conclusions. Analyze each section separately before requesting a consolidated critique or cross-examination outline.

Sample prompt:
Using only the attached expert report, analyze the methodology section. Identify stated assumptions, potential weaknesses, and areas for cross-examination. Quote or cite to the report where possible. Do not infer facts not stated. End with a Verification Checklist.

Productions and Document Sets
Start with targeted subsets rather than entire productions. Focus on key custodians, date ranges, or document families, then expand outward as needed.

Sample prompt:
Review the attached documents from Custodian A dated January–June 2025. Summarize recurring themes and identify documents likely relevant to contract formation. List document identifiers. If relevance is uncertain, flag it for review.

Medical Records
Chunk medical records by provider or time period. Analyze each segment for chronology, gaps, and red flags before combining results.

Sample prompt:
Using only the attached medical records from Provider X (2019–2021), prepare a treatment chronology noting dates, diagnoses, and procedures. Identify gaps in care or inconsistencies. Cite record pages. End with a Verification Checklist.

Chronologies and Timelines
Build chronologies incrementally, confirming accuracy at each stage before extending the timeline.

Sample prompt:
Based solely on the attached documents covering March–May 2025, prepare a dated chronology of events relevant to notice and response. Cite each entry to document and page. Do not speculate beyond the record.

Large Spreadsheets and Databases
Process data in logical slices—by date range, category, or variable set—before seeking broader conclusions.

Sample prompt:
Review the attached spreadsheet for transactions from Q1 2025. Identify unusual patterns, outliers, or missing fields. Describe findings in plain English and note any assumptions requiring verification.

Practical Applications for Litigation and Trial Practice

Large language models are most useful in litigation when they perform tasks lawyers already do—just faster and with fewer false starts. Used properly, they assist with framing, organization, and first-pass drafting, while leaving judgment and verification where they belong: with counsel.

Drafting Discovery Requests

AI tools excel at translating narrative descriptions into actionable discovery requests. You can describe the dispute, the issues in play, and the information you seek, and the model can generate targeted requests tailored to those needs. This is particularly effective for drafting requests for production, interrogatories, and ESI-specific requests that reflect proportionality and scope.

In ChatGPT, this works best when you provide context and constraints in the prompt. In Harvey, it is most effective when requests are generated within a matter using defined custodians, systems, and date ranges already present in the vault.

Summarizing Voluminous Data Sets

LLMs can sift through testimony, expert reports, or document subsets to produce summaries, chronologies, or issue-focused digests. Rather than reading everything end-to-end, lawyers can direct the model to surface themes, flag anomalies, or extract testimony on discrete topics.
This is especially valuable for depositions, medical records, and expert materials. The key is scoping: summaries are most reliable when limited to defined witnesses, documents, or time periods, and when citations back to the source are required.

Structuring Information

ChatGPT and Harvey are adept at converting unstructured narrative into **tables, timelines, matrices, and outlines**. This includes witness charts, issue grids, chronology tables, and comparison charts for contracts or expert opinions.

Structured outputs are often more useful than prose. Asking for a table or outline forces clarity, exposes gaps in the record, and creates work product that can be refined or repurposed for motions, mediation, or trial preparation.

Search Term Generation and Query Assessment

Turning plain-English descriptions into search terms or Boolean queries for discovery is another area where LLMs perform well. They can analyze relevant documents and propose keyword lists, proximity connectors, and alternative formulations based on how people actually communicate rather than how lawyers guess they might.

LLMs can also critique existing search terms, identifying overbreadth, redundancy, or likely blind spots. This is useful both for improving your own searches and for evaluating the adequacy of an opponent's proposed terms.

Drafting Pleadings and Arguments

AI tools can assist in drafting and organizing pleadings, motions, and briefs by suggesting structure, sequencing arguments, and refining language. They are particularly useful for first drafts, issue outlines, and transitions.

They are less reliable for unsupported legal authority and should never be trusted to supply citations without verification. Used properly, they help you get from a blank page to a workable draft more quickly, leaving you free to focus on substance and strategy.

Demonstrative Evidence and Trial Themes

LLMs can assist in **conceptualizing and planning demonstrative evidence**, but their usefulness—and their limits—vary by platform. Neither ChatGPT nor Harvey will produce trial-ready, admissible exhibits. What they can do, if properly directed, is help lawyers think through how complex information might be **organized, sequenced, and explained**.

ChatGPT, particularly in its consumer and enterprise forms, can generate **visuals, diagrams, and mock-ups**, or describe how a demonstrative might look and function. While some might be extraordinary, these outputs are best treated as sketches or talking points rather than finished products. They are helpful for brainstorming timelines, flowcharts, or comparative graphics, but may require translation into proper demonstratives by counsel or a graphics professional.

Harvey, by contrast, is not designed to generate images or visual mock-ups. Its strength lies in **structuring the underlying content** that demonstratives rely on—chronologies, tables, issue matrices, and witness summaries—within the confines of a specific matter. Harvey can help ensure that demonstrative concepts remain grounded in the record and consistent across documents, but the visual execution occurs elsewhere.

In both platforms, the value lies upstream. LLMs can help identify what should be shown, in what order, and for what purpose. The judgment about how—and whether—to show it to a jury remains firmly in the lawyer's hands.

They are also effective at stress-testing trial themes—identifying weaknesses, alternative framings, or juror-friendly explanations that can be refined by counsel.

Deposition Preparation and Impeachment

LLMs are particularly effective at preparing topic-organized deposition outlines, identifying internal inconsistencies, and surfacing potential impeachment material across multiple transcripts or documents. When constrained to the record, they can generate contradiction tables keyed to page-and-line citations, helping lawyers see where testimony shifts over time.

Meet-and-Confer and Proportionality Advocacy

AI tools can assist in drafting meet-and-confer correspondence that explains discovery positions in plain English: why certain custodians matter, why others do not, and how proposed limits reflect proportionality. This is especially useful when translating technical search methodology into language suitable for opposing counsel—or the court.

Used carefully, LLMs help sharpen the *explanation* of discovery positions, not the positions themselves. They are best employed to clarify rationale, not to invent it.

Expert Report Deconstruction and Cross-Examination Planning

Beyond summarizing expert reports, LLMs can break opinions into assumptions, inputs, methods, and conclusions, exposing where an expert relies on disputed facts or unexplained leaps. This lends itself naturally to cross-examination planning and Daubert-style challenges.

Privilege Log Assistance and Consistency Checks

Within well-defined constraints, AI can assist in drafting privilege log descriptions, normalizing entries, and identifying inconsistencies across large logs. It can also flag entries that appear under- or over-described.

This use should be framed carefully: AI assists with form and consistency, not privilege determinations. Final judgment remains squarely with counsel.

Voir Dire and Jury Instruction Framing

LLMs can help generate neutral, open-ended voir dire questions designed to surface attitudes about damages, corporations, law enforcement, or experts without argument or indoctrination. They can also help reorganize jury instructions into juror-friendly thematic groupings for counsel's internal use.

Settlement and Mediation Preparation

AI tools can assist in preparing internal case assessments, summarizing strengths and weaknesses, identifying factual disputes, and drafting concise mediation narratives. They can also help battle-test positions by articulating how an opponent might frame the same facts.

This use is most effective when the model is instructed to argue *both sides*—a task lawyers often avoid but benefit from confronting early.

Ten Examples with Prompts

An LLM can be a gamechanger in litigation and trial preparation, going well beyond simple document drafting. Having discussed what these tools can do at a high level, the examples below illustrate practical, effective ways to deploy them. These are not "magic words." Treat them as starting points and adapt the prompts to the facts, forum, and posture of your case.

1. Drafting Deposition Questions:

- **Prompt Example:** "You are a Texas trial lawyer. Using the facts I provide, generate a topic-organized deposition outline for a corporate designee in a products liability case involving an allegedly defective medical device. Focus on design specifications, quality control, complaint handling, and failure reports. Do not assume facts not supplied. Flag areas requiring document support."
- **Application:** Build a disciplined outline tied to the pleaded liability theories and the areas of inquiry fairly encompassed by the witness's role, designation, and knowledge under the Texas Rules of Civil Procedure. In Harvey, constrain the task to documents in the matter vault to keep the outline grounded in the record.

2. Discovery and ESI Protocols:

- **Prompt Example:** "Draft a meet-and-confer letter proposing a proportional ESI protocol for a federal employment discrimination case in Texas. Address data sources, custodians, date ranges, search methodology (with examples), metadata fields, production format, and privilege clawback. Explain the rationale for each choice in plain English."

  Tip: you might try uploading my free primer on ESI Protocols along with your prompt: http://www.craigball.com/ESIProtocol.pdf

- **Application:**
  Accelerate preparation of defensible ESI proposals and meet-and-confer correspondence. This is particularly effective in Harvey, where custodians and systems may already be defined within the matter.

3. Trial Preparation and Cross-Examination:

- **Prompt Example:** "Using only the materials I provide, prepare a cross-examination outline for a medical expert offering causation opinions in a personal injury case. Organize questions by assumptions, methodology, and conclusions. Identify potential bias or financial incentives. Cite to specific report pages or testimony where possible."

- **Application:** Develop targeted cross-examination outlines and issue-spot weaknesses without drifting beyond the record.

4. Case Summaries for Motion Work

- **Prompt Example:** "Using only the cases I provide, summarize the holdings relevant to preservation and spoliation standards. Identify the legal rule, the court's reasoning, and limits on the holding. Do not supply citations beyond the provided materials. End with a verification list."
- **Application:** Generate concise, record-bound summaries for motion drafting while constraining hallucinated authority.

5. Privilege Logs and Document Review

- **Prompt Example:** "Create a privilege log template suitable for federal court, including fields for document date, author, recipients, privilege asserted, and description. Then review the attached draft entries for consistency and over- or under-description. Do not determine privilege; flag issues for attorney review."
- **Application:** Assist with formatting, consistency, and quality control in large privilege logs—never with privilege determinations themselves.

6. Voir Dire Preparation:

- **Prompt Example:** "Prepare a voir dire outline for a wrongful death case involving a commercial truck collision in Texas. Draft neutral, conversational questions designed to surface attitudes toward commercial drivers, corporate responsibility, and damages. Include follow-up questions and a brief note distinguishing cause challenges from strikes."
- **Application:** Develop voir dire questions that probe bias without argument or indoctrination.

7. Opening and Closing Argument Assistance:

- **Prompt Example:** "Based on the facts I provide, draft a preliminary opening statement outline for a breach of contract case involving delayed delivery of construction materials. Focus on chronology, responsibility, and damages. Write in plain language suitable for a jury. This is a draft for refinement, not a final script."
- **Application:** Generate themes, structure, and transitions—not finished courtroom rhetoric.

8. Crafting Jury Instructions:

- **Prompt Example:** "Using Texas Pattern Jury Charge concepts, organize the elements of a negligence claim—duty, breach, causation, and damages—into plain-English explanations

for internal use. Do not draft instructions for filing. Identify areas requiring verification against the current TPJC."

- **Application:** Help counsel understand and organize jury issues before drafting or requesting formal instructions.

9. Case Chronologies:

- **Prompt Example:** "Using only the attached documents, build a dated chronology of key events relevant to notice and response. Cite each entry to a document and page. If dates or facts are uncertain, flag them."
- **Application:** Create reliable chronologies for motion practice, mediation, or trial preparation, built incrementally and verified.

10. Building Better Prompts:

- **Prompt Example:** "Given the following task and constraints, help me draft a prompt that will extract information from uploaded interrogatories and produce a table identifying, for each interrogatory, likely information sources, custodians, and appropriate objections under the Texas Rules of Civil Procedure. Include self-check instructions and a verification list."

- **Application:** Use the LLM to design disciplined prompts for complex tasks. LLMs are often remarkably effective at structuring their own instructions—provided you supervise the result.

Ten Tips for Improved Prompts

The examples above show what LLMs can do in practice. The tips below focus on **how to ask** so the tool works the way a trial lawyer thinks. These are not rigid rules. You can get good results without following every one. But when outputs disappoint, the fixes are usually found here.

1. Precision Matters: Frame with Factual and Legal Specificity

Avoid generalities. Prompts grounded in concrete facts, posture, and governing law yield more useful results.

Example:

"Draft voir dire questions tailored for a Texas medical malpractice case involving an alleged delay in cancer diagnosis, where the plaintiff contends a breach of the applicable standard of care."

Specificity clarifies intent and reduces irrelevant or generic output.

2. Direct the LLM to Perform Defined Legal Tasks

Frame prompts around what you would ask a junior lawyer to do, not abstract research goals.

Example:
"Prepare a checklist for an FRCP Rule 26(f) conference in a complex products liability case, emphasizing ESI preservation, proportionality, and phased discovery."

Task-focused prompts align outputs with litigation workflow rather than academic exposition. In Harvey, this is most effective when tied to the documents and custodians already present in the matter vault.

3. Break Complex Requests Down into Logical Steps
Multifaceted prompts often fail because the model tries to do too much at once.

Example:
"First, list the elements of defamation under Texas law. Next, identify which elements are most commonly challenged at the pleading stage. Finally, outline arguments for dismissal based on those elements."

Stepwise instructions reduce oversimplification and make it easier to spot errors.

4. Instruct the Model to Assume a Role—With Limits
Assigning a role can improve focus and tone, but it should reflect how lawyers actually think, not fictional authority.

Example:
"You are a Texas trial lawyer evaluating competing positions on the scope of ESI production in a commercial contract dispute. Draft a neutral analysis suitable for a meet-and-confer."

Role prompts improve relevance without implying that the model replaces judgment or authority.

5. Request Multiple Variations for Strategic Flexibility
When optionality matters, ask for alternatives.

Example:
"Provide three alternative opening statement outlines for a wrongful death case: one emphasizing responsibility, one focusing on chronology, and one centered on damages."

Variations expand strategic choices and often reveal framing you might not have considered.

6. Use Iterative Refinement—But Restate What Matters
Refinement works best when you re-anchor the model to what matters most.

Example:
"Revise the following deposition outline to emphasize contradictions between the witness's testimony and prior internal emails. Maintain a neutral tone and cite source documents."

Iteration sharpens outputs, but remember that earlier context may fall out of scope as sessions grow.

7. Anchor Responses to Legal Frameworks—Not Speculation
Encourage analysis grounded in recognized legal standards rather than free-floating commentary.

Example:
"Summarize the elements of promissory estoppel under Texas law and identify common evidentiary proof offered for each element. Flag areas requiring independent verification."

This reduces conjecture and keeps the analysis tethered to doctrine.

8. Clarify the Desired Level of Detail and Tone
Tell the model who the audience is and how polished the output should be.

Example:
"Explain comparative responsibility in Texas premises-liability cases for a CLE audience—accurate, formal, but accessible to non-specialists."

Clear tone instructions prevent over-lawyering or oversimplification.

9. Use Conditional ("If–Then") Prompts for Nuanced Analysis
Conditional framing encourages balanced reasoning rather than one-sided conclusions.

Example:
"If a party fails to timely designate an expert under the Texas Rules of Civil Procedure, explain the risks of exclusion and the circumstances under which a court may allow late designation."

This approach surfaces contingencies and exceptions that matter in practice.

10. Require Verification and Source Limits
Do not assume the model will police itself.

Example:
"Explain spoliation standards applicable in Texas courts. Identify the governing principles and flag where independent review of current case law is required. Do not fabricate citations."

Explicit verification instructions reduce hallucinations and reinforce your supervisory role. In Harvey, pair this with instructions to rely only on materials within the matter vault.

Understanding Limitations and Ethical Guardrails

While powerful, LLMs are not substitutes for legal expertise or professional judgment. Neither ChatGPT nor Harvey provides authoritative legal research, and neither should be treated as a source of verified law or fact. These tools do not reliably access live court dockets or proprietary research services such as Lexis, Westlaw, or PACER, and—as discussed—they can generate incorrect, incomplete, or fabricated information.

Of the more than 700 *reported* cases to date where generative AI produced hallucinated content (typically fake citations), there is a startling commonality to all: *the lawyer simply failed to verify the authorities cited.* The common thread is—let's not mince words—lazy, careless lawyering.

Accordingly, every legal citation, quotation, and factual assertion must be read, verified, and owned by counsel before use. Courts have made clear they will not accept "the AI did it" as an excuse. Think of these tools as junior lawyers: fast, tireless, and often insightful—but requiring supervision, skepticism, and correction.

Ethical use also requires attention to confidentiality and privilege. In ChatGPT, that means understanding and configuring privacy controls, avoiding public share links, and being deliberate about what information is uploaded. In Harvey, governance is stronger by design—matter-centric workspaces, access controls, retention policies, and audit trails—but those safeguards do not relieve lawyers of their independent duties of competence, supervision, and confidentiality under Texas ethics rules. Technology can assist compliance; it does not outsource responsibility.

Finally, remember that LLMs reason statistically, not legally. They recognize patterns in language, not truth. When the record is thin, ambiguous, or contested, the risk of confident but wrong output increases. The antidote is disciplined prompting, scoped inputs, and explicit verification requirements—habits already familiar to good trial lawyers.

Final Thoughts for Trial Lawyers

As a law professor, longtime trial lawyer, and special master, I've found AI tools genuinely useful for early-stage drafting, ideation, organization, and breaking through inertia. I also concede the obvious: AI is no substitute for a skilled lawyer. It will not replace judgment, credibility, or the art of advocacy.

What it can do—when properly constrained—is accelerate routine work, surface patterns you might otherwise miss, and help you explore arguments more efficiently. Whether drafting deposition outlines, structuring chronologies, summarizing testimony, or preparing ESI protocols, LLMs can act as force multipliers that amplify what good lawyers already do well.

The lawyers who benefit most are not those who chase novelty, but those who treat AI the way they treat every other tool in practice: with curiosity, discipline, and skepticism. Embrace AI for

efficiency and organization, and you may find it becomes an indispensable part of your litigation toolkit. Just remember: *every tool needs a craftsman*, and no amount of technology can replace the judgment and insight of a seasoned trial lawyer.

Appendix: AI Prompt to Critique and Improve Keyword Searches

Adapted from Ball in Your Court, August 19, 2024

For more than two decades, I've wished for a tool that could critique keyword searches the way an experienced e-discovery professional would—flagging stop words, poor logic, overbreadth, and avoidable noise, and explaining *why* a query is likely to fail. Despite years of effort, nothing like it ever materialized.

Large language models finally make that idea practical.

Used properly, an LLM can function as a search-strategy reviewer—not a replacement for human judgment, but a disciplined assistant that explains why certain keywords or Boolean queries are inefficient, risky, or poorly suited to the platform in use. For lawyers who rely on lexical search (and most still do), this can be a meaningful upgrade.

How to Use the Prompt

1. Open ChatGPT (or an equivalent LLM).
2. Paste the prompt below into the message field **without submitting it yet**.
3. Use the **file-attachment ("+") control** next to the prompt field to upload a discrete list of keywords and Boolean queries for assessment (plain text, Word, or PDF).
4. Submit the prompt. If the response truncates, instruct the model to continue.

For best results, specify the **review platform** (e.g., Relativity/dtSearch, DISCO, Everlaw) and upload any relevant reference material (search syntax guides, stop word lists, ESI protocols, or primers).[96]

Prompt: AI Review of Keyword and Boolean Search Strategy

Role:
You are an e-discovery search consultant assisting litigation counsel.

Objective:
Evaluate the effectiveness of the attached keyword list and Boolean queries for use in litigation discovery. The goal is to improve recall and precision while avoiding unnecessary noise, overbreadth, and platform-specific pitfalls.

Scope and Constraints:
• Analyze only the uploaded keywords and queries.
• Do not assume facts not provided.
• Do not fabricate platform rules—flag uncertainty where platform behavior varies.
• Focus on defensibility and practical improvement.

[96] I saw improved results when I uploaded pertinent additional information, such as my own Primer on Processing in E-Discovery found here.

Analysis Framework (apply to each term or query):

1. **Stop-Word and Noise Risk**
   Identify terms likely to be ignored, diluted, or overly common, noting platform sensitivity where relevant.
2. **Synonyms, Variants, and Misspellings**
   Identify obvious linguistic variants, abbreviations, and common misspellings that may affect recall.
3. **Industry or Case-Specific Jargon**
   Flag missing domain-specific terminology suggested by the context of the queries.
4. **Boolean Logic and Structure**
   Assess operator usage, grouping, and proximity logic for internal consistency and likely effectiveness.
5. **Syntax and Platform Compatibility**
   Identify syntax issues that may fail or behave differently across common e-discovery platforms.
6. **Overbreadth and False-Hit Risk**
   Flag terms or constructions likely to generate disproportionate noise.
7. **Refinement Suggestions**
   Propose specific, defensible improvements to each query.

Output Requirements:
• Present results in a table with one row per keyword or query and one column per analysis category above.
• After the table, provide a short narrative summarizing common issues and best-practice guidance for refining lexical searches.
• End with a **Verification Checklist** identifying assumptions that require human confirmation.

Tone:
Analytical, conservative, and suitable for use in a defensible discovery workflow.

Practice Notes

This prompt is not intended to "optimize" searches automatically or replace testing against real data. Its value lies in explaining why searches fail, surfacing avoidable mistakes, and improving lawyers' understanding of lexical search mechanics.

As with all AI-assisted work, outputs must be reviewed, tested, and owned by counsel. Keyword search remains a blunt instrument—but until something better is universally affordable, reliable, and repeatable, we should at least sharpen it.



Exercise 16: Legal Hold

This is a two-part drafting exercise requiring you to read, research, think, and write. Your work should reflect a sensible and defensible balancing of legal duties, litigation strategy, budgetary considerations, and operational disruption. Estimated time: 3-4 hours. Do not devote so many additional hours that it interferes with other obligations.

Why did I position it after the article on using LLMs in litigation? Because you are encouraged to ethically use an LLM (such as Harvey or ChatGPT) to assist you. If you do, include the prompts you used with your submission. You are responsible for ensuring your submission meets all requirements and that any factual assertions or authorities are verified as authentic and reliable.

Scenario
Date: November 19, 2025[97]
Your Role: Associate attorney at Bevo ✯ Orange ✯Tower, P.C. in Austin, Texas
Client: Artemis Energy Solutions, Inc. (Defendant)
Supervising Partner: Tex S. Tower
The Situation: Mr. Tower hands you the materials in Appendix A *as they existed on November 19, 2025*. He asks you to draft a plan for a defensible and cost-effective legal hold in response to the lawsuit.
Important: Do not use information that would not have been available on November 19, 2025. You only know what your firm knew on that date.

Deliverables
You must submit **two separate documents**:
PART 1: Legal Hold Memorandum (Maximum 10 pages)
A succinct memo to Mr. Tower describing your recommended legal hold plan. The memo must address **all** of the following issues:

#	Issue to Address
1	**Identification of ESI Sources:** How should Artemis identify potentially responsive sources of ESI and preserve them?

[97] **This date is important.** Don't avail yourself of information that is not yet available to you according to the timeline of events. You only know what was known to your firm on November 19, 2025.

#	Issue to Address
2	**Relevance Determination:** What categories of information are potentially relevant to the claims and defenses in this case?
3	**Key Custodians:** By role (or name, if known from the materials), who are the key players and custodians subject to the hold?
4	**Custodian Obligations:** What must key players and/or custodians do? Do they all do the same thing in the same way?
5	**Custodial vs. Centralized Hold:** Should you rely on custodial-directed hold alone? If not, what additional measures are warranted?
6	**IT Department Role:** What must IT do to implement and maintain the hold?
7	**Backup and Retention Policies:** Must Artemis suspend backup tape rotation or its document retention policy? If so, how?
8	**Forensic Imaging:** Must Artemis image drives and phones? If so, by what means and whose devices?
9	**Employee Transitions:** Are special procedures needed for new hires and departing employees?
10	**Third-Party Data:** What about information held by third parties (cloud providers, vendors, etc.)?
11	**Communication with Opposing Counsel:** Should Artemis reach out to Lost Creek's counsel on any preservation issues? If so, how and when?

The memo must also include:

- A **timetable** for implementation
- **Methodology** for implementation
- **Follow-up and audit strategies** (if advisable)
- **Reasonable projections** of costs, resources, and personnel required

Practical constraints to address: Mr. Tower emphasized that Artemis "can't shut down to do this" and "won't spend disproportionately on the legal hold versus what the case is worth." Make every dollar count. He adds, "*Keep the memo succinct—it MUST be no more than ten pages—and we'll talk further once I've reviewed your checklist, advice, recommendations and exemplar notices. I'm not good with this e-discovery stuff, so I'm counting on you to steer me through it like a champ.*"

PART 2: Legal Hold Notice(s) (Maximum 3 pages each)

Draft the actual legal hold notice(s) that Montgomery Bonnell (Artemis CEO) should send to key custodians and IT personnel.

Requirements:

- Each notice should be **no longer than 3 pages** (1-2 pages preferred)
- If the same notice applies to multiple recipients, identify them by name or role
- You may create different notices for different recipient groups if appropriate (e.g., separate notices for executives, engineers, IT personnel)
- Notices should be practical and actionable for the recipients

Grading Criteria

Your submission will be evaluated based on:

- **Clarity** – Is your writing clear and easy to follow?
- **Succinctness** – Have you been concise while still being complete?
- **Propriety of Scope** – Is the hold appropriately tailored to the claims and defenses?
- **Proportionality** – Are your recommendations proportional to the case value and burden?
- **Practical Direction** – Can your plan actually be implemented?
- **Creativity** – Have you shown thoughtful problem-solving?
- **Consideration of Recipients** – Have you accounted for who will receive notices, their roles, what they likely possess, and whether they can be relied upon to comply?

Permitted Resources

You may consult **any** of the following:

- Written and online resources (including Harvey, Google, PACER filings, journal articles, Lexis, Westlaw, form books)
- Practicing attorneys, judges, professors, law students
- IT personnel, e-discovery consultants, vendors
- Anyone else who may provide useful input

However:

- Do **not** present anyone else's work as your own
- You may borrow from published forms and examples, but you **must provide full and proper attribution**
- Presenting someone else's work without attribution constitutes plagiarism
- Mindless use of boilerplate (whether from you, Harvey, or ChatGPT) is strongly discouraged

---

Submission Checklist

Before submitting, confirm:

- **Part 1 (Memo)** is no more than 10 pages
- **Part 1** addresses all 11 issues listed above

- **Part 1** includes timetable, methodology, follow-up/audit strategies, and cost projections
- **Part 2 (Notices)** – each notice is no more than 3 pages
- **Part 2** identifies recipients by name or role
- All sources are properly attributed
- All factual assertions and legal authorities have been verified
- If you used an LLM, you have included your prompts
- Your submission reflects the information available as of **November 19, 2025** only

Mobile Data in E-Discovery

Do you live two lives, one online and the other off? Millions lead lives divided between their physical presence in the real world and a deeply felt presence in virtual worlds, where they chat, post, friend, like and lurk. They are constantly checking themselves in and checking others out in cyberspace. In both worlds, they leave evidence behind. They generate evidence in the real world that comes to court as testimony, records and tangible items. Likewise, they generate vast volumes of digital evidence in cyberspace, strewn across modern electronic systems, sites, devices and applications.

Trial lawyers who know how to marshal and manage evidence from the real world are often lost when confronted with cyber evidence. Here, we take an introductory look at discovery from mobile devices.

The Blessing and Curse of ESI

Even if you don't know that data volume is growing at a compound annual rate of 42 percent, you probably sense it. This exponential growth suggests there's little point feeling overwhelmed by data volumes *today* because we are facing volumes *ten times as great in five years*, and *fifty times* as great in ten years. Today is tomorrow's "good old days."

There's going to be a lot more electronic evidence; but there's still time to *choose* how you deal with it.

A lawyer can curse electronic evidence and imagine he or she is preserving, collecting and requesting all they need without cell phones, the Cloud and all that other 'e-stuff.'

Or, the lawyer can see that electronic evidence is powerful, probative and downright amazing, and embrace it as the best thing to happen to the law since pen and ink. Never in human history have we enjoyed more or more persuasive ways to prove our cases.

Mobile Miracle

According to the U.S. Centers for Disease Control and Prevention (CDC), roughly 73% of U.S. adults now live in wireless-only households (no landline), and among adults ages 25–29 the wireless-only rate exceeds 80%. Studies also show that many people start using their smartphones within minutes of waking up and, for most, it's the very first thing they do, ahead of brushing their teeth or answering nature's call. For those in the lowest economic stratum, mobile phones remain the principal and often sole source of internet connectivity.

Apple reports more than 2.2 billion active Apple devices as of 2024, and the App Store offers roughly 2 million iOS apps globally, cumulatively downloaded and installed tens of billions of times.

Worldwide, Android accounts for roughly 70% of the mobile OS market, versus about 30% for iOS. The United States Supreme Court summed it up handily a decade ago: "*Today many of the more than 90% of American adults who own cell phones keep on their person a digital record of nearly every aspect of their lives*."[98]

Our digital lives spawn a cornucopia of probative evidence, gathered through an ever-expanding array of sensors, radios, and embedded capabilities. Modern smartphones incorporate microphones; biometric sensors (fingerprint and facial recognition); barometers; accelerometers; magnetometers; gyroscopes; proximity, touch, ambient light, and moisture sensors; high-resolution, multi-lens still and video cameras; and global navigation satellite systems. They also include multiple radio technologies—cellular, Wi-Fi, Bluetooth, ultra-wideband, and near-field communications—continually interacting with surrounding devices and networks. As well, users generate vast quantities of content and metadata through texts, encrypted and ephemeral messaging, email, social and collaboration platforms, voice assistants, payment systems, and continuous requests for web- and app-based services.

Smartphones may serve as a source of the following categories of data:
- Subscriber and SIM/eSIM data
- Device and user files
- Wi-Fi and Bluetooth connection history
- Call logs and telephony metadata
- Photographs, videos, and associated metadata
- Contacts and contact interaction history
- Geolocation and movement data (GPS, Wi-Fi, cellular, and app-derived)
- Email content and metadata (note: standard computer backups generally do not preserve mail)
- Voicemail and voice recordings
- Chat and messaging data (SMS, MMS, RCS, and app-based messaging like WhatsApp)
- Application data, including caches and local databases
- Web browsing history and search activity
- Calendar entries and reminders
- Bookmarks and saved links
- Task lists and notes
- Health, fitness, and sensor-derived activity data

[98] *Riley v. California*, 573 U.S. 373 (2014).

• Payment, wallet, and transaction artifacts
• Music, video, and other rich media

Mustering Mobile

For the past decade and more, lawyers have learned—sometimes grudgingly—to cope with electronic evidence. We know how to acquire the contents of hard drives. We understand forensic imaging and targeted collection. We've grown more competent at culling, filtering, and processing PC and server data. Most corporate information still resides within standardized file systems, email platforms, and collaboration tools, and even the once-fearsome databases tend to be built atop a relatively small number of familiar architectures. Better still, we have mature tools and an experienced cadre of practitioners to deploy them.

Now, let's talk mobile.

Let's talk interfaces. For decades, forensic acquisition from storage media rested on a small and stable set of physical interfaces. PATA gave way to SATA, but the transition was evolutionary, not chaotic. USB has been with us for more than twenty years, and despite generational speed changes, USB devices have remained physically interoperable and backward compatible.

Mobile devices tell a different story. Connectors have come and gone—30-pin docks, micro-USB, Lightning, USB-C—with regulatory pressure now forcing a measure of standardization, at least externally. Internally, however, the protocols, security layers, and access controls evolve relentlessly. Secure enclaves, hardware-backed encryption, and sealed storage architectures mean that physical access no longer equates to evidentiary access. The cable may fit, but the data can remain firmly out of reach.



Let's talk operating systems. For much of modern computing history, two operating systems—Windows and macOS—have dominated the personal computer landscape, providing a degree of predictability essential to forensic practice. In the mobile world, market share has consolidated around Android and iOS, but uniformity remains illusory. Android exists in countless vendor-specific forks, versions, and security postures. iOS, though tightly controlled, evolves rapidly, with frequent updates that can materially alter acquisition feasibility from one release to the next. Add wearables, tablets, and companion devices into the mix, each with their own operating variants and data stores, and the mobile ecosystem becomes less a platform and more a moving target.



In short, while traditional e-discovery matured around stability and standardization, mobile evidence demands that lawyers confront volatility, opacity, and technical fragility—often with higher stakes and far less margin for error.

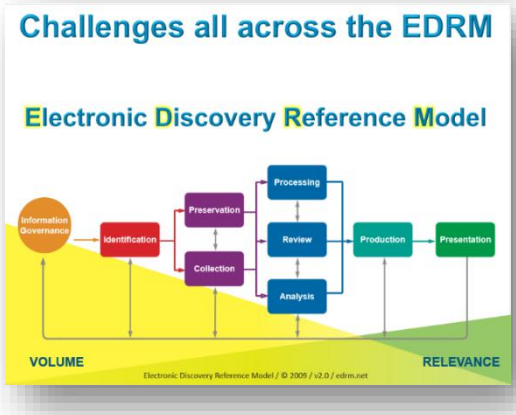
Let's talk encryption. There is content on phones and tablets—especially communications and application data—that cannot be acquired without user cooperation or the fortuity of an exploitable vulnerability. Apple now asserts that encryption is so deeply and deliberately integrated into its devices and services that it lacks the technical ability to access certain user content even when served with lawful process. Hardware-backed key storage, secure enclaves, and end-to-end encryption have shifted evidentiary access from a matter of legal authority to one of technical happenstance. In practice, law enforcement and civil litigants must rely on commercial forensic tools whose success depends on undisclosed vulnerabilities and transient exploits. That a system of justice committed to transparency, consistency, and repeatability should hinge on the discovery of software flaws is not a feature of modern encryption—it is a defect.

Let's talk tools. Anyone can move information off a PC. Forensic disk-imaging software is inexpensive, widely available, and simple enough for routine use. Write blockers suitable for forensically sound acquisition are commodity hardware. Mobile devices are different. What,

exactly, do you have that will reliably preserve the contents of an iPhone or iPad? A local device backup? A cloud sync? Each captures only a curated subset of data, excludes others entirely, and obscures what is omitted. Even so-called "full" backups leave behind secure storage, certain application artifacts, ephemeral messages, and device-resident secrets by design.

And assuming you do obtain a backup, what then? Mobile data does not arrive as emails and loose files neatly ready for review. It arrives as proprietary databases, encrypted containers, binary blobs, and fragments of state scattered across apps and services. Normalizing that material into something reviewable requires specialized tooling, expert interpretation, and lossy transformations. Those tools are neither free nor incidental: mobile forensic software licenses routinely cost orders of magnitude more than disk-imaging tools, require continual paid updates to remain viable, and often impose per-device or per-extraction constraints. Most e-discovery platforms remain document-centric, and mobile evidence must be flattened, translated, or reduced to fit, often at the cost of context, metadata, and meaning. There is still no turnkey path from mobile acquisition to defensible review—and no app that makes it so.

Let's talk time. Depending on device, OS, security posture, and extraction method, acquiring and rendering mobile data can take many hours and sometimes longer—often far longer than imaging conventional storage. And unlike drive imaging, mobile acquisition is harder to parallelize because it is device- and session-dependent, frequently requiring supervision and specialized tooling. Scalability—the defining virtue of modern e-discovery—largely evaporates when the evidence lives in pockets and purses.



Challenges All Across the EDRM

The Electronic Discovery Reference Model or EDRM is an iconic workflow schematic that depicts the end-to-end e-discovery process. It's a handy context in which to address the ways that mobile devices pose challenges in e-discovery.

Information Governance:

Organizations embrace **Bring Your Own Device (BYOD)** policies to enhance mobility and productivity, but BYOD blurs the line between personal and enterprise data. Permitting employees to connect personal phones and tablets to corporate systems does not, by itself, secure access to those devices for e-discovery. Meaningful governance requires advance, informed consent—typically embedded in employment agreements or acceptable-use policies—together with realistic disclosures about what data may be accessed, when, and by whom. Even then, consent does not guarantee technical access, and policies routinely outpace what devices will actually yield. Consent also does not supplant statutory limits (e.g., the

Stored Communications Act, state privacy laws, or GDPR for EU residents); scope, notice, minimization, and segregation remain essential.

Identification:
Mobile devices are replaced, upgraded, lost, and repurposed far more frequently than laptops and desktops; accordingly, it's harder to maintain an up-to-date data map for mobile sources. And the good news is that remote collection is no longer science fiction. Service providers and vendors now market remote mobile collections using a mix of approaches—cloud-side acquisition (e.g., iCloud/Google with valid credentials and 2FA), targeted artifact harvesters running on a custodian's computer while the phone remains in their possession, and guided in-app exports or screen-sharing workflows—allowing custodians to keep physical possession of their devices while data is captured for discovery.

But the fine print matters. Much of what passes for "remote collection" is necessarily targeted and method-specific: it may capture a defined subset of artifacts (messages, attachments, photos, app data) rather than a comprehensive, forensic snapshot of the device; it may depend on what is synced to the cloud versus what resides locally; and it may vary materially by device model, operating system version, and app configuration. Add to this the app-centric architecture of mobile platforms—dozens of apps, each with unique data stores and retention behaviors, many tethered to third-party cloud services—and the task of "identifying what exists" becomes less an inventory exercise and more a moving-target inquiry.

Preservation:
It remains common for organizations and individuals to possess mobile devices without any practical, defensible means to preserve their contents when litigation is reasonably anticipated. While cloud backups, screenshots, and in-app exports are often pressed into service, these methods are selective, user-dependent, and opaque as to what they omit. Even seemingly simple preservation tasks—such as retaining text or chat messages—can defeat well-intentioned custodians, who quickly discover that modern smartphones offer no native, comprehensive, and verifiable way to export or lock down message content, metadata, and context. Auto-deletion settings, ephemeral messaging, routine device upgrades, and account-level sync behaviors further undermine preservation unless affirmative, informed steps are taken early.

Collection:
Because there are few reliable ways to preserve mobile data *in situ*, preservation of mobile evidence generally entails collection from the device itself or from associated cloud accounts. Whether performed locally or remotely, collection typically requires specialized tools, technical expertise, and device- or account-specific workflows. Compared to PC and server environments, mobile collection remains harder, slower, and costlier, with greater variability in completeness depending on device model, operating system version, encryption state, and user configuration.

Processing:
The unpacking, parsing, indexing, and reduction of mobile data—collectively referred to as "processing"—is complicated by heterogeneous operating systems, app-centric storage, and

pervasive encryption. Mobile evidence rarely arrives as discrete, review-ready files; instead, it consists of proprietary databases, application artifacts, caches, and state information that must be interpreted before it can even be searched. Data that cannot be decrypted cannot be processed at all, and processing outcomes often depend as much on tool capability and versioning as on the underlying evidence.

Review:
Review of electronic evidence typically occurs in document-centric review platforms such as Relativity and similar systems. While these platforms continue to evolve, they remain ill-suited to many forms of mobile evidence. Messages, reactions, edits, deletions, attachments, location histories, and app-specific metadata are frequently flattened, reordered, or stripped of context to fit legacy review paradigms. As a result, mobile evidence is often reviewed as a pale abstraction of its native form, with consequential loss of meaning.

Analysis:
Much mobile data—particularly the terse, conversational messaging that dominates mobile use—does not lend itself well to advanced analytics such as predictive coding. The brevity, informality, and contextual dependence of chats, texts, and reactions frustrate traditional relevance-ranking models trained on longer, document-like content. While analytics may assist in clustering or prioritization, they are far less reliable substitutes for human review in mobile-heavy datasets.

Production:
Finally, production poses its own challenges. How does one produce data unique to a particular app in a form that can be viewed, searched, and understood by recipients who lack both the device and the application? Too often, production entails lossy conversions—PDFs, spreadsheets, or static transcripts—that sacrifice metadata, threading, timing, and interactivity. Standards for mobile production remain immature, and significant work remains to ensure that mobile evidence can be produced in a manner that preserves integrity, completeness, and utility as it moves from proprietary device and app environments into conventional discovery workflows.

Where messaging is central, stipulate production specifications up front (e.g., Relativity Short Message Format (RSMF) or a structured JSON/CSV deliverable) preserving conversation IDs, participant roles, reactions, edit/delete indicators, time zone normalization, and attachment linkage. Provide conversation-level metadata and, where feasible, renderings that visually reflect native threading to aid review.

Challenges end-to-end
Mobile evidence does not merely complicate one or two discrete stages of the EDRM; it stresses the entire model at once. Breakdowns in identification propagate into preservation failures, incomplete preservation constrains collection, limited collection undermines processing, and lossy processing distorts review, analysis, and production. Unlike enterprise systems, mobile evidence resists clean, sequential handoffs between phases. The comforting assumption that deficiencies can be cured downstream is misplaced.

Compounding these technical challenges is an inescapable human one. Mobile devices are personal in a way corporate systems are not. Users are often reluctant—sometimes justifiably so—to surrender physical possession of their phones or to permit access to personal data commingled with work content. Privacy concerns, fear of overcollection, and simple attachment to the device itself introduce delay, resistance, and negotiation at precisely the moment when prompt, decisive action is required. With mobile data, what is missed early is often lost forever, and what access is delayed may never be meaningfully recovered.

So, What's a Lawyer to Do?

Mobile evidence is unlike anything we previously encountered in e-discovery. Although the ecosystem of tools and services has improved, preserving and processing mobile data remains more constrained, more variable, and more expensive than traditional PC or server evidence. None of that relieves lawyers of the duty to preserve mobile data when litigation is reasonably anticipated or to produce it when discoverable.

The first hurdle is often persuading the device's user to permit access at all. Mobile devices are unique in their intimacy and indispensability. Unlike computers, phones are constant companions—rarely out of arm's reach and often essential to daily functioning. For many users, it is inconceivable to part with a phone for even a few hours, much less overnight or indefinitely. Most would be unable to contact their spouse, children or closest friends without access to the data stored on their phones. Their mobile phone number may be the only way they can be contacted in the event of an emergency. Their phones wake them up in the morning, summon their ride to work, buy their morning

Geolocation

Cell phones have always been locatable by virtue of their constant communication with cellular towers. In the United States, federal law also requires that phones support precise, GPS-class location to enable Enhanced 9-1-1 emergency services. Modern smartphones routinely determine location with single-digit meter accuracy by blending signals from global navigation satellite systems (GPS, GLONASS, Galileo, BeiDou), cellular triangulation, Wi-Fi positioning, Bluetooth beacons, and onboard motion sensors. Location is not a single data point but a continuously inferred narrative.

Your phone is not merely tracking itself; it is constantly negotiating location with the world around it. It polls nearby Wi-Fi routers, exchanges identifiers with Bluetooth devices, records cell handoffs, logs motion, and timestamps it all. Turning off location services does not silence this activity; it simply changes how location is inferred. Any time a phone can communicate—by voice, text, or data—it generates geolocation artifacts. And when that capability is interrupted—airplane mode, power loss, SIM removal—that interruption itself leaves a record.

What's new is that cellular coverage is no longer the outer boundary. Recent smartphones now support direct satellite communication for emergency SOS, limited messaging, and location sharing beyond terrestrial networks. These satellite interactions create additional, highly precise, time-stamped location records maintained by devices and service providers.

Phones, however, are only the beginning. The Internet of Things has become a sprawling geolocation engine. Smart thermostats infer presence and absence. Home cameras record comings and goings and store them in the cloud. Door locks, lighting systems, vehicles, fitness trackers, and virtual assistants all generate location-adjacent data, often automatically and persistently. Convenience is purchased with persistence.

Geolocation today is not something devices occasionally reveal; it is something our environments continuously infer. "Hey Siri, can you be my alibi?" is no longer a joke so much as a caution.

coffee and serve as an essential link to almost every aspect of their social and business lives. Phones have become cognitive prosthetics, the other half of our brain.

Accordingly, when you advise a custodian that their phone must be accessed—or worse, temporarily surrendered—for discovery, you should expect resistance, anxiety, and sometimes concealment or preemptive deletion. A plan is essential. You need a credible timeline, clear scope limits, and, where feasible, a substitute device configured to preserve continuity of access without altering evidence. Credentials are critical: passcodes, biometric unlock and 2-factor cooperation, cloud account access, and backup passwords should be obtained promptly and tested. Be alert to the risk of remote wipes, auto-deletion, or "helpful" factory resets undertaken out of panic or misunderstanding. Mobile users are often far more proprietary and protective of their devices than they ever were of their computers.

Options for Mobile Preservation

In civil cases, before you do anything with a mobile device, it's good practice to back it up using the native application (e.g., a computer backup via Finder on macOS or the Apple Devices app/iTunes on Windows for iPhones and iPads, and preserve the backup). This gives you a path back to the data and a means to provision a substitute device, if needed. Then, you have five options when it comes to preserving data on mobile devices, selected based on feasibility, proportionality, cost and risk:

1. Prove You Don't Have to Do It.

If you can demonstrate that no unique, relevant information resides on the mobile device—and that all potentially responsive content is preserved from other, more accessible sources—you may avoid collecting from the device itself. Given modern mobile usage, this justification is increasingly difficult to sustain and must be supported with proof, not assumption.

2. Sequester the Device.

Where immediate analysis is unnecessary, the device itself may be preserved by isolation: replacing the phone, disabling network connectivity, and maintaining the original intact. Credentials should be obtained and tested in advance, and chain of custody documented.

3. Use Targeted Software or Services.

When preservation obligations are narrow, targeted tools or managed services may be used to capture defined categories of data (e.g., texts, chats, photos). These approaches are selective by design and must be evaluated for completeness, metadata integrity, and defensibility.

4. Hire a Pro and Forensically Acquire What You Can.

Professional mobile forensic acquisition remains the most robust technical approach, though increasingly constrained by encryption and platform security. Logical and file-system extractions,

often combined with cloud acquisition, are now the norm; physical imaging is rare and device-specific. Unobtainable data should be identified and documented early.

5. Preserve by Command, Configuration, and Oversight.

Where technical acquisition is impracticable or disproportionate, preservation may instead be achieved by controlling user behavior and device settings. This includes issuing device-specific hold instructions; disabling auto-deletion, ephemeral messaging, and retention timers; suspending OS and app updates where feasible; preserving cloud-side data contemporaneously; and requiring periodic attestations of compliance. Though inherently fragile, this behavioral preservation—when prompt, documented, and monitored—has become an unavoidable component of reasonable mobile preservation.

Coda: Preservation Without Illusions

Mobile preservation rarely admits of a single, clean solution. Each of the foregoing options trades completeness for feasibility, speed, cost, or cooperation. None—alone or in combination—can guarantee a perfect evidentiary record. Modern smartphones are encrypted, app-centric, continuously in motion, and deeply personal; they do not submit readily to the kind of decisive, one-time capture that characterized earlier generations of electronic evidence.

Accordingly, preservation of mobile data should be understood not as a discrete event but as a series of reasoned choices made under real technical and human constraints. What matters is not whether every artifact is captured, but whether counsel acted promptly, scoped the effort sensibly, selected methods proportionate to the stakes, and documented both the steps taken and the limits encountered. In the mobile context, defensibility turns less on technical completeness than on informed judgment, transparency, and good faith.

Simple, Scalable Solutions to iOS-Device Preservation

We've spent the last chapter mapping the mobile evidence landscape: interfaces, operating systems, encryption, tools, time, and the EDRM pinch points. Now let's put that groundwork to work with a practical, defensible workflow you can deploy without prying phones from custodians' hands.

Can anyone doubt the changes wrought by the modern "smart" cellphone? All day, we see drivers looking at their phones, some so engrossed they fail to move when the light turns green. Phones have altered the progress of traffic in every community. We've turned into distracted digital zombies behind the wheels of our cars.

Set aside the hyperbole and the headlines. The point is not to relitigate the statistics cataloged earlier but to acknowledge the obvious: *phones now mediate our movements, our messages, our memories, and our money*. Look around you (if you can pry your eyes from your screen). How many are using their phones? At a concert, how many are experiencing it through the lens of their phone cameras? How many selfies? How many texts? How many apps?

Smartphones have changed us, and lawyers ignore this sea change at their peril.

Today, if you fail to advise clients to preserve relevant and unique mobile data when under a preservation duty, you're committing malpractice.

Yes, I used the "M" word, and not lightly.

Two things have changed such that we can't hide our heads in the sand anymore when it comes to mobile evidence. First, the data on phones and tablets is not a copy of information held elsewhere. It's unique and often the most probative evidence in the case. Second, the locking down of phone content has moved routine preservation from the esoteric realm of digital forensics to the practical world of backups and targeted app exports. These developments mean that, notwithstanding the comforting rationales lawyers trot out for ignoring mobile, *the time has come to accept that mobile is routinely within the scope of preservation obligations.*

Once, it was convenient to ignore mobile in e-discovery. Mobile was a black hole. You had to hire technical experts to use expensive tools to preserve the contents of phones, and it was like pulling teeth to get users to surrender their devices for the time it took to image them. Users protested, "My mobile phone is the only way the kids' school can reach me in an emergency, and I can't use another phone because everyone texts now, and anyway, WHO REMEMBERS PHONE NUMBERS ANYMORE?"

A few years ago, mobile phones shared some of the characteristics of personal computers in that they held latent data that could be recovered using specialized tools sold for princely sums by a couple of shadowy tech companies. So, the preservation of mobile devices slipped into the shadows, too. Phones and tablets were forensic evidence, and only forensic examiners could collect their contents.

Although everyone uses mobile devices all day, the contents of mobile devices were often deemed "not reasonably accessible" because it was too costly and burdensome to preserve phones. That excuse has expired.

There are now straightforward, low-cost ways to preserve relevant mobile content without taking phones away from users. Because it's become quick, easy, and inexpensive to preserve, mobile content is readily accessible, and its preservation (when potentially relevant) is commonly proportionate to the needs identified in the prior "Identification" and "Preservation" discussions.

In e-discovery today, the forensic-level preservation of phones---the sort geared to deleted content and esoteric artifacts--is increasingly impractical for routine civil matters. Modern smartphones are locked down hard. Content is encrypted, and even the keys to access encrypted content are themselves protected in hardware. Phone forensics isn't what it used to be. More and more, we can't reach the recoverable artifacts, and chasing it can waste time and money better spent elsewhere.

Yet it's quick and simple for a user to generate a full computer backup of an iPhone without surrendering possession. On macOS, you use Finder; on Windows, you use the Apple Devices app (or iTunes where still installed). The custodian can create the backup and place it in a designated location for safekeeping by counsel or IT. Will this be a "forensic image"? Strictly speaking, no. But as phone security strengthens, "forensic imaging" is overkill and increasingly unlikely to yield data of the sort encompassed by a routine e-discovery preservation obligation. Not every case is a job for CSI.

A full unencrypted computer backup of an iPhone will not encompass everything a forensics lab might glean using specialized tools under ideal conditions. But so what? The standard for ESI preservation isn't perfect (but it isn't lousy either). Preserving by backup isn't perfect; it's good enough for many matters, and far superior to what is too often done instead, *i.e.,* doing nothing to preserve mobile evidence.

Preservation of mobile device content must become a standard component of a competent preservation effort except where the mobile content can be shown to be beyond scope. Mobile content has become so relevant and unique, and the ability to preserve it so undemanding, that the standard must be preservation.

The need is obvious. You may be reading this on your phone or tablet. If not, it's a virtual certainty that your phone or tablet is nearby. Few of us separate from our mobile devices for more than minutes a day. It took thirty years for e-mail to displace paper as our primary target in discovery. It's taken barely ten for mobile data, especially texts and chats, to unseat e-mail as the Holy Grail of probative electronic evidence. *Mobile is where evidence lives now*; yet in far too many cases, mobile data remains "off the table" in discovery. It's infrequently preserved, searched, or produced.

No one can say that mobile data isn't likely to be relevant, unique, and material. Today, the most candid communications aren't e-mail; they're messages. Mobile devices are our principal conduit to online information, eclipsing use of laptops and desktops. Texts and app data often reside primarily-and sometimes exclusively-on mobile devices or in personal clouds tied to those devices.

No one can say that mobile data isn't reasonably accessible. We use phones continuously, for everything from games to gossip to geolocation. Messages can be remarkably durable on-device, and mobile content readily replicates as computer backups and cloud syncs. The mobile preservation burden pales compared to what we take for granted in preserving potentially relevant ESI on servers and personal computers.

That's what this chapter is about: modest burden. Building on the obstacles outlined under "Challenges All Across the EDRM," the preservation burden is minimal when it comes to capturing the most common and relevant mobile data. I'll go so far as to say that the burden of preserving mobile device content, even at an enterprise scale, can be less than that of preserving a comparable volume of data on laptops or desktops. Too, the workflows are as defensible and auditable as any we accept as reasonable in meeting other ESI preservation duties.

The following three principles anchor a sensible approach:

When mobile data may be unique and relevant, it should be preserved in anticipation of litigation-especially when the preservation burden is trivial, as it is with computer backups. You can demonstrate the absence of relevant data by, *e.g.*, sampling the contents of devices; but standing alone, a policy barring the use of a device to store relevant data is not proof that the device wasn't used to store data; practice too often belies policy, particularly in messaging.

Mobile preservation should be a customary feature of a defensible litigation hold; absent issues of spoliation, few matters warrant the added cost of forensic-level preservation or the disruption of separating users from their phones.

Legitimate concerns respecting personal privacy and privilege do not justify a failure to *preserve* relevant mobile data, although they will dictate how data is protected, searched, reviewed, and produced.

Several provisos are also in order:

The method outlined here is one simple, scalable, defensible way to preserve iPhone content. It's not the only way or even the optimum way.

Preservation isn't production; lawyers' abilities to search, review, and produce mobile content in utile, complete forms hasn't kept pace with our need to do so. This chapter is about routine *preservation*; it addresses downstream processes only insofar as ensuring the information preserved remains readily amenable to all methods of search, review, and production that follow.

The duty to preserve is real and immediate; yet, there's room for honest debate about what depth of mobile preservation is warranted, matter by matter. If you reject a preservation method because you deem it flawed, *is the alternative a superior method-or nothing at all?* "Doing nothing" is rarely the proper choice with mobile evidence. Preserving "most" is better than "none," but considerations of risk may dictate preserving "all" over "most." In turn, proportionality may elevate "most" over "all." Ask: *Is the incremental cost of forensic-level preservation justified by relevant and unique content? If not, might "good" be good enough?*

Defensibility is the lodestar

Ignoring mobile evidence isn't the path taken by competent, ethical attorneys. We must employ methods of preservation that aren't unduly costly or burdensome yet pose little risk that a judge will find them unreasonable. The essence of defensibility is being able to show that an action was prudent in light of what was known-or should have been known-when the action occurred. If mobile content required to be preserved is lost, the Court will ask whether the method employed was reasonably calculated to guard against loss or corruption. That turns on the method, its deployment, and its oversight. Hence, audit and verification.

Defensible Custodial Hold

The predominant approach to preservation in e-discovery remains the legal hold: custodian-directed preservation. It has well-known flaws. It asks custodians to judge relevance and privilege, to run searches without the right tools or training, and to spend effort without time, resources, or guidance. It doesn't deter the custodian keen to destroy or alter inculpatory or embarrassing data.

So, if custodian-directed preservation is problematic, why put custodians in charge of preserving their own devices instead of handing the devices over to digital forensics experts for imaging? Isn't that inviting the fox to guard the henhouse?

The signal challenge to preserving mobile devices is persuading custodians to part with them. By empowering custodians to preserve the data themselves, custodians need never surrender custody of their devices. Accordingly, users are less threatended by the process and less inclined to fight or subvert it. Backing up an iPhone is simple and quick; and crucially, the process affords the custodian neither the need nor the practical ability to select or omit content. Compare that to tasking a custodian to collect e-mail or documents, where it's easy to overlook or deliberately omit material with little chance of detection.

How It's Done

For iOS devices, the custodian-directed backup method marries speed, familiarity, and control. If the earlier "Let's talk tools" section told you what's hard, this is what's easy. On macOS, the custodian connects the device and creates a computer backup using Finder. On Windows, the custodian uses the Apple Devices app (or iTunes where still installed). For preservation workflows that rely on broad accessibility downstream, I favor an unencrypted computer backup. Encrypted backups include more sensitive data (e.g., Health, Keychain, Apple Pay settings), but many discovery tools cannot process encrypted backups, even with credentials; none can process them without. Collect unencrypted data to avoid credential dead-ends; then protect the resulting backup by compressing it to a container and encrypting *the container* using a matter-level passphrase. This yields one consistent credential for counsel and providers and avoids harvesting end-user passwords into risky spreadsheets.

Step-by-step: Create a computer backup on macOS (Finder)

If you've just come from the "Mustering Mobile" and "Preservation" sections, this is the concrete, repeatable workflow that operationalizes those principles.

Connect the iPhone/iPad to the Mac with a reliable data-capable cable.

Open Finder and, under Locations in the left sidebar, select the device when it appears. If prompted, tap "Trust This Computer" on the device and enter the device passcode.

In the General tab, under Backups, select "Back up all of the data on your iPhone to this Mac." Ensure "Encrypt local backup" is unchecked for an unencrypted backup (check only if you have planned encrypted-backup processing and credential management).

Click "Back Up Now" and wait for completion; do not disconnect the cable or use the device during backup.

When finished, note the “Latest backup” timestamp shown in Finder. Do not click “Restore” or “Sync.”

Locate the backup folder at: ~/Library/Application Support/MobileSync/Backup/ (use Finder > Go > Go to Folder to paste the path). Identify *the most recent folde*r by modified time.

Compress the most recent backup folder into a single archive (e.g., .zip) and then encrypt that container with a matter-level passphrase per your legal hold protocol.

Step-by-step: Create a computer backup on Windows (Apple Devices app or iTunes)

Connect the iPhone/iPad to the Windows PC with a reliable data-capable cable.

Open the Apple Devices app (or iTunes, if present). Click the device icon when it appears. If prompted, tap “Trust This Computer” on the device and enter the device passcode.

In the Summary/General view, under Backups, choose “This computer.” Ensure “Encrypt local backup” is unchecked for an unencrypted backup (check only if you have planned encrypted-backup processing and credential management).

Click “Back Up Now” and wait for completion; avoid disconnecting the cable or using the device during backup.

Confirm completion by noting the “Latest backup” timestamp in the Apple Devices app/iTunes.

Locate the backup folder at: C:\Users[username]\AppData\Roaming\Apple Computer\MobileSync\Backup\ (enable viewing hidden items if necessary). Identify the most recent folder by modified time.

Compress the most recent backup folder into a single archive (e.g., .zip) and then encrypt that container with a matter-level passphrase per your legal hold protocol.

Scope: What You Won’t Get with a Backup

An iPhone backup is not a forensic image. It is a curated export, deliberately designed to exclude entire categories of data. Most notably, an iPhone backup will not preserve email content stored on the device when that email is synchronized from a server (e.g., Exchange, Gmail, IMAP). This omission is by design: Apple assumes the authoritative copy resides elsewhere.

Even more is excluded depending on how the backup is created. Per Apple, an **unencrypted** local backup (Finder or iTunes) will not include, among other things:

• Content from the App Store or media stores (apps, music, movies, books), or PDFs downloaded directly into Apple Books

- Content synced from a computer (e.g., imported MP3s, videos, books, or photos)
- Photos and videos already synced to iCloud Photos (only references are retained)
- Saved passwords and Keychain items
- Health, Fitness, and Activity data
- Apple Pay cards, transaction history, and settings
- Face ID and Touch ID configuration data
- Wi-Fi passwords and VPN credentials
- Secure Enclave–protected data of any kind

Why Not an Encrypted Backup?

An encrypted local backup preserves more categories of data than an unencrypted one—Apple will not store items such as passwords, browsing history, Health data, or Wi-Fi settings in an unencrypted backup. However, encrypted backups introduce their own risks. Many tools cannot process encrypted backups reliably, even when credentials are available, and no tool can process them at all without credentials.

Accordingly, collecting an unencrypted local backup is a compromise balancing completeness against usability without dependency on user-controlled passwords. To mitigate the security tradeoff, the backup set can be compressed and encrypted immediately using credentials established for the legal hold or matter—credentials controlled by counsel, not by individual custodians. This shifts encryption from a device-specific constraint to a project-level safeguard.

Why Compress the Backup Data?

One reason we compress the data to a Zip file is to make it easier to copy to new media. Smaller data volumes move faster. However, depending upon the composition of the data backed up, the compressed Zip file may be much smaller or hardly changed. If much of the data on an iPhone consists of, *e.g.,* JPEG photos already in a compressed format, there's little 'space' to squeeze out by further compression.

So why bother compressing the backup files?

Two reasons. First, placing the preserved data in a Zip file guards against overwriting the data by a subsequent backup of the device. Second, depending upon the Zip tool employed to compress the file, the Zip process affords a means to securely encrypt the data without having to install an encryption tool.



Exercise 17: Preserving a Mobile Device

GOALS: The goals of this exercise are for the student to:

1. Assess the cost and burden of preserving a mobile device like a smartphone or pad;
2. Gain experience useful to help guide litigants in meeting preservation duties; and
3. Assess issues of data integrity and defensibility when custodians manage preservation.

OUTLINE: Students will preserve the contents of their personal phones, applying the methodology and software suited to their device using the following Exemplar Phone Backup Instructions.

Assignment: Complete all the tasks outlined in the **Exemplar Phone Backup Instruction for Custodian-Directed Backup** that follow for either your iPhone or Android phone. Submit the data described in the last step of the Backup Instructions via Canvas. You are NOT expected to submit any of your personal data. Submit *only* the **name, date/time and size of the zip file you create**.

NOTE: If you don't have enough space on your computer to hold the Phone image, look at the article, "Redirecting iPhone Backup Files to External Media" that follows the exemplar backup instructions or, for Android phones, change the location where CoolMuster lands the backup data to an external storage device or site.

Exemplar iPhone Backup Instruction for Custodian-Directed Backup (Updated for 2026)

[NOTE: This draft directive is offered to assist counsel in formulating language suited to the needs of the case and controlling law. It is not a form to be deployed without counsel. This example omits optional steps to encrypt the data set and transfer same to a distal repository for preservation, as such steps are frequently unnecessary to meet preservation duties.]

Dear [Custodian]:

You recently acknowledged your obligation to preserve information relevant to a dispute between our company and _____. Please see the _____ hold notice for further details.

Within *48 hours of your receipt of this notice, y*ou must preserve the contents of your company-issued iPhone. If you cannot comply, please advise me at once by e-mail or phone. *Time is of the essence.*

You must make an unencrypted backup using iTunes and compress the backup folder per the instructions below. *Do not assume that you have been automatically making an unencrypted backup or preserving what's required using iCloud. You must carefully follow the procedures set out below.*

What you will need:

- **Your company-issued iPhone**
- **A *data-capable* USB-C or Lightning charge/sync cable (charge-only cables will not work)**
- **Your company-issued desktop or laptop computer**

The computer must have **available free storage space on its startup drive equal to at least twice the storage capacity of the iPhone.**
– To find iPhone capacity: *Settings > General > About > Capacity*
– To find available disk space:
– **Windows:** File Explorer > This PC
– **Mac:** Apple menu > About This Mac > Storage

Time Required

Plan for **1–3 hours**, most of which will be unattended "machine" time. Larger phones or phones with substantial photo/video content may take longer than earlier devices. Do not interrupt the process once it begins.

Follow These Steps

If You Are Using a Mac (macOS Catalina or Later)

1. Connect your iPhone to the Mac using a data-capable cable.
2. Open **Finder**.
3. Select your iPhone under **Locations** in the Finder sidebar.
4. If prompted, unlock your iPhone and select **Trust This Computer**.
5. In the General tab:
    - Select **“Back up all of the data on your iPhone to this Mac.”**
    - **Ensure “Encrypt local backup” is NOT checked.**
6. Click **Back Up Now**.
7. Do **not** disconnect the phone or actively use it until the backup completes.

Finder will display the **date and time of the most recent backup** once the process finishes.

If You Are Using Windows

1. Open **iTunes** and install updates if prompted.
2. Connect your iPhone using a data-capable cable.
3. Unlock your iPhone and select **Trust This Computer** if prompted.
4. Select the iPhone icon, then click **Summary**.
5. Under Backups:
    - Select **This Computer**
    - **Uncheck “Encrypt iPhone Backup”**
6. Click **Back Up Now** and wait for completion.

After the Backup Completes

Once the backup finishes successfully, you may disconnect your phone.

Locate the backup folder as follows:

Windows:
Using File Explorer, navigate to:
C:\Users\[your account name]\AppData\Roaming\Apple Computer\MobileSync\Backup\

Mac:
Using Finder, select **Go > Go to Folder** and enter:
~/Library/Application Support/MobileSync/Backup/

In both Windows and Mac environments, the Backup folder will contain one or more subfolders with long alphanumeric names (for example: 12da34bf5678900386c48267658d340eb34007f8). If more than one folder is present, identify the folder with the **most recent modification date and time**, corresponding to the backup you just completed.

Compress the Backup Folder

Windows: Right-click the identified folder and select **Send to > Compressed (zipped) folder**
Mac: Right-click the identified folder and select **Compress**

Do **not** shut down, restart, or disconnect your computer while compression is underway. The process may take up to an hour or more depending on data volume.

Confirm and Report

Once compression completes, confirm that a file with the same name as the backup folder now exists with a **.zip** extension.

By reply email, report:
• The name of the compressed file
• The date and time created
• The file size

Do **not** open, rename, move, or delete this file. It must be preserved without alteration until further notice.

Your supervisor is copied to ensure you are afforded the time, oversight, and support necessary to comply promptly. Thank you for your cooperation. Please contact me at _____ if you have any questions.

iPhone Backup – Frequently Asked Questions (FAQ)

1. I don't see iTunes on my Mac. Is that a problem?
No. On newer Macs, iTunes has been replaced by **Finder**. Finder performs the same backup function. Follow the Mac instructions in the directive exactly.

2. I connected my phone, but it doesn't appear on my computer.
First, unlock your iPhone and confirm that you selected **"Trust This Computer."**
Second, confirm you are using a **data-capable cable**. Some cables charge only and will not support backups.

3. The backup is taking much longer than expected. Did something go wrong?
Not necessarily. Newer phones store large volumes of photos, videos, and app data. Backups can take **several hours**. As long as the computer remains powered on and the cable connected, allow the process to finish.

4. Can I use my phone during the backup?
Please avoid doing so. Active use—especially messaging, calls, or taking photos—can delay the backup or cause it to fail. Leave the phone connected and idle until completion.

5. I can't turn off "Encrypt local backup." What should I do?
Stop and contact us immediately. Do **not** proceed with an encrypted backup unless instructed. We will help you resolve this or provide an alternative method.

6. I forgot my iPhone passcode or Apple ID password.
Contact us immediately. Do not attempt repeated guesses or reset the device. Failed attempts can lock the phone or trigger data loss.

7. I don't have enough free space on my computer.
Do not delete files or move data without approval. Contact us so we can arrange an alternative computer or preservation method.

8. My phone says it is backing up to iCloud. Is that enough?
No. iCloud backups are **not a substitute** for the required local backup and should not be relied upon for preservation.

9. I see multiple backup folders. Which one do I compress?
Select the folder with the **most recent date and time**, corresponding to the backup you just completed.

10. Should I open the zip file to make sure it worked?
No. Do **not** open, rename, move, or modify the file. Simply report the file name, date/time, and size as instructed.

11. Will anyone read my personal messages or photos?
The backup is preserved solely to meet legal obligations. Access is restricted, and review—if any—will be governed by applicable legal protections and scope limitations.

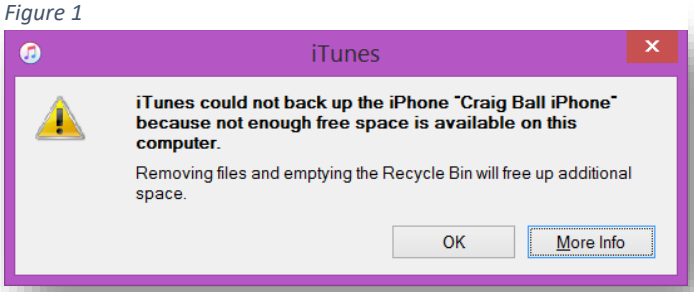
12. What if something goes wrong or I'm unsure?
Stop and contact us immediately. Do **not** attempt fixes on your own.

Redirecting iPhone Backup Files to External Media

Q. What if I don't have enough space on my Windows C: drive to hold the backup?

A. Smart phones have evolved to capture a *lot* of data. Ten years ago, you couldn't store more than 8GB of data on an iPhone. Today, they store up to 1TB, 125 times as much . So, an iTunes backup may fail to complete because not enough free space is available on the computer performing the backup. You may be able to resolve this by, *e.g.,* emptying the Recycle Bin; but, if you simply can't garner enough space on the boot drive where Apple stores the backup by default, you may need to "trick" your Windows machine into storing the backup on a sufficiently-sized alternate or external storage medium.

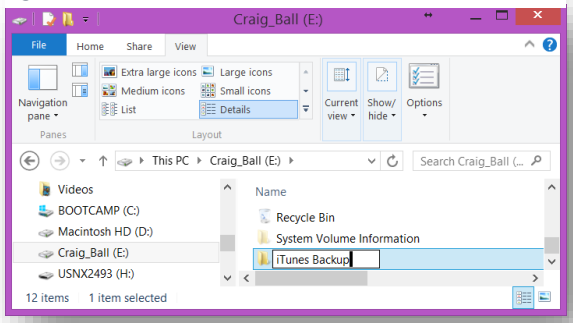
Figure 1



How to Redirect an iTunes Backup Location in Windows

Step 1. Create a new folder on a disk with sufficient free space—at least twice the capacity of the iPhone. An external SSD or hard drive is preferred over a USB flash drive, as backups involve sustained write activity. In Figure 4, the new backup location is on the **E:** drive and named iTunes_Backup. You may name the folder as you wish.

Figure 2



Step 2. Rename the current iTunes backup folder

Using Windows File Explorer, navigate to your current iTunes "Backup" folder. By default, it's:
C:\Users*your account name*\AppData\Roaming\Apple Computer\MobileSync
where "*your account name*" is the name of your Window's User ID on the machine.

Right click on the "Backup" folder and rename it. I called mine "Old_Backup;" but here again, call it whatever you like.

3. Redirect the Old Backup Folder Address to the New One

Here, it gets a tad tricky because you must use a Windows Command line interface. Make it easier on yourself by writing down the full paths to the old and new backup folders. *You must get both right for the redirection to work.*

The old one *should* be:
C:\Users*your account name*\AppData\Roaming\Apple Computer\MobileSync\Backup

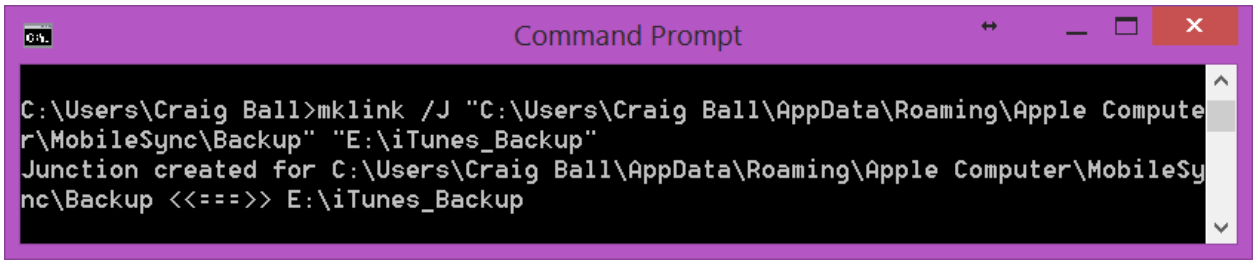
The new path is on whatever storage medium you chose, using whatever path and folder name you gave it in step 1, above (mine was "**E:\iTunes_Backup**").

Open a command prompt window by pressing the Windows key on your keyboard, then typing **CMD**, then right-click **Command Prompt** and select **Run as administrator** *(On Windows 11, you may instead see "Open in Terminal"; ensure Command Prompt is selected.)*

At the command line, carefully type the following command:
mklink /J "*path to old backup location*" "*path to new backup location*"
where you substitute the old and new paths you've written down. *Be sure to enclose each path in quotation marks, as shown.*

For example:
mklink /J "C:\Users\Craig\AppData\Roaming\Apple Computer\MobileSync\Backup" "E:\iTunes_Backup"

On my machine, the command and response looked like Figure 5:
Figure 3

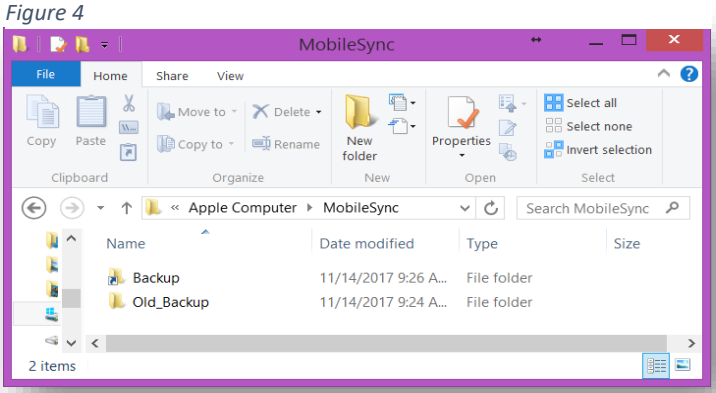


The "junction created" refers to a Windows **Directory Junction,** that will serve to redirect any actions that would have been performed on the old backup folder to be redirected to the new one.

Note: The **mklink /J** command creates a link to the new folder from the old one. It's like creating a shortcut of D:\Backup from the original MobileSync\Backup folder. You can test the effect by double-clicking on the Backup folder in MobileSync. It will take you to the new Backup folder.

Now, if you look in your MobileSync folder:
(C:\Users*your account name*\AppData\Roaming\Apple Computer\MobileSync)
you will see a folder shortcut named "Backup" alongside your renamed former backup folder as mine appears in Figure 8.

Figure 4



4. Move your Old Backups

If desired, you can move your old iTunes backup files from your old renamed Backup folder to your new backup folder and delete them from the old location.

5. Run your iTunes Backup

Ensure the external storage device remains connected. Open iTunes and run the backup as usual. If the redirection is functioning properly, the backup will be created in the new location.

Preserve the Contents of your Android Phone Using CoolMuster Android Assistant

CoolMuster is a commercial tool enabling the backup and collection of content (media, contacts, SMS, call logs, apps, etc.) from phones and other devices using the Android operating system.

If AT&T is your cell carrier, you should temporarily disable In Advanced Messaging before backup in order to collect all your messages. See instructions here.

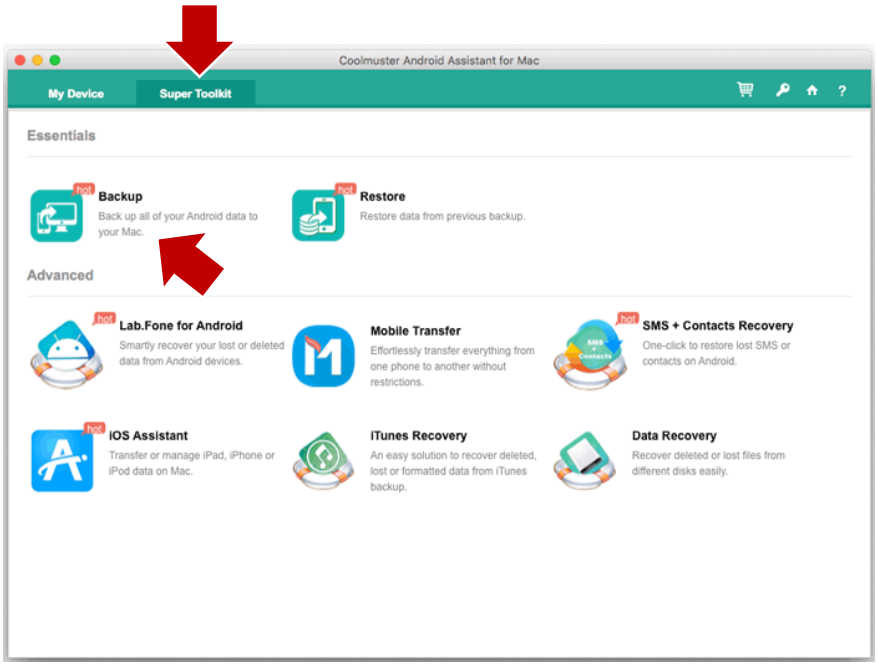
Before getting started: E-mail me at craig@ball.net to obtain the license needed to register the software. You DO NOT have to pay to use the software. I will purchase and supply a license for you.

Step 1: Download and install the CoolMuster Android Assistant software suited to your computer (PC or Mac) from the appropriate link below:
For Windows: https://www.coolmuster.com/downloads/cool-android-assistant.exe
For Mac: https://www.coolmuster.com/downloads/cool-android-assistant-for-mac.dmg

For Windows installation assistance, click here
For Mac help, click here.

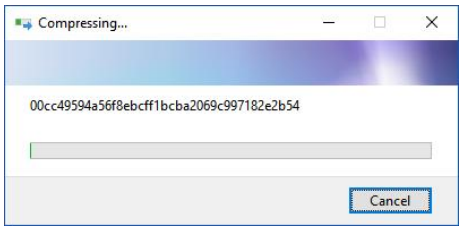
Step 2: Start Android Assistant and Connect to your Phone via USB cable. If asked to activate USB debugging mode on your Android device, follow the onscreen instructions to do so. If your phone appears in the main interface screen of the program, great. If not, go here to troubleshoot your connection.

Step 3: Go to "Super Toolkit" section where you can see the option to "Backup" or "Restore" (right):



Step 4: Click the "Backup" button on Super Toolkit, and select all contents to backup EXCEPT Music. Click "Browse" to choose an output location on your system to save your backup, then click "Back Up" to begin the backup process.

Make note of the target folder you specified.



Step 5: Navigate to the target folder holding your backup data and create a compressed file holding the target folder. of the backup files. In Windows, right click on the subfolder just identified and select "***Send to>Compressed (zipped) folder***." A progress panel like the one at right should appear. On a Mac, right click on the subfolder and select "Compress." Do not turn off your computer or reboot. Allow the compression process to complete. It could take less than an hour to finish depending upon the type and volume of data backed up.

Step 6: Once compression has completed, Windows users should again navigate to the backup folder to confirm the presence of a file with the same name as the subfolder you identified but with the file extension .zip. **Record the name, date/time and size of the zip file.** *[If you cannot see file extensions on your Windows machine, open "My Computer," click "Tools" and click "Folder Options" or click "View" and then "Options" depending on your version of Windows. In the Folder Options window, click the "View" tab. Uncheck the box that says, "Hide file extensions for known file types." This should make file extensions visible.]*

By reply e-mail, send me the **name, date/time and size of the zip file you just created**. ***DO NOT send me the file! It holds your private data***. *Do not delete or open the compressed file. It should be preserved without alteration until the end of the semester please.*

Note: These tools generally work without a hitch; but sometimes they push back. ***I'm here to help you.*** Don't suffer in frustration for hours. Get in touch with me and let's get to the bottom of the problem. Glitches are endemic to e-discovery and computer forensics; so, working through technical issues goes with the territory. E-mail me at craig@ball.net or text me at 713-320-6066. You can voice call me anytime between 8:30am and midnight.

Obtaining and Preserving Social Media Content as Evidence

Social Media Content (SMC) remains a rich source of evidence. Photos, posts, and messages can illuminate claims of disability and damages, establish intent or motive, and support challenges to parental fitness—to say nothing of the steady stream of criminals who post selfies at crime scenes or with stolen goods, drugs, or weapons. SMC may reveal mental instability, propensity to violence, hate speech, racial animus, or misogyny. It is also increasingly a medium for business communications and, for many users, the primary channel for cross-border messaging. In short, social media and messaging have become heirs-apparent to email in their importance to e-discovery. Competence demands **swift identification and timely preservation** of SMC.

Collecting the Client's SMC

Collecting SMC is typically a two-step process: **requesting the data** and, once prepared by the platform, **downloading the data**. Depending on the service, minutes, hours, or even days may elapse between request and availability. Because having a client perform the download weakens chain-of-custody documentation, it is preferable to have the client forward download links to counsel or a designated service provider—or, where feasible, to conduct the collection directly under client authorization.

Obtain the client's written consent to collect and preserve account data. In some cases, credentials or supervised access may be required, but multi-factor authentication (MFA) and security alerts are now the norm and should not be disabled unless absolutely necessary and carefully documented. Download data promptly once available; many platforms make archives accessible only for a short period.

Collection Steps for Seven Social Media Platforms

(Interfaces change frequently; verify paths before execution.)

Facebook / Instagram (Meta):

Log in and navigate to **Settings & Privacy > Accounts Center > Your Information and Permissions > Download Your Information**. Select the relevant account(s), date ranges, and data categories (e.g., posts, messages, photos, comments). Meta will notify the account holder when the archive is ready for download.

X (formerly Twitter):

Log in and go to **Settings and Privacy > Your Account > Download an Archive of Your Data**. Re-authentication is required. X will notify the account holder when the archive is available. Archive frequency and content vary by account type.

Google (Gmail, YouTube, etc.):
Log in at https://accounts.google.com and select **Data & Privacy > Download Your Data (Google Takeout)**. Select only the data expressly authorized for collection. Google will notify the account holder when the archive is ready.

LinkedIn:
Log in and select **Me > Settings & Privacy > Data Privacy > Get a Copy of Your Data**. If unsure, select the larger archive. LinkedIn will notify the account holder when the data is ready.

Snapchat:
Log in at https://accounts.snapchat.com, select **My Data**, and submit a request. Data is delivered as a compressed archive.

Tumblr:
Log in and navigate to **Account Settings > Privacy > Request Privacy Data**. Data is delivered in compressed JSON format.

Review and Authentication
SMC exports are often voluminous and encoded in formats unfamiliar to lawyers, such as JSON. While such files can be inspected with basic tools, meaningful review, searching, and tagging are best accomplished using established e-discovery platforms (e.g., Relativity, Everlaw, DISCO), which can preserve metadata and context while supporting defensible workflows.

Although not strictly required, calculating a hash value for preserved SMC is prudent to demonstrate data integrity. See, *e.g.*, FRE 902(13) and (14). Hash values should be calculated upon receipt and recorded alongside collection logs. Be mindful that re-zipping or re-packaging data can change hash values and should be avoided.

Caveat
There are no “guest passes” to social media accounts. When you log in as the account holder, you act in that user’s stead. Maintain careful records of access, dates, times, and actions taken. Never seek or accept access to an opponent’s social media account using their credentials; doing so invites allegations of spoliation, alteration, or improper access. Transparency and restraint are essential.

Search Science and The Streetlight Effect in e-Discovery

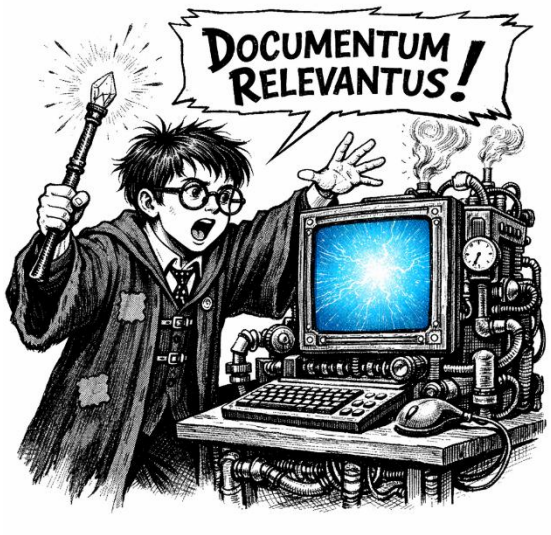


In the wee hours, a beat cop sees a drunken lawyer crawling around under a streetlight, searching for something. The cop asks, "What's this now?" The lawyer looks up and says, "I've lost my keys." They both search for a while, until the cop asks, "Are you sure you lost them here?"
"No," the tipsy lawyer explains, "I lost them in the park—but the light's better here."

I shared that dad joke in court, trying to explain why opposing counsel's insistence that we blindly supply keywords to be run against the email archive of a Fortune 50 insurance company wasn't a reasonable or cost-effective approach to e-discovery. The **Streetlight Effect**, described by David H. Freedman in his 2010 book *Wrong*, is a species of observational bias: people look for answers where it's easiest, not where they're most likely to be found. It neatly describes how lawyers approach electronic discovery. We look for responsive ESI only where and how it's easiest, with little consideration of whether our methods are calculated to find it.

Easy is wonderful when it works. But looking where it's easy when failure is assured is something no sober-minded counsel should accept—and no sensible judge should indulge.

The Myth of the Enterprise Search



Counsel inside and outside corporations—and lawyers on both sides of the docket—continue to believe that organizations can simply "run keywords" across their data. Email systems, archives, local drives, network shares, portable devices, removable media, collaboration platforms, and databases are imagined to yield up responsive ESI if only the right incantation were uttered. ***Documentum Relevantus.***

That fantasy persists despite decades of contrary experience. Data repositories may share networks, but they rarely share search engines, indexing rules, or

query syntax. Some support Boolean logic; others don't. Some allow proximity searching; others don't. **Stemming**, **lemmatization**, fielded searches, wildcard behavior, and treatment of special characters vary widely. Searching databases may require specialized query languages, elevated privileges, or schema knowledge. Different systems extract text differently—or not at all. The same query that finds a document in one repository may miss it entirely in another.

Yet nowhere is the Streetlight Effect more pernicious than in the reliance on keyword searches against email archives and other indexed collections of ESI.

A Familiar Corporate Fable

That Fortune 50 company—I'll call it *All Farm Indemnity*—captured *over a billion* email and attachments through message journaling. Journaling copies every message traversing in and out of the system into an archive where messages are indexed for search. Keyword searches interrogate the index, *not* the underlying messages or attachments. If it's not an entry in the index, it might as well not exist.

All Farm gets sued every day. When a request for production arrives, they run keyword searches against their mail archive using a tool we'll call *Truthiness*. Hundreds of large organizations use Truthiness—or something just like it—and confidently assume their systems will find all documents containing the chosen keywords.

They're wrong. Or worse, they're studiously incurious.

If requesting parties don't force companies like All Farm to confront the limits of their search infrastructure, incomplete productions will remain the norm. To prompt the necessary epiphany, consider serving discovery *about* discovery.

A Useful Interrogatory

Interrogatory: For each electronic system, archive, index, or search technology that will be used to respond to discovery, please state:

1. The rules employed by the system to tokenize data for indexing and search;
2. The stop words or noise words applied when data was indexed;
3. The number and nature of documents or communications in the system that are not text-searchable, and why; and
4. Any limitation in the system, index, or query syntax that tends to limit or impair the effectiveness of keyword, Boolean, proximity, semantic, or similarity-based search in identifying responsive materials.

The responding party's lawyer will blow a gasket, but courts will permit this kind of "discovery about discovery" when a party can show that search adequacy is genuinely in dispute. Each inquiry targets a real and recurring failure point.

Tokenization Rules

Machines don't read documents; they consult indices. Characters are converted into searchable "tokens" according to rules that vary by system. Many indices ignore short words or acronyms. None index single letters or isolated numerals. Tokenization rules govern punctuation ("wind-driven"), case sensitivity, diacritics ("Rene" vs. "René"), hyphenation, and numeric strings ("Clause 4.3"). Users routinely assume these searches will work. Often, they don't work—unless the system has been configured with those use cases in mind.

Stop Words

Many systems exclude common words—*stop words*—from indices entirely. Searches containing them silently fail because the words never existed in the index. These omissions aren't always trivial. In the Truthiness index, "all" and "farm" were stop words. A search for "All Farm" would miss documents containing the company's *own name.* Other common stop words include *part*, *side*, *down*, *necessary*, *general*, *state*, and *years*. Users are rarely warned that searches containing stop words are doomed, so they mistakenly conclude no responsive documents exist when they simply weren't indexed.

Non-Searchable Content

A substantial portion of ESI remains invisible to text search. Scans, faxes, TIFF images, and many PDFs are not searchable absent Optical Character Recognition (OCR) and indexing of the recognized text. Password-protected, encrypted, or compressed files may be excluded from indexing altogether. Chat data, emojis, reactions, edits, and deletions may be indexed selectively—or not at all. If adjusters at All Farm attach scanned bids and receipts sent via email or snail mail, keyword searches will miss them unless OCR has been applied competently and consistently.

Other Limitations (Including "AI")

Today's search landscape includes TAR, predictive codingvand large language model-assisted retrieval. These tools are powerful—but they are not magical. They are trained, tuned, and constrained by the same underlying data quality issues: what was indexed, how it was normalized, and what was excluded. If a system's limitations are known, they should be disclosed. If they're unknown, they should be tested. Hope is not a strategy.

The Streetlight Effect, Now with LLMs

Large language models promise to free lawyers from the tyranny of keywords. They don't require Boolean logic. They don't care about word order. They can "understand" concepts, summarize themes, cluster conversations, and answer questions in plain English. Compared to traditional search, LLM-assisted retrieval feels like stepping out of the streetlight and into daylight.
Or so the sales pitch goes.

In truth, LLMs do not abolish the Streetlight Effect; they risk **automating it**. An LLM can only reason over the data it is allowed to see, and it only "understands" what has been extracted, indexed, normalized, and presented to it. If responsive ESI was never ingested, never indexed, never OCR'd, or never surfaced because of system limitations, an LLM will miss it. Fluency is not fidelity.

LLMs do not search original documents. They operate on representations—text extractions, embeddings, summaries, or truncated context windows—each shaped by design choices and constraints that are often invisible to lawyers. If the underlying corpus is incomplete, skewed, or biased by convenience, the LLM's output will be equally so, only more persuasive. The danger is not that LLMs get answers wrong; it's that they make wrong answers sound so right.

Nor do LLMs solve the problems of tokenization, stop words, or non-searchable content. An email attachment that was never OCR'd remains invisible. A scanned exhibit remains silent. Emojis, reactions, edits, deletions, and metadata stripped in preprocessing remain lost. LLMs reason exquisitely about what they are given—but they cannot conjure what was never captured.

There is also a subtler streetlight at work: **prompt bias**. Asking an LLM a question is itself an act of framing. The model will search for meaning consistent with the prompt supplied, not for meaning the proponent failed to imagine. In that respect, LLMs resemble the drunken lawyer under the lamp: they search diligently where instructed, not where the truth may lie.

None of this is to disparage LLM-assisted search. Properly deployed, LLMs can mitigate much of the bluntness of keywords, surface conceptual relationships, and accelerate understanding of large corpora. They can help lawyers ask better questions and recognize patterns that might otherwise go unnoticed. But they do not excuse ignorance of the underlying systems, nor do they relieve counsel of the duty to understand what data was searched, how it was processed, and what was excluded.

The lesson of the Streetlight Effect endures. Whether the tool is a keyword index, a TAR engine, or a large language model, **search is only as good as the science beneath it and the judgment guiding its use**. When lawyers mistake elegance for completeness, or convenience for competence, the light grows brighter—and the keys remain in the dark.



Exercise 18: Honing Your Search Skills

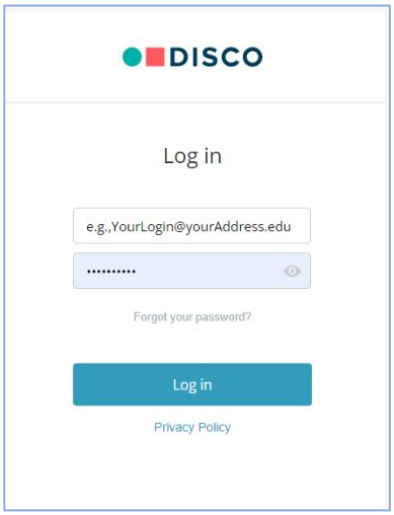
GOALS: The goals of this exercise are for the student to:

1. Understand some of the limitations of lexical- and indexed search, seeing why lexical search works well in some settings but very poorly in others, particularly e-discovery; and

2. Learn to test and refine keyword and Boolean lexical searches to improve precisicontract or payment on and recall

These exercises employ the John Podesta E-Mail Collection published by WikiLeaks and freely downloadable as a large compressed file from https://file.wikileaks.org/file/podesta-emails/podesta-emails.mbox-2016-11-06.gz. **NOTE: You DO NOT need to download this data! It's loaded for you already.**

We use the Podesta e-mail collection because there are few publicly available, contemporary corpora of messages and attachments in their native (or near-native) forms. The Podesta E-Mail Collection is large without being unwieldy (about 5GB of uncompressed data comprising over 50,000 messages and thousands of attachments). Better still, it concerns issues, events and personalities about which many Americans have some familiarity.[99]



Students will access the Podesta E-Mail Collection using the DISCO online review to conduct and refine searches. The processed and indexed collection is reached by navigating to

https://login.csdisco.com/Account/Login

and entering the Username and Password you set up for access to DISCO.

In the Matter named **UT Spring 2026**, click the button labeled "Ediscovery" and select the database called "**Podesta Email**."[100]

Stick to Keyword Search: These exercises assume that all students possess a working knowledge of Boolean search syntax and a basic familiarity with the complement of metadata attendant to e-mail messages and common productivity files like Microsoft Word,

[99] By contrast, the massive Enron e-mail collection widely employed in e-discovery is decades old, concerns persons and events unfamiliar to most students and has been filtered, culled and sterilized so as to be almost unrecognizable as a useful native corpus.

[100] Please do not use the matter called UT Processing Exercise 2025. That's for a different exercise.

PowerPoint, Excel and Adobe PDF. These exercises are geared to understanding routine pitfalls of indexed search and improving your skill in framing, refining and testing lexical search. The capabilities and limitations of the tools used here are common to most e-discovery search and review tools. In these exercises, we seek to better utilize *basic* keyword search tools because despite its drawbacks, ***keyword search remains the most common approach to e-discovery***. Accordingly*,* ***no advanced search features*** are to be employed in working through these exercises.

In the 25+ years I've studied lexical search of ESI, I've learned that:

1. Lexical search is a crude tool that misses much more than it finds and leads to review of a huge volume of non-relevant information. That said, *even crude tools work wonders in the hands of skilled craftspeople who chip away with care to produce masterpieces.* The efficacy of lexical search increases markedly in the hands of adept practitioners who meticulously research, test and refine their search strategies.

2. Lawyers embrace lexical search despite knowing almost nothing about the limits and capabilities of search tools and without sufficient knowledge of the datasets and indices under scrutiny. Grossly overestimating their ability to compose effective search queries, lawyers blithely proffer untested keywords and Boolean constructs. Per Judge John Facciola a generation ago, *lawyers think they're experts in search "because they once used Google to find a Chinese restaurant in San Francisco that served dim sum and was open on Sundays."*

3. *Without exception*, every lexical search is informed and improved by the iterative testing of queries against a substantial dataset, *even if that dataset is not the data under scrutiny*. Iterative testing is *invaluable* when queries are run against representative samples of the target data. Every. Single. Time.

4. Hit counts alone are a poor measure of whether a lexical search is "good" or "bad." A "good" query may simply be generating an outsize hit count when run against the wrong dataset in the wrong way (*e.g.*, searching for a person's name in their own email). Lawyers are too quick to exclude queries with high perceived hit counts before digging into the causes of poor precision.

5. A query's success depends on how the dataset has been processed and indexed prior to search, challenging the assumption that search mechanisms just 'work,' as if by magic.

6. Lexical search is a sloppy proxy for language; and language is replete with subtlety, ambiguity, polysemy and error, all serving to frustrate lexical search. Effective lexical search adapts to accommodate subtlety, ambiguity, polysemy and error by, *inter alia*, incorporating synonyms, jargon and industry-specific language, common misspellings and alternate spellings (*e.g.*, British vs. American spellings) and homophones, acronyms and initializations.

7. Lexical search's utility lies equally in filtering out irrelevant data as it does in uncovering relevant information; so, it demands meticulous effort to mitigate the risk of overlooking pertinent documents.

Search Syntax: A competent search requires that the syntax be suited to the tool. That may seem obvious, but the brainstorming, exchange and negotiation of search queries and syntactic variations across search tools leads to improperly structured queries. It's important to understand that the syntax of search varies across tools. For example, when proximity searching, Relativity and dtSearch use the common "w/n" to denote a search for two terms or phrases within the number "n" words of one another. OpenText Insight uses the syntax "NEAR/n" for the same purpose and DISCO uses, simply, "/n" when searching within n words in any order and "+n" when searching within n words where the first term must precede the second. DISCO's search syntax manual is available here: https://support.csdisco.com/hc/en-us/articles/213049583-DISCO-search-syntax-manual and a list of basic operators and their usage follows below.

Processing Data for Search: As prior exercises addressed, electronically stored information is stored using various schemes comprising multiple encoded "layers" that must be properly decoded to yield the intelligible information sought in discovery. The Microsoft Word document attached to an e-mail message is encoded in Extensible Markup Language (XML) that has been further encoded by a Zip compression algorithm to comprise what we see as the Word DOCX format. As an e-mail attachment, the DOCX file is encoded as base-64 within the transmitting message. The message itself will be encoded within the mail application that houses it (*e.g.,* PST, EDB, MBOX, NSF, etc.). It may be further encoded depending upon whether it is collected from backup media, a live server or a forensic image. Thus, ESI is like a set of Russian matryoshka nesting dolls in terms of its encoding within encoding within encoding.

"Processing" in e-discovery describes the operations performed on ESI to extract its information and metadata and render the extracted data amenable to culling and search. Effective processing must be **recursive,** thoroughly cycling through all the levels of encoding and applying the correct decoding methods to harvest all desired content.

Processing also entails, *inter alia,* flagging files that cannot be fully accessed or understood ("**exceptions**"), cataloguing the processed items, **de-NISTing** (excluding operating system and application files lacking evidentiary value), digitally "fingerprinting" files ("**hashing**") and suppressing duplicates ("**de-duping**"). Files (like scanned paper documents and photos) that depict but don't store text may be subjected to **optical character recognition** to enable electronic search.

Processing culminates in the creation of an index (called a "**concordance**") of extracted text and metadata which can be searched to find matching text and culled by parameters like date range, file type, author, custodian and the like.

The salient point is that ***when you search for information in an e-discovery tool, you are <u>not</u> searching the source data; you search a collection of information that has been <u>extracted</u> from the source data and indexed.*** The accuracy and completeness of culling and search in e-discovery is only as good as the accuracy and completeness of the index and the capabilities of the search tools and their operator—you.

When you search for information in an e-discovery tool, you are <u>not</u> searching the source data; you search a collection of information that has been <u>extracted</u> from the source data and indexed.

The Pros and Cons of Indexed Search: Again, search in electronic discovery doesn't entail examination of the files comprising the evidence; all searches are directed against the index of extracted text and metadata. The index can be no more complete than its source data, and by design or error, it is frequently less complete in ways perilous to the unwary.

The advantage of indexed search is <u>speed</u>. It's much faster to query a database of extracted text and metadata than to repeatedly burrow down into and across the source data.

DISCO Search basics

DISCO ignores most punctuation and non-alphanumeric characters when searching. Periods, colons, semicolons, and apostrophes within a word are not removed. As an example, periods in a name or email address are indexed and searchable. Word operators (*and*, *or*, and *not*) can be searched if placed in quotation marks, *"contract and payment"*.

Operators	Description	Example
&, and	Includes results with both terms	**contract & payment** **contract and payment** — Documents that contain both *contract* and *payment*
[space] or	Includes results with either term or field	**contract payment** **contract or payment** — Documents that contain *contract* or *payment*
%, not	Excludes term or field from results	**contract % payment**

Operators	Description	Example
		contract not payment — Documents that contain *contract* but do not contain *payment*
+family	Includes family members as search hits for the entire query or any portion of a query; learn more here	**to("Jane.Smith@email.com")+family & "risk factors"** — Returns emails, and family members of those emails, sent to Jane.Smith@email.com that include the phrase "risk factors" **tag(Responsive)+family % tag(Attorney-Client)+family** — Returns documents tagged Responsive and documents whose family members are tagged Responsive, but removes documents tagged Attorney-Client and family members of documents tagged Attorney-Client.
" "	Exact phrase intended	**"contract payment"** — Documents that contain the exact phrase *contract payment* **"Matt.Motley@email.com"** — Documents that contain the email address *Matt.Motley@email.com* Be sure to include quotes when searching for an email address.

Operators	Description	Example
!	Truncation search or root expander; can be used at the beginning of a term, end of a term, or both. Also can be used to return all documents.	**contract!** — Documents that contain any term starting with *contract,* for example, *contract, contracts, contractual, contracting,* or *contracted* **!contract** — Documents that contain any term ending with *contract,* for example, *subcontract* ***filename(!contract!)*** — Documents with the term *contract* anywhere in the file name **!**— All documents **! % type(email)**— All documents except those of type email
*****	Wildcard search for single character	**contract*** — Searches for words that have one, but only one, character after *contract,* such as *contracts*, but not *contracted, contractual*, or *contract*
/n	Proximity search, searching within *n* words, in any order	**contract /10 payment** — Documents that contain *contract* within ten words of *payment*, where the order is irrelevant
+n	Proximity search, searching within *n* words in prescribed order	**contract +10 payment** — Documents that contain *contract* within ten words of *payment*, where *contract* must precede *payment*
~	Fuzzy or approximate word search. Fuzzy searching allows for the addition, deletion, or	**guaranty~** — Searches for *guaranty, guarantee, garanty,* and similar (mis)spellings

Operators	Description	Example
	substitution of up to two letters in a word.	
.	Period in a name or email address	**Matt.Motley** — Documents that contain *Matt.Motley* (with the period), but not *MattMotley* (without the period)
( )	Grouping syntax	**(failure & consideration) & (contract agreement)** — Documents must contain both *failure* and *consideration*, and must also contain either *contract* or *agreement*
sample(n, search)	Returns *n* documents randomly selected from results of search. If *n* is less than 1, this number is treated as a percentage.	**sample(.5, contract);** **sample(700, contract)** — Returns 50% and 700 of the search results for *contract* at random, respectively
field(terms)	Field searching (see below for standard DISCO fields)	**custodian(Holcombe)** — Documents with *Holcombe* in the *custodian* field

Order of operations

DISCO performs a search using the following order of operations:
Term modifiers: *!, *, ~*
Exact phrases: *“ ”*
Groupings: *( )*
Proximity: */n, +n*
Family subsearch: *+family*
&, *and*, *%*, *not*
[space], *or*

Essential Tips for Effective Lexical Search in Civil Discovery
From "Lessons from Lousy Lexical Search," Ball in Your Court, February 26, 2024

Pre-Search Preparation:

1. **Understand the Dataset**
    - Identify data sources and types, then tailor the search to the data.
    - Assess the volume and organization of the dataset. Can a search of fielded data facilitate improved precision?
    - Review any pre-processing steps applied, like normalization of case and diacriticals or use of stop words in creating the searchable indices.
2. **Know Your Search Tools**
    - Familiarize yourself with the tool's syntax and keyword search capabilities.
    - Understand the tool's limitations, especially with non-textual data and large documents.
3. **Consult with Subject Matter Experts (SMEs)**
    - Engage SMEs for insights on relevant terminology and concepts.
    - Use SME knowledge to refine keyword selection and search strategies.

Search Term Selection and Refinement:

4. **Develop Comprehensive Keyword Lists**
    - Include synonyms, acronyms, initializations, variants, and industry-specific jargon.
    - Consider linguistic and regional variations.
    - Account for misspellings, alternate spellings and common transposition errors.
5. **Utilize Boolean Logic and Advanced Operators**
    - Apply Boolean operators and proximity searches effectively.
    - Experiment with wildcards and stemming for broader term inclusion.
6. **Iteratively Test and Refine Search Queries**
    - Conduct sample searches to evaluate and refine search terms.
    - Adjust queries based on testing outcomes and new information.

Execution and Review:

7. **Provide for Consistent Implementation Across Parties and Service Providers**
    - Use agreed-upon terms where possible. The most defensible search terms and methods are those the parties choose collaboratively.
    - Ensure consistency in search term application across the datasets, over time and among multiple parties.

8. **Sample and Manually Review Results**
    - Randomly sample search results to assess precision and recall.
    - Adjust search terms and strategies based on manual review findings.
9. **Negotiate Search Terms with Opposing Counsel**
    - Engage in discussions to agree on search terms and methodologies.
    - Document agreements to preempt disputes over discovery completeness.
    - Make abundantly clear whether a non-privileged document hit by a query must be produced or whether (as most producing parties assume) the items hit may nevertheless be withheld after a review for responsiveness.

Post-Search Analysis:

10. **Validate and Document the Search Process**
    - Maintain comprehensive documentation of search terms, queries, exception items and decisions. Never employ a set of queries to exclude items from discovery without the ability to document the queries and process employed.
    - Ensure the search methodology is defensible and compliant with legal standards.
11. **Adapt and Evolve Search Strategies**
    - Remain flexible to adapt strategies as case evidence and requirements evolve.
    - Leverage lessons from current searches to refine future discovery efforts.
12. **Ensure Ethical and Legal Compliance**
    - Adhere to privacy, privilege, and ethical standards throughout the discovery process.
    - Review and apply discovery protocols and court orders accurately.

Exercise 18: Scoping the Collection Under Scrutiny

It's said, "What you don't know won't hurt you," and "Ignorance is bliss." Certainly, those platitudes and attitudes afflict e-discovery, as legions of lawyers remain blissfully ignorant of what's absent or unsearchable in the collections under scrutiny. Notwithstanding, it is the duty of competent e-discovery counsel (or counsel working with competent support) to identify custodians holding responsive data, select sources and lay out the proper parameters used to identify, preserve, collect, cull, process and search electronically stored information in a reasonably diligent and defensible way.

The most ingenious searches won't find what isn't there; so, the threshold component of a competent search strategy is looking at the right information (custodians, sources and files) and, just as crucially, ensuring that its content is amenable to search. Too, understanding the composition of the collection permits searches to be limited to data fields and file types most likely to yield responsive information without excessive recall of non-responsive material.

What's in the Collection? The Podesta email collection consists of the contents of an archive of John Podesta's purloined and published email. But considering all the attachments and encoding and such, what are you really looking at? Establishing reliable metrics, *e.g.*, file counts, processing exceptions and file types, is essential for keeping an e-discovery effort from spinning out of control. ***NOTE: there are 17 questions to this exercise (Questions 18.1 through 18.17 with subparts). Be sure to answer all subparts. An optional answer sheet may be found at the end of this exercise for your convenience.***

Question 18.1: Determine the item count reported in DISCO for each of the following file types in the Podesta email collection:

a. All Items in Collection: _____

b. Top Level (Parent) E-Mail Messages only: _____

c. Adobe Acrobat PDF files: _____

d. MS Office Files (Word, PowerPoint and Excel): _____

Resolving Exceptions: It's common for data to fail to process correctly when ingested in an e-discovery processing tool. Sometimes files are corrupted, encrypted or encoded in ways the tool can't resolve. As well, the indexing tool may be unable to extract searchable text from the file. The latter is common with scanned documents saved in TIFF or other image formats. Sometimes PDFs are created without searchable text. To resolve these exceptional items, they must first be identified, resolved (*e.g.,* decrypted or subjected to optical character recognition) and added back

to the collection. Whether to do so (and the cost of same) often depends on the volume of exception items to be resolved.

Question 18.2: Determine the item count reported in DISCO for:

a. Processing exceptions (partial failures): _____

b. Unsupported File failures _____

c. Password protected files: _____

d. Exception Items that were Attachments _____

Hint: DISCO lists exception items in its Ingest Reports. Look at the Menu.

De-NISTing a Collection:

De-NISTing is a technique used in e-discovery and computer forensics to reduce the number of files requiring review by excluding standard components of the computer's operating system and off-the-shelf software applications like Word, Excel and other parts of Microsoft Office. Everyone has this digital detritus on their systems, things like screen saver images, document templates, clip art, system sound files and so forth. It's the stuff that comes straight off the installation disks, and it's just noise in a discovery review.

It's called "de-NISTing" because those noise files are identified by matching their hash values (*i.e.,* digital fingerprints) to a huge list of software hash values maintained and published by the National Software Reference Library, a branch of the National Institute for Standards and Technology (NIST). The NIST list is free to download, and pretty much everyone who processes data for e-discovery and computer forensic examination uses it or a customized exclusion list including NIST hash values.

But the NIST list isn't magical, and it's useful to grasp its limitations. The NIST list is of limited utility in reducing the volume of irrelevant documents obtained by a targeted collection (versus a wholesale collection like that obtained when preserving ESI as drive images). A sensible targeted collection won't grab the sorts of system files that the NIST list excels at excluding. NISTing is still somewhat useful to cull a targeted collection, but don't expect many files will be excluded.

QUESTION 18.3: What is the version number and date (month and year) of the current NIST NSRL Reference Data Set (RDS hashes) available for download? _____

Exercise 18: Understanding the Constraints on Search:

Adjusting the Indexed Alphabet: A law firm client asked me to search a large collection using a Boolean query including the term "20%." The query was:

"20%" AND ("payment" OR "amount" OR "check" OR "pay")

That request is problematic in several respects, but the parties had agreed to the query after testy talks and the judge had signed off on the search protocol. My client didn't want to upset the judge, so asked that I find a way to make the search work.

First, I had to address three questions:

1. Did the search tool index numbers?

2. Did the search tool *index* the percent sign, treat it as a space (*i.e.*, a word break) or ignore it?

3. Is the percent sign a Special Character or Operator in the tool?

A problem with such an insidious search is that it tends to return a lot of truly relevant items along with a ton of junk.

Running the query against the Podesta e-mail using dtSearch hit on **13,430 files.**

Running the query against the Podesta e-mail in Nuix hit on **5,478 items**

QUESTION 18.4: How many items are returned when you run the query against the Podesta e-mails in DISCO? ***"20%" AND (payment OR amount OR check OR pay)***

How Many? _____

QUESTION 18.5: Apart from any sense that the hit count is excessive, how do you ascertain that the results don't meet expectations in terms of returning items that reference "20%" (meaning twenty percent)?

_____

_____

QUESTION 18.6: How many items are returned searching for

"20" AND ("payment" OR "amount" OR "check" OR "pay")

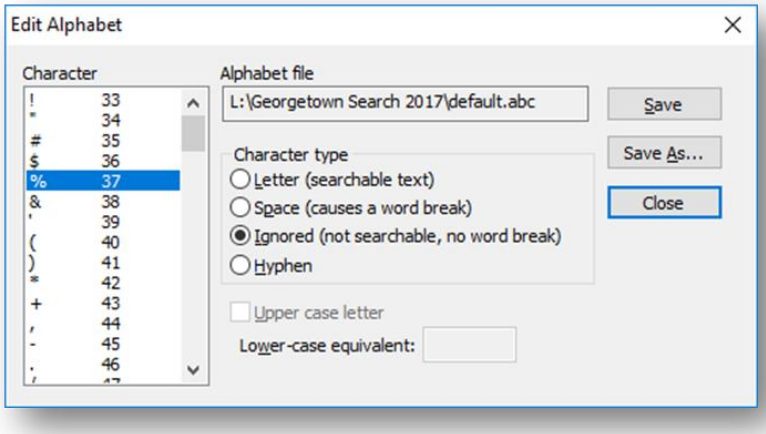
How Many? _____

QUESTION 18.7: Did DISCO index the percent sign, treat it as whitespace (*i.e.,* a word break) or ignore it altogether? What is the percentage sign reserved to do in DISCO?

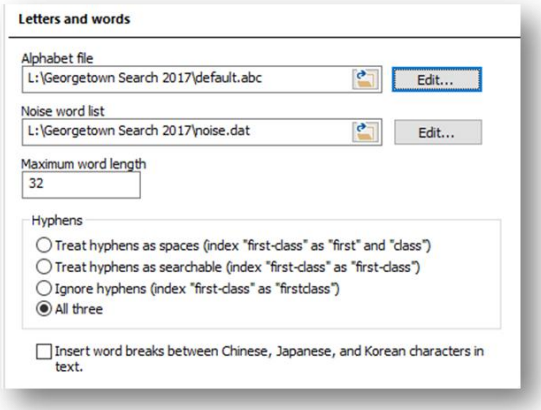
_____

_____

Do you see the risk attendant to agreeing to a query without understanding the syntax and limitation of the search tool? In solving the problem for my client, the solution lay in changing alphabet parameters for the search tool, Nuix, and re-indexing the data. In dtSearch or Relativity, that change would be made by editing the alphabet file (default.abc) that determines whether the file parser treats a character as searchable text, as a word break or hyphen or ignores the character altogether. By default, dtSearch ignores the percent sign and does not treat it as a word break (figure right).[101]



Text search tools tend to treat punctuation as spaces. Some tools allow you to index characters in multiple ways (*e.g.* as a searchable character, as a space and as something to be ignored). This ensures that all instances are found at the cost of over- inclusive results. Note how dtSearch allows hyphens to be indexed in three distinct ways in its "Letters and words" preferences menu (figure left).



The treatment afforded punctuation, particularly hyphens and periods in initializations, can prompt surprising outcomes. Depending upon the tokenization rules of the indexing tool—rules which determine whether a group of characters will be indexed as a searchable term—the hyphen in the term "e-discovery" may be treated as a space and the now-detached "e" treated as too brief to index. For efficiency, search tools

[101] The picture is further complicated because dtSearch and Relativity employ the percent sign to denote a variable character in fuzzy search. The bottom line is that you simply cannot use the percent sign as an effective search term using these tools because, as a fuzzy operator, % can't be deployed to pull up *only* the percent sign.

commonly do not index single letters as searchable words.

QUESTION 18.8: Run searches for "e-discovery" (in quotes), "e discovery" (in quotes) e-discovery (no quotes) and e OR discovery. Record the hit counts for each below:

"e-discovery" _____

"e discovery" _____

e-discovery _____

e OR discovery _____

Do you see what's happening in the treatment of the hyphen as an indexed character?

Question 18.9: Draft a query to identify documents in the Podesta email collection that contain variations on the phrase, 'third-party candidates.' Be thorough without being over-inclusive.

_____

_____

Dealing with Diacritics:

Consider the following English sentence:

Zoë Budinger (née Baird) enjoyed a soupçon of pâté and rosé before she rose to strike the pate of the piñata.

Six words feature seven diacritical marks.[102] Two words (***pâté and rosé)*** have markedly different meanings with and without diacritics. E-discovery tools can be configured to distinguish spellings using diacritics or to treat letters with diacritics in the same way as their unaccented counterparts.

QUESTION 18.10: How many documents in the Podesta e-mail collection reference the person known as *Zoë Baird?*

_____

QUESTION 18.11: Draft a query to identify documents in the Podesta e-mail collection that reference the person known as *Zoë Baird*. **Consider diacritics, variations and aliases.**

_____

QUESTION 18.12: How many documents did your query recall? _____

[102] In order, the umlaut or diaeresis, acute accent, cedilla, circumflex, two more acute accents and a tilde.

Controlling for Case: Typically, we want search tools to equate upper and lowercase letters in queries to secure maximum recall; yet sometimes we need to find BAT, not bat, and Ball, not ball. Search tools may allow you to limit the scope of search by case, but so doing requires that processing be configured to differentiate indexed items by case when the index is created or that the source data be reprocessed to update the index. Even then, *you must test the system to ascertain whether the hoped-for case differentiation functions as you desire*.

QUESTION 18.13: You've been asked to search the Podesta email collection for references to former Attorney General Eric Holder. Draft a query to identify responsive documents while excluding those containing the term "holder" when NOT used as a surname.

Noise Words and Stop Words: Search tools seek ways to increase the efficiency and reduce the size of the index. One shortcut is the exclusion of so-called "noise words," also called "stop words" from the index. By default, dtSearch excludes the 86 words in the following table from its indices.

dtSearch Noise Word List

a	**at**	**even**	**i**	**not**	**see**	**there**	**was**	**who**
about	**be**	**for**	**if**	**now**	**she**	**these**	**way**	**will**
after	**been**	**from**	**in**	**of**	**some**	**they**	**we**	**with**
all	**but**	**get**	**into**	**on**	**take**	**this**	**well**	**would**
also	**by**	**got**	**it**	**only**	**than**	**those**	**were**	**you**
an	**can**	**he**	**its**	**or**	**that**	**to**	**what**	**your**
and	**come**	**her**	**just**	**other**	**the**	**too**	**when**	
any	**could**	**him**	**like**	**our**	**their**	**under**	**where**	
are	**did**	**his**	**me**	**out**	**them**	**up**	**which**	
as	**do**	**how**	**my**	**over**	**then**	**very**	**while**	

This is a modest list. Some tools exclude a much larger list of stop words. Relativity excludes 112 stop words by default and won't index single letters or numbers.

Remember: If a noise word isn't indexed, it cannot be searched by querying the index. The word is simply not there to be found. Noise word lists should always be checked to determine if any proposed queries incorporate a noise word. Stop words may seem unlikely to mess you up; but, where there's a will, there's a way. Wait, can't search that! "Where," "will," "there," "a" and "way" are all stop words!

QUESTION 18.14: Search for the assertion "They are there!" with and without quotes and with and without the exclamation point. What hit counts do you get for each?

"They are there!" _____ They are there! _____

"They are there" _____ They are there _____

QUESTION 18.15: Can you find any documents in the Podesta e-mail collection containing Hamlet's famous question, "to be or not to be?" On what date? _____

If you'd used DISCO before April 29, 2019, you'd have gotten very different results because, DISCO was unable to search for the following stop words in any matter created before that date:

a an and are as at be by for if in is it

of on or that the their then there these they to was with

During the first battle of the Texas Revolution against Mexico, the Texians[103] fashioned a defiant flag bearing the phrase "come and take it" along with a black star and an image of the cannon the Mexican forces had been ordered to capture. The Texians prevailed. Now, imagine trying to find documents about the Battle of Gonzales flag using search tools like dtSearch or Relativity that treat ALL the words "Come and Take It" as noise words? How would you make the words searchable?



To modify the list of words defined as noise words in dtSearch, you would click Options>Preferences> Letters and Words, then click the Edit button next to the name of the noise word list you want to revise. The data must then be re-indexed. Don't do it; just know HOW to do it.

Takeaway: So far, we've concerned ourselves with the integrity and completeness of the index, the upshot being that how you process data into an index determines what you can get out through search. The algorithms that parse and tokenize data reflect compromises balancing effectiveness and efficiency. What you see is not governed by what you've got so much as by what you choose to index. So, we must be prepared to ask and respond to questions about the parameters of processing when it matters. Questions like:

- **What stop words have been excluded from the index?**
- **Can we constrain searches by upper and lowercase?**
- **Can numbers and single letters be searched?**
- **Are there characters that cannot be searched or are treated as spaces or ignored?**
- **How are diacritics resolved?**
- **What processing exceptions were seen, and how were they resolved?**

[103] Fun fact: Those who fought for Texas independence are called Texians, not Texans.

Arrogant and ignorant opponents will deride these questions as distractions. Perhaps they imagine they are searching through the evidence instead of a shadow of same. Perhaps they don't grasp how minor changes in processing parameters have a major impact on whether what is sought will be found or not. Either way, opponents should know the answers, and litigants have a right to know.

Exercise 18: Setting the Parameters for Search:

"*Parameters* for search?!?! Doesn't it just search *everything*?"

No, and often you don't want to search everything.

We use keyword search because we expect that the queries will lead us to relevant content. Reviewers want keyword hits to be highlighted and make a beeline to the places in the body of a document where keywords occur to assess responsiveness or privileged character.

So, what does it mean when a document is hit but you can't find the keywords? It *could* mean that the search tool has been configured to return all items in a *family* of documents (*i.e.,* a message and all attachments) or it *could* signify that the search tool found the hit beyond the body of the document, in a filename, -path or -property. Anytime a document is returned by keyword search and the keywords aren't visible, you should understand *why* it was returned.

By default, DISCO treats any word you put into the search bar as a keyword search term. Search terms can be of words or phrases (when enclosed by quotation marks) located in the body of a document, or within a specific field of information a document possesses, such as its metadata, or applied tags. It is important to note that search queries that do not specify a field in the search syntax automatically search the following:

- Document text
- Document notes
- Custodians
- Subject line
- Author information
- User defined fields

It's useful to be able to limit and target the scope of a search to reduce noise hits. For example, non-responsive documents may reside in a folder or file path titled with a search term, resulting in all contents of all subfolders being returned because the search term shows up in each document's properties. Alternatively, you may wish to search *only* file names or file properties.

In addition to limiting the scope of search, you may want to limit your searches to only particular fields of data for items in the collection. The most frequent application of fielded search is culling by date range, but e-discovery tools typically offer a broad range of discretely searchable fields.

QUESTION 18.16: In searching for resumés in the Podesta email, you crafted a query that included the term "bio" and found 586 hits for "bio." You've been asked to determine the number of items where the word bio appears in the document text of an email message but NOT in the subject line of the message. In responding, supply the query you used AND the number of messages identified meeting the criteria. *Hint: try using the DISCO Search Builder to construct searches within fields, exclude items and limit searches to file types. To find the Search Builder, click within the search box and select Search Builder.*

_____

_____

Culling Before Search

Some years ago while visiting Australia, skulking around the mummies in a small-but-fine museum on the University of Sydney campus, I learned that mystery writer Agatha Christie was married to archaeologist Max Mallowan and that she'd assisted him in Syrian digs. Dame Agatha even used her cold cream and knitting needles to clean rare ivory artifacts. The experience found its way into her work. An exhibit of Christie-cleaned carvings included a quote from the author's fictional detective, Hercule Poirot, in Death on the Nile (1937):

> Once I went professionally to an archaeological expedition–and I learnt something there. In the course of an excavation, when something comes up out of the ground, everything is cleared away very carefully all around it. You take away the loose earth, and you scrape here and there with a knife until finally your object is there, all alone, ready to be drawn and photographed with no extraneous matter confusing it. That is what I have been seeking to do–clear away the extraneous matter so that we can see the truth–the naked shining truth.

This naturally got me thinking about the way we approach search in electronic discovery. Most lawyers use keywords to find responsive documents despite their propensity to sweep in too much. We get lots of the documents we seek with keywords; unfortunately, the results come caked with the loose earth of documents that are "hit" but have no connection to the case. Testing confirms this occurs with a ratio of about 20% responsive matter to 80% extraneous. That's a lot of loose earth!

If most of the material culled by keyword search is extraneous matter, any technique that pulls away chaff (*e.g.,* non-responsive sources, custodians, file types, data ranges, etc.) without losing wheat translates to significant savings of time and money while improving quality.

Seems obvious, right? Why search data in ways that can't possibly yield responsive hits? But lawyers do it all the time by failing to cull non-responsive content before running queries and by failing to tailor searches to the data under scrutiny. The most frequent and outrageous example is searching for a custodian's name or e-mail address within the custodian's own data. You've either got to cull the custodian's data when running the search or exclude the search when combing through the data. Running all search terms against all data won't fly.

Testing and Refining Queries:

Test, Test, TEST!

The most important step you can take to assess keywords is to test search terms against representative data from the universe of machines and data under scrutiny. No matter how well you think you know the data or have refined your searches, testing will open your eyes to the unforeseen and save time and money.

Keyword search is the art of finding documents containing words and phrases that signal relevance (usually) followed by page-by-page (linear) review of those documents. It's often called the "gold standard" of electronic discovery.

That's ironic, because extracting and refining gold relies less on finding precious aurum than it does on dispersing all that isn't golden. Prospectors use water and chemicals to flush away all but the gold left behind. So, a true "gold standard" for keyword search must incorporate both precise inclusion (smart queries) and defensible exclusion (smart culling).

To illustrate, in one e-discovery dispute over search, the plaintiff submitted keywords to be run against the defendant's e-mail archive for a three-month interval. Unfortunately, the archive held all e-mail for all custodians, and the defendant adamantly refused to segregate by key custodian or deduplicate before running searches. The interval was narrow, but the collection was vast and redundant.

The defendant tested the agreed-upon keywords but shared only aggregate hit rates for each. Thinking the numbers too high, but unwilling to look at the hits in context, the defendant rejected the search terms. The plaintiff agreed the hit counts were daunting but asked to see examples of hits on irrelevant documents before furnishing exclusionary (AND NOT) modifications to flush away more of what wasn't golden.

The defendant refused, insisting it wasn't necessary to see the noise hits in context to generate more precise queries. The parties were at an impasse, with one side grousing "too many hits" and demanding different search terms and the other side uncertain how to exclude irrelevant documents without knowing what caused the noisy results.

A lawyer who dismisses a search because it yields “too many hits” is as astute as the Emperor Joseph II dismissing Mozart’s *Il Seraglio* as an opera with “too many notes.” Mozart replied, “There are just as many notes as there should be." Indeed, if data is properly processed to be susceptible to text search and the search tool performs appropriately, a keyword search generates just as many hits as there should be. Of course, few lawyers craft queries with the precision Mozart brought to music; so, when the terms used seem well chosen for relevance, it’s crucial to scrutinize the results to learn what tailings are cropping up with the gilt-edged, relevant documents.

Keyword search is just a crude screen: “Show me items that contain these words, and don’t show me items that contain those.” High hit counts don’t always signal a bad screen. If search terms merely divide the collection into one pile holding relevant documents and one without, you’re closer to striking gold. Then, you look at what you can reliably exclude with the next screen and the next, drawing ever closer to that elusive quarry, *documentum relevantus*.

But you must see hits in context to refine queries by exclusion. That seems so manifestly obvious; it’s astounding how often it’s not done.

When lawyers delegate keyword search, they often get back only aggregate hit counts and mistakenly conclude that’s enough information to judge searches noisy or not. If, instead, counsel got their hands dirty with the data—as by personally exploring representative samples using desktop or hosted tools—the parties could work quickly, effectively and cooperatively to zero in on relevant material. Good queries are best refined by knowledgeable people testing them against pertinent, small collections. Lousy outcomes spring from lawyers thinking up magic words and running them against everything.

The nature and sample size of representative data will vary with each case. The goal in selection isn’t to reflect the average employee’s collection but to fairly mirror the collections of employees likely to hold responsive evidence. Don’t select a custodian in marketing if the key players are in engineering.

Often, the optimum custodial choices will be obvious, especially when their roles made them a nexus for relevant communications. Custodians prone to retention of ESI are better candidates than those priding themselves on empty inboxes. The goal is to flush out problems *before* deploying searches across broader collections, so opting for uncomplicated samples lessens the value.

It’s amazing how many false hits turn up in application help files and system logs; so early on, I like to test for noisy keywords by running searches against data having nothing whatsoever to do with the case or the parties (*e.g.,* the contents of a new computer). Being able to show many hits in wholly irrelevant collections is compelling justification for limiting or eliminating unsuitable keywords.

Similarly, you might wish to test search terms against data samples collected from employees or business units having nothing to do with the subject events to determine whether search terms are too generic.

Finally, test against known responsive items, especially when seeking to identify privileged material. A competent search must pick up the material you *already know* to be responsive or privileged.

Incorporate Misspellings, Variants and Synonyms

Did you know Google got its name because its founders couldn't spell googol? Whether due to typos, transposition, IM-speak, misuse of homophones or ignorance, electronically stored information fairly crawls with misspellings that complicate keyword search. Merely searching for "management" will miss "managment" and "mangement."

To address this, you must either include common variants and errors in your list of keywords or employ a search tool that supports fuzzy searching. The former tends to be more efficient because fuzzy searching (also called *approximate string matching*) mechanically varies letters, often producing an unacceptably elevated level of false hits.

How do you convert keywords to their most common misspellings and variants? A linguist could help, or you can turn to the web. The optimum approach is examining an alphabetized list of all words in the search tool's index. Many tools offer such a list; unfortunately, DISCO does not, so we must explore alternatives.

You could begin by running alternate spellings through the search tool to identify alternate spellings in the index, or you might try a site like https://www.rankwatch.com/free-tools/typo-generator that generates misspelled variants of keywords you supply, or consult Wikipedia's list of common misspellings (Wikipedia shortcut: **WP:LCM**).

To identify synonyms, pretend you are playing the board game Taboo. Searches for "car" or "automobile" will miss documents about someone's "wheels" or "ride." Consult a thesaurus for likely alternatives for critical keywords, but don't go hog wild with Dr. Roget's list. Question key players about internal use of alternate terms, abbreviations or slang

QUESTION 18.17: You've served a Request for Production upon John Podesta as an opposing party to civil litigation seeking, "Any and all documents touching or concerning coordinated attacks upon U.S. government personnel in Libya on or about September 11, 2012."

Counsel for Mr. Podesta proposes the following query be used to identify potentially responsive material from the Podesta email collection:

Bengahzi AND (attack OR security) AND (embassy)

What query or queries do you counter-propose, if any? How many responsive items did you find in the Podesta using your query (NOT by manually reviewing documents)?

Start with the Request for Production

It's against the backdrop of the Request for Production (RFP) that your production efforts will be judged, so the RFP warrants careful analysis to transform its often expansive and bewildering demands to a coherent search protocol.

The structure and wording of most RFPs are relics from a bygone time when information was stored on paper. You'll first need to hack through the haze, getting beyond the "any and all" and "touching or concerning" legalese. Try to rephrase the request in plain English to get closer to the terms most likely to appear in the ESI. Incorporate terms of art from the RFP to your list of keyword candidates. Have several persons do the same, insuring you include multiple interpretations of the requests and obtain keywords from different points of view.

If a request isn't clear or is hopelessly overbroad, push back promptly. Request a clarification, move for protection or specially except if your Rules permit same. Don't assume you can trot out boilerplate objections and ignore the request. If you can't make sense of it, or implement it in a reasonable way, tell the other side how you'll interpret the demand and approach the search for responsive material. Wherever possible, you want to be able to say, "We told you what we were doing, and you didn't object."

Seek Input from Key Players

Custodians are THE subject matter experts on their own data. Proceeding without their input is foolish. Ask key players, "If you were looking for responsive information, how would you go about searching for it? What terms or names would likely appear in the messages we seek? What kinds of attachments? What distribution lists would have been used? What intervals and events are most significant or triggered discussion?" Invite custodians to show you examples of responsive items, and carefully observe how they go about conducting their search and what they offer. You may see them take steps they neglect to describe or discover a strain of responsive ESI you didn't know existed.

Emerging empirical evidence underscores the value of key player input. Higher precision and recall closely correlate with the amount of time devoted to questioning persons who understand the documents and why they are relevant. The need to do so is obvious, but lawyers routinely dive into search without benefit of the insight of subject matter experts.

Communicate and Collaborate

Engaging in genuine, good faith collaboration is the most crucial step you can take to ensure successful, defensible search. Cooperation with the other side is not a sign of weakness, and courts demand it in e-discovery. Treat cooperation as an opportunity to show competence and readiness, as well as to assess your opponent's mettle. What do you gain from wasting time and money on searches the other side didn't seek and can easily discredit? Won't you benefit from knowing if they have a clear sense of what they seek and how to find it?

Tell the other side the tools and terms you're considering and seek their input. They may balk or throw out hundreds of absurd suggestions, but there's a good chance they'll highlight something you overlooked, and that's one less do over or ground for sanctions. Don't position cooperation as a trap nor blindly commit to run all search terms proposed. "We'll run your terms if you agree to accept our protocol as sufficient" isn't fair and won't foster restraint. Instead, ask for targeted suggestions, and test them on representative data. Then, make expedited production of responsive data from the sample to let everyone see what's working and what's not.

Importantly, frame your approach to accommodate at least two rounds of keyword search and review, affording the other side a reasonable opportunity to review the first production before proposing additional searches. When an opponent knows they'll get a second dip at the well, they don't have to push Draconian demands.

Filter and Deduplicate First

Always filter out irrelevant file types and locations before initiating search. Music and images are unlikely to hold responsive text, yet they'll generate vast numbers of false hits because their content is stored as alphanumeric characters. The same issue arises when search tools fail to decode e-mail attachments before search. Here again, you must know *how* your search tool handles encoded, embedded, multibyte and compressed content.

Filtering irrelevant file types can be accomplished in many ways, including culling by binary signatures, file extensions, paths, dates or sizes and by de-NISTing for known hash values. Again, the National Institute of Standards and Technology maintains a registry of hash values for commercial software and operating system files that can be used to reliably exclude known, benign files from e-discovery collections prior to search. http://www.nsrl.nist.gov.

The exponential growth in the volume of ESI doesn't represent a leap in productivity so much as an explosion in duplication and distribution. Much of the data we encounter are the *same* documents, messages and attachments replicated across multiple backup intervals, devices and custodians. Accordingly, the efficiency of search is greatly aided—and the cost greatly reduced—by *deduplicating* repetitious content *before* indexing data for search or running keywords. Employ a

method of deduplication that tracks the origins of suppressed iterations so that repopulation can be accomplished on a per custodian basis.

Applied sparingly and with care, you may even be able to use keywords to exclude irrelevant ESI. For example, the presence of keywords "Cialis" or "baby shower" in an e-mail may reliably signal the message isn't responsive; but *testing and sampling must be used to validate such exclusionary searches*.

Search Tips

Defensible search strategies are well-documented. Record your efforts in composing, testing and tweaking search terms and the reasons for your choices along the way. Spreadsheets are handy for tracking the evolution of your queries as you add, cut, test and modify them.

When searching for names, it's wise to use the NEAR/2 or W/2 connector between first and last names to account for the use of middle names or initials. When searching e-mail for recipients, it's almost always better to search by e-mail address than by name. In a company with dozens of Bob Browns, each must have a unique e-mail address. Be sure to check whether users employ e-mail aliasing (assigning idiosyncratic "nicknames" to addressees) or distribution lists, as these can thwart search by e-mail address or name.

Keyword Search is Here to Stay

These exercises will help you wring more quality and trim the fat from text retrieval. It will be some time before everyone embraces technology-assisted review, and even those using predictive coding tools use keyword search to compile "seed sets" of relevant documents to train their tools and aid in document review. Despite serious shortcomings, clunky-but-comfy keyword search will be with us for a long time to come.

OPTIONAL ANSWER SHEET for Exercise 18, subparts 1-17 (use if convenient for you)

QUESTION 18.1: Determine the item count reported in DISCO for each of the following file types in the Podesta collection (not on your Evidence Drive, in *the full* Podesta collection):

e. All Items in Collection: _____

f. Top Level (Parent) E-Mail Messages only: _____

g. Adobe Acrobat PDF files: _____

h. MS Office Files (Word, PowerPoint and Excel): _____

QUESTION 18.2: Determine the item count reported in DISCO for:

e. Processing exceptions (Partial failure): _____

f. Unsupported File failures _____

g. Password protected files: _____

h. Exception Items that were Attachments _____

QUESTION 18.3: What is the version number and date (month and year) of the current NIST NSRL Reference Data Set (RDS) available for download? **_____**

QUESTION 18.4: How many items are returned when you run the following query against the Podesta e-mails in DISCO? ***"20%" AND ("payment" OR "amount" OR "check" OR "pay")***

How Many? _____

QUESTION 18.5: Apart from any sense that the hit count is excessive, how do you ascertain that the results don't meet expectations in terms of returning items that reference "20%"?

_____

_____

_____

QUESTION 18.6: How many items are returned searching for

"20" AND ("payment" OR "amount" OR "check" OR "pay")

How Many? _____

QUESTION 18.7:

Did DISCO index the percent sign, treat it as whitespace (*i.e.,* a word break) or ignore it altogether? What is the percentage sign reserved to do in DISCO?

_____

QUESTION 18.8: Run searches for "e-discovery" (in quotes), "e discovery" (in quotes), e-discovery (no quotes) and e OR discovery. Record the hit counts for each below:

"e-discovery" _____ **"e discovery" _____**

e-discovery_____ **e OR discovery _____**

QUESTION 18.9: Draft a query to identify documents in the Podesta email collection that contain variations on the phrase, 'third-party candidates.' Be thorough without being over-inclusive.

QUESTION 18.10: How many documents in the Podesta e-mail collection reference the person known as *Zoë Baird?*

_____

QUESTION 18.11: Draft a query to identify all documents in the Podesta e-mail collection that reference the person known as *Zoë Baird*. Consider diacritics, variations and aliases.

_____

QUESTION 18.12: How many documents did your *Zoë Baird*. query recall? _____

QUESTION 18.13: You've been asked to search the Podesta email collection for references to former Attorney General Eric Holder. Draft a query to identify responsive documents while excluding those containing the term "holder" when NOT used as a surname.

_____

QUESTION 18.14: Search for the assertion "They are there!" with and without quotes and with and without the exclamation point. What hit counts do you get for each?

"They are there!" _____ They are there! _____

"They are there" _____ They are there _____

QUESTION 18.15: Can you find any documents in the Podesta e-mail collection containing Hamlet's famous question, "to be or not to be?" On what date? _____

QUESTION 18.16: In searching for resumés in the Podesta email, you crafted a query that included the term "bio" and found 586 items with hits for "bio." You've been asked to determine the number of items where the word bio appears in the document text of an email message but NOT in the subject line of the message. In responding, supply BOTH the query you used AND the number of messages identified meeting the criteria.

_____

QUESTION 18.17: You've served a Request for Production upon John Podesta as an opposing party to civil litigation seeking, "Any and all documents touching or concerning coordinated attacks upon U.S. government personnel in Libya on or about September 11, 2012."

Counsel for Mr. Podesta proposes the following query be used to identify potentially responsive material from the Podesta email collection: **Bengahzi AND (attack OR security) AND (embassy)**

What query or queries do you counter-propose, if any? How many responsive items did you find in the Podesta collection using your query (NOT by manually reviewing documents)?

_____



Exercise 19A: Negotiating Search Protocols

GOALS: The goals of this exercise are for the student to:

1. Assess proposed search terms for efficacy and efficiency, considering the limitations of indexed search tools. and

2. Learn to test and refine keyword and Boolean lexical searches to improve them.

In a (now resolved) case in Iowa involving allegations of data theft brought by Nutra Blend against a former employee (and alleged data thief) named Andy Noah and his new employer, Consumer's Supply, plaintiff's counsel made the following demand of counsel for Consumers Supply:

In our view, an appropriate search to locate responsive documents would include the following 75 search terms to these fifteen custodians' electronic files for the time period August 1, 2018 to the present. By electronic files, this refers to Outlook and any electronic repository of documents and files, including Word and Excel documents:

Custodians:

Andy Noah	Lawrie Music	Mychael Irish
William Millikin	Ron Rottman	Cecily Johnston
Dave Patee	Logan Ginkens	Katie Schmidt
Dan Patee	Mike Gregoricka,	Mark Regnier
Jerod Johnson	Payson Hedlund	Dan Regnier

Search Terms:

1. Big V	39. Bluebonnet
2. Simmons	40. Farmer's Elevator Pilgrims
3. Keith Smith	41. Bryant Grain Company
4. OK Foods	42. Bryant Grain
5. Cal-Maine	43. Bryant
6. OMP	44. BGC
7. Ozark Mountain Poultry	45. North East Texas Farmers Coop
8. Peco Farms	46. NE Texas Farmers Coop
9. Peco	47. NETFC
10. TFC	48. Arhberg Milling
11. TN Farmers Coop	49. Arhberg
12. TN Farmers	50. LCN
13. Wayne Farms	51. NTB

14. Wayne	52. NB
15. Mid-America Feeds	53. Nutra Blend
16. Mid-America	54. Nutrablend
17. Mid America Pet Food	55. Nutra Blend basket analysis
18. MAPF	56. Basket
19. George's	57. Customer opportunity analysis
20. Kent Nutrition Group	58. RFR
21. Kent Nutrition	59. Andy
22. Kent	60. Noah
23. KNG	61. Big V Items to Target.xlxs
24. Stockman's Mill & Grain	62. Mid-America Feeds - Talala, OK. (Dax) Potential Items.xlsx
25. Stockman's	
26. SMG	63. Keith Smith Targeted Items-Premixes.xlsx
27. Martindale Feeds	64. Simmons Processing.xlsx
28. Martindale	65. Wayne Farms Potential Items.xlsx
29. MF	66. TN Farmers Co-op Potential Items.xlsx
30. Crescent Feeds	67. Peco Farms Potential Items.xlsx
31. Crescent	68. George's Potential Items.xlsx
32. CF	69. OMP (Ozark Mountain Poultry) Potential Items.xlsx
33. So-Mo Ag	70. Cal-Maine Potential Items.xlsx
34. So-Mo	71. OK Foods Potential Items.xlsx
35. Hanor	72. Price Quote (1313448) KENT NUTRITION GROUP, INC
36. HANOR	73. 1313448
37. AC Nutrition	74. Accts to call on with Ron.xlsx
38. AC	75. bravado@ruralinet.net

Assignment (approximately 30-45 minutes): You are the defendant's counsel. Applying what we have learned in class and through our readings on processing and search, critique the discovery demand and queries from the standpoint of their effectiveness using the search platform Relativity (with default settings). For example: Are any queries likely to be incapable of identifying the information sought? Are any likely to be especially noisy or overinclusive? It's understood that you don't know anything more than is set out above, and there is no need for you to seek out further information about the lawsuit, parties or custodians to respond. I do not expect you to critique every query; just *at least six* you identify as problematic, more are welcome.



Exercise 19B: AI Search Terms Analysis

GOALS: The goals of this exercise are for the student to:

1. Apply a Proposed AI Methodology to test and refine a proposed set of search terms; and

2. Tweak and/or critique the methodology to assess and improve its effectiveness.

OUTLINE: Students will use a suitable AI tool of their choice (ChatGPT-4 encouraged but not required) to test and refine a proposed set of search terms.

In Exercise 19A, you were presented with a proposed set of 75 search terms that were sought to be run against ESI in the care, custody and control of fifteen custodians of the defendant. In that exercise, you used your knowledge of search techniques, pitfalls and limitations to critique the proposal. Here, you will use a Generative AI Large Language Model to undertake the same task.

Exercise 19B: Review the Appendix to the chapter entitled, The Leery Lawyer's Guide to AI and LLMs in Trial Practice and apply the methodology it describes to the search effort sought in Exercise 19A:

How to Approach this Exercise

Step 1: Apply the proposed methodology (or such applicable parts of same as you see as appropriate to include) to the proposed queries and custodians from Workbook Exercise 19 and share your prompt and LLM response(s).

Step 2: Compare the AI-assisted outcome to your prior manual assessment submitted for Exercise 19. In a few sentences at most: What did you miss? What did the AI miss?

Step 3: What changes would you make to the prompt or methodology you employed in this exercise to achieve superior results going forward?

Submit your work via Canvas.



Exercise 20: Processing, Culling, Search and Export

GOALS: The goals of this exercise are for the student to:

1. Become acquainted with image mounting, ingestion, processing, culling and search in the context of a commercial e-discovery processing and review tool; and

2. Generate a Bates-labeled production set with accompanying load files.

OUTLINE: Students will mount the Evidence Drive forensic image created in a prior exercise as a read-only virtual drive, then create a Zip container file of the image contents and ingest it into DISCO, a cloud-based e-discovery platform for processing, culling, search, and creation of a Bates-stamped production set with load files. Allow 90-120 minutes to complete this exercise, though most of that time will be spent waiting for the software to finish tasks.

Students will use DISCO, a commercial SaaS hosted ESI review platform. Each student has been supplied with their own user ID and should have set up a personal password for access. *Do not share these.*

Exercise 20: Before Getting Started

Be sure you have the following at hand:

1. The forensic image (.E01 or .dd file(s)) of the Evidence Thumb Drive you created in Imaging Exercise 12;
2. Your DISCO login credentials setup link (emailed to you; check your spam folder) and
a. **WINDOWS USERS ONLY**: a downloaded copy of the AccessData FTK Imager application located in the Files>Software Installer Files>FTK Imager folder of the class Canvas site. Also, you will need a zip archival tool like the free application 7-Zip available for your use (https://www.7-zip.org). You will NOT be able to use Windows' native zipping feature to do this, so take the time to install 7-Zip (or WinZip or another standalone archiver).

Step 1: Mount the Forensic Image as a Virtual Drive on your Computer

In this step you will mount the forensic image you collected so that your machine sees it as a virtual drive, making its contents accessible for compression into a Zip container.

A. Mac Users Do This:

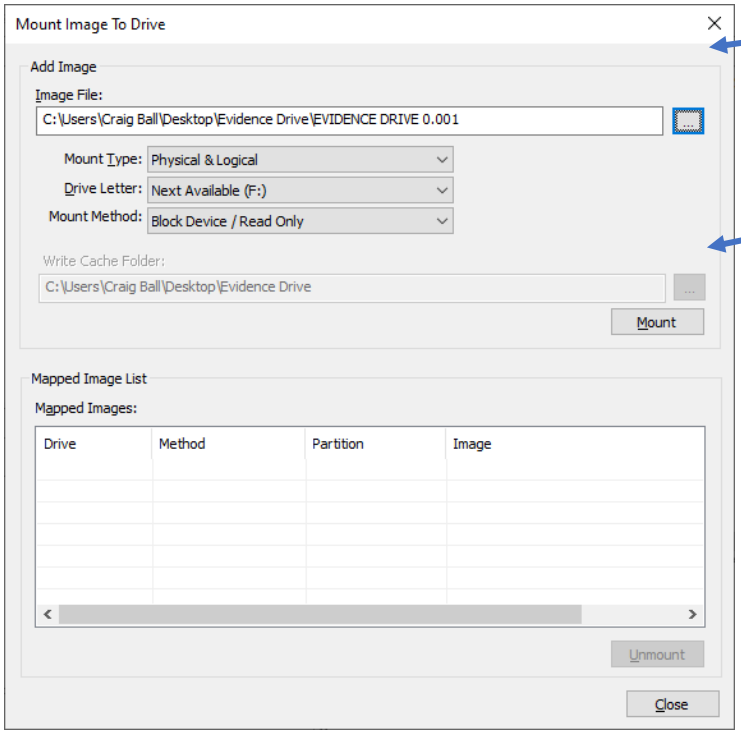
i. Make a working copy of the Evidence Drive image file EVIDENCE.dd and change the extension of your working copy to .dmg. (I typically do this using the File Info screen.) Recall that changing a file's extension does not alter its contents—we are simply "tricking" macOS into opening the image file as if it were one of Apple's own Disk Image files.

ii. Double click on the renamed image file to mount its contents. You should see the evidence file appear on your desktop (and within Finder) as a drive attached to your machine.

B. Windows Users Do This:

i. **Run FTK Imager**: If not still installed, install Exterro® FTK® Imager ver. 8.2.0.33 on your PC and launch the application.

ii. **Mount the Image as a drive:** On the program's menu bar, click File>Image Mounting. When the Mount Image to Drive menu appears, locate the "Image File" field and click the adjacent button with three dots. Navigate to your Evidence Drive image file, select the image file (.E01 extension) then click "Open." Click "Mount" and **note the assigned drive letter**.



Click here to navigate to your Evidence Drive image file.

Click here to Mount the image file as the next available drive letter (note the drive letter assigned on YOUR machine—mine is F:-- yours may be different).

Step 2 (All Platforms): Compress the Contents of the Mounted Image to a Zip Archive

Unfortunately, DISCO cannot ingest RAW (dd), DMG, or E01 files. **DISCO requires that files reside in a Zip file for ingestion.** Using 7-Zip or the standalone compression application of

your choice, add the complete contents of the mounted evidence drive to a Zip archive called "Evidence Drive.zip" and ***store this archive on your Desktop*** (or other sufficient storage). Remember to make a .ZIP, NOT a .7Z archive.
IMPORTANT: You're compressing the contents of the mounted image, NOT the image itself. Expect the compression step to take about ten minutes.

Windows users: Do NOT use the "Send to > Compressed (zipped) folder" context menu option—Windows attempts to create the archive *in the same folder* as the files being compressed, and you cannot write to the mounted read-only virtual drive. Use 7-Zip (https://www.7-zip.org/) and direct the output to your Desktop instead. Even if you could write to the evidence media, you must *never overwrite evidence media!* Doing so destroys potentially recoverable files in unallocated clusters—a major forensic *faux pas*! If using 7-Zip, ensure you create a ZIP archive, not an alternate format.

Mac users: CTRL-click the mounted Evidence Drive folder on your Desktop and select Compress "Evidence Drive" from the menu. This creates "Evidence Drive.zip" on your desktop. If this doesn't work, install a free standalone Zip tool and create the Zip of the mounted drive's contents. Remember, you are zipping the contents of the mounted image, NOT the image itself!

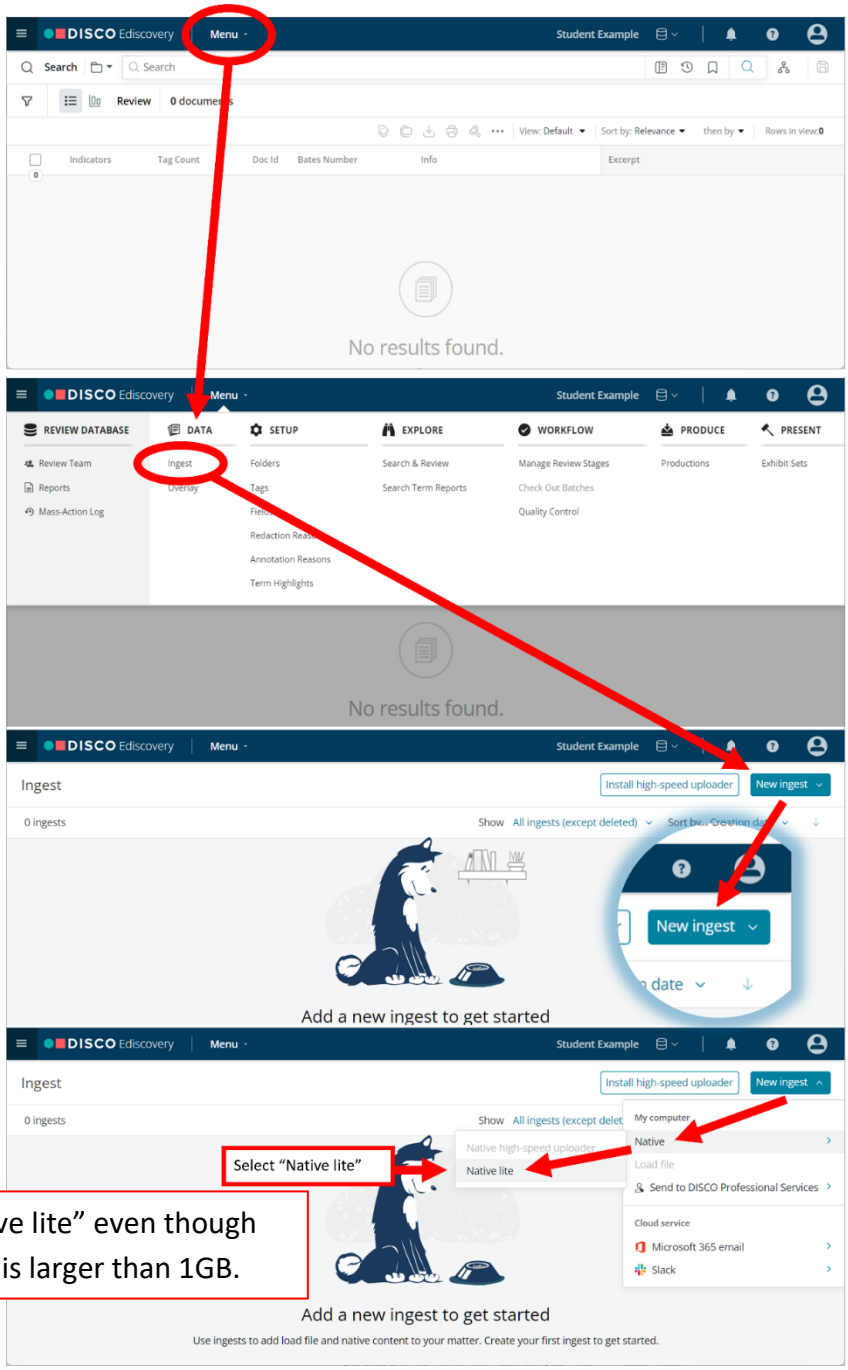
At this stage, you should have a file called "Evidence Drive.zip" on your Desktop.

Step 3: Ingest Evidence

A. Using your browser (ideally, a Chrome browser but try your preferred browser first), go to https://login.csdisco.com and log in using your credentials.

B. Select the Matter named "**UT Processing Exercise 2026**" and the review database in **your name**. You should see the DISCO eDiscovery screen (the screenshots that follow were made from a different database called "Student Example," **but YOU will use your personal account under "UT Processing Exercise 2026**").

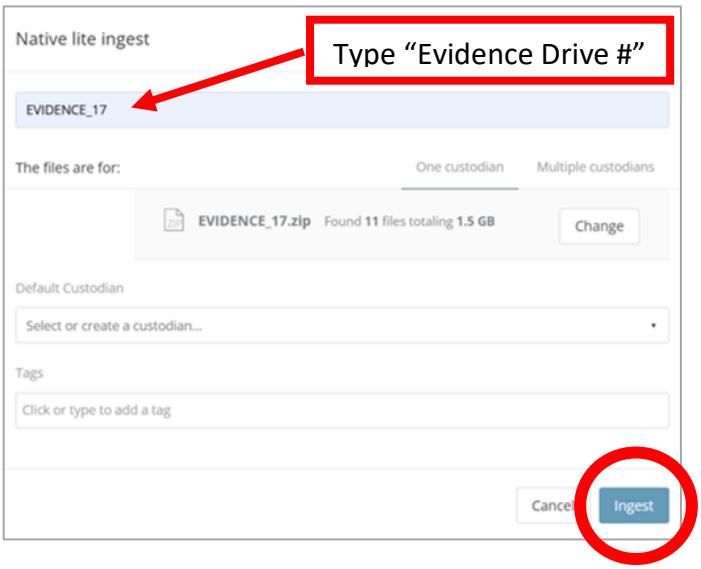
C. Select “**Menu”** (on the blue bar) and locate the “DATA” column in the menu screen. Click “**Ingest**,” then from the Ingest screen, select “**New Ingest,”** then **Native** and **“Native Lite.” (**See illustrated workflow below).



Select “Native lite” even though your Zip file is larger than 1GB.

d. On the next screen (right), select "browse" and navigate to the Zip file you created in Step 2. Select that Zip file for ingestion. *You may ignore the caution against Zip files greater than 1GB.*





DISCO will validate the Zip file and present the screen at left. Your file count may be slightly different.

Enter "Evidence Drive #" in the "Name or Describe this ingest" blank.

Click the "Ingest" button.

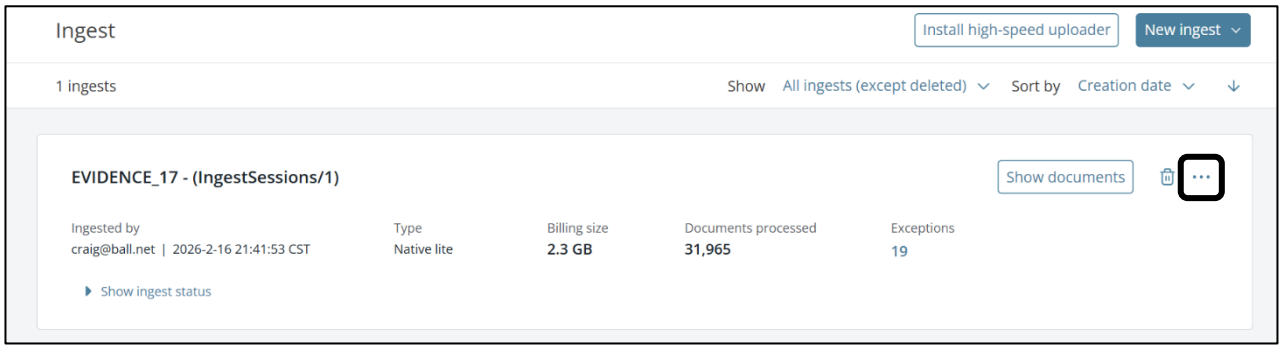
You should see an "Uploading Natives" progress menu (below). Uploading typically takes 10-20 minutes, longer on slower connections.



Step 4: Take a break (~45-75 minutes) while DISCO unpacks and processes the ingested evidence. If you want to review what DISCO is doing while you wait, you can click on "Show Ingest Status" or feel free to re-read the processing chapters of this Workbook 😉 …or not.

Step 5: Download your Ingest and Exception Reports

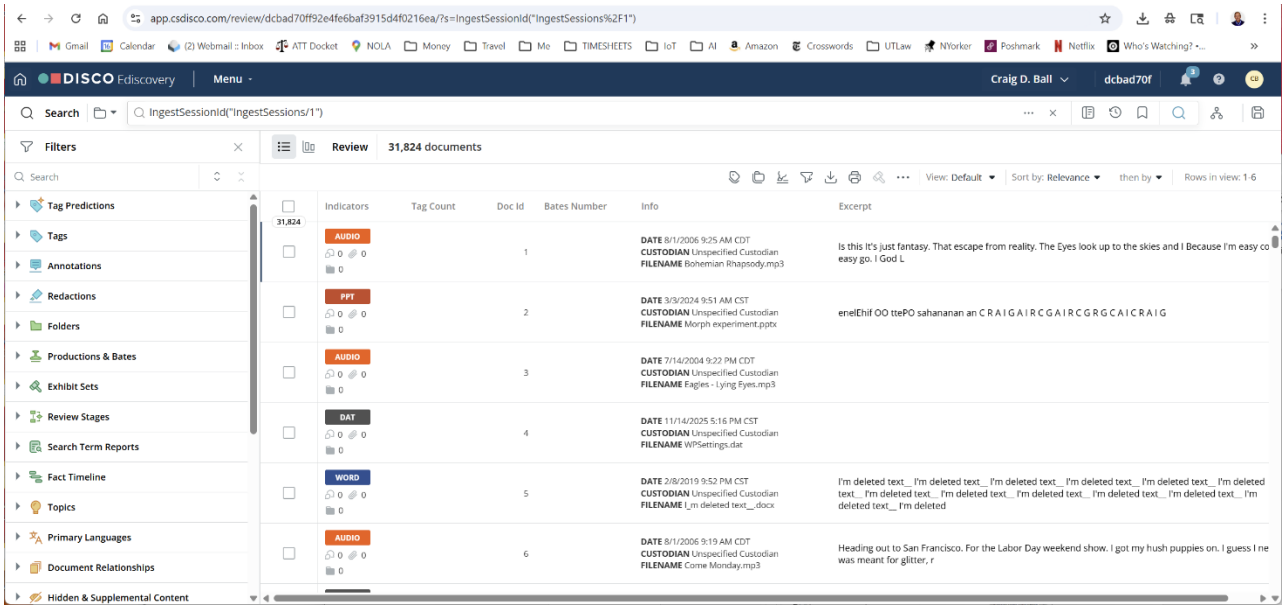
When ingestion and processing is complete, don't hit "Show Documents" just yet! Instead, click the box with three dots to the right of the trash can (see screenshot next page) and **download BOTH your Ingest Report AND your Exception Report. <u>You will submit both of these to me via Canvas</u> as part of your responses to this Exercise.**



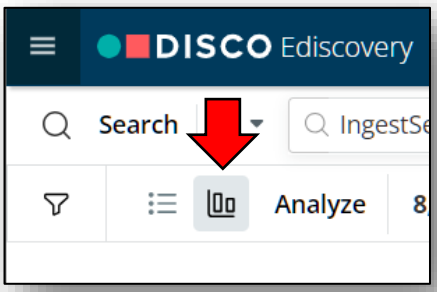
Now, click the "Show Documents" button.

Step 6: Perusing the Evidence

You should see the DISCO Search and Review screen displaying some of the files identified within the contents of the ingested Zip container. The Zip container you uploaded contained ~11 files but generated ~31,824 discrete items after processing, recursion and folder suppression.

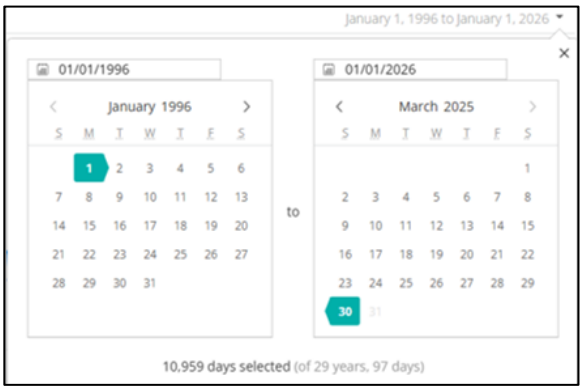


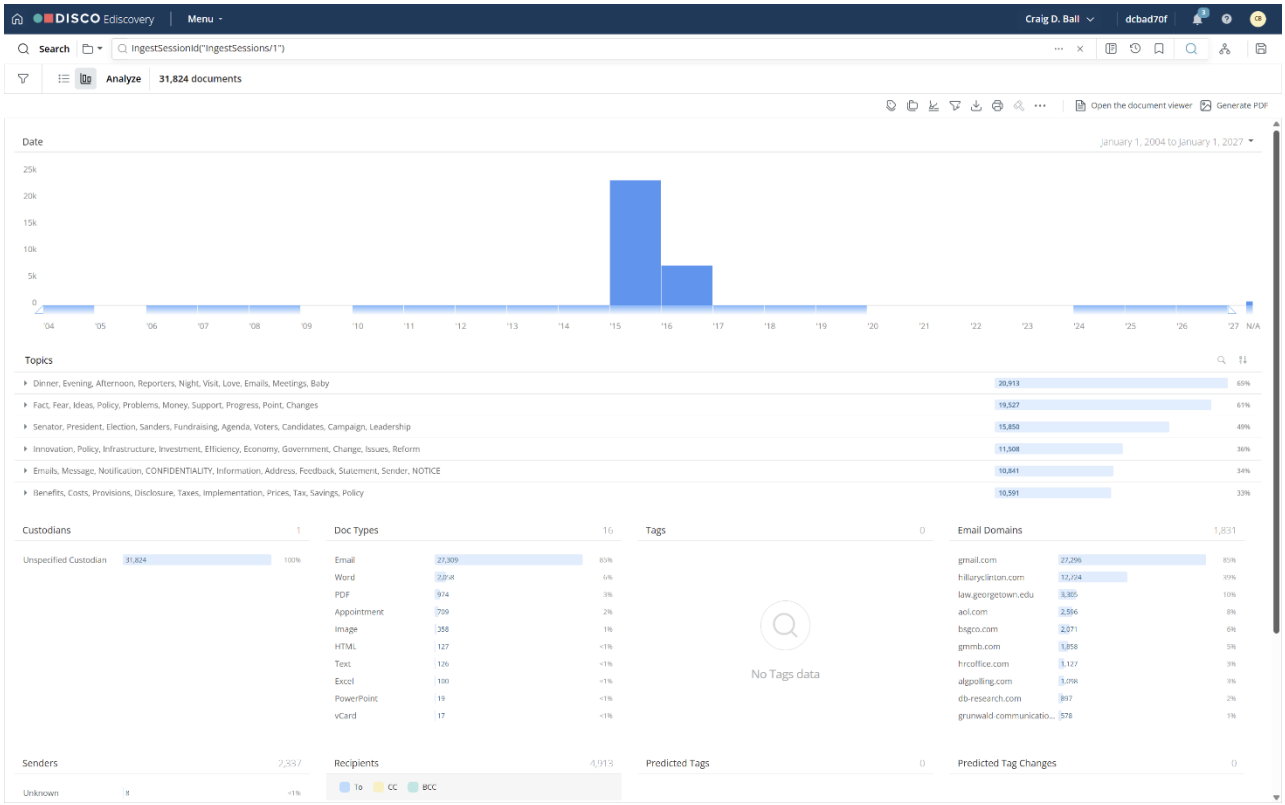
Let's gain a sense of what data comprises our collection by clicking on the bar chart icon near the upper left of the screen to open the Search Visualization screen (see figure at right). You will call up an item volume timeline broken out by file type, email domain, senders and recipients (figures below). You can access this bar graph breakdown, called a



"**histogram**," for any subset of data on your screen as you deploy searches and filters

Note that most everything onscreen is hyperlinked such that clicking on a link displays just the hyperlinked content for that element of the histogram. Within the Search Visualization screen, you can filter content for a date range by dragging across the vertical bars comprising the histogram or you can filter more precisely by clicking on the date range displayed in the upper right corner of the timeline to display a calendar interface (seen at right).

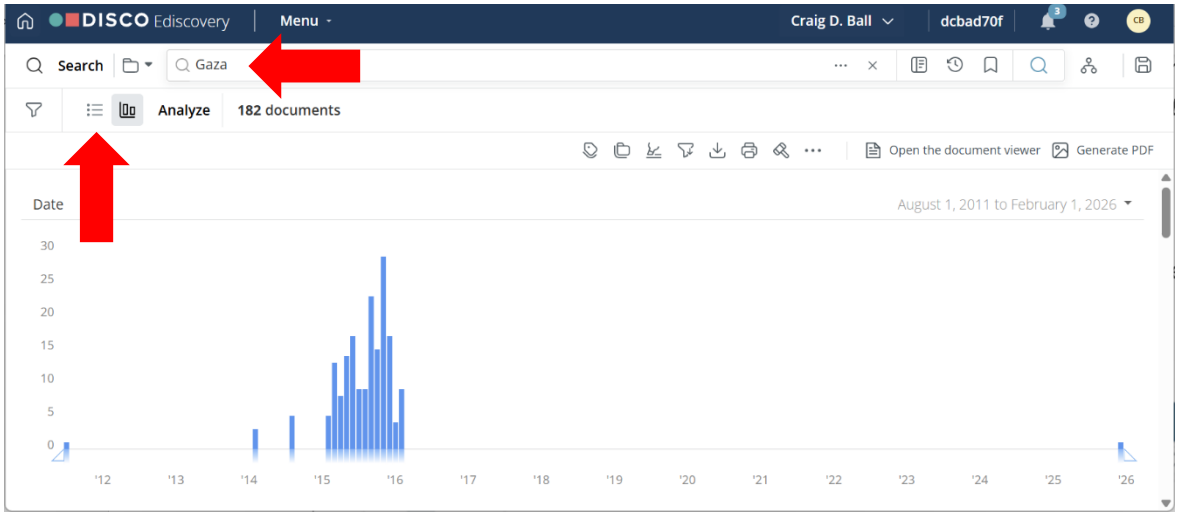




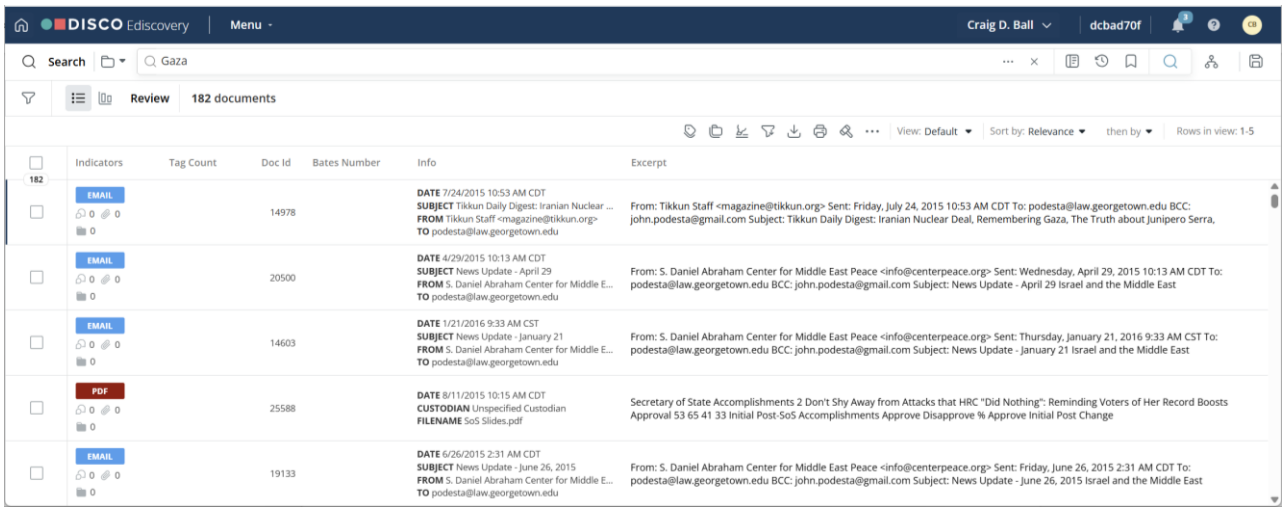
Try filtering for a date range like 1/1/2016 to 12/31/16 by entering the dates manually using the calendar. I got 7,471 documents—the same result I get by clicking on the 2016 bar in the histogram. *Be sure to clear your date filter before proceeding to Step 7. Note the "Clear Filters" link near the top right of your screen appears whenever a filter is applied.*

Step 7: Running a Keyword Search

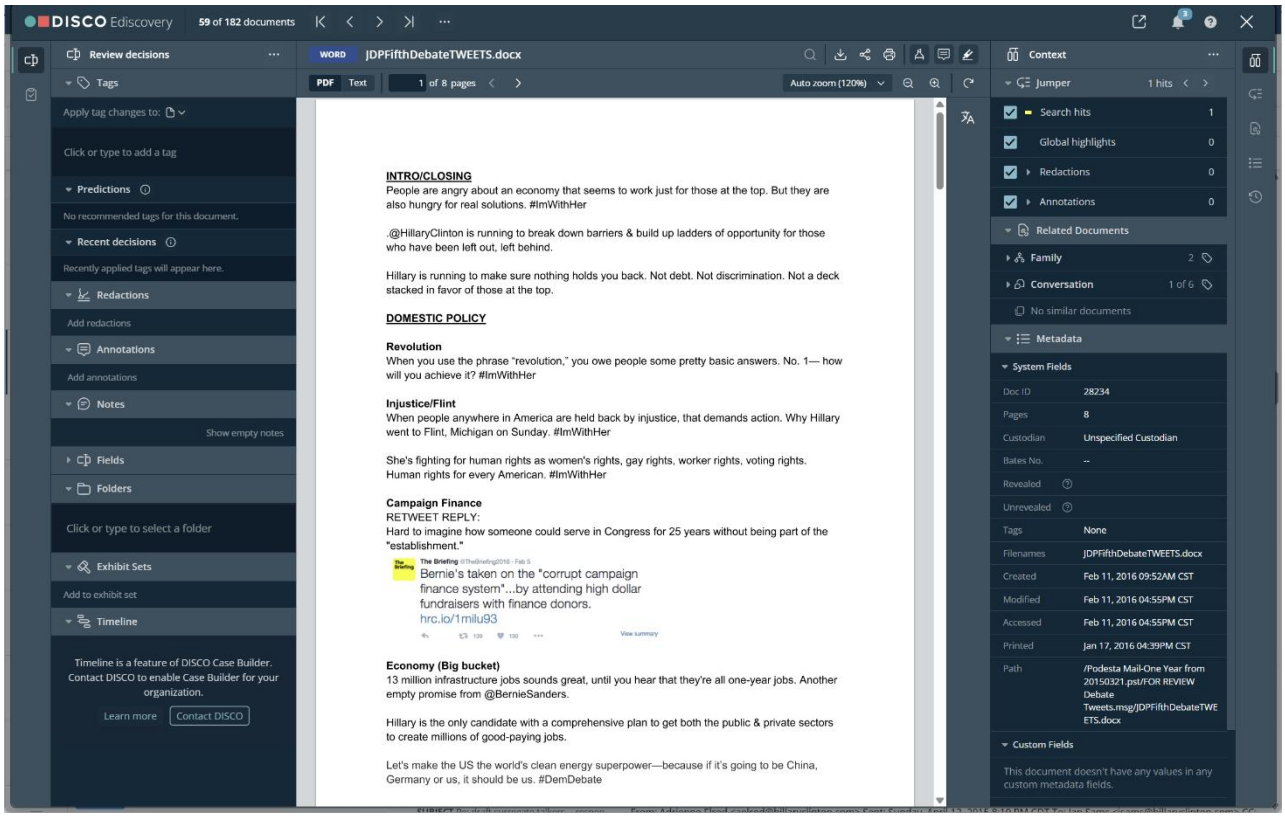
Locate the search entry field and type “Gaza” in the blank, then hit enter. As seen below, DISCO reports 182 documents with search hits (spanning 99 email messages, 65 Word documents, 9 PDF files and 9 Excel spreadsheets).



To view a list of those 182 documents, switch to List View (lower red arrow, above left), closing the Search Visualization screen, and scroll through the results.



Clicking any listed item launches it in a viewer, where we can also see metadata, add tags and move forward and backward through the collection of items hitting on the search query “Gaza.”



DISCO is an example of a "hosted" eDiscovery "Review Tool" or "Platform" running "in the Cloud." Not so many years ago, tools like these ran on the processors of desktop machines displaying documents stored locally on file servers. Today, most review work takes place via hosted tools like DISCO, Relativity, Everlaw, Nuix Discover, Logikcull, Catalyst iConect and many others.

Creating a Production Set

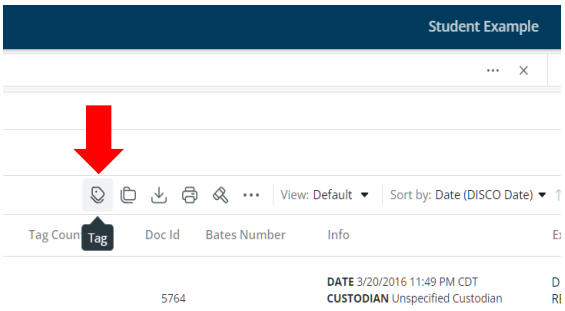
Step 8: Set up a "Responsive" Tag and Bulk Tag the Search Results

Return to the Search Visualization screen and clear all searches and filters so that all 31,000+ documents comprise the histogram. Create an empty folder on your desktop called "**Gaza Production Set**." Run the following search in DISCO:

(Gaza OR Hamas OR Palestin!) /10 Netanyahu AND date(after 1/1/2015)

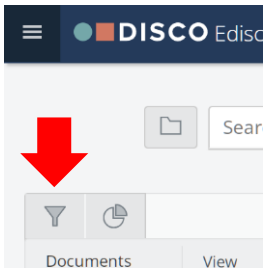
The /10 connector means "within ten words of" and the exclamation point is a "root expander" allowing the term to hit on, e.g., Palestine and Palestinian. I got 53 results: 42 emails, 10 Word Documents and one PDF file.

Close the Search Visualization screen to return to the document list of 53 results. Find the Bulk Tag button on the menu bar (figure at right). Click the Bulk Tag button and create a tag. In the blank named "Apply," click on the words "Select one or more tags" and choose "Responsiveness" then "Responsive." Now, click "Update." Your results are now tagged as Responsive. Clear your search query to return to the view of all 31,824 documents.



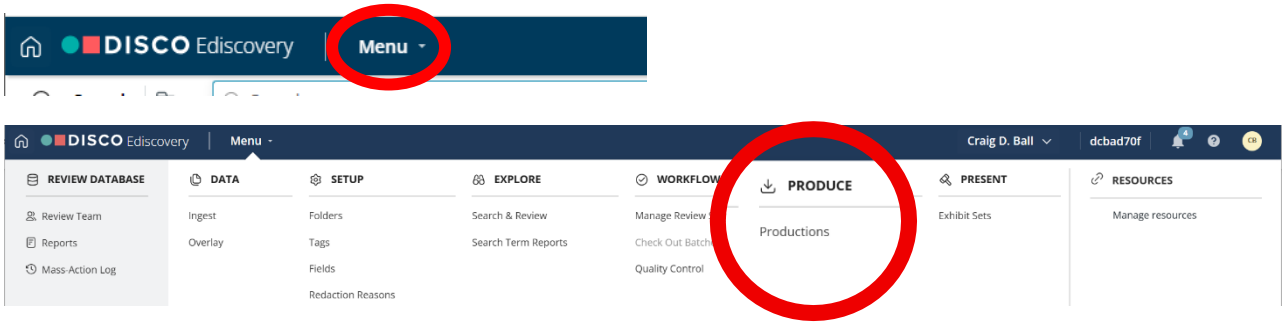
Step 9: Select Tagged Items

Locate and click the funnel icon alongside the Search Visualization button to reveal a list of filters. Click on Tags and drill down through the list of tags until you see "Responsive." Check the box for "Responsive." You should see 53 items listed.

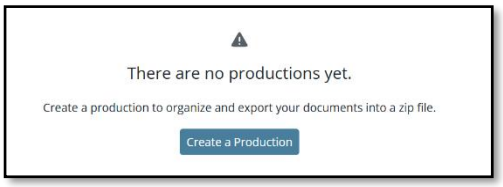


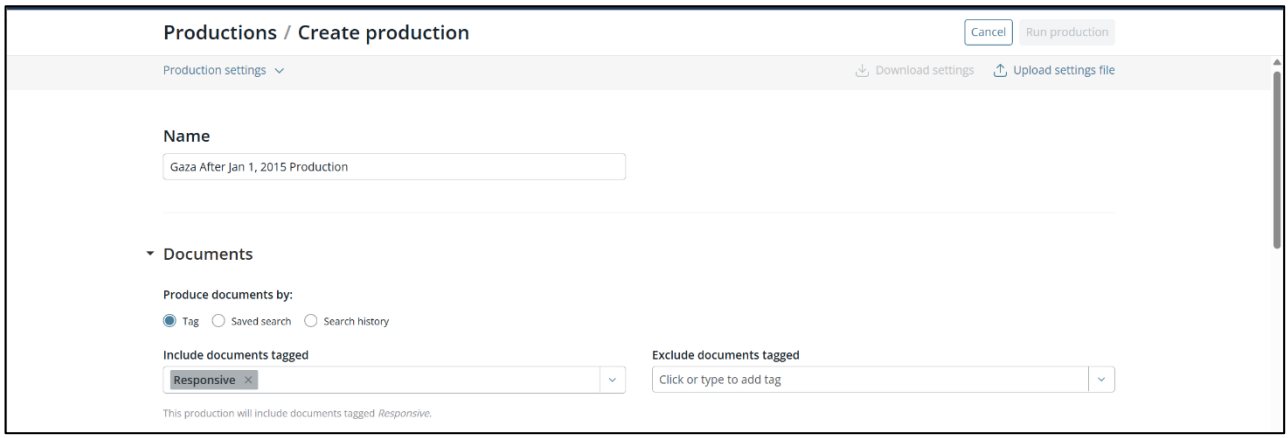
Step 10: Configure the Production Set

Return to the Menu option on the blue bar. Click "Menu" then select "Productions."

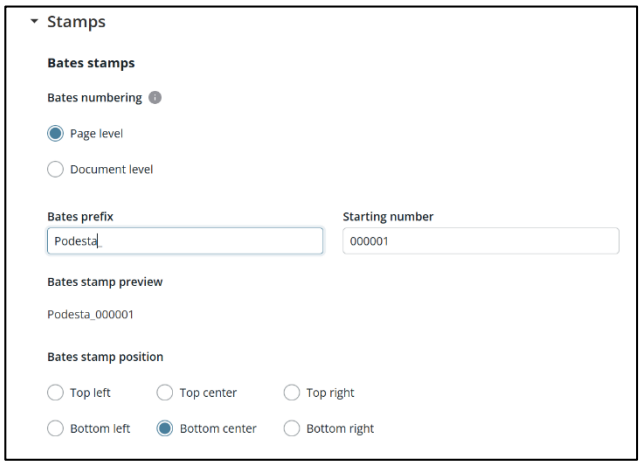


On the next screen, click "Create a Production" and configure your production as follows: Name your production "Gaza After Jan 1, 2015 Production" and ensure that "Tag" is selected under "Produce documents by:" and that "Responsive" is chosen for "Include documents tagged," as seen below:

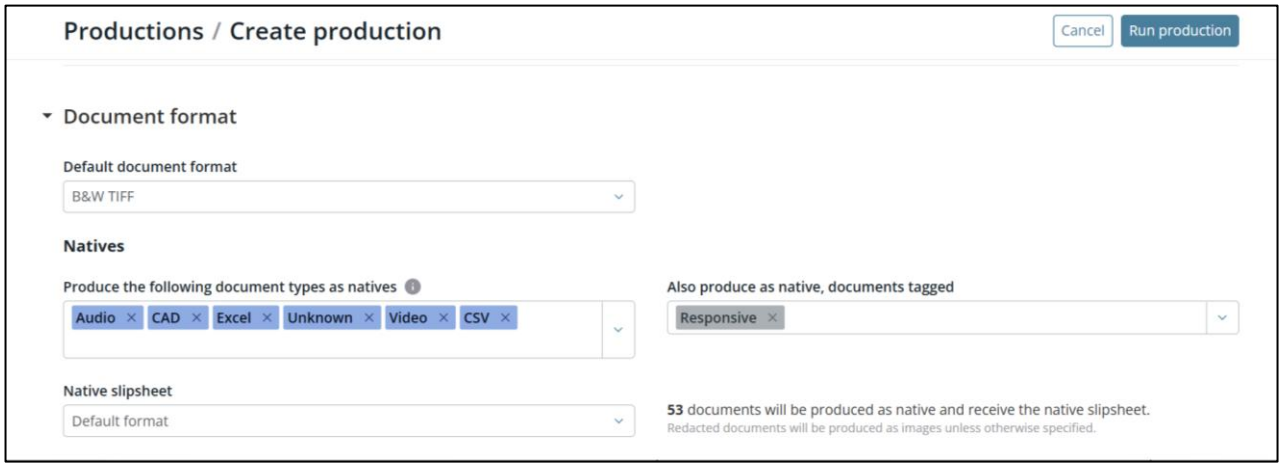




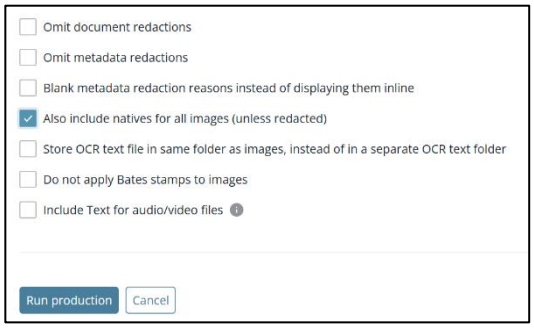
Under “Bates Stamps,” set the Bates prefix to Podesta_” and leave the starting number as 000001. Select “Bottom center” as Bates stamp position.

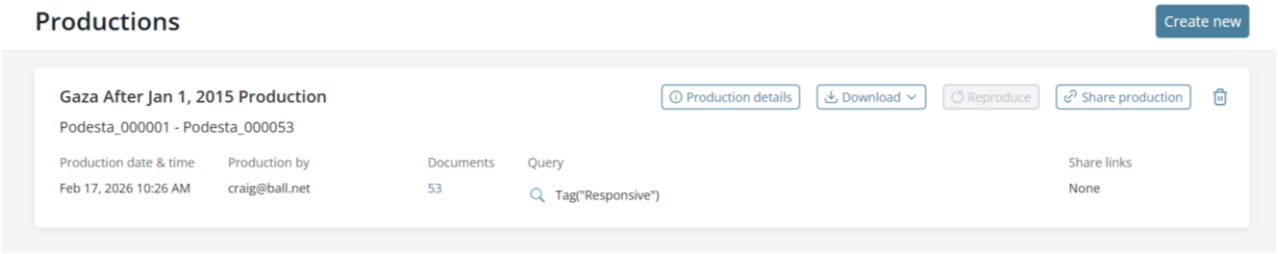


Under “Document Format,” select “B&W TIFF” as the Default document format. Your configuration should look like the screenshot below:



Step 11: Check your settings and leave the balance of settings below at their defaults EXCEPT check “Also include natives for all images (unless redacted)” as seen at right. If everything looks good, scroll to the bottom and click “Run Production.” The production process should take about five to ten minutes to complete.





When the production process completes, click "Download" and choose "Load files only." Save these load files to the "Gaza Production Set" folder on your desktop. It will likely be a compressed archive file named, "Podesta_000001-Podesta_000053_loadfiles.zip;" however, the final Bates number may vary somewhat. If it varies markedly, you may have missed a step when configuring production.

For this exercise, you will submit the **three** following items via Canvas:

1. **Your Ingestion AND Exception Reports from Step 5, and**
2. **The 35KB file "Podesta_000001-Podesta_000053_loadfiles.zip" from Step 11. You do NOT need to turn in the 31MB file called Podesta_000001-Podesta_000053.zip.**

Collectively, these three files should be about 9 MB in size. If substantially larger or smaller than that, please be certain you're sending the correct files (and maybe reach out to me for help).

You're done!

UT Processing Exercise – Condensed Instructions

STEP 1: MOUNT THE FORENSIC IMAGE

Mac:

1. Make a working copy of your image file (e.g., Evidence.001).
2. Change the file extension to .dmg (use Get Info > Name & Extension).
3. Double-click the .dmg to mount it. It should appear on your Desktop as a drive.

Windows:

1. Launch FTK Imager.
2. Go to File > Image Mounting.
3. Click the "..." button to select your .001 image file.
4. Click Mount and note the drive letter.

STEP 2: CREATE A ZIP ARCHIVE

Important: Do not zip the image file. Zip the contents of the mounted drive.

Mac:

• Right-click (or Ctrl-click) the mounted volume > Compress "Evidence Drive".

Windows:

• Use 7-Zip (not built-in Windows zip): Select contents of the mounted drive > Right-click > 7-Zip > Add to archive > **choose ZIP format (important!)**.
• Save to Desktop (not to the mounted image!).

STEP 3: UPLOAD TO DISCO

1. Go to https://login.csdisco.com and log in.
2. Choose UT Processing Exercise 2026 and your personal review database.
3. Menu > Data > Ingest > New Ingest > Select Native > Native Lite.
4. Browse to your Evidence Drive.zip and upload it.
5. Name the ingest (e.g., “Evidence Drive #”).

STEP 4: WAIT WHILE DISCO PROCESSES (45–75 MIN)

STEP 5: DOWNLOAD REPORTS

• Click the 3-dot menu > Download:
  - Ingest Report
  - Exception Report

STEP 6: EXPLORE THE EVIDENCE

• Click the bar chart icon to open Search Visualization.
• Apply date filter: 1/1/2016 – 12/31/2016.

STEP 7: SEARCH FOR “GAZA”

• Enter Gaza in search bar. Review and scroll through the 182 hits.
• Switch to List View to browse.

STEP 8: BULK TAG RESPONSIVE DOCUMENTS

1. Clear all filters and searches.
2. Search: (Gaza OR Hamas OR Palestin!) /10 Netanyahu AND date(after 1/1/2015)
3. In List View, click Bulk Tag > Choose Responsiveness > Responsive.

STEP 9: FILTER BY TAG

• Click funnel icon > Tags > Select Responsive.

STEP 10: CREATE A PRODUCTION SET

1. Menu > Productions > Create a Production.
2. Name: Gaza After Jan 1, 2015 Production.
3. Produce by Tag: Responsive.

4. Bates: Podesta_, start at 000001, Bottom center.
5. Format: B&W TIFF + Check “Include natives.”

STEP 11: DOWNLOAD LOAD FILES

- Download Load files only and save to: Gaza Production Set folder on Desktop.
- File name should be like: Podesta_000001-Podesta_000053_loadfiles.zip

SUBMIT VIA CANVAS:

- Ingest Report
- Exception Report
- Load file Zip (~9 MB total)



Exercise 21: Culling, Search and Export Part II

GOALS: The goals of this exercise are for the student to:

1. Generate a Review Set from a Request for Production; and
2. Apply the skills acquired and evidence ingested in Exercise 20

OUTLINE: Students will convert a request for production into searches and filters serving to cull a collection of potentially responsive email and attachments into a much smaller subset likely to contain all responsive items but stripped of as many non-responsive items as a reasonably conceived and -executed search and filtering strategy allows. Once the culled set is identified, students will use skills acquired in the previous exercise to generate a PDF production set for review by senior counsel.

Problem: *You are an associate attorney working as John Podesta's counsel. A Justice Department civil subpoena requires that Podesta produce documents, communications and other ESI found on the Evidence Drive that touch or concern Russian President Vladimir Putin and Ukraine and/or the Ukrainian people for the one-year period ending on June 30, 2016. Not just Putin alone, and not just Ukraine, but both the Russian leadership and Ukraine and/or its people. Do not concern yourself with filtering out privileged content; HOWEVER, exclude any responsive messages and attachments sent to John Podesta via his Georgetown Law School email address from the production*

Though modest review of some items in the collection may help you identify search terms and filters to employ, this exercise is not designed to require much manual review. I'm seeking instead to assess your insight into effective conventional methods of automated search and culling. Accordingly, I wouldn't expect this exercise to require more than an hour or two of student effort to complete.

You will submit two items via Canvas:

1. Your production set (PDF production ONLY-no TIFFs or native files); and
2. A succinct description of the queries and filters employed to generate the result set.
Note that I'm not seeking a description of how you tagged and exported the data but, a description of the searches, filters, settings and/or methods you used that would enable anyone else with DISCO and an identical Evidence Drive to *identify and isolate* the same items in your production set by following your strategy and methodology, Replicability and effective documentation of process are key features of defensible e-discovery.

Forms that Function

This chapter discusses how to request and produce electronically stored information (ESI) in *forms that function*—that is, in more utile and complete forms of production that preserve the integrity, efficiency and functionality of digital evidence. It explains the advantages of securing production in native and near-native forms and supplies exemplar language crafted to convey forms of production and metadata values sought.

BACKGROUND

Historically, the law little concerned itself with "forms" of production because there were few alternatives to paper. Later, evidence became digital: documents, pictures, sounds, text messages, e-mail, spreadsheets, presentations, databases and more were created, communicated and recorded as sequences of "ones" and "zeroes." Flat forms of information acquired new dimension and depth, described and supplemented by **metadata,** *i.e.,* data *about* data supporting the ability to find, use and trust digital information.

Digital photographs hold EXIF data revealing where they were taken and by what camera, spreadsheets carry formulae supporting complex calculations and Word documents store editorial histories and are laced with conversations between collaborators. Presentations feature animated text and rich media, including sound, video and dynamic connections to other data. Databases don't "store" documents as much as assemble them on demand. Even conversations—once the most ethereal of interactions—now linger as text messages and data packets traversing the internet and cellular networks.

Today, the forms in which information is supplied determine if evidence is intelligible, functional and complete.

FORMS OF PRODUCTION IN THE FEDERAL RULES

The Federal Rules of Civil Procedure further the goals that lawyers understand the forms of ESI in their cases and discuss and resolve forms disputes *before* requests for production are served so that unresolved forms disputes be brought to the Court before expending the cost and time of misbegotten production.

Rule 26(f)(3)(C) requires the parties to submit a discovery plan to the Court prior to the first pretrial conference. The plan must address "any issues about disclosure or discovery of electronically stored information, including the form or forms in which it should be produced."

Rule 34(b)(1)(C) permits requesting parties to "specify the form or forms in which electronically stored information is to be produced," yet it's common for requests for production to be wholly silent on forms of production, despite pages of detailed definitions and instructions.

Practice Tip: Requesting parties should supply a clear and practical written specification of forms sought *before* the initial Rule 26(f) conference, affording opponents the opportunity to assess the feasibility, cost and burden of producing in specified forms. Even parties who do not know the forms in which an opponent's data natively resides can anticipate the *most common* forms suited to, *e.g.*, e-mail, word processed documents, presentations and spreadsheets.

The Federal Rules lay out **five steps** to seeking and objecting to forms of production:

1. Before the first pretrial conference, parties must hash out issues related to "the form or forms in which [ESI] should be produced." FRCP 26(f)(3)(C)

2. Requesting party specifies the form or forms of production for each type of ESI sought: paper, native, near-native, imaged formats or a mix of same. FRCP 34(b)(1)(C)

3. If the responding party will supply the specified forms, the parties proceed with production. If not, the responding party must object and designate the forms in which it intends to make production. If the requesting party fails to specify forms sought, responding party must state the form or forms it intends to produce. FRCP 34(b)(2)(D)

The Notes to Rule 34(b) add: "A party that responds to a discovery request by simply producing electronically stored information in a form of its choice, without identifying that form in advance of the production . . . runs a risk that the requesting party can show that the produced form is not reasonably usable and that it is entitled to production of some or all of the information in an additional form."

4. If requesting party won't accept the forms the producing party designates,the requesting party must confer with the producing party in an effort to resolve the dispute. FRCP 37(a)(1)

5. If the parties can't agree, requesting party files a motion to compel, and the Court selects the forms to be produced.

Practice Tip: Even when producing parties use native and near-native forms when reviewing for responsiveness and privilege, the last step before production is often to downgrade the evidence to static images. Accordingly, requesting parties shouldn't wait until the response date to ascertain if an opponent won't furnish the forms sought. Press for a commitment; and if not forthcoming, move to compel ahead of the response date. Don't wait to hear the Court ask, "Why didn't you raise this earlier?"

WHAT ARE THE OPTIONS FOR FORMS OF PRODUCTION?

It's rarely necessary or feasible to employ a single form of production for all ESI produced in discovery; instead, tailor forms to the data. Options for forms of production include:

- Paper [where the source is paper and the volume small]
- Page Images [best for items requiring redaction and scanned paper records]
- Native [spreadsheets, electronic presentations and word processed documents]
- Near-native [e-mail and database content]
- Hosted production

Paper

Converting searchable electronic data to paper is rarely a reasonable form of production for ESI, but paper remains an option where the items to be produced are paper documents and so few in number that electronic searchability isn't essential.

Page Images

Parties produce digital "pictures" of documents, e-mails and other electronic records, typically furnished in Adobe's Portable Document Format (PDF) or as Tagged Image File Format (TIFF) images. Converting ESI to TIFF images strips its electronic searchability and metadata. Accordingly, TIFF image productions are accompanied by load files holding searchable text and selected metadata (a so-called "TIFF+ production"). Searchable text is obtained by extraction from an electronic source or, for scanned paper documents, by use of optical character recognition (OCR). Load files are composed of *delimited text*, *i.e.,* values following a predetermined sequence and separated by characters like commas, tabs or quotation marks. The organization and content of load files must be negotiated, and is often pegged to review software like Relativity, Everlaw or DISCO.

> ***Pros:*** Imaged formats are ideal for production of scanned paper records, microfilm and microfiche, especially when OCR serves to add electronic searchability.
>
> ***Cons:*** Imaged production breaks down when ESI holds embedded information (*e.g.,* collaborative content like comments or formulae in spreadsheets) or non-printable information (*e.g.,* voice mail, video or animation and structured data). Imaged productions may also serve to degrade evidence when the information is fielded (*e.g.,* structured data and messaging) or functional (*e.g.,* animations in presentations, table relationships in structured data or threads in e-mail).

Native Production

Parties produce the actual data files containing responsive information, *e.g.,* Word documents in their native .DOC or .DOCX formats, Excel spreadsheets as .XLS and .XLSX files and PowerPoint presentations in native .PPT and .PPTX. Native production is cheaper and superior in competent hands using tools purpose-built for native review.

> ***Pros:*** The immediate benefits *to the producing party* are speed and economy—little or nothing must be spent on image conversion, text extraction or OCR.
>
> The benefits *to the requesting party* are substantial. Using native review tools or applications like those used to create the data (Careful here! —see *Cons* below), requesting parties see the evidence as it appeared to the producing party. Embedded commentary and metadata aren't stripped away, deduplication is facilitated, e-mail messages can be threaded into conversations, time zone irregularities are normalized and costs are reduced and utility enhanced every step of the way. Moreover, native file sizes tend to be many times more compact than their counterparts converted to static images, making native forms much less costly to ingest for processing and host for review.
>
> ***Cons:*** Applications needed to view rare and obscure data formats may be prohibitively expensive (*e.g.,* specialized engineering applications or enterprise database software). If native applications are (unwisely) tasked to review, *e.g.,* Microsoft Word for reviewing Word documents, copies must be used to avoid altering evidence.

Near-Native Production

When some ESI cannot be feasibly or prudently tendered in true native formats, *near*-native forms preserve the essential utility, content and searchability of native forms but are not, strictly speaking, native forms. Examples:

- **Enterprise e-mail** – Enterprise email systems store messages in monolithic container formats like ab Exchange Server's EDB format; so, exported messages tend to be stored in container- or single message formats not native to the mail server. These replicate the pertinent content and essential functionality of the source, but again, are not, strictly speaking, native forms.
- **Databases** - Exports from databases are often produced in delimited formats not native to the database yet supporting the ability to interpret the data in ways faithful to the source.
- **Social networking content** - Content from social networking sites won't replicate the precise way the content is stored in the cloud, so near-native forms seek to replicate its essential utility, completeness and searchability.

Hosted Production

Hosted production is more a delivery medium than a discrete form of production. Hosted production resides on a secure website. Requesting parties access data using their web browser, searching, viewing, annotating and downloading data. The electronic forms of production above are the grist ingested by hosting providers (service providers) to comprise the hosted collection.

Load Files Explained

Some years back, I got a call from a lawyer who reported that he'd received production of ESI from a bank and spent the weekend going through it. He'd found images of the pages of electronic documents, but they weren't PDFs and he couldn't search them. He also found a lot of what he called "Notepad documents." He'd specified native production, so thought it odd that the other side produced so many pictures of documents and plain text files.

As it's unlikely a bank would rely on Windows Notepad as its word processor, I probed further and learned that that the production included folders of TIFF images, folders of .TXT files (those "Notepad documents") and folders of files with extensions like .DAT and .OPT. My caller didn't know what to do with these.

Unbeknownst to my caller, he'd received a TIFF+ production from an opponent who ignored his demand for native forms and simply printed everything to electronic paper. The producing party no doubt expected the requesting party to buy or own a review tool capable of cobbling together page images with extracted text and metadata produced in load files. Lacking such tools, the caller found the production to be wholly unsearchable and largely unusable. Should my caller protest, the other side will tell him how all those other files represent the very great expense and trouble they've gone to make the page images searchable, as if furnishing load files to add crude searchability to page images of inherently searchable electronic documents constitutes some great favor.

It brought to mind that classic Texas comeback, "Don't piss in my boot and tell me it's raining."

It also reminded me that not every lawyer knows about load files, those unsung digital sherpas tasked to tote metadata and searchable text otherwise lost when ESI is converted to TIFF images. Grasping the fundamentals of load files is important to fashioning a workable electronic production protocol, whether you're dealing with TIFF images, native file formats or a mix of the two.

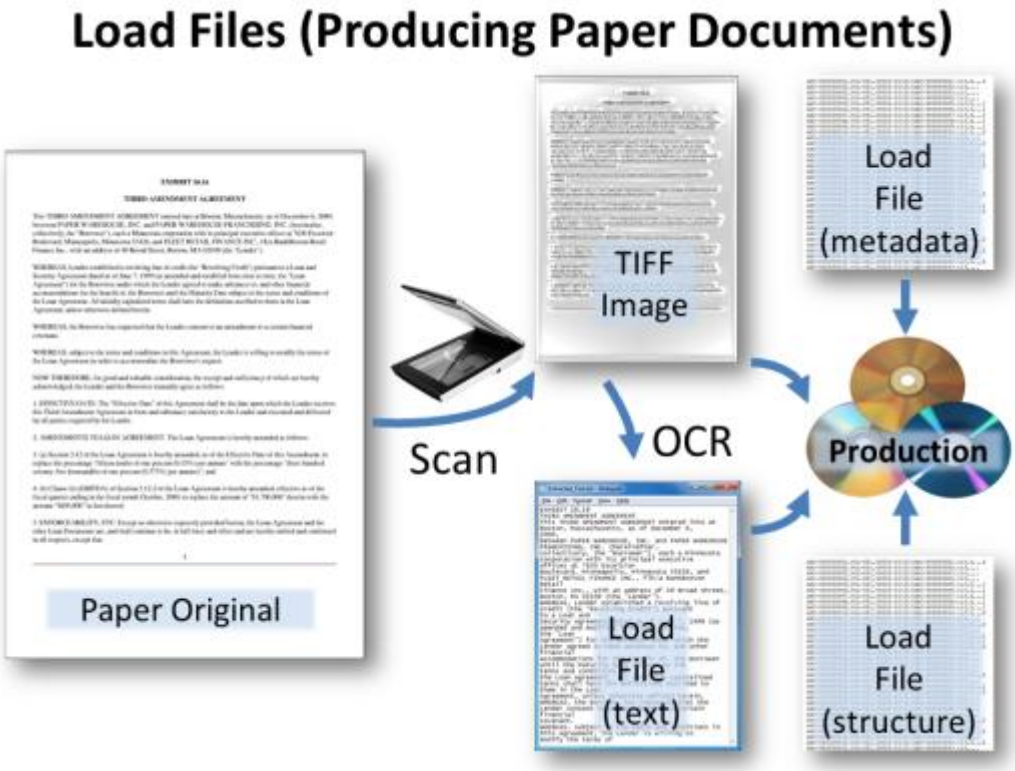
In simplest terms, load files carry data that has nowhere else to go. They are called load files because they are used to load data into, *i.e.,* to "*populate*" a database. Load files first appeared in civil discovery in the 1980s to add electronic searchability to scanned paper documents. Then as now, paper documents were scanned to TIFF image formats and the images subjected to optical character recognition (OCR). Unlike Adobe PDF images, TIFF images weren't designed to integrate searchable text; consequently, the text garnered using OCR was stored in simple ASCII text files named with the Bates number of the corresponding page image. Think pants with pockets versus skirts without pockets. When you use TIFF images for production, text must go *somewhere* and, since TIFFs have no "pockets," the text goes into a "purse" called a "load file." Compared to

unsearchable paper documents, imaging and OCR added functionality. It was 20th century information technology improving upon 19th century printing technology, and if you were a lawyer in the Reagan-era, this was futuristic stuff.

Metadata is "data about data." While we tend to think of metadata as a feature unique to electronic documents, paper documents have metadata, too. They come from custodians, offices, files, folders, boxes and other physical locations that must be tracked. Still more metadata takes the form of codes, tags and abstracts reflecting reviewers' assessments of documents. Then as now, all this metadata needs somewhere to lodge as it accompanies page images on their journey to document review database tools (a/k/a "review platforms. This data goes into load files.

Finally, we employ load files as a sort of road map and as assembly instructions laying out, *inter alia*, where document images and their load files holding their searchable text and metadata are located on disks or other media used to store and deliver productions and how the various pieces relate to one-another.

So, to review, some load files carry extracted text to facilitate search, some carry metadata about



the documents and some carry information about how the pieces of the production are stored and how they fit together. Load files are used because neither paper nor TIFF images are suited to carrying the same electronic content; and if it weren't supplied electronically, you couldn't load it into review platforms to search it using computers.

Different review platforms use different load file formats to order and separate information according to guidelines called "load file specifications." Load files employ characters called delimiters to field (separate) the various information items in the load file.

Load File Structure

Imagine creating a table to keep track of documents. You might use the first two columns of your table to number the beginning (first) and ending (last) pages of each document. The next column holds the document's file name and then each succeeding column carries information about the document (*e.g.,* Date, Author, Type). To tell one column from the next, you'd draw lines to delineate the rows and columns, like so:

The lines serve as delimiters—literally delineating one field of data from the next. Vertical and horizontal lines are excellent visual delimiters for humans, but computers work well with characters like commas or tabs. So, if the tabular data were a load file, it might be delimited as:

```
BEGDOC,ENDDOC,FILENAME,MODDATE,AUTHOR,DOCTYPE
0000001,0000004,Contract,01/12/2013,J. Smith,docx
0000005,0000005,Memo,02/03/2013,R. Jones,docx
0000006,0000073,Taxes_2013,04/14/2013,H. Block,xlsx
0000074,0000089,Policy,05/25/2013,A. Dobey,pdf
```

Each comma replaces a column divider, each line signifies another row and the first or "header" row is used to define the data that follows and the way it's delimited.

Load files using commas to separate values are called "comma separated value" or CSV files. Legacy load file formats adhere to conventions established by early review tools like Concordance and Summation, using unique delimiters that remain common today.

In e-discovery, there are three principal functions for delimiters:

- **Document Delimiter:** A Document Delimiter signals a switch from one document to the next. In the most common load file format, a carriage return/line feed serves this purpose.
- **Field Delimiter:** A Field Delimiter signals a change from one field to the next.
- **Quote Delimiter**: A Quote Delimiter permits a delimiting character to be used within the fielded data without it being treated as a field delimiter. When, for example, a comma serves as a delimiter in A CSV file, the Quote Delimiter enables the comma to be treated as a comma in the text rather than indicating a shift from one field to the next.

Concordance load files use the file extension .DAT and the **þ** (thorn, ALT-0254, Unicode 00FE) as the Quote Delimiter and the **¶** (pilcrow, ALT-0182, Unicode 00B6) character as the Field Delimiter. Note how each line denotes a single document:

Concordance Load File

```
þBEGDOCþ¶þenddocþ¶þfilenameþ¶þMODDATEþ¶þAUTHORþ¶þDOCTYPEþ
þ0000001þ¶þ0000004þ¶þContractþ¶þ01/12/2013þ¶þJ. Smithþ¶þdocxþ
þ0000005þ¶þ0000005þ¶þMemoþ¶þ02/03/2013þ¶þR. Jonesþ¶þdocxþ
þ0000006þ¶þ0000073þ¶þTaxes_2013þ¶þ04/14/2013þ¶þH. Blockþ¶þxlsxþ
þ0000074þ¶þ0000089þ¶þPolicyþ¶þ05/25/2013þ¶þA. Dobeyþ¶þpdfþ
```

Summation load files use the file extension .DII, and separate each record like so:

Summation Load File

```
; Record 1
@T 0000001
@DOCID 0000001
@MEDIA eDoc
@C ENDDOC 0000004
@C PGCOUNT 4
@C AUTHOR J. Smith
@DATESAVED 01/12/2013
@EDOC \NATIVE\Contract.docx
; Record 2
@T 0000005
@DOCID 0000005
@MEDIA eDoc
@C ENDDOC 0000005
@C PGCOUNT 1
@C AUTHOR R. Jones
@DATESAVED 02/03/2013
@EDOC \NATIVE\Memo.docx
@C AUTHOR A. Dobey
@DATESAVED 05/25/2013
@EDOC \NATIVE\Policy.pdf
```

Two more load files:

Opticon load files (file extension .OPT) are used in conjunction with Concordance load files to pair Bates numbered pages with corresponding page images and to define the **unitization** of each document; that is, where they begin and end. Documents may be unitized *physically*, as when constituent pages are joined by clips, staples or bindings, or *logically*, where constituent pages

belong together even if not physically unitized (as when documents are bulk scanned or transmittals reference enclosures). Logical unitization is also a means to track family relationships between container files and contents and e-mail messages and attachments.

Opticon Load File

```
0000001_0001,,TIFF\001\0000001_0001.tif,Y,,,4
0000002_0002,,TIFF\001\0000002_0002.tif,,,,
0000003_0003,,TIFF\001\0000003_0003.tif,,,,
0000004_0004,,TIFF\001\0000004_0004.tif,,,,
0000005_0001,,TIFF\001\0000005_0001.tif,Y,,,1
0000006_0001,,TIFF\001\0000006_0001.tif,Y,,,68
0000007_0002,,TIFF\001\0000007_0002.tif,,,,
0000008_0003,,TIFF\001\0000008_0003.tif,,,,
0000009_0004,,TIFF\001\0000009_0004.tif,,,,
0000010_0005,,TIFF\001\0000010_0005.tif,,,,
```

Opticon load files employ a simple seven-field, comma-delimited structure:

1. Page identifier,
2. Volume label (optional),
3. Path to page image,
4. New document marker (the letter "Y" for "yes" in the illustration above),
5. Box identifier (optional),
6. Folder identifier (optional),
7. Page count (optional).

Overlay load files are used to update or correct existing database content by replacing data in fields in the order in which the records occur. Thus, it's crucial that the order of data within the overlay file match the order of data replaced. Data must be sorted in the same way, and the overlay must not add or omit fields.

Making the Case against Imaged Production

Parties don't print their e-mail before reading it or emboss a document's name on every page. Parties communicate and collaborate using tracked changes and embedded comments. Parties use native forms because they are the most utile, complete and efficient forms in which to store and access data.

Lawyers come along and convert native forms to images, Bates label each page and purge tracked changes and embedded comments without disclosing the destruction.

Converting a client's ESI from its native state as kept "in its ordinary course of business" to TIFF images **injects needless expense in at least half a dozen ways**:

1. You pay to convert native forms to TIFF images and emboss Bates numbers;
2. You pay to generate load files;
3. You must produce multiple copies of documents (like spreadsheets) that are virtually incapable of production as images;
4. TIFF images and load files are <u>much</u> "fatter" files than their native counterparts (*i.e.,* bloated *5-40 times* as large), so you pay more for vendors to ingest and host them;
5. It's difficult to reliably de-duplicate documents once converted to images; and
6. You must reproduce everything when opponents recognize that imaged productions fall short of native productions.

REBUTTING THE CASE AGAINST NATIVE

When producing parties insists on converting ESI to TIFF despite a timely request for native production, they often rely on Federal Rules of Civil Procedure 34(b)(2)(E)(ii), which obliges parties to produce ESI in "the form or forms in which it is ordinarily maintained or in a reasonably usable form or forms." This reliance is misplaced because "[i]t is only if the requesting party declines to specify a form that the producing party is offered a choice between producing in the form 'in which it is ordinary maintained'—native format—or 'in a reasonably useful form or forms.' Fed. R. Civ. P. 34(b)(2)(E)(i)-(ii)"). *The Anderson Living Trust v. WPX Energy Production,* LLC, No. CIV 12-0040 JB/LFG. (D. N.M. March 6, 2014).

Producing parties usually assert **FOUR JUSTIFICATIONS** for refusing to produce ESI in native and near-native forms. None withstand scrutiny:

1. *You can't Bates label native files.* Making the transition to modern forms of production requires acceptance of three propositions:

- Printouts and images of ESI are not "the same" as ESI;
- Most items produced in discovery aren't used in proceedings; and
- Names of electronic files can be simply changed without altering contents of files.

Native documents carry more information than their imaged counterparts, and are inherently functional, searchable and complete. Moreover, native documents are described by more and different metadata—information invaluable in identifying, sorting and authenticating evidence.

Though you can't *emboss* Bates-style identifiers on discrete pages of a native file until printed or imaged, many native forms (*e.g.,* spreadsheets, social networking content, video, and sound files) don't lend themselves to paged formats and would not be Bates labeled. When Bates-style identifiers are needed on pages for use in proceedings, simply require that file identifiers and page numbers be embossed on images or printouts. In practice, that impacts only a small subset of production.

Practice tip: It's simple and cheap to replace, prepend, or append an incrementing Bates-style identifier to a filename. One free file renaming tool is Bulk Rename Utility, available at www.bulkrenameutility.co.uk. You can even include a protective legend like "Subject to Protective Order." **Renaming a file does not alter its content, hash value or last modified date.**

2. *Opponents will alter evidence*. Evidence tampering is not a new fear or a hazard unique to e-discovery. Page images, being black and white pictures of text, are simple to manipulate (and Adobe Acrobat has long allowed extensive revision of PDF files).

Though any form of production is prey to unscrupulous opponents, native productions support quick, reliable ways to prevent and detect alteration. Producing native files on read-only media guards against inadvertent alteration. Alterations are easily detected by comparing hash values (digital fingerprints) of suspect files to the files produced.

Counsel savvy enough to seek native production should be savvy enough to refrain from evidence handling practices prone to alter the evidence.

3. *Native production requires broader review*. Native forms routinely hold user-generated content (*e.g.,* collaborative comments in Word documents, animated "off-screen" and layered text in presentations and formulae in spreadsheets) that is rarely visible on page images or intelligible in extracted text. Imaged productions often obliterate such matter *without review and without disclosure, objection or logging.* Review is only "broader" because this user-contributed content has long been furtively and indefensibly stripped away.

4. *Redacting native files changes them*. Change is the sole purpose of redaction. The form of production for items requiring redaction should be the form or forms best suited to efficient removal of privileged or protected content without rendering the remaining content wholly unusable.

Some native file formats support redaction brilliantly; others do not. In the final analysis, the volume of items redacted tends to be insignificant. Accordingly, the form selected for redaction shouldn't dictate the broader forms of production when, overall, native forms have decided advantages for items not requiring.

Practice Tip: **Don't let the redaction tail wag the production dog**. If an opponent wants to redact in TIFF or PDF, *let them*, but only for the redacted items and only when they restore searchability after redaction.



UPDATING YOUR REQUESTS FOR PRODUCTION

The first step in getting the information you seek in the forms you desire is to ask for it, applying the rules and eschewing dated boilerplate. Clear, specific requests are the hardest to evade and the easiest to enforce. See **Appendix: Exemplar Production Protocol, *infra.***

Most digital evidence—including e-mail—exists as data within databases. So, stop thinking about discovery as the quest for "documents" and start focusing on what you really seek: *information in utile and complete forms.*

The definition of "document" must give way to an alternate term like "information" or "information items." Instead of the usual thesaurus-like litany of types of information, consider:

> "Information items" as used here encompass individual documents and records (including associated metadata) whether on paper or film, as discrete "files" stored electronically, optically or magnetically or as a record within a database, archive or container file. The term should be read broadly to include e-mail, messaging, word processed documents, digital presentations, spreadsheets and database content.

Next, **cut junk prose** like "including, but not limited to" and "any and all." They don't add clarity. If you must incorporate examples of responsive items in a request, just say "including" and add an instruction that says, "Examples of responsive items set out in any request should not be construed to limit the scope of the request." If drafting a request without "any and all" makes you quake, add the instruction, "Requests for production should be read so as to encompass any and all items responsive to the request."

Before you serve discovery, **check your definitions** to be sure you've defined only terms you've used and used terms only in ways consistent with your definitions.

Specify the forms you seek

The most common error seen in requests for production is the failure to specify the forms sought for ESI production. Worse, requests often contain legacy boilerplate specifying forms the requesting party *doesn't* want.

Every request for production should specify forms of production sensibly and precisely. Don't assume that "native format" is clear or sufficient; instead, specify the formats sought for common file types, *e.g.:*

Information that exists in electronic form should be produced in native or near-native formats and should not be converted to imaged formats. Native format requires production in the same format in which the information was customarily created, used and stored in the ordinary course. The table below supplies examples of the native or near-native forms in which specific types of electronically stored information (ESI) should be produced.

Source ESI	**Native or Near-Native Form or Forms Sought**
Microsoft Word documents	.DOC, .DOCX

Microsoft Excel Spreadsheets	.XLS, .XLSX
Microsoft PowerPoint Presentations	.PPT, .PPTX
Microsoft Access Databases	.MDB, .ACCDB
WordPerfect documents	.WPD
Adobe Acrobat Documents	.PDF
Images	.JPG, .JPEG, .PNG
E-mail	Messages should be produced in a form or forms that readily support import into standard e-mail client programs; that is, the form of production should adhere to the conventions set out in the internet e-mail standard, RFC 5322. For Microsoft Exchange or Outlook messaging, .PST format will suffice. Single message production formats like .MSG or .EML may be furnished with folder data. If your workflow requires that attachments be extracted and produced separately from transmitting messages, attachments should be produced in their native forms with parent/child relationships to the message and container(s) preserved and produced in a delimited text file.
Databases	Unless the entire contents of a database are responsive, extract responsive content to a fielded and electronically searchable format preserving metadata values, keys and field relationships. If doing so is infeasible, please identify the database and supply information concerning the schema and query language of the database along with a detailed description of its export capabilities so as to facilitate crafting a query to extract and export responsive data.
Documents that do not exist in native electronic formats or which require redaction of privileged content should be produced in searchable .PDF formats or as single page .TIFF images with unredacted OCR text furnished and logical unitization and family relationships preserved.	

Practice Tip: In settling upon a form of production for e-mail, use this inquiry as a litmus test to distinguish "native" forms from less functional forms: ***Can the form produced be imported into common e-mail client or server applications?*** If the form of the e-mail is so degraded that e-mail programs cannot recognize it as e-mail, that's a strong indication the form of production has strayed too far from functional.

Specify the Load File Format

Every electronic file has a complement of descriptive information called *system metadata* residing in the file table of the system or device storing the file. Different file types have different metadata. Every e-mail message has *"fields"* of information in the message "*header*" that support better searching, sorting and organization of messages. This may be data probative in its own right or simply advantageous to managing and authenticating electronic evidence. Either way, you want to be certain to request it sensibly and precisely. Simply demanding "the metadata" reveals you don't fully understand what you're seeking.

Develop a comprehensive production protocol tailored to the case and serve same with discovery. Always specifically request the metadata and header fields you seek, *e.g.:*

Produce delimited load file(s) supplying relevant system metadata field values for each information item by Bates number. Typical field values supplied include:

a. **Source file name** (original name of the item or file when collected from the source custodian or system);
b. **Source file path** (fully qualified file path from the root of the location from which the item was collected);
c. **Last modified date and time** (last modified date and time of the item);
d. **UTC Offset** (The UTC/GMT offset of the item's modified date and time*, e.g.,* -0500).
e. **Custodian or source** (unique identifier for the original custodian or source);
f. **Document type**;
g. **Production File Path** (file path to the item from the root of the production media);
h. **MD5 hash** (MD5 hash value of the item as produced);
i. **Redacted flag** (indication whether the content or metadata of the item has been altered after its collection from the source custodian or system);
j. **Embedded Content Flag** (indication that the item contains embedded or hidden comments, content or tracked changes); and
k. **Deduplicated instances** (by full path)**.**

The following additional fields shall accompany production of e-mail messages:

l. **To** (e-mail address(es) of intended recipient(s) of the message);
m. **From** (e-mail address of the person sending the message);
n. **CC** (e-mail address(es) of person(s) copied on the message);
o. **BCC** (e-mail address(es) of person(s)blind copied on the message);
p. **Subject** (subject line of the message);
q. **Date Received** (date the message was received);
r. **Time Received** (time the message was received);
s. **Attachments** (beginning Bates numbers of attachments);
t. **Mail Folder Path** (path of the message from the root of the mail folder);and
u. **Message ID** (unique message identifier).

Hybrid productions mixing imaged and native formats also require that paths to images and extracted text be furnished, as well as **logical unitization data** serving as the electronic equivalent of paper clips and staples.

De-duplication and Redaction

You may wish to specify whether the production should or should not be de-duplicated, *e.g.:*

> Documents should be vertically de-duplicated by custodian using each document's hash value. Near-deduplication should not be employed so as to suppress different versions of a document, notations, comments, tracked changes or application metadata.

Because redaction tends to impact just a small part of most productions, it's important that it not co-opt the forms of production.

> Information items that require redaction shall be produced in static image formats*, e.g.,* single page TIFF or multipage PDF images with logical unitization preserved. The unredacted content of each document should be extracted by optical character recognition (OCR) or other suitable method to a searchable text file produced with the corresponding page image(s) or embedded within the image file. Redactions should not be accomplished in a manner that serves to downgrade the ability to electronically search the unredacted portions of the item.

A TIFF-OCR redaction method works reasonably well for text documents, but often fails when applied to complex and dynamic documents like spreadsheets and databases. Unlike text, you can't spellcheck numbers, so the inevitable errors introduced by OCR make it impossible to have confidence in numeric content or reliably search the data. Moreover, converting a spreadsheet to a TIFF image strips away its essential functionality by jettisoning the underlying formulae that distinguishes a spreadsheet from a table.

Specify the medium of production

A well-crafted request should address the *medium* of ESI production; that is, the mechanism used to convey the electronic production to the requesting party. If you're receiving 100GB of data, you don't want it tendered on a dozen thumb drives.

> Production of ESI should be made using appropriate electronic media of the producing party's choosing that does not impose an undue burden or expense upon a recipient.

Conclusion

It's time to take a hard look at the language of the definitions and instructions accompanying requests for production. Most are boilerplate borrowed from someone who borrowed it from

someone who drafted it in 1947. It's hand-me-down verbiage long past retirement age; so, retire it and craft modern requests for a modern digital world.

We will never be less digital than we are today. Isn't it time we demand modern evidence and obtain it in the forms in which it serves us best? We must move forms of production upstream, from depleted images and load files to functional native and near native forms retaining the content and structure that supports migration into any form. *Utile* forms. *Complete* forms. *Forms that function.*

Exemplar Production Protocol

This Appendix is an example of a *production protocol*, sometimes called a *data delivery standard*. Geared to civil litigation and seeking the lowest-cost approach to production of ESI, it seeks native production of common file types and relieves parties of the burden of converting ESI to imaged formats except when needed for redaction. This exemplar protocol specifies near-native alternatives for production of native forms when near-native forms are preferable. For an example of a U.S. Government data delivery standard, see:
http://www.sec.gov/divisions/enforce/datadeliverystandards.pdf

Appendix: Exemplar Production Protocol

1. "Information items" as used here encompass individual documents and records (including associated metadata) whether on paper or film, as discrete "files" stored electronically, optically or magnetically or as a record within a database, archive or container file. The term should be read broadly to include e-mail, messaging, word processed documents, digital presentations, spreadsheets and database content.

2. Information that exists in electronic form should be produced in native formats and should not be converted to imaged formats. Native format requires production in the same format in which the information was customarily created, used and stored in the ordinary course.

3. If it is infeasible to produce an item of responsive ESI in its native form, it may be produced in an agreed-upon near-native form; that is, in a form in which the item can be imported into the native application without a material loss of content, structure or functionality as compared to the native form. Static image production formats serve as near-native alternatives only for information items that are natively static images (*i.e.*, photographs and scans of hard-copy documents).

4. The table below supplies examples of agreed-upon native or near-native forms in which specific types of ESI should be produced:

Source ESI	Native or Near-Native Form or Forms Sought
Microsoft Word documents	.DOC, .DOCX
Microsoft Excel Spreadsheets	.XLS, .XLSX
Microsoft PowerPoint Presentations	.PPT, .PPTX
Microsoft Access Databases	.MDB, .ACCDB
WordPerfect documents	.WPD
Adobe Acrobat Documents	.PDF
Photographs	.JPG, .PDF
E-mail	Messages should be produced in a form or forms that readily support import into standard e-mail client programs; that is, the form of production should adhere to the conventions set out in the

	internet e-mail standard, RFC 5322. For Microsoft Exchange or Outlook messaging, .PST format will suffice. Single message production formats like .MSG or .EML may be furnished with folder data. If your workflow requires that attachments be extracted and produced separately from transmitting messages, attachments should be produced in their native forms with parent/child relationships to the message and container(s) preserved and produced in a delimited text file.
Databases	Unless the entire contents of a database are responsive, extract responsive content to a fielded and electronically searchable format preserving metadata values, keys and field relationships. If doing so is infeasible, please identify the database and supply information concerning the schema and query language of the database along with a detailed description of its export capabilities so as to facilitate crafting a query to extract and export responsive data.
Documents that do not exist in native electronic formats or which require redaction of privileged content should be produced in searchable .PDF formats or as single page .TIFF images with OCR text of unredacted content furnished and logical unitization and family relationships preserved.	

5. Absent a showing of need, a party shall produce responsive information reports contained in databases through the use of standard reports; that is, reports that can be generated in the ordinary course of business and without specialized programming efforts beyond those necessary to generate standard reports. All such reports shall be produced in a delimited electronic format preserving field and record structures and names. The parties will meet and confer regarding programmatic database productions as necessary.

6. Information items that are paper documents or that require redaction shall be produced in static image formats scanned at 300 dpi (e.g., single-page Group IV.TIFF or multipage PDF images). If an information item employs color to convey information (versus purely decorative use), the producing party shall not produce the item in a form that does not display color. The full content of each document will be extracted directly from the native source where feasible or, where infeasible, by optical character recognition (OCR) or other suitable method to a searchable text file produced with the corresponding page image(s) or embedded within the image file. Redactions shall be logged along with other information items withheld on claims of privilege.

7. Parties shall take reasonable steps to ensure that text extraction methods produce usable, accurate and complete searchable text.

8. Individual information items requiring redaction shall (as feasible) be redacted natively, produced in .PDF format and redacted using the Adobe Acrobat redaction feature or redacted and produced in another reasonable manner that does not serve to downgrade the ability to electronically search the unredacted portions of the item. Bates identifiers should be endorsed on the lower right corner of all images of redacted items so as not to obscure content.

9. Upon a showing of need, a producing party shall make a reasonable effort to locate and produce the native counterpart(s) of any .PDF or .TIF document produced. The parties agree to meet and confer regarding production of any such documents. This provision shall not serve to require a producing party to reveal redacted content.

10. Except as set out in this Protocol, a party need not produce identical information items in more than one form. The content, metadata and utility of an information item shall all be considered in determining whether information items are identical, and items reflecting different information shall not be deemed identical.

11. Production of ESI should be made using appropriate electronic media of the producing party's choosing that does not impose an undue burden or expense upon a recipient. Label all media with the case number, production date, Bates range and disk number (1 of X, if applicable). Organize productions by custodian, unless otherwise instructed. All productions should be encrypted for transmission to the receiving party. The producing party shall, contemporaneously with production, separately supply decryption credentials and passwords to the receiving party for all items produced in an encrypted or password-protected form.

12. Each information item produced shall be identified by naming the item to correspond to a Bates identifier according to the following protocol:

    i. The first four (4) characters of the filename will reflect a unique alphanumeric designation identifying the party making production;

    ii. The next six (6) characters will be a designation reserved to the discretionary use of the party making production for the purpose of, e.g., denoting the case or matter. This value shall be padded with leading zeroes as needed to preserve its length;

    iii. The next nine (9) characters will be a unique, consecutive numeric value assigned to the item by the producing party. This value shall be padded with leading zeroes as needed to preserve its length;

    iv. The final six (6) characters are reserved to a sequence consistently beginning with a dash (-) or underscore (_) followed by a five digit number reflecting pagination of the

item when printed to paper or converted to an image format for use in proceedings or when attached as exhibits to pleadings.

v. By way of example, a Microsoft Word document produced by Acme in its native format might be named: ACMESAMPLE000000123.docx. Were the document printed out for use in deposition, page six of the printed item must be embossed with the unique identifier ACMESAMPLE000000123_00006. Bates identifiers should be endorsed on the lower right corner of all printed pages, but not so as to obscure content.

vi. This format of the Bates identifier must remain consistent across all productions. The number of digits in the numeric portion and characters in the alphanumeric portion of the identifier should not change in subsequent productions, nor should spaces, hyphens, or other separators be added or deleted except as set out above.

13. Information items designated Confidential may, at the Producing Party's option:

a. Be separately produced on electronic production media prominently labeled to comply with the requirements of the **[DATE]** Protective Order entered in this matter; or, alternatively,

b. Each such designated information item shall have appended to the file's name (immediately following its Bates identifier) the following protective legend:
~CONFIDENTIAL-SUBJ_TO_PROTECTIVE_ORDER

When any item so designated is converted to a printed or imaged format for use in any submission or proceeding, the printout or page image shall bear the protective legend on each page in a clear and conspicuous manner, but not so as to obscure content.

14. Producing party shall furnish a delimited load file supplying the metadata field values listed below for each information item produced (to the extent the values exist and as applicable):

Field Name	Sample Data	Description
BegBates	ACMESAMPLE000000001	First Bates identifier of item
EndBates	ACMESAMPLE000000123	Last Bates identifier of item
AttRange	ACMESAMPLE000000124 - ACMESAMPLE000000130	Bates identifier of the first page of the parent document to the Bates identifier of the last page of the last attachment "child" document
BegAttach	ACMESAMPLE000000124	First Bates identifier of attachment range
EndAttach	ACMESAMPLE000000130	Last Bates identifier of attachment range
Parent_Bates	ACMESAMPLE000000001	First Bates identifier of parent document/e-mail message. ***This Parent_Bates field should be populated in each record representing an attachment "child" document. ***
Child_Bates	ACMESAMPLE000000004; ACMESAMPLE000000012; ACMESAMPLE000000027	First Bates identifier of "child" attachment(s); may be more than one Bates number listed depending on number of attachments.

		***The Child_Bates field should be populated in each record representing a "parent" document. ***
Custodian	Houston, Sam	E-mail: mailbox where the email resided. Native: Individual from whom the document originated
Path	E-mail: \Deleted Items\Battles\ SanJac.msg Native: Z:\TravisWB\Alamo.docx	E-mail: Original location of e-mail including original file name. Native: Path where native file document was stored including original file name.
From	E-Mail: Davy@Crockett.net Native: D. Crockett	E-mail: Sender Native: Author(s) of document **semi-colons separate multiple entries **
To	Genl. A.L. de Santa Anna [mailto: sa@sa.mx]	Recipient(s) **semi-colons separate multiple entries **
CC	Jim.Bowie@bigknife.com	Carbon copy recipient(s) **semi-colons separate multiple entries **
BCC	AustinSF@state.tx.gov	Blind carbon copy recipient(s) **semi-colons separate multiple entries **
Date Sent	03/18/2024	E-mail: Date the email was sent
Time Sent	11:45 AM	E-mail: Time the message was sent
Subject/Title	Remember the Alamo!	E-mail: Subject line of the message
IntMsgID	<A1315BC17ABD4774BF779CB3 E3E62B9B@gmail.com>	E-mail: For e-mail in Microsoft Outlook/Exchange, the "Unique Message ID" field; For e-mail in Lotus Notes, the UNID field. Native: empty.
Date_Mod	02/23/2024	E-mail: empty. Native: Last Modified Date
Time_Mod	01:42 PM	E-mail: empty Native: Last Modified Time
File_Type	XLSX	E-mail: empty Native: file type
Redacted	Y	Denotes that item has been redacted as containing privileged content (yes/no).
File_Size	1,836	Size of native file document/email in KB.
HiddenCnt	N	Denotes presence of hidden Content/Embedded Objects in item(s) (Y/N)
Confidential	Y	Denotes that item has been designated as confidential pursuant to protective order (Y/N).
MD5_Hash	eb71a966dcdddb929c1055ff2f1 ccd5b	MD5 Hash value of the item.
DeDuped	E-mail: \Inbox\SanJac.msg Native: Z:\CrockettD\Alamo.docx	Full path of deduped instances. **semi-colons separate multiple entries **

15. Each production should include a cross-reference load file that correlates the various files, images, metadata field values and searchable text produced.

16. Parties shall respond to each request for production by listing the Bates identifiers/ranges of responsive documents produced, and where an information item responsive to these discovery requests has been withheld or redacted on a claim that it is privileged, the producing party shall furnish a privilege log.



Exercise 22: Forms of Production and Cost

GOALS: The goals of this exercise are for the student to:

1. Generate a production set in native and TIFF+ image formats with extracted text; and

2. Assess the impact of alternate forms of production (TIFF versus Native) on the cost of ingestion and hosting over the life of a case and on the utility of the evidence.

OUTLINE: Students will use DISCO and the Evidence Drive database created in Exercise 20 to generate a production, compare the sizes of each form of production and look at examples of diminished utility and intelligibility.

Producing parties frequently seek to convert native file formats used by and collected from custodians into static image formats like PDF, or more commonly, TIFF images plus load files holding extracted text or text generated through use of optical character recognition. The latter static production sets are called **TIFF+** productions. Proponents of static image productions assert claims of superior document security and point to the ability to emboss page numbers and other identifiers on page images. Too, page images can be viewed using any browser application, affording users ready accessibility to some content, albeit sacrificing other content and utility.

Often overlooked in the debate over forms of production is the impact on ingestion, processing, storage and export costs engendered by use of static image formats. Most e-discovery service providers charge to ingest, process, host (store) and export electronically stored information on a per-gigabyte basis. As a result, when items produced occupy more space (measured in bytes), they cost the recipient more to use. This exercise invites students to consider what, if any, increase in cost may flow from the production of static imaged formats as forms of production versus native forms.

The Myth of Page Equivalency

It's comforting to quantify electronically stored information as some number of pieces of paper or bankers' boxes. Paper and lawyers are old friends. But you can't reliably equate a volume of data with a corresponding page count unless you know the composition of the data. Even then, it's a leap of faith.

If you search the Internet for page equivalency claims, you'll be astounded by how widely they vary, though each is offered with utter certitude. A gigabyte of data is variously equated to an absurd 500 million typewritten pages, a naively accepted 500,000 pages, the popularly cited 75,000 pages and a laggardly 15,000 pages. The other striking aspect of page equivalency claims is that they're blithely

accepted by lawyers and judges who wouldn't concede the sky is blue without a supporting string citation.

In testimony before the committee drafting the federal e-discovery rules, Exxon Mobil representatives twice asserted that one gigabyte yields 500,000 typewritten pages. The National Conference of Commissioners on Uniform State Laws included that value in its "Uniform Rules Relating to Discovery of Electronically Stored Information." The Conference of Chief Justices cites the same equivalency in its "Guidelines for State Trial Courts Regarding Discovery of Electronically-Stored Information." Scholarly articles and reported decisions pass around the 500,000 pages per gigabyte value like a bad cold. Yet, 500,000 pages per gigabyte isn't right. It's not even particularly close to right.

Many years ago, e-discovery guru Kenneth Withers wrote a section of the fourth edition of "The Manual on Complex Litigation" that equated a terabyte of data to 500 billion typewritten pages. It was supposed to say million, not billion. Eventually, the typo was noticed and corrected; but the echoes of that innocent thousand-fold mistake still reverberate. Anointed by the prestige of the manual, the 500-billion-page equivalency was embraced as gospel. Even when the value was "corrected" to 500 million pages per terabyte—equal to 500,000 pages per gigabyte—we're still talking about equivalency with all the credibility of an Elvis sighting.

So, how many pages are there in a gigabyte? It's the answer lawyers love: "*It depends.*"

Page equivalency is a myth. *One must always look at individual file types and quantities to gauge page equivalency*, and there is no reliable rule of thumb geared to how many files of each type a typical user stores. It varies by industry, by user and even by the life span of the media and the evolution of applications. A reliable page equivalency must be expressed with reference to both the quantity and form of the data, *e.g.,* "*a gigabyte of single page TIFF images of 8-1/2-inch x 11- inch documents scanned at 300 dots per inch equals approximately 18,000 pages*."

The TIFF+ Markup

When I tell lawyers that converting native Microsoft Office documents to TIFF images for production inflates the size of the collection by big multiples (*five* times larger, *ten* times larger), they don't believe me--even lawyers experienced in e-discovery; *especially* lawyers experienced in e-discovery. "That can't be right," they insist. "***You do the math***," I challenge them.

Producing parties can't seem to let go of static images as their preferred form of production. They have trouble accepting that alternate approaches to Bates numbering work seamlessly despite identifying data at the file level rather than in a paginated way. Too, they mistakenly believe that imaging ESI makes it more secure, despite all evidence to the contrary. Producing parties want to

do things as they've always done them, and if it happens to make matters harder and costlier for their opponents, well, bring on the crocodile tears!

But Rule 1 of the Federal Rules of Civil Procedure requires the court and parties to construe, administer and employ the Rules to "*secure the just, speedy, and inexpensive determination of every action and proceeding.*" So, when a producing party refuses a demand for native forms of production and seeks to substitute TIFF images plus load files ("TIFF+"), how much additional expense is too much? *At what point does a form of production become so costly and burdensome that it cannot fairly be proportionate or reasonably usable?*

How much is too much? Let's do the math.

First, a few facts to establish common ground.

FACT 1: Requesting Parties May Specify a Form or Forms of Production

When you pursue discovery, you may call what you seek "documents" and mentally equate them to paper records, but it's electronically stored information (ESI). ESI must be produced in specified *forms of production,* either in *native forms* (being the form stored and used in the ordinary course of business) or in a *static image* format (a black and white screenshot of each page called a ***Tagged Image File Format*** or *TIFF image* plus a *load file* or files holding text and metadata). There are also *near-native* forms of production, such as when mailboxes are produced as individual messages called MSGs or EMLs.

The Federal Rules of Civil Procedure and most states' rules empower a requesting party to specify the form or forms in which electronically stored information is to be produced. *FRCP Rule 34(b)(1)(C).* If a requesting party *fails* to specify a form of production (and you should NEVER fail to specify), a producing party must supply ESI in the form or forms in which it is ordinarily maintained or in a reasonably usable form or forms. *FRCP Rule 34(b)(2)(E)(ii)*

Fact 2: Most E-Discovery Service Providers "In the Cloud" Charge by Data Volume

Whether in native or static image format, ESI must be *processed* ("ingested") and *hosted* to be searchable and reviewable. Native forms are processed to extract their text and metadata, then indexed for search. TIFF and load file productions are indexed for search and processed to pair the page images with text and metadata. Either way, you pay a vendor to prepare the production for viewing and then pay a recurring "hosting" charge for online access to the production. **The fees charged are based on the volume of data processed and/or hosted, month after month.** More

data costs more money. If you receive 10 times as much data, you pay a commensurate amount more to ingest and host. Vendors usually assess hosting fees as a monthly subscription, so the more data they host for you, the more you pay every month for the life of the case. It's no accident that vendor charges are opaque and vary wildly; but, at the bottom line, the rule is more data, more dollars.

Fact 3: TIFF images of native files are much larger than the native files.

More data isn't the same thing as more information because not all electronic forms of information are equally efficient. When you convert native forms to static images and load files you explode the size of production by many multiples, and static productions come burdened by the further cost of impaired searchability, diminished functionality and lost color, animation and rich media. ***With TIFF, you get less and pay more.*** Not 50% or 100% more; but multiples more and beyond. This is notably the case for Word documents, PowerPoint presentations, Excel spreadsheets and collections of e-mail messages and attachments–the native forms at the heart of electronic discovery. The difference is genuine, material and carries a big bottom-line cost.

That's a categorical statement, and some will immediately search for an exception. They will wonder, *is it possible to fashion a native file larger than its TIFF counterpart?* You could certainly construct a PowerPoint or Word document so laden with high-resolution color photos, sound and video that, once you strip away the rich content, a static black & white image would occupy a size smaller than the native. But is a TIFF shorn of sound, video and color truly *comparable?* Is such a diminished file representative of most collections produced in e-discovery? An emphatic "no," on both counts, and it's not an "apples-to-apples" comparison.

A production must be reasonably usable. The TIFF without sound and video isn't. When you add back the rich media and produce with extracted sound and video files, the TIFF production is indeed larger than the native, and more unwieldy.

You Do the Math

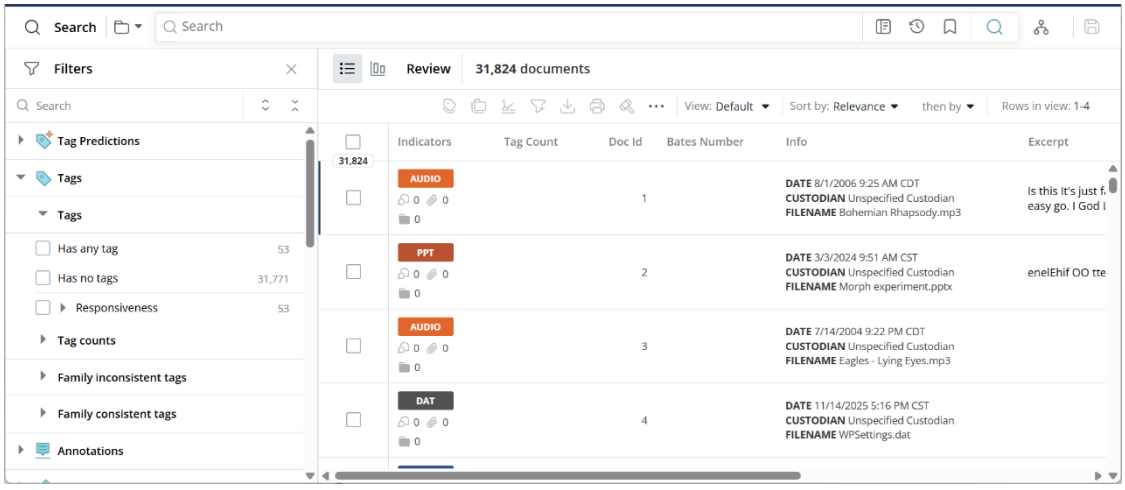
Using the DISCO online toolset, we can generate alternate production sets for the same evidence to determine the volumetric impact attendant to native versus imaged forms of production. In each instance, we want the production sets to fairly mirror industry standards; so, we will generate a set of native files and a TIFF+ set of the same evidence. The most-commonly seen specification for TIFF image production calls for **single-page monochrome Group IV images at 300 dpi resolution**. Breaking that down, it means that, unlike a PDF where the entirety of a document typically occupies one file, a single-page TIFF specification demands that each page of every

document be rendered as a single file. "Monochrome" specifies that the image be devoid of color, *i.e.,* rendered in black and white (which reduces the byte size of the image but sacrifices appearance and intelligibility when color is used to convey content, like color coding and highlighting).[108] "Group IV" is a bitmap image compression specification, and 300dpi (for "300 dots per inch") is a measure of print resolution. The higher the number of dots per inch, the crisper and more detailed the image. TIFF+ production specifications typically deal with the issue of color by requiring that documents with color be produced as costlier JPEG renderings; but you can imagine the cost and complexity of doing so when, for example, the only color content in an email is a corporate logo in a signature block.

Exercise 22: Calculate the Cost Difference Flowing from Alternate Forms of Production

Step 1: Identify and Tag Items for Production

a. Using your browser, go to https://login.csdisco.com and log in using the credentials supplied to you.
b. Select the Matter named "**UT Processing Exercise 2026**" and the database in **your name**. This is the database you created and used for Exercises 20 and 21. DO NOT use the Podesta Email database used for Exercise 18 (Search) because it's much larger and you only have Reviewer access for that database. Only in the matter named "**UT Processing Exercise 2026**" do you have the Admin rights required to generate productions in your own database. You should see the DISCO eDiscovery screen. It should look like the image below, with about 31,824 documents listed:

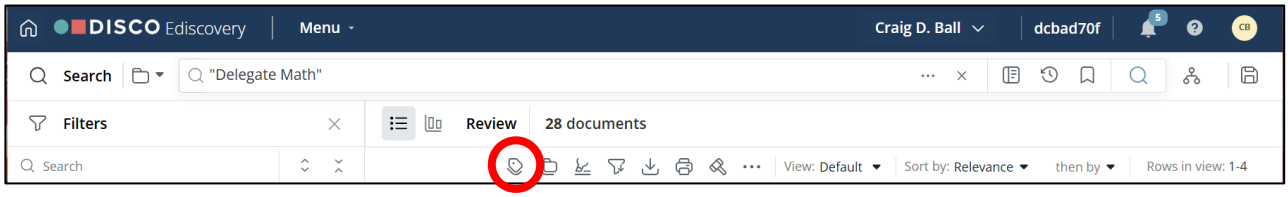


[108] A monochrome image requires only one bit per image element (dot or pixel) because that one or zero is sufficient to indicate black or white. When you introduce color, each bit requires color information, enough "bit depth" to store each of the available colors. So, eight bits can encode up to 256 different colors and 24 bits can encode over 16 million colors (256 x 256 x 256 colors). Accordingly, adding bit depth for color serves to significantly increase file size because color information must be stored for every image element.

c. Run a search for **"Delegate Math"**
Be sure to include the quotes around "Delegate Math".

You should retrieve 28 documents (comprising 19 emails, 6 Word documents, two PowerPoint presentations and one Adobe PDF document). If you're not getting these 28 items, check your search and settings carefully.

c. Click the **Bulk Tag** button (circled below):

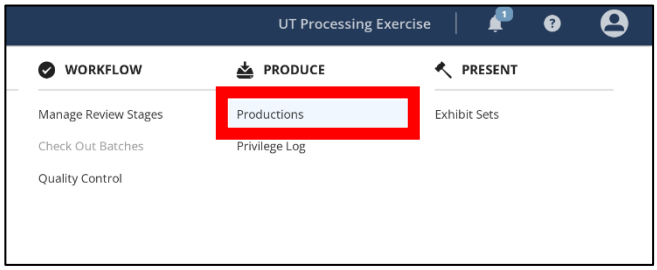


d. Click in the Apply box and under "**Importance**," select "**Hot**" then click "**Update**."

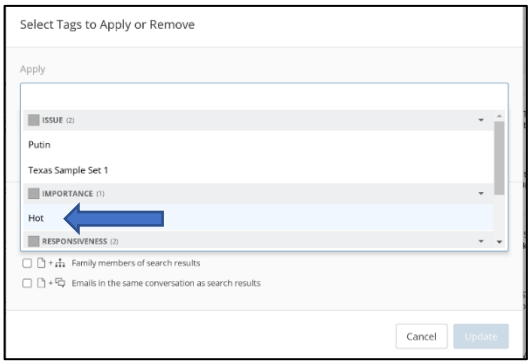
All 28 items should now be tagged as "Hot" documents.

Step 2: Generate a TIFF+ Production Set

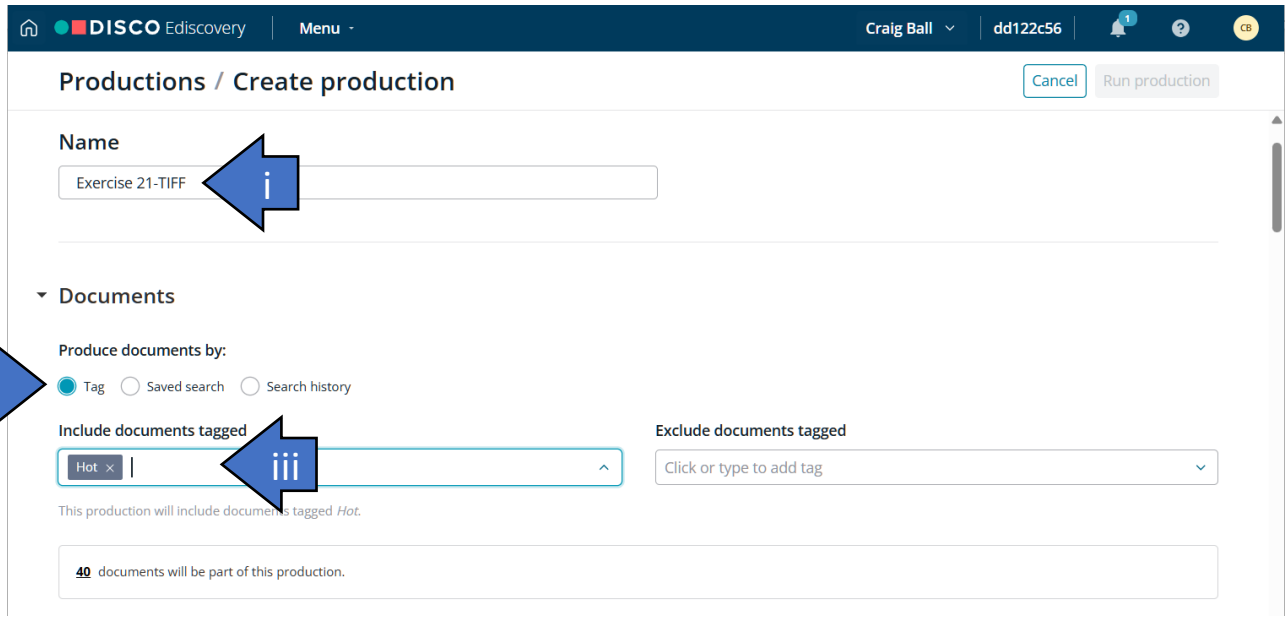
a. Click "**Menu**" (top of DISCO screen), then select "**Productions**." At the Productions screen, click "**Create New**" at upper right of screen.



b. Configure your production using the following settings:
   i) Name your production set "**Exercise 22-TIFF**"
   ii) Produce documents by "Tag"



iii) Include documents tagged “Hot.” **Important**: remove any other tag listed except “Hot.”



iv) Set the Bates prefix to “**Exercise 22-TIFF_**”

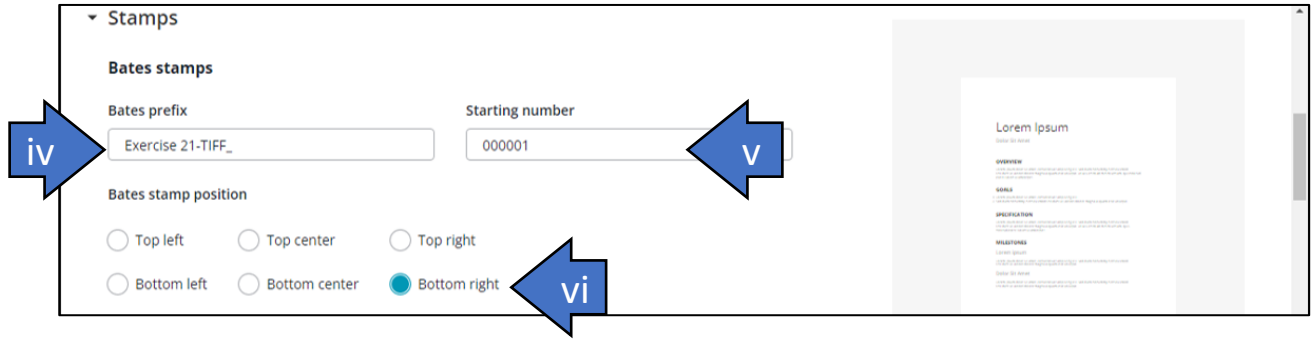
v) Set the starting number to **000001**

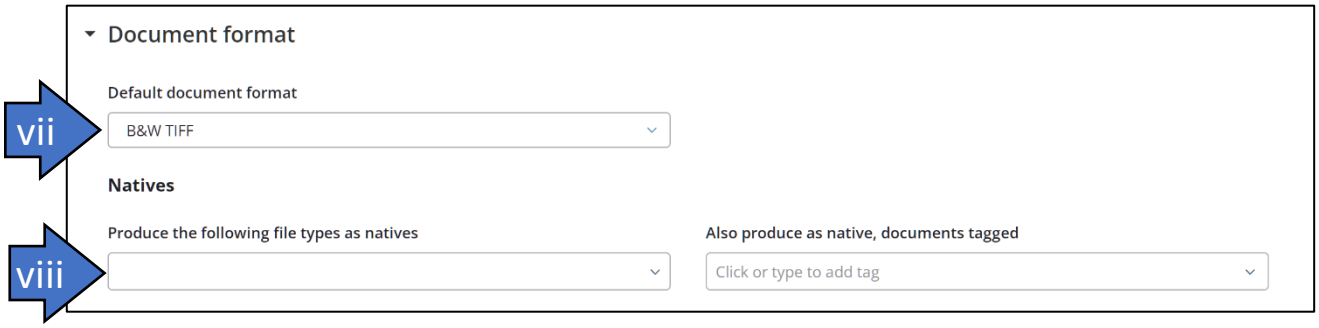
vi) Position the Bates stamp at Bottom right.

vii) Select **B&W TIFF** as the Default document format using the dropdown menu

viii) Clear <u>all</u> settings from the Natives as seen below. NOTHING in the Natives boxes!

ix) Leave all other settings at defaults. Click **Run Production** (button top/bottom of page)



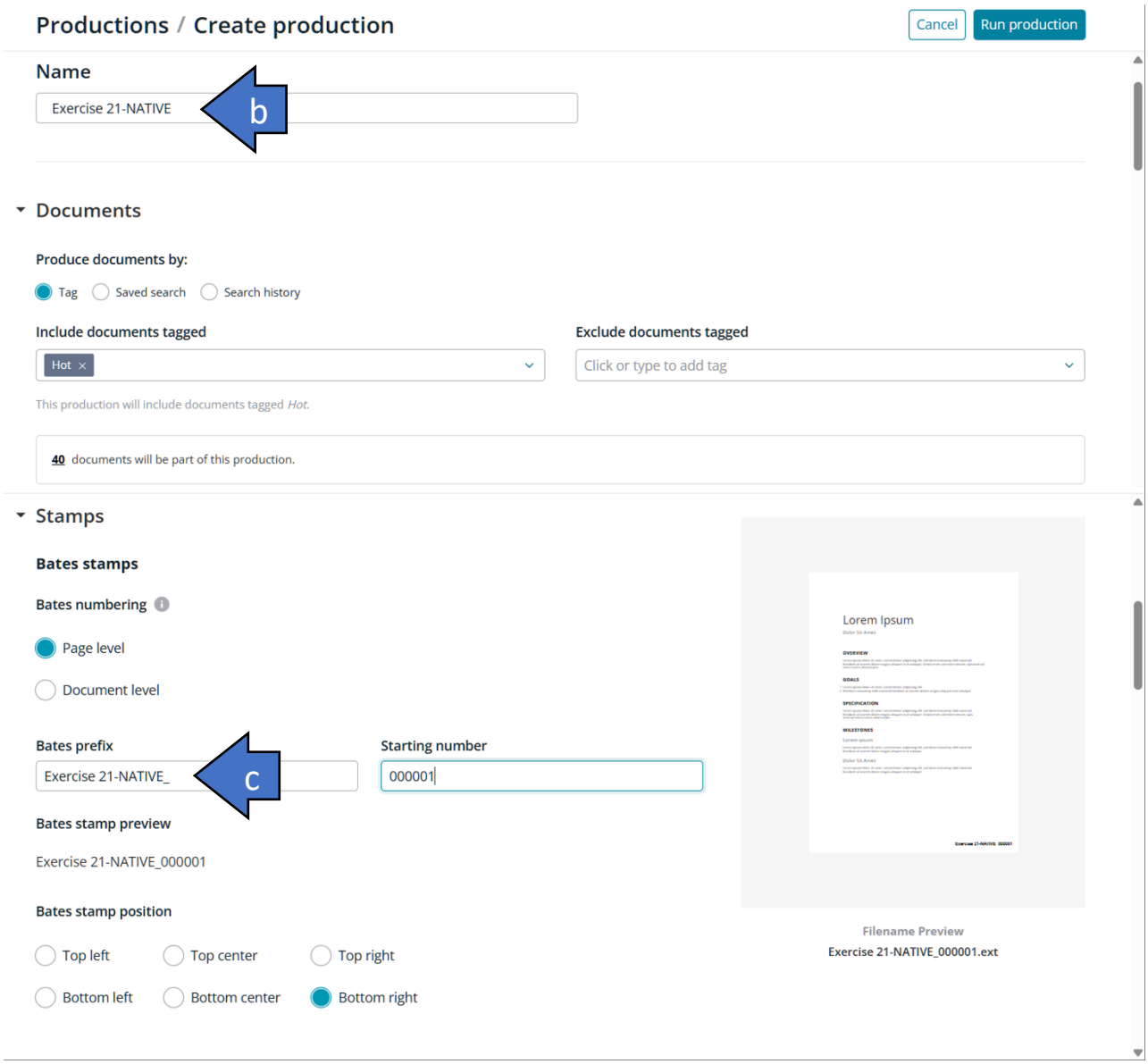


Step 3: Download and Unzip the TIFF Production Set

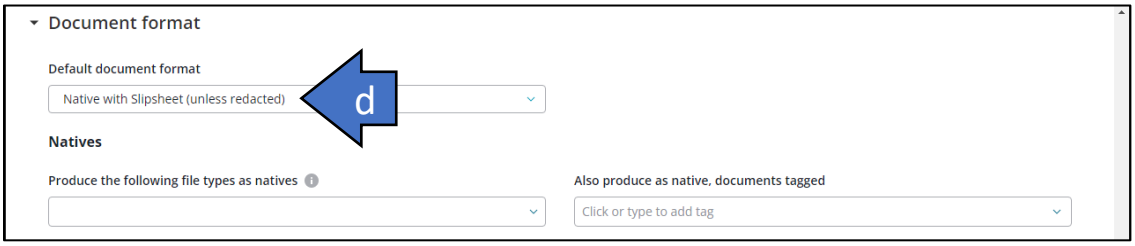
a) It should take 10-15 minutes for your production set to finish. When it completes, locate your production named "Exercise 22-TIFF" and click download. Be sure to select "Production" for download and DO NOT select "Load Files Only" as you will need to examine parts of the production later in this Exercise. Download the Zip file to your Desktop or another convenient location. The download may take a few minutes to complete depending upon your connection speed.
b) On your Desktop, create a folder named "**TIFF Production**" and unzip (decompress) the contents of the Zip file just downloaded into this folder. When successfully completed, the "TIFF Production" folder will contain a single subfolder named **VOL0001**.

Step 4: Generate a Native Production Set

a) Go back to Step 2(a). Follow the same process as just completed with **three differences**:
b) For Step 2(b)(i), name your production set "**Exercise 22-NATIVE**."
c) For Step 2(b)(iv), set the Bates prefix to "**Exercise 22-NATIVE_**"



d) For Step 2(b)(vii), select "**Native with Slipsheet (unless redacted)**" and no other form.



e) Leave all other settings at defaults. Click **Run Production** (button at bottom of page)

Step 5: Download and Unzip the NATIVE Production Set

a) It should take only a couple of minutes for your native production set to finish. When it completes, locate the production named "Exercise 22-NATIVE" and click download. Once more, select "Production;" DO NOT select "Load Files Only" as you will need to examine parts of the production later in this Exercise. Download the Zip file to your Desktop or another convenient location.
b) On your Desktop, create a folder named "**NATIVE Production**" and unzip (decompress) the contents of the Zip file just downloaded into this folder. When successfully completed, the folder will contain a single folder named **VOL0001**.

Step 6: Compare the Two Forms of Production Volumetrically and Functionally

If you've done all steps correctly, you have two folders on your Desktop: **TIFF Production** and **Native Production**. Each holds a production set, including extracted text and load files, for the same 28 evidence items responsive to our search query.

b) Determine the size of the contents of each folder and record them below:

i) TIFF Production Folder Size _____

ii) Native Production Folder Size: _____

iii) How many times larger than the other is the larger of the two folders? _____

c) In the Native Production folder, find the second, larger (3MB) PowerPoint presentation called "JDP Path to Victory 3.7.16 V4.pptx." It should be the PowerPoint file in **Native Production\VOL0001\NATIVES\0001 with its first slide called "Path to Victory."** Make a working copy of this PowerPoint file and **open the working copy only** in PowerPoint.[109] Examine the ninth (last) slide in the presentation. Note the ease with which you can identify the states color coded to correspond to the dates of primary contests.
d) Now, locate the corresponding TIFF image for slide 9 of the same PowerPoint presentation in the TIFF Production set.[110] Compare it to the native version. **In completing this Exercise,**

[109] Good practice dictates that we never open original evidence produced natively in its associated application to minimize any possibility of unwittingly altering the evidence or impacting hash values.
[110] In my production set, slide 9 was the image named "Exercise 22-TIFF_0001310.tif," but it may carry a different number in your set. It's challenging to locate the corresponding TIFF image, I used the file called "VOL0001.xlsx" in the Tiff Production set to identify the corresponding Bates number range of the TIFF images for the PowerPoint file called "JDP Path to Victory 3.7.16 V4.pptx." You may find it expedient to scan the TIFF IMAGES to identify the Powerpoint slide imagery. I found it that way in group 0003.

you are to submit a copy of the TIFF image *only* for slide 9 of the PowerPoint along with your answers to the questions above (*i.e.,* Step 6(a)(i-iii).

There are many variables that go into computing the cost of vendor services for e-discovery; the charges for ingestion, processing, hosting and export are just parts of a costly, complicated puzzle. The purpose of this exercise is to gauge the difference that forms of production can make as a consequential contributing component of overall cost and utility.

Cost is cause enough to demand production in native forms, but when an opponent produces in native formats, you're receiving what the other side used in the ordinary course of business. *It's the real evidence.* It's a form witnesses recognize. It's complete and utile. Crucially, you can convert native forms to other forms–including static image formats–for those times you may want alternative formats.

But it doesn't work both ways. You can't convert TIFF images back to native originals. Not really. You can't slim bloated static images down to svelte native forms. You can't restore animations, color, formulas, tracked changes and comments, application metadata or hash values.

With TIFF productions, you're stuck. You must pay vendors to ingest and host at grossly inflated data volumes. You have no choice. It's like buying a car and the dealer delivers it encased in a block of concrete. You're not going anywhere.

A Hidden Cost: Impaired Search

In the "ordinary course of business," none but litigators "ordinarily maintain" TIFF images as substitutes for native evidence When requesting parties seek production in native forms, responding parties counter with costly static image formats by claiming they are "reasonably usable" alternatives. However, the drafters of the 2006 Rules amendments were explicit in their prohibition:

> *[T]he option to produce in a reasonably usable form does not mean that a responding party is free to convert electronically stored information from the form in which it is ordinarily maintained to a different form that makes it more difficult or burdensome for the requesting party to use the information efficiently in the litigation. If the responding party ordinarily maintains the information it is producing in a way that makes it searchable by electronic means, the information should not be produced in a form that removes or significantly degrades this feature.*

FRCP Rule 34, Committee Notes on Rules – 2006 Amendment.

I contend that substituting a form that costs many times more to load and host counts as making the production more difficult and burdensome to use. But what rarely acknowledged is that so-called TIFF+ productions wreak havoc on searchability, too. It boggles the mind, but when I share what you've just proven with opposing counsel, they immediately retort, "That's not true." They deny the reality without checking its truth, without *caring* whether what they assert has a basis in fact. And I'm talking about lawyers claiming deep expertise in e-discovery. It's disheartening, to say the least.

A little background: We all know that ESI is inherently electronically searchable. There are quibbles to that statement, but please take it at face value for now. When parties convert evidence in native forms to static image forms like TIFF, the process strips away all electronic searchability. A monochrome screenshot replaces the source evidence. Since the Rules say you can't remove or significantly degrade searchability, the responding party must act to restore a measure of searchability. They do this by extracting text from the native ESI and delivering it in a "load file" accompanying the page images. This is part of the "plus" when people speak of TIFF+ productions.

E-discovery vendors then seek to pair the page images with the extracted text in a manner that allows some text searchability. Vendors index the extracted text to speed search, a mapping process intended to display the page where the text was located when mapped. This is important because *where* the text appears in the load file dictates what page will be displayed when the text is searched and determines whether features like proximity search and even predictive coding work as well as we have a right to expect. ***Upshot: The location and juxtaposition of extracted text in the load file matters significantly for accurate searchability.***

Now, let's consider the structure of modern electronic evidence. We could talk about formulae in spreadsheets or speaker notes in presentations, but those are not what we fight over when it comes to forms of production. Instead, I want to focus on Microsoft Word documents and those components of Word documents called Comments and Tracked Changes; particularly Comments because these aren't "metadata" by any stretch. Comments are *user-contributed content*, typically communications between collaborators. Users see this content on demand and it's highly contextual and positional because it is nearly always a comment *on adjacent body text*. It's NOT the body text, and it's not much use when it's separated from the body text. Accordingly, Word displays comments as marginalia, giving them the power of place but not enmeshing them with the body text.

But what happens to these contextual comments when you extract the text of a Word document to a load file and then index the load files?

There are three ways I've seen vendors handle comments, and all three significantly degrade searchability:

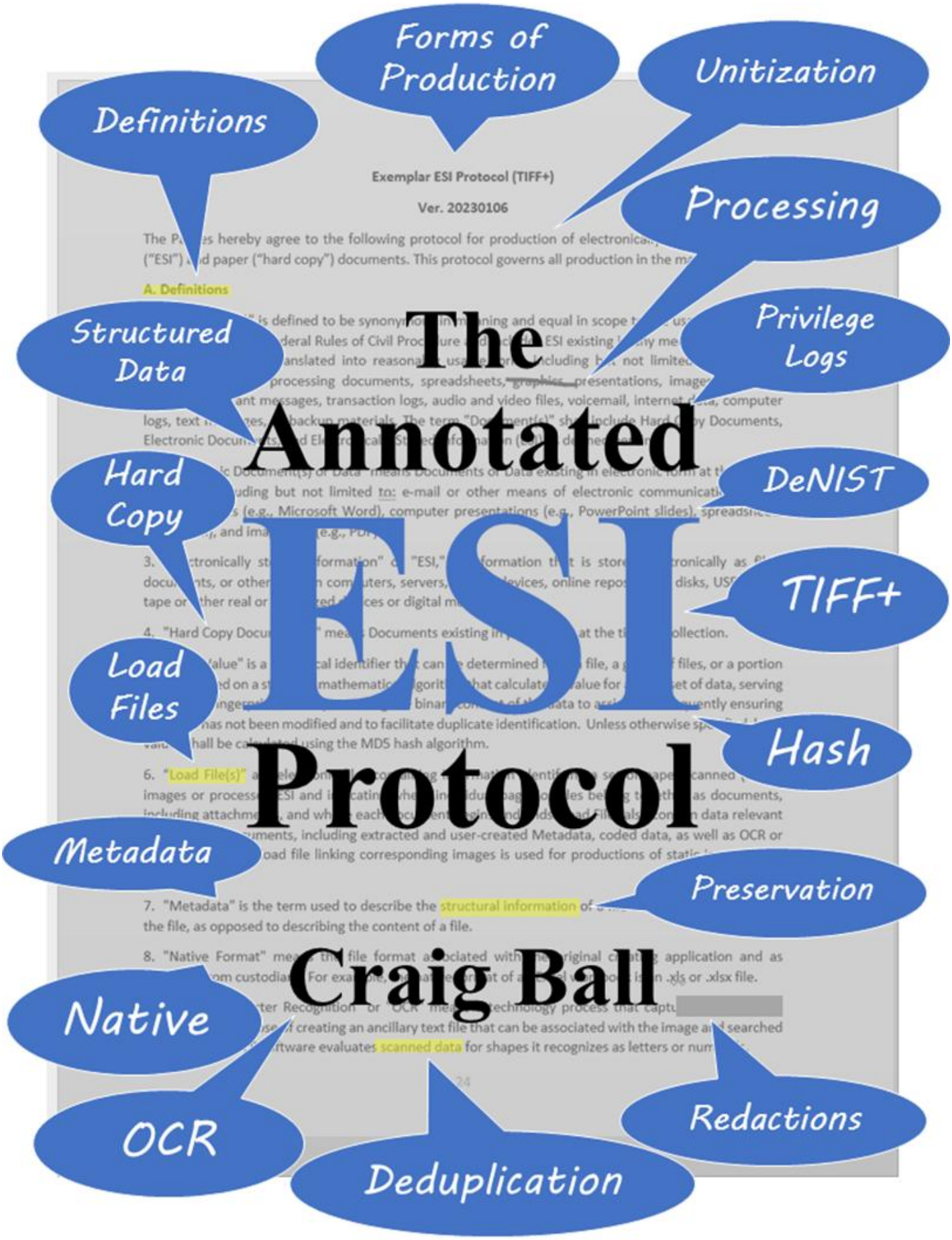
First, they suppress comments altogether and do not capture the text in the load files. This is content deletion. It's as if the content was never there, and you can't find the text using any method of electronic search. Responding parties don't disclose this deletion, nor is it grounded on any claim of privilege or right. Spoliation is just S.O.P.

Second, they merge the comments into the adjacent body text. This has the advantage of putting the text more-or-less on the same page where it appears in the source, but it also frustrates proximity search and analytics. The injection of comment text between a word combination or phrase causes searches for that combination or phrase to fail. For example, if your search is for *ignition w/3 switch* and a four-word comment comes between "ignition" and "switch," the search fails.

Third, and frequently, vendors aggregate comments and dump them at the end of the load file with no clue as to the page or text they reference. No links. No pointers. Every search hitting on comment text takes you to the wrong page, devoid of context.

Some of what I describe reflects challenges inherent to dealing with three-dimensional data using two-dimensional tools. Native applications deal with Comments, speaker notes and formulae three-dimensionally. We can reveal that data as needed, and it appears in exactly the way witnesses use it outside of litigation. But flattening native forms to static images and load files destroys that multidimensional capability. Vendors do what they can to add back functionality; but we should not pretend the results are anything more than a pale shadow of what's possible when native forms are produced. I'd call it a tradeoff, but that implies requesting parties know what's being denied them. How can requesting party's counsel know what's happening when responding parties' counsel don't have a clue what their tools do, yet misrepresent the result?

But now *you* know. Check it out. *Look* at the extracted text files produced to accompany documents. These text files are the sole means by which a meager measure of searchability is restored for TIFF images in production, and the caliber of the content is often shockingly poor.


Forms of Production
Unitization
Definitions
Exemplar ESI Protocol (TIFF+)
Ver. 20230106
Processing
A. Definitions
Structured Data
The
Privilege Logs
Annotated
Hard Copy
DeNIST
ESI
TIFF+
Load Files
Protocol
Hash
Metadata
Preservation
Craig Ball
Native
OCR
Redactions
Deduplication

The Annotated E-Discovery Protocol: A Primer on ESI Protocols
Craig Ball ©2023

An ESI or E-Discovery Protocol is an agreement or order that answers common questions encountered when dealing with electronically stored information (ESI) in discovery, questions like:

- What forms of production should be employed?
- What metadata must be collected and produced?
- How are document "family relationships" and "unitization" handled?
- How do parties protect privileged data from and rectify inadvertent disclosure?
- What processes may producing parties use to suppress duplicates?
- How must items produced be named and labeled?
- How is information on paper integrated with ESI production?
- How is information conveyed via color to be presented?
- How are productions efficiently transmitted and protected in transit?
- What must be made searchable by optical character recognition (OCR)?
- What must be done to resolve evidence processing exceptions and errors?
- Who serves as liaison counsel when discovery questions and disputes arise?

Ambitious ESI protocols encompass more nuanced and nettlesome issues like:

- The execution and scope of preservation duties
- Search queries and strategies
- Issues attendant to discovery from databases and other structured data sources
- Use and validation of advanced analytics
- Issues involving documents and data in foreign languages
- Confidentiality designations/legends and handling of confidential data
- The use and timing of rolling productions
- Alternative approaches to logging items withheld as privileged
- Mechanisms and timetables for dispute resolution

While it's prudent and competent to deploy an ESI protocol, anticipating consensus across too-broad a range of issues is unrealistic. *Routine ESI protocols should focus on matters of technical consistency and expediency*; that is, they should address the geeky details that ensure that what the parties exchange in discovery will be complete and utile. Yet, some parties stonewall and nitpick the most basic points of an ESI protocol in recognition that many judges—like most lawyers—are discomfited by technical disputes and retreat to solutions suited to simpler times and simpler, paper-centric discovery.

The fault for that failure lies less with Luddite judges than with advocates who can't distinguish the essential features of an ESI protocol from the merely desirable ones or articulate the "why" of either. Certainly, it's human nature to fear what we don't understand, so acceding to a different way of doing something feels risky when you don't grasp the rationale. This paper seeks to lay out the core provisions of ESI protocols, explaining their purpose and highlighting the impact of alternatives. I'll use the Federal Rules of Civil Procedure as a frame of reference, recognizing that

few state courts have procedural rules entirely identical to the Federal Rules (*e.g.*, not all states have a rule mirroring the FRCP's Rule 26(f) 'meet and confer' duty).[119]

A "clean" version of the exemplar protocol follows as an appendix. The example defaults to clunky TIFF+ static images as the principal form of production, so it's less efficient and economical than it could be. If you're interested in a superior protocol with lower cost and higher functionality, simply swap in the alternative native production language discussed in the Forms of Production section below.

Are ESI Protocols Compulsory?

Effectively, yes; explicitly, no. The Rules do not *expressly* require that the range of ESI-related topics on which counsel must engage be memorialized in an ESI Protocol; but where consensus exists, agreements should be memorialized as part of a discovery plan. So, *effectively* the Rules require an ESI Protocol to emerge, whether we call it that or not.

The Federal Rules of Civil Procedure require that parties confer regarding, *inter alia*:

- Issues about preservation of ESI (*Rule 26(f)(3)(C)*)
- Issues about the form or forms in which ESI should be produced (*Id*.)
- Issues about claims of privilege or of protection as trial-preparation materials (*Rule 26(f)(3)(D)*)

Additionally, Rule 34(b)(1)(C) permits parties seeking production to specify the form or forms in which electronically stored information is to be produced, and it allows a party to whom the request is made to object and state the form or forms it intends to use. The 2006 Advisory Committee Notes to Rule 34 underscore that a party is not free to convert ESI to forms that make it more difficult or burdensome for the requesting party to use efficiently in the litigation or to forms that remove or significantly degrade searchability by electronic means.

[119] You'll see this language again at the end, but I'm putting the takeaway here in case you don't get to the end: *Modern* evidence is *electronic* evidence and demands the use of electronic review tools. The *raison d'être* of an ESI Protocol is to *make productions work*, ensuring that responsive electronic evidence produced in discovery is as complete, utile and accessible as reasonably possible without exposing privileged and protected content. Modern electronic evidence resides in rich and complex information taxonomies, on systems, machines and media, in databases, accounts, folders, containers and files. Only through the meticulous management and production of data and metadata can this architecture be understood in ways essential to proving authenticity and admissibility. *These technical details matter*, and failure to attend to them thoroughly and competently prompts pernicious consequences ranging from inaccurate searches to brutally inflated review costs to losing the case because you missed probative evidence. That's the takeaway: *ESI protocols are worth fighting for, and the better both sides understand their application and purpose, the less there is to fight about.*

These obligations can be met by means other than an ESI Protocol, and parties are not-duty bound to agree on anything. Yet, FRCP Rule 1 mandates the Rules "be construed, administered, and employed by the court and the parties to secure the just, speedy, and inexpensive determination of every action and proceeding," and judges expect lawyers to manage discovery primarily through agreement and cooperation. Isn't it smarter for parties to nail down basic discovery issues and ensure those agreements coalesce as a well-crafted ESI Protocol?

Should the Protocol be Court-Ordered?

Civil discovery was conceived as a party- and lawyer-directed process, which works well until it doesn't, at which point the Court must step in to keep discovery abuse from derailing the case. My view is, if I agree to something, I'm content to put in writing; and if I'm willing to agree to it in writing, I'm content for it to be memorialized in an order. But there's a school of thought that lawyers should afford their clients ample wiggle room in agreements, and court-ordered protocols make it difficult to adapt to the unforeseen and change direction when discovery becomes riskier, more disruptive or more costly than expected. Whether a court-ordered protocol is a guardrail or tripwire depends upon whose ox is gored.

In the final analysis, judges guard their authority more jealously than litigants' rights; accordingly, courts tend to enforce their orders more rigorously than party agreements. If you want an ESI Protocol with teeth, get it entered as an order.

Eschew Blather and Boilerplate

Are ESI Protocols improved by stating the obvious? Many lawyers must think so because ESI Protocols can teem with blather and boilerplate. Pertinent definitions and aspirational statements defining the goals of the protocol may guide courts called on to divine the parties' intent, but paragraphs asserting that the applicable Rules apply or that discovery must be "reasonable" or "proportional" are pointless. A protocol reciting that parties must act in "good faith" or "cooperate" is no more likely to prompt salutary conduct than one silent on same. Likewise, though definitions of terms of art are helpful, defining terms never used in the protocol is sloppy. Some protocols reference e-discovery glossaries like those published periodically by The Sedona Conference. If you take that approach, be sure you can live with *all* the positions advocated by the glossary because it may contain language that will bite you in court. Also, specify the edition of the glossary agreed upon since they change over time, sometimes significantly and diametrically (*e.g.,* compare Sedona's positions on metadata across the First, Second and Third editions of The Sedona Principles). It's safer to incorporate only the definitions you need and avoid referencing materials beyond the four corners of the protocol.

Absent from the exemplar protocol language below is the customary litany of promises to meet and confer about matters left unresolved or in the face of conflicts and unforeseen complications. Certainly, parties should seek a framework for dispute resolution short of going to court, but the obligation to confer before filing motions already exists in federal practice and most states. If the parties see benefit in adding mandates to meet and confer respecting, *inter alia*, production of structured data, keyword search or technology-assisted review, there's no harm (albeit little benefit) in including them.

The Annotated ESI Protocol

What follows is exemplar language of the sort often seen in ESI Protocols, culled and adapted piecemeal from dozens of examples. It's certainly not "The Perfect ESI Protocol" but one crafted in the hope of achieving both a representative assemblage of protocol provisions and a measure of coherence and consistency. There are no "magic words." A suitable protocol may require tweaking to adapt to the issues and evidence in the case and, most often, to the software and capabilities of the technical staff and service providers charged to collect, process, host and produce electronic evidence.

Exemplar Protocol Language		**Explanation and Commentary**
<u>Definitions</u> **11. "Document(s)" is defined to be synonymous in meaning and equal in scope to the usage of the term in Rule 34(a) of the Federal Rules of Civil Procedure and includes ESI existing in any medium from which information can be translated into reasonably usable form, including but not limited to email and attachments, word processing documents, spreadsheets, graphics, presentations, images, text files, databases, instant messages, transaction logs, audio and video files, voicemail, internet data, computer logs, text messages, and backup materials. The term "Document(s)" shall include Hard Copy Documents, Electronic Documents, and Electronically Stored Information (ESI) as defined herein.** **12. "Electronic Document(s) or Data" means Documents or Data existing in electronic form at the time of collection, including but**		Definitions artfully deployed in a protocol can serve to streamline and simplify the language of the Protocol and Requests for Productions that follow. Accordingly, care should be taken to ensure that boilerplate definitions in requests conform to definitions contained in applicable protocols. Because the term "document" hearkens back to a paper-centric era of discovery, it's sensible to clarify that the term must be read expansively to include information in all its myriad forms, particularly data stored electronically, magnetically, optically and otherwise, and that "documents" encompass not only routine records (like memos, reports, presentations and ledgers) but also stored communications, like email, text messaging and collaborative communications (*e.g.,* comments as tracked change and Slack) and relevant rich media, like video and audio recordings or social networking content.

not limited to: e-mail or other electronic communications, word processing files (*e.g.,* Microsoft Word), computer presentations (*e.g.,* PowerPoint slides), spreadsheets (*e.g.*, Excel), and image files (*e.g.*, PDF). **13. "Electronically stored information" or "ESI," is information that is stored electronically as files, documents, or other data on computers, servers, mobile devices, online repositories, disks, USB drives, tape or other real or virtualized devices or digital media.** **14. "Hard Copy Document(s)" means Documents existing in paper form at the time of collection.** **15. "Hash Value" is a numerical identifier that can be determined from a file, a group of files, or a portion of a file, based on a standard mathematical algorithm that calculates a value for a given set of data, serving as a digital fingerprint, and representing the binary content of the data to assist in subsequently ensuring that data has not been modified and to facilitate duplicate identification. Unless otherwise specified, hash values shall be calculated using the MD5 hash algorithm.** **16. "Load File(s)" are electronic files containing information identifying a set of paper scanned (static) images or processed ESI and indicating where individual**		

pages or files belong together as documents, including attachments, and where each document begins and ends. Load Files also contain data relevant to individual Documents, including extracted and user-created Metadata, coded data, as well as OCR or Extracted Text. A load file linking corresponding images is used for productions of static images (*e.g.,* TIFFs) **17. "Metadata" is the term used to describe the structural information of a file that contains data about the file, as opposed to describing the content of a file.** **18. "Native Format" means the file format associated with the original creating application and as collected from custodians. For example, the native format of an Excel workbook is an .xls or .xlsx file.** **19. "Optical Character Recognition" or "OCR" means a technology process that captures text from an image for the purpose of creating an ancillary text file that can be associated with the image and searched in a database. OCR software evaluates scanned data for shapes it recognizes as letters or numerals.** **20. "Searchable Text" means the native text extracted from an Electronic Document or, when extraction is infeasible, by Optical**		

Character Recognition text ("OCR text") generated from a Hard Copy Document or electronic image.		Metadata remains among the most misunderstood topics in ESI discovery, encompassing not only *system metadata*, the contextual information computing devices keep about electronically stored information and stored without the file, but also *application metadata*, content about the file and stored within the file, moving with the file when copied. Examples of system metadata are a file's name and the date the file was last modified. Examples of application metadata for a word-processed document are the date a file was last printed and tracked changes and comments.
<u>Preservation</u> **The Parties represent that they have issued litigation hold notices to those custodians with data, and persons or entities responsible for maintenance of non-custodial data, which, based upon then-current information available, are reasonably likely to contain discoverable information.**		ESI protocols often incorporate preservation clauses that do no more than enunciate the parties' common law duties. Unless the purpose of the provision is to narrow or expand the duty of preservation beyond the common law obligation, the provision can be dispensed with. A preservation clause may be used to identify the classes of custodians or sources that will *not* be routinely preserved, such as backup media dedicated to disaster recovery, web cache, server log files and other items that deemed not reasonably accessible or unduly burdensome.

The Parties agree there is no need to preserve potentially relevant materials from the following sources: **.** **6. Deleted, fragmented, or data in unallocated clusters of storage media that is only accessible by computer forensics.** **7. Volatile random-access memory (RAM), temp files, or other ephemeral data that is difficult to preserve without disabling the operating system or through the use of computer forensics.** **8. Temporary internet files, browser history files, cache files, and cookies.** **9. Back-up data that a party knows to be duplicative of ESI, documents, data or tangible things, including metadata about such information, verified to have been retained. and** **10. Server, system, or network logs.**		
<u>eDiscovery Liaison</u> **The parties agree to designate one or more competent persons to serve as liaisons for purposes of meeting, conferring and attending court hearings regarding discovery of ESI.**		Though even the best ESI liaisons must sometimes reply, "I'll get back to you," communication and efficiency really suffer when questions filter through counsel unschooled in eDiscovery. Working through skilled liaisons that "speak geek" won't guarantee harmony but fosters focused, dispassionate diplomacy.

<u>Databases and Structured Data</u> **If ESI in commercial or proprietary database formats can be produced in an existing and reasonably usable, delimited report format (*e.g.,* Excel or CSV), the Parties will produce the information in such format.** **If an existing report format is not reasonably available or usable, the Parties will meet and confer to attempt to identify a mutually agreeable form of production based on the specific needs and the content and format of data within such structured data source.**		Much data sought in discovery is structured data; it resides within and is retrieved from databases. Email is a database. Social networks are databases. Financial records, health records, payroll records, customer and sales records all tend to be structured data in databases. A distinguishing feature of structured data is that it's fielded; that is, information is stored in locations dedicated to holding just that information. Fielding data serves to separate and identify information so you can search, sort and cull using just that information. It's a capability we take for granted in digital applications but can be crippled or eradicated when data is produced in e-discovery without preserving its fielded ("delimited") character. For more on databases in eDiscovery: http://www.craigball.com/Ball_DB_2010.pdf
<u>Hard Copy Documents</u> **Hard Copy Documents shall be scanned to single page Group IV TIFF format, 300 dpi quality or better with corresponding searchable OCR text. Image file names will be identical to the corresponding Bates numbered images, with a ".tif" file extension.[120] The file name of each text file should correspond to the file name of the first image file of**		Although there's no legal duty that Hard Copy Documents be digitized, sound practice dictates that legacy paper records meld with modern digital evidence. ESI Protocols specify the form and quality of scanned items and whether and how paper records must be made text searchable. TIFF is an initialism for Tagged Image File Format, a long-used file format for storing page images as black & white pictures. "Single page" requires that each page of a document be produced as a single image file dedicated to each page. Where a 100-page file

[120] Bates numbering has historically been employed as an organizational method to label and identify legal documents, especially those produced in discovery. "Bates" is capitalized because the name derives from the Bates Manufacturing Company, which patented and sold auto-incrementing, consecutive-numbering stamping devices. Bates numbering serves the dual function of sequencing and uniquely identifying documents.

the document with which it is associated.		produced as a PDF would consist of a single file holding 100 pages, the same document produced in single page TIFF would consist of 100 individual files, each an image of a single page of the document. "Group IV" refers to the way the scanned image is compressed to speed transmission and optimize storage space. 300 dpi speaks to the "dots per inch," a measure of scanning and printing resolution. The higher the dots per inch, the clearer and more detailed the image; however, higher resolutions require more image data and produce larger files per page. Hard Copy Documents are inherently unsearchable electronically, so searchability may be achieved by subjecting the page images to optical character recognition (OCR). TIFF images do not store the associated text of the imaged document, so the OCR text is supplied in an accompanying file, typically a single file of text for the entire document rather than a single text file corresponding to each page. In this provision, the text file name pairs with the image file name of the first page of the document. *Note however,* Hard Copy Documents are inherently unsearchable; thus, there is no legal duty under the Rules to add searchability. The obligation to supply OCR is one the parties *choose* to take on, so apart from redacted documents, no party is *obliged* to supply OCR text absent an agreement or order. Because this provision demands an image be produced for each page, Bates numbering ensures filenames are unique and pages are produced sequentially. This requires that page images be created (or renamed) using software that supports Bates numbering and careful attention paid to avoid reusing sequences from prior productions.

		Comment: This provision is as close to an enduring, industrywide standard as exists despite serious shortcomings. We are captive to 80's era technology when it comes to scanned hard copies. TIFF images tend to be much larger files than the same document supplied as a PDF image, making TIFF productions more expensive to host online and slower to appear onscreen. Unlike PDFs, TIFFs convert color data to black and white, a sometimes-serious downgrading of the evidence. The 300-dpi resolution works well enough for letters and reports but may be insufficient to adequately display technical drawings and fine details.
<u>Unitizing Documents</u> **In scanning Hard Copy Documents, distinct documents should not be merged into a single record, and single documents should not be split into multiple records (*i.e.*, paper documents should be logically unitized). For example, Hard Copy Documents stored in a binder, folder, or similar container should be produced in the same order as they appear in the container. The front cover of the container should be produced immediately before the first document in the container. The back cover of the container should be produced immediately after the last document in the container. The Parties will undertake reasonable efforts to, or have their vendors, logically unitize documents correctly, and will**		"Unitization" refers to the organization of pages into a document, chapter or volume. Paper documents are physically unitized by means of, *e.g.,* clips, staples, bindings and folders. Multiple documents may comprise a "family" unit; for example, a transmittal and its attachments or a report and its exhibits/appendices comprise a parent/child relationship. When unitized paper records are scanned, metadata supplies a logical unitization of files mirroring the physical unitization of the physical document or volume scanned. For documents that contain affixed notes, pages may be scanned once with the notes as they appear on the page and again without the notes, so all content is captured. The relationship of documents in a document collection should be maintained throughout scanning, and processing (*e.g.,* cover letter and enclosures, e-mail and attachments, binder holding multiple documents, folder and other compilations where a parent-child relationship exists between the documents).

commit to address situations of improperly unitized documents.		For ESI, the keys to preserving unitization lie in both the ordering of documents by Bates numbers and the metadata supplied in load files.
Parent-Child Relationships **The Parties agree that if any part of a Document or its attachments is responsive, the entire Document and attachments will be produced, except any attachments that must be withheld or redacted and logged based on privilege or work-product protection.** **The Parties shall take reasonable steps to ensure that parent-child relationships within a document family (the association between an attachment and its parent document) are preserved. The child document(s) should be consecutively produced immediately after the parent document. For further clarification, this shall not require a party to produce documents merely referenced in responsive documents; provided, however, that documents sent via a link within an email should be produced.**		Few things are as frustrating in a production review as being unable to pair a "parent" transmittal with its "child" attachments. This provision reflects the custom of extracting child attachments from the parent transmittal and supplying them *seriatim*. Too, it touches on potentially-fractious *scope* of discovery issues by requiring producing parties to treat a document family as a single item to be produced if any component is responsive (although any part may be withheld or redacted on claim of privilege). A producing party may resist, arguing that discovery allows for granular treatment of the family and does not require production of non-responsive attachments or transmittals. Note that the exemplar language obliges the parties to produce hyperlinked files or so-called "modern attachments." The parties must appreciate what this obligation entails in the context of their messaging environment. Some Cloud systems (*e.g.,* Microsoft 365) make it easy to collect documents transmitted as hyperlinked files versus embedded attachments, whereas others may demand manual collection with attendant uncertainty as to whether the item collected remains faithful to the item transmitted. As phrased, the operative distinction is whether the hyperlink in the transmittal points to a resource readily available to anyone with the link (that is, "documents merely referenced") or whether the modern attachment item is unavailable to the requesting party if not produced with the transmittal.
Hard Copy Document Metadata		Paper documents have metadata, too, some of it essential for proper unitization and management. In

The following metadata fields should be provided for Hard Copy Documents when reasonably available: **9. Beginning Bates number** **10. Ending Bates number** **11. First attachment Bates number** **12. Last attachment Bates number** **13. Source location/custodian** **14. Confidentiality designation** **15. Redacted (Y/N) and** **16. Extracted/OCR text file path.**		the example, note that the eight data points required are not usually found within a document. Instead, these metadata values are either collected (like source location/custodian) or (like Bates numbers), assigned as part of an ESI processing and production workflow.
<u>Forms of Production</u> ***Alternative 1: Native Production*** **The Parties will produce Electronic Documents, Data and ESI in Native** **Formats with the metadata specified in ADDENDUM A. Redacted ESI may be redacted natively, as feasible, or produced as redacted TIFFs with applicable, non-privileged metadata and OCR searchable text.** **Electronic Documents, Data and ESI will be Bates numbered by substituting, prepending or appending the Bates number for/to the file name. When any party prints produced ESI for use in a filing or proceeding, such party shall ensure that the Bates number of the item, any required confidentiality notices and**		Establishing the form or forms of production is the centerpiece of any ESI protocol, and the feature with the greatest influence on the cost of processing and hosting the data. Here, alternative clauses specify native or TIFF+ as the default form of production for ESI. Note that each approach borrows from the other in that native productions provide that redacted data be supplied in TIFF formats, and TIFF+ productions contemplate that ESI that doesn't lend itself to static imaging be produced natively. Native forms ensure a level playing field between producing and requesting parties in that a native production will faithfully mirror the ways in which the custodians view and work with evidence. Colors and functional features are preserved, along with tracked changes and comments appearing in original files. Above all, native forms are massively smaller in size versus TIFF images created from the native file. Consequently, native productions are many times less costly to load and host when eDiscovery vendors

pagination are embossed on the face of the printed item without obscuring its content.[121] **<u>OR</u>** ***Alternative 2: TIFF+ Production*** **The Parties will produce Electronic Documents, Data and ESI as single page Group IV TIFF images, 300 dpi quality or better, and 8.5"x11" page size, except for documents requiring different resolution or page size with the metadata specified in ADDENDUM A. *However,* the Parties will produce the following forms of ESI in native formats:** **8. Spreadsheets** **9. PowerPoint presentations** **10. Access databases** **11. Delimited text files** **12. Photographs**		price services based on the byte volume of the data.[122] Parties favoring TIFF+ point to a diminished potential for fraudulent or inadvertent alteration of the evidence and the ability to emboss a Bates number on the face of a page image versus naming the produced files to their Bates numbers. Also, TIFF images may be viewed in any browser, though they won't be text searchable doing so. When converting electronic documents to static images, parties must consider the wealth of information users see in the native application like tracked changes and comments between collaborators in word processed documents and speaker notes in presentations. Do you require these items be made visible on the page images or leave them out of the production? The exemplar language takes the first path, but each approach has its pitfalls. Producing the document both ways doubles volume and expense. Native productions solve this issue as a native production affords requesting parties

[121] A common question is, "How do we Bates number native productions?" Because electronic files often have the same file names, the best practice is to replace the native filename with a unique Bates number and supply the original filename, paired with its Bates number, in the accompanying load file. An alternative is to ensure the filenames are unique by prepending or appending the Bates number to the filename. To facilitate page level references by Bates number when a party prints a native document for use in a deposition or proceeding, the Protocol requires that parties emboss the native file's Bates numbers and pagination on the printed document, just as with TIFF+ productions. Thus, when parties *change* the form of the evidence post-production (*e.g.,* native-to-paper), the party changing the evidence is obliged to preserve the connection between the native source and the paginated printout.

[122] Whether in native or static image format, ESI must be processed ("ingested") and hosted to be searchable and reviewable. Native forms are processed to extract their text and metadata, then indexed for search. TIFF and load file productions are indexed for search and processed to pair the page images with text and metadata. Either way, you pay a vendor to prepare the production for viewing and then pay a recurring "hosting" charge for online access to the production. The fees charged are based on the volume of data processed and/or hosted. More data costs more money. If you receive 10 times as much data, you pay a commensurate amount more to ingest and host. Vendors usually assess hosting fees as a monthly subscription, so the more data they host for you, the more you pay every month for the life of the case. More data isn't the same thing as more information because not all electronic forms of information are equally efficient. When you convert native forms to static images and load files you explode the size of production by many multiples, and static productions come burdened by the further cost of impaired searchability, diminished functionality and lost color, animation and rich media.

13. Audio and video files **14. Documents of a type which cannot be reasonably converted to useful TIFF images.** **All images of documents which contain tracked changes such as comments, deletions and revision marks (including the identity of the person making the deletion or revision and the date and time thereof), speaker notes, or other user-entered data that the source application can display to the user will be processed such that all that data is visible in the image.** **<u>File Names</u>** **Each TIFF image should have a unique file name corresponding to the Bates number of that page with a ".tif" file extension. The file name should not contain any blank spaces and should be zero-padded (e.g., DEF-000001), taking into consideration the estimated number of pages to be produced. If a Bates number or set of Bates numbers is skipped in a production, Producing Party will so note in a cover letter or production log accompanying the production. Bates numbers will be unique across the entire production and**		comparable access to content as the custodian of the evidence. When parties convert evidence in native forms to static image forms like TIFF, that process strips away all electronic searchability. A monochrome screenshot replaces the source evidence. Since the Federal Rules of Civil Procedure say parties can't remove or significantly degrade searchability, responding parties must act to restore a measure of searchability. They do this by extracting text from the native ESI and delivering it in a "load file" accompanying the page images. This (and metadata) is the "plus" when people speak of "TIFF+" productions. To search a TIFF+ production, page images and load files must be hosted in an eDiscovery review "platform" capable of pairing the extracted text with the corresponding page images.[123]

[123] This process can operate to materially impair accurate search as in https://craigball.net/2020/01/15/degradation-how-tiff-disrupts-search/

prefixes will be consistent across all documents produced. **Producing Party will brand all TIFF images in the lower right-hand corner with its corresponding Bates number without obscuring any part of the underlying image.** **<u>Extracted Text Files</u>** **For each document, a single Unicode text file containing extracted text shall be provided along with the image files and metadata. The text file name shall be the same as the Bates number of the first page of the document. File names shall not have any special characters or embedded spaces. Electronic text must be extracted directly from the native electronic file to the extent reasonably feasible unless the document is an image file or contains redactions, in which case, a text file created using OCR should be produced in lieu of extracted text.**		Once more—and unlike native files and PDFs--TIFF images are merely black-and-white pictures of pages and cannot be searched for words or phrases. They hold no text. To facilitate searchability, the text of documents must be produced in separate load files meant to be loaded into review software. Searches are then run against the text file data (more accurately, an index of created from that text) and, because the Bates numbered text files share names with the Bates numbered image files, search hits within text ties to page images. This is only possible when naming conventions are adhered to, hence attendant language of the protocol. Also, because the text of a document may include foreign languages and specialized characters, the provision requires that the text be produced as Unicode text, meaning that it must be encoded to support a wide array of

		international characters versus the paltry 256 characters of the once-ubiquitous ASCII encoding.[124]
<u>Load Files</u> **Productions will, as applicable, include image load files in Opticon or IPRO format as well as Concordance format data (.dat) files with the applicable metadata fields identified in ADDENDUM A. All metadata will be produced in UTF-16LE or UTF-8 with Byte Order Mark format.** **All native format files shall be produced in a folder named "NATIVE,"** **All TIFF images shall be produced in a folder named "IMAGE," which shall contain sub-folders named "0001," "0002," etc. Each sub-folder shall contain no more than 10,000 images. Images from a single document shall not span multiple sub-folders.** **All extracted Text and OCR files shall be produced in a folder named "TEXT."**		Load files are used to import image, native, and text files and their corresponding metadata and production information into a document database or "review tool". Load files carry indispensable information, such as file names, file locations (both their origination and within a production), sources, custodians and dates. The information in load files enables search, sorting, tracing, authentication, unitization and much more. They are the Rosetta Stones of ESI production. The references to Opticon, IPRO and Concordance do not oblige a party to use a particular vendor or software; instead, those are shorthand ways to designate the *structure* of the load files and of the delimiters ("character separators") employed to distinguish one field of metadata from the next. "UTF" stands for Unicode Transformation Format, a universal way to encode alphanumeric character sets for worldwide consistency and intelligibility. For more on load files: https://craigball.net/2013/07/17/a-load-file-off-my-mind/

[124] ASCII is an acronym for American Standard Code for Information Interchange and describes one of the oldest and simplest standardized ways to use numbers—particularly binary numbers expressed as ones and zeroes–to denote a basic set of English language alphanumeric and punctuation characters.

All load files shall be produced in a folder named "DATA" or at the root directory of the production media.		
<u>Color</u> **Paper documents or redacted ESI that contain color used to convey information (*e.g.*, color coding and highlighting versus merely decorative use) shall be produced as single-page, 300 DPI JPG images with JPG compression set to its highest-quality setting so as not to not degrade the original image.** **<u>OR</u>** **Where .TIFF images are illegible due to color content (such as colored text on a colored background) or where color is material to the interpretation of a document, JPG image files shall be provided upon reasonable request.**		JPG images and native productions show color, but TIFF images are black and white renderings, so an unsuitable form of production when color is used to convey information. Some protocols address the problem by allowing requesting parties to make *ad hoc* requests for reproduction of items in forms supporting color. The obvious problem is that it's often impossible to discern the use of color working from a black and white image. Some eDiscovery software tools offer the ability to detect the use of color in a file and can programmatically pivot the form of production between TIFF and JPG formats. As a rule, JPG images should *always* be produced when the source evidence is a JPG image (*e.g.*, a photograph). Email transmittals frequently contain decorative color (in logos), so best lend themselves to *ad ho*c requests for color reproduction. PowerPoint presentations and Excel spreadsheets should never be produced in anything but native formats (where color is natively supported).
<u>Redactions</u> **Any redacted material must be clearly labeled on the face of the document as having been redacted and shall be identified as such in**		ESI documents can contain both apparent and non-obvious content. For example, PDFs often include an image layer and a textual layer such that altering the image won't change the searchable text. Accordingly, ESI poses unique challenges when a document contains privileged and non-privileged

the load file provided with the production. Each redacted document shall be produced with an OCR *.txt file containing unredacted text. A document's status as redacted does not relieve the producing party from providing all the metadata required herein unless the metadata withheld is contains privileged content.		information. Although many forms of ESI are easy to redact reliably in their native formats and privileged content can be expurgated without impairing the searchability of non-privileged content, lawyers tend not to trust native redaction. Instead, they demand that "blacked out" TIFF images be used for redaction even when all other documents are produced natively. This requires searchability be restored for the unredacted content; and since text extraction might grab privileged content, OCR is used instead.
<u>Privilege Logs</u> **With each production, Producing Party shall supply a log of the documents withheld or redacted under a claim of privilege and/or work product with sufficient information to allow the Receiving Party to understand the basis for the claim.** **Communications involving trial counsel that post-date the filing of the complaint need not be placed on a privilege log.**		The obligation to furnish a privilege log is governed by the applicable Rules of Civil Procedure, *e.g.,* Fed. R. Civ. P. 26(b)(5)(A). Privilege logs don't implicate unique technical concerns except to the extent that a Producing Party seeks a "metadata privilege log" or a "categorical privilege log," to avoid the description duties required in the Rules. The exemplar language includes a categorical exemption for post-suit communications with trial counsel. Commentary: Though ESI protocols often address privilege logs, the timing and scope of privilege logs is best addressed in an agreement incorporating a liberal clawback and non-waiver provision and, in federal court, a Federal Rule of Evidence 502(d) order governing inadvertent production of privileged information.[125]
<u>Deduplication</u> ***Vertical Deduplication*** **Producing Party may vertically de-duplicate documents based on**		Parties should endeavor to produce a single copy of each responsive document while identifying unproduced duplicates via their metadata values in load files. In this way, Receiving Parties are not burdened by production of duplicates yet can

[125] A clawback provision governs what parties must do when there's been an inadvertent disclosure of privileged information: issues such as disclosure, sequestration, return, destruction, non-use and non-waiver. Such provisions are designed to minimize the harm flowing from unwitting disclosure and, crucially, to forestall the dread "subject matter waiver" whereby the release of even a narrow range of privileged material may serve to "open the door" to all privileged material touching on the subject matter of the inadvertent disclosure.

MD5 or SHA-I hash values at the document level, by Message ID, EDRM MIH[126] or other standard methodology for email deduplication within the collection of a custodian or a data source. Attachments to parent documents may not be deduplicated against a duplicate standalone version of the attachment exists, and standalone versions of documents may not be suppressed if a duplicate version exists as an attachment. **OR** ***Horizontal Deduplication*** **Producing Party may horizontally (globally) de-duplicate documents based on MD5 or SHA-I hash values at the document level or by Message ID, EDRM MIH or other standard methodology for email deduplication within the collection of a custodian or a data source. Attachments to parent documents may not be deduplicated against a duplicate standalone version of the attachment exists, and standalone versions of documents may not be suppressed if a duplicate version exists as an attachment.**		determine which custodians possessed duplicates and, *inter alia,* know the unique dates, names and locations of deduplicated instances. *Vertical deduplication* refers to deduplication within the collection of a single source or custodian, differentiated from *horizontal or global deduplication* where deduplication spans the collections of multiple sources or custodians. MD5 and SHA-1 are standard cryptographic hash algorithms, mathematical formulas that calculate a fixed length value for a given binary input of any size. These hash values serve as digital fingerprints of the binary content of files to facilitate duplicate identification. E-Discovery service providers employ varying methods to calculate a hash value for email messages and attachments. The exemplar language provides that, whatever method is used, it will not be implemented in a way that would make it difficult to distinguish documents made attachments to email transmittals from the same documents existing as standalone files.

[126] The EDRM MIH (for Message Identification Hash) is a unique identifier enabling cross platform email duplicate identification. Specifications and information about the EDRM MIH are found at https://edrm.net/edrm-projects/dupeid-2/ and a white paper describing the MIH is here: https://edrm.net/download/161805

Producing Party will track all deduplicated files and provide the names of all custodians of these duplicates of in the load file. If the duplicates are e-mails, the producing party must detail the process of creating the hash value, *e.g.*, the names and order of concatenated fields by which the deduplication hash was calculated.		
<u>De-NISTing</u> **System and application files without user created content (as identified by matching to the NIST National Software Reference Library database) need not be processed, reviewed or produced.**		The National Software Reference Library, part of the U.S. National Institute for Standards and Technology, compiles and distributes digital signatures for software, including the files comprising most operating systems and commercial applications. Because the constituents of commercial software are seldom relevant evidence in civil cases, excluding these from eDiscovery fosters efficiency.
<u>Email Threading</u> **To reduce the volume of entirely duplicative content within email threads, the parties may, but are not required to, use email threading. A party may use industry standard message threading technology to remove email messages where the content of those messages, and any attachments, are wholly contained within a later email message in the thread; *provided, however*, that the use of threading must not serve to obscure whether a recipient received an attachment.**		When email messages are produced as static images, email threading simplifies review by presenting all messages that comprise an email conversation as a continuous, temporally-ordered "thread." The objection most often voiced is that threading may serve to suppress a message or attachment.

<u>Production Media</u> **The producing party will use the appropriate electronic media (DVD, thumb drive, hard drive or secure FTP transfer) for its ESI production and will endeavor to use the highest capacity suitable media. The producing party will label the production media with the name of the producing party, production date, media volume name, and Bates number range(s).** **Productions on physical media should be encrypted for transmission to the Receiving Party. At the time of production and under separate cover, Producing Party shall furnish decryption credentials to Receiving Party.**		ESI protocols specify both the *form* of production and the *medium* of production, the former being the file types to be supplied and the latter the type of storage device used to hand off the data. Production media should be selected to minimize the number of disks or drives required for transfer, although that's a concern tied to the era of floppy disks and optical disks and not an issue with today's huge hard drives. Parties should ensure that the contents of production media are encrypted, both to protect against loss in transit and to guard against unauthorized access. Care should be taken not to transmit encrypted data with decryption passwords and to never label or store encrypted media with its decryption credentials.
<u>Processing</u> **The Parties will use reasonable efforts and standard industry practices to address and resolve exception issues for items that present processing, imaging or form of production problems (including encrypted, corrupt and/or protected files identified during the processing of ESI). The Parties will meet and confer regarding procedures that will be used to identify, access, and process and resolve exception issues.**		For more about processing: http://www.craigball.com/Ball_Processing_2019.pdf

Parties shall normalize times and dates to conform to [UTC] or [specified local time zone]. **For archive files (zip, jar, rar, gzip, TAR, etc.), all contents should be** **extracted from the archive with source pathing and family relationships preserved and produced. The fully unpacked archive container file does not need to be included in the production.**		
<u>Non-Waiver</u> **This Protocol is solely intended to address the format of document productions and does not limit the temporal or substantive scope of discovery. Nothing in this Protocol is intended to affect the right of any party to object to a request for production or to operate as a waiver of any party's right to promulgate, object to, or seek relief from a request for discovery.**		

<u>Metadata Production Fields</u>

The exemplar ESI protocol above contemplates that the parties will agree upon the metadata fields that will be extracted or populated and produced in the load file. Different forms of ESI hold different application metadata, and some metadata isn't collected with or extracted from the ESI but must be assigned or calculated when the data is processed. Custodians are typically determined at collection and designated when their data is ingested by eDiscovery software for processing. A hash value is calculated for each file. A Bates number is assigned to each file or page image. Not every eDiscovery vendor can supply every field below, and some use different field names for the same data.

ADDENDUM A

Field Name	Description
BegBates	First Bates identifier of item
EndBates	Last Bates identifier of item
PgCount	Number of pages in the document
FileSize	Size of native file document/email in KB
FileName	Original name of file as it appeared in location where collected
Path	E-mail: Original location of e-mail including original file name Native: Originating path where native file document was collected including original file name
NativeLink	Relative path and filename for native file on production media
TextLink	Relative path and filename for text file on production media
AttRange	Bates identifier of the first page of the parent document to the Bates identifier of the last page of the last attachment "child" document
BegAttach	First Bates identifier of attachment range
EndAttach	Last Bates identifier of attachment range
AttachCount	Number of attachments to an e-mail
AttachName	Names of each individual Attachment, separated by semicolons
ParentBates	First Bates identifier of parent document/e-mail message (will not be populated for documents that are not part of a family)
ChildBates	First Bates identifier of "child" attachment(s); may be more than one Bates number listed depending on number of attachments (will not be populated for documents that are not part of a family)
Custodian	E-mail: mailbox where the email resided Native: Individual from whom the document originated
OtherCustodians	Custodians whose file/message has been de-duplicated; separated by semicolons
From	E-mail: Sender Native: Author(s) of document; separated by semicolons

To	E-mail: Recipient(s); separated by semicolons
CC	E-mail: Carbon copy recipient(s); separated by semicolons
BCC	E-mail: Blind carbon copy recipient(s) separated by semicolons
DateSent (mm/dd/yyyy hh:mm:ss AM or PM)	E-mail: Date and time the email was sent
Subject	E-mail: Subject line of email
Title	Document: Title provided by user within the document
MsgID	E-mail: "Unique Message ID" field
EDRM_MIH	Identifier enabling cross platform email duplicate identification. Specifications and information about the EDRM MIH are found at https://edrm.net/edrm-projects/dupeid-2/
ModifiedDate (mm/dd/yyyy hh:mm:ss AM or PM)	Document: Last Modified Date and time
CreationDate (mm/dd/yyyy hh:mm:ss AM or PM)	Document: Create Date and time
FileExt	Document: file extension
FileType	Document: file type
Redacted	Denotes that item has been redacted as containing privileged content (yes/no)
Hash	MD5 Hash value of the item
HiddenContent	Denotes presence of Tracked Changes/Hidden Content/Embedded Objects in item(s) (Y/N)
Confidential	Denotes that item has been designated as confidential pursuant to confidentiality agreement or protective order (Y/N)
DeDuped	Full path of deduplicated instances; separated by semicolons

Takeaway

By now, you may be marveling at the persnickety technical details requiring precise management to enable lawyers to view and search ESI productions. Alternatively, you may be bored and irritated at having to deal with any of this…stuff. If it strikes you as fussy, then you're probably not the person responsible for making it work.

Modern evidence is *electronic* evidence and demands the use of electronic review tools. The *raison d'être* of an ESI Protocol is to *make productions work*, ensuring that responsive electronic evidence produced in discovery is as complete, utile and accessible as reasonably possible without exposing privileged and protected content. Modern electronic evidence resides in rich and complex information taxonomies, on systems, machines and media, in databases, accounts, folders, containers and files. Only through the meticulous management and production of data and metadata can this architecture be understood in ways essential to proving authenticity and admissibility. *These technical details matter*, and failure to attend to them thoroughly and competently prompts pernicious consequences ranging from inaccurate searches to brutally inflated review costs to losing the case because you missed probative evidence. That's the takeaway: *ESI protocols are worth fighting for, and the better both sides understand their application and purpose, the less there is to fight about.*

Exemplar ESI Protocol (TIFF+)

Ver. 20231010

The Parties hereby agree to the following protocol for production of electronically stored information ("ESI") and paper ("hard copy") documents. This protocol governs all production in the matter.

A. Definitions

1. "Document(s)" is defined to be synonymous in meaning and equal in scope to the usage of the term in Rule 34(a) of the Federal Rules of Civil Procedure and includes ESI existing in any medium from which information can be translated into reasonably usable form, including but not limited to email and attachments, word processing documents, spreadsheets, graphics, presentations, images, text files, databases, instant messages, transaction logs, audio and video files, voicemail, internet data, computer logs, text messages, or backup materials. The term "Document(s)" shall include Hard Copy Documents, Electronic Documents, and Electronically Stored Information (ESI) as defined herein.

2. "Electronic Document(s) or Data" means Documents or Data existing in electronic form at the time of collection, including but not limited to e-mail or other means of electronic communications, word processing files (e.g., Microsoft Word), computer presentations (e.g., PowerPoint slides), spreadsheets (e.g., Excel), and image files (e.g., PDF).

3. "Electronically stored information" or "ESI," is information that is stored electronically as files, documents, or other data on computers, servers, mobile devices, online repositories, disks, USB drives, tape or other real or virtualized devices or digital media.

4. "Hard Copy Document(s)" means Documents existing in paper form at the time of collection.

5. "Hash Value" is a numerical identifier that can be determined from a file, a group of files, or a portion of a file, based on a standard mathematical algorithm that calculates a value for a given set of data, serving as a digital fingerprint, and representing the binary content of the data to assist in subsequently ensuring that data has not been modified and to facilitate duplicate identification. Unless otherwise specified, hash values shall be calculated using the MD5 hash algorithm.

6. "Load File(s)" are electronic files containing information identifying a set of paper scanned (static) images or processed ESI and indicating where individual pages or files belong together as documents, including attachments, and where each document begins and ends. Load Files also contain data relevant to individual Documents, including extracted and user-created Metadata, coded data, as well as OCR or Extracted Text. A load file linking corresponding images is used for productions of static images (e.g., TIFFs)

7. "Metadata" is the term used to describe the structural information of a file that contains data about the file, as opposed to describing the content of a file.

8. "Native Format" means the file format associated with the original creating application and as collected from custodians. For example, the native format of an Excel workbook is an .xls or .xlsx file.

9. "Optical Character Recognition" or "OCR" means a technology process that captures text from an image for the purpose of creating an ancillary text file that can be associated with the image and searched in a database. OCR software evaluates scanned data for shapes it recognizes as letters or numerals.

10. "Searchable Text" means the native text extracted from an Electronic Document or, when extraction is infeasible, by Optical Character Recognition text ("OCR text") generated from a Hard Copy Document or electronic image.

B. Preservation

The Parties represent that they have issued litigation hold notices to those custodians with data, and persons or entities responsible for maintenance of non-custodial data, which, based upon then-current information available, are reasonably likely to contain discoverable information.

The Parties agree there is no need to preserve potentially relevant materials from the following sources:

1. Deleted, fragmented, or data in unallocated clusters of storage media that is only accessible by computer forensics.

2. Volatile random-access memory (RAM), temp files, or other ephemeral data that is difficult to preserve without disabling the operating system or through the use of computer forensics.

3. Temporary internet files, browser history files, cache files, and cookies.

4. Back-up data that a party knows to be duplicative of ESI, documents, data or tangible things, including metadata about such information, verified to have been retained; and

5. Server, system, or network logs.

C. eDiscovery Liaison

The parties agree to designate one or more competent persons to serve as liaisons for purposes of meeting, conferring and attending court hearings regarding discovery of ESI.

D. Databases and Structured Data

If ESI in commercial or proprietary database formats can be produced in an existing and reasonably usable, delimited report format (*e.g.*, Excel or CSV), the Parties will produce the information in such format.

If an existing report format is not reasonably available or usable, the Parties will meet and confer to attempt to identify a mutually agreeable form of production based on the specific needs and the content and format of data within such structured data source.

E. Hard Copy Documents

Hard copy documents shall be scanned to single page Group IV TIFF format, 300 dpi quality or better with corresponding searchable OCR text. Image file names will be identical to the corresponding Bates numbered images, with a ".tif" file extension. The file name of each text file should correspond to the file name of the first image file of the document with which it is associated.

F. Unitizing Documents

In scanning hard copy documents, distinct documents should not be merged into a single record, and single documents should not be split into multiple records (i.e., paper documents should be logically unitized). For example, hard copy documents stored in a binder, folder, or similar container should be produced in the same order as they appear in the container. The front cover of the container should be produced immediately before the first document in the container. The back cover of the container should be produced immediately after the last document in the container. The Parties will undertake reasonable efforts to, or have their vendors, logically unitize documents correctly, and will commit to address situations of improperly unitized documents.

G. Parent-Child Relationships

The Parties agree that if any part of a Document or its attachments is responsive, the entire Document and attachments will be produced, except any attachments that must be withheld or redacted and logged based on privilege or work-product protection.

The Parties shall take reasonable steps to ensure that parent-child relationships within a document family (the association between an attachment and its parent document) are preserved. The child document(s) should be consecutively produced immediately after the parent document. For further clarification, this shall not require a party to produce documents merely referenced in responsive documents; provided, however, that documents sent via a link within an email should be produced.

H. Hard Copy Document Metadata

The following metadata fields should be provided for hard copy documents when reasonably available:

1. Beginning Bates number
2. Ending Bates number
3. First attachment Bates number
4. Last attachment Bates number
5. Source location/custodian
6. Confidentiality designation
7. Redacted (Y/N) and

8. Extracted/OCR text file path.

I. Forms of Production

TIFF+ Production

The Parties will produce Electronic Documents, Data and ESI as single page Group IV TIFF images, 300 dpi quality or better, and 8.5”x11” page size, except for documents requiring different resolution or page size with the metadata specified in Addendum A. However, the Parties will produce the following forms of ESI in native formats:

1. Spreadsheets
2. PowerPoint presentations
3. Access databases
4. Delimited text files
5. Photographs
6. Audio and video files
7. Documents of a type which cannot be reasonably converted to useful TIFF images.

All images of documents which contain tracked changes such as comments, deletions and revision marks (including the identity of the person making the deletion or revision and the date and time thereof), speaker notes, or other user-entered data that the source application can display to the user will be processed such that all that data is visible in the image.

J. File Names

Each TIFF image should have a unique file name corresponding to the Bates number of that page with a “.tif” file extension. The file name should not contain any blank spaces and should be zero-padded (e.g., DEF-000001), taking into consideration the estimated number of pages to be produced. If a Bates number or set of Bates numbers is skipped in a production, Producing Party will so note in a cover letter or production log accompanying the production. Bates numbers will be unique across the entire production and prefixes will be consistent across all documents produced.

Producing Party will brand all TIFF images in the lower right-hand corner with its corresponding Bates number without obscuring any part of the underlying image.

K. Extracted Text Files

For each document, a single Unicode text file containing extracted text shall be provided along with the image files and metadata. The text file name shall be the same as the Bates number of the first page of the document. File names shall not have any special characters or embedded spaces. Electronic text must be extracted directly from the native electronic file to the extent reasonably feasible unless the document is an image file or contains redactions, in which case, a text file created using OCR should be produced in lieu of extracted text.

L. Load Files

Productions will, as applicable, include image load files in Opticon or IPRO format as well as Concordance format data (.dat) files with the applicable metadata fields identified in Addendum A. All metadata will be produced in UTF-16LE or UTF-8 with Byte Order Mark format.

All native format files shall be produced in a folder named "NATIVE,"

All TIFF images shall be produced in a folder named "IMAGE," which shall contain sub-folders named "0001," "0002," etc. Each sub-folder shall contain no more than 10,000 images. Images from a single document shall not span multiple sub-folders.

All extracted Text and OCR files shall be produced in a folder named "TEXT."

All load files shall be produced in a folder named "DATA" or at the root directory of the production media.

M. Color

Paper documents or redacted ESI that contain color used to convey information (e.g., color coding and highlighting versus merely decorative use) shall be produced as single-page, 300 DPI JPG images with JPG compression set to its highest-quality setting so as not to not degrade the original image.

N. Redactions

Any redacted material must be clearly labeled on the face of the document as having been redacted and shall be identified as such in the load file provided with the production. Each redacted document shall be produced with an OCR *.txt file containing unredacted text. A document's status as redacted does not relieve the producing party from providing all the metadata required herein unless the metadata withheld contains privileged content.

O. Privilege Logs

With each production, Producing Party shall supply a log of the documents withheld or redacted under a claim of privilege and/or work product with sufficient information to allow the Receiving Party to understand the basis for the claim.

Communications involving trial counsel that post-date the filing of the complaint need not be placed on a privilege log.

P. Deduplication

Global Deduplication
Producing Party may horizontally (globally) de-duplicate documents based on MD5 or SHA-I hash values at the document level or by Message ID, EDRM MIH or other standard methodology for email deduplication within the collection of a custodian or a data source. Attachments to parent documents may not be deduplicated against a duplicate standalone version of the attachment

exists, and standalone versions of documents may not be suppressed if a duplicate version exists as an attachment.

Producing Party will track all deduplicated files and provide the names of all custodians of these duplicates of in the load file. If the duplicates are e-mails, the producing party must detail the process of creating the hash value, e.g., the names and order of concatenated fields by which the deduplication hash was calculated.

Q. De-NISTing

System and application files without user created content (as identified by matching to the NIST National Software Reference Library database) need not be processed, reviewed or produced.

R. Email Threading

To reduce the volume of entirely duplicative content within email threads, the parties may, but are not required to, use email threading. A party may use industry standard message threading technology to remove email messages where the content of those messages, and any attachments, are wholly contained within a later email message in the thread; provided however, that the use of threading must not serve to obscure whether a recipient received an attachment.

S. Production Media

The producing party will use the appropriate electronic media (DVD, thumb drive, hard drive or secure FTP transfer) for its ESI production and will endeavor to use the highest capacity suitable media. The producing party will label the production media with the name of the producing party, production date, media volume name, and Bates number range(s).

Productions on physical media should be encrypted for transmission to the Receiving Party. At the time of production and under separate cover, Producing Party shall furnish decryption credentials to Receiving Party.

T. Processing

The Parties will use reasonable efforts and standard industry practices to address and resolve exception issues for items that present processing, imaging or form of production problems (including encrypted, corrupt and/or protected files identified during the processing of ESI). The Parties will meet and confer regarding procedures that will be used to identify, access, and process and resolve exception issues.

Parties shall normalize times and dates to conform to [UTC] OR [specified local time zone].

For archive files (zip, jar, rar, gzip, TAR, etc.), all contents should be extracted from the archive with source pathing and family relationships preserved and produced. The fully unpacked archive container file does not need to be included in the production.

U. Non-Waiver

This Protocol is solely intended to address the format of document productions and does not limit the temporal or substantive scope of discovery. Nothing in this Protocol is intended to affect the right of any party to object to a request for production or to operate as a waiver of any party's right to promulgate, object to, or seek relief from a request for discovery.

ADDENDUM A

Field Name	**Description**
BegBates	First Bates identifier of item
EndBates	Last Bates identifier of item
PgCount	Number of pages in the document
FileSize	Size of native file document/email in KB.
FileName	Original name of file as appeared in location where collected
Path	E-mail: Original location of e-mail including original file name. Native: Originating path where native file document was collected including original file name.
NativeLink	Relative path and filename for native file on production media
TextLink	Relative path and filename for text file on production media
AttRange	Bates identifier of the first page of the parent document to the Bates identifier of the last page of the last attachment "child" document
BegAttach	First Bates identifier of attachment range
EndAttach	Last Bates identifier of attachment range
AttachCount	Number of attachments to an e-mail
AttachName	Names of each individual Attachment, separated by semicolons
ParentBates	First Bates identifier of parent document/e-mail message (will not be populated for documents that are not part of a family).
ChildBates	First Bates identifier of "child" attachment(s); may be more than one Bates number listed depending on number of attachments (will not be populated for documents that are not part of a family).
Custodian	E-mail: mailbox where the email resided. Native: Individual from whom the document originated

OtherCustodians	Custodians whose file/message has been de-duplicated; separated by semicolons
From	E-mail: Sender Native: Author(s) of document; separated by semicolons
To	E-mail: Recipient(s); separated by semicolons
CC	E-mail: Carbon copy recipient(s); separated by semicolons
BCC	E-mail: Blind carbon copy recipient(s) separated by semicolons
DateSent (mm/dd/yyyy hh:mm:ss AM)	E-mail: Date and time the email was sent
Subject	E-mail: Subject line of email.
Title	Document: Title provided by user within the document
MsgID	E-mail: "Unique Message ID" field
EDRM_MIH	Identifier enabling cross platform email duplicate identification. Specifications and information about the EDRM MIH are found at https://edrm.net/edrm-projects/dupeid-2/
ModifiedDate (mm/dd/yyyy hh:mm:ss AM)	Document: Last Modified Date and time
CreationDate (mm/dd/yyyy hh:mm:ss AM)	Document: Create Date and time
FileExt	Document: file extension
FileType	Document: file type
Redacted	Denotes that item has been redacted as containing privileged content (yes/no).
Hash	MD5 Hash value of the item
HiddenContent	Denotes presence of Tracked Changes/Hidden Content/Embedded Objects in item(s) (Y/N)

Confidential	Denotes that item has been designated as confidential pursuant to confidentiality agreement or protective order (Y/N).
DeDuped	Full path of deduplicated instances; separated by semicolons

Preparing for Meet and Confer

Federal Rule of Civil Procedure 26(f) requires parties to confer about preserving discoverable information and to develop a proposed discovery plan addressing discovery of electronically stored information and the form or forms in which it should be produced. This conference[128], and the overall exchange of information about electronic discovery, is called "meet and confer." [129]

Meet and confer is more a process than an event. Lay the foundation for a productive process by communicating your expectations. Send a letter to opposing counsel a week or two prior to each conference identifying the issues you expect to cover and sharing the questions you plan to ask.

E-discovery duties are reciprocal. At meet and confer, be prepared to answer many of the same questions you'll pose. And while the focus will be on large data stores of ESI, don't forget that even if your client has little electronic evidence, you must nonetheless act to preserve and produce it.

If you want client, technical or vendor representatives in attendance, say so. If you're bringing a technical or vendor representative, tell them. Give a heads up on forms of production you'll seek or are prepared to offer. Study up on any load file specification you want used and keywords to search, if only to let the other side know you've done your homework. True, your requests may be ignored or even ridiculed, but it's not an empty exercise. A cardinal rule for electronic discovery, indeed for any discovery, is to tell your opponent what you seek or possess, plainly and clearly. They may show up empty-handed, but not because you failed to set the agenda.

The early, extensive attention to electronic evidence may nonplus lawyers accustomed to the pace of paper discovery. Electronic records are ubiquitous. They're more dynamic and perishable than their paper counterparts, require special tools and techniques to locate and process and implicate

[128] The Fed. R. Civ. P. 26(f) conference must occur "as soon as practicable and in any event at least 21 days before a scheduling conference is held or a scheduling order is due under Rule 16(b)…."

[129] *Hopson v. Mayor of Baltimore*, 232 F.R.D. 228, 245 (D. Md. 2006) details some of counsel's duties under Fed, R. Civ. P. 26(f):"[C]ounsel have a duty to take the initiative in meeting and conferring to plan for appropriate discovery of electronically stored information at the commencement of any case in which electronic records will be sought….At a minimum, they should discuss: the type of information technology systems in use and the persons most knowledgeable in their operation; preservation of electronically stored information that may be relevant to the litigation; the scope of the electronic records sought (i.e. e-mail, voice mail, archived data, back-up or disaster recovery data, laptops, personal computers, PDA's, deleted data) the format in which production will occur (will records be produced in "native" or searchable format, or image only; is metadata sought); whether the requesting party seeks to conduct any testing or sampling of the producing party's IT system; the burdens and expenses that the producing party will face based on the Rule 26(b)(2) factors, and how they may be reduced (i.e. limiting the time period for which discovery is sought, limiting the amount of hours the producing party must spend searching, compiling and reviewing electronic records, using sampling to search, rather than searching all records, shifting to the producing party some of the production costs); the amount of pre-production privilege review that is reasonable for the producing party to undertake, and measures to preserve post-production assertion of privilege within a reasonable time; and any protective orders or confidentiality orders that should be in place regarding who may have access to information that is produced."

daunting volumes and multifarious formats. These differences necessitate immediate action and unfamiliar costs. Courts judge harshly those who shirk their electronic evidence obligations.

Questions for Meet and Confer

The following exemplar questions illustrate the types and varieties of matters discussed at meet and confer. They're neither exhaustive nor unique to any type of case but are offered merely as talking points to stimulate discussion.

1. **What's the case about?**

Relevance remains the polestar for discovery, no matter what form the evidence takes. The scope of preservation and production should reflect both claims *and* defenses. Pleadings only convey so much. Be sure the other side understands your theory of the case and the issues you believe should guide their retention and search.

2. **Who are the key players?**

Cases are still about *people* and what they did or didn't say or do. Though there may be shared repositories and databases to discover, begin your quest for ESI by identifying the *people* whose conduct is at issue. These *key players* are *custodians* of ESI, so determine what devices and applications they use and target their relevant documents, application data and electronic communications. Too, determine whether assistants or secretaries served as proxies for key players in handling e-mail or other ESI.

Like so much in e-discovery, identification of key players should be a collaborative process, with the parties sharing the information needed for informed choices.

3. **What events and intervals are relevant?**

The sheer volume of ESI necessitates seeking sensible ways to isolate relevant information. Because the creation and modification dates of electronic documents tend to be tracked, focusing on time periods and events helps identify relevant ESI, but only if you understand what the dates signify and when you can or can't rely on them. The Created Date of a document doesn't necessarily equate to when it was written. For ESI, the "last modified" date tends to be the most reliable.

4. **When do preservation duties begin and end?**

The parties should seek common ground concerning when the preservation duty attached and whether there is a continuing preservation duty going forward. The preservation obligation generally begins with an expectation of litigation, but the facts and issues dictate if there is a continuing obligation to preserve throughout the course of the litigation. Sometimes, events like plant explosions or corporate implosions define the endpoint for preservation, whereas a continuing tort or loss may require periodic preservation for months or years after the suit is filed.

Even when a defendant's preservation duty is fixed, a claimant's ongoing damages may necessitate ongoing preservation.

5. What data are at greatest risk of alteration or destruction?

ESI is both tenacious and fragile. It's hard to obliterate but easy to corrupt. Once lost or corrupted, ESI can be very costly or impossible to reconstruct. Focus first on fragile data, like storage media slated for reuse or messaging subject to automatic deletion and ensure its preservation. Address backup tape rotation intervals, disposal of legacy systems (e.g., obsolete systems headed for the junk heap), and re-tasking of machines associated with new and departing employees or replacement of aging hardware.

6. What steps have been or will be taken to preserve ESI?

Sadly, there are holdouts who believe all they have to reveal about ESI preservation is, "We're doing what the law and the Rules require." But that's a risky tack, courting spoliation liability by denying you an opportunity to address problems before irreparable loss. More enlightened litigants see that reasonable disclosures serve to insulate them from sanctions for preservation errors.

7. What nonparties hold information that must be preserved?

ESI may reside with former employees, attorneys, agents, accountants, outside directors, Internet service providers, contractors, Cloud service providers, family members and other nonparties. Some of these non-parties may retain copies of information discarded by a party. Absent a reminder, litigants may focus on their own data stores and fail to take steps to preserve and produce data held by others over whom they have rights of direction or control.

8. What data require forensically sound preservation?

"Forensically sound" preservation of electronic media preserves, in a reliable and authenticable manner, an exact copy of all active and residual data, including remnants of deleted data residing in unallocated clusters and slack space. When there are issues of data loss, destruction, alteration or theft, or when a computer is an instrumentality of loss or injury, computer forensics and attendant specialized preservation techniques may be required. Though skilled forensic *examination* can be expensive, forensically-sound *preservation* is relatively inexpensive. So, talk about the need for such efforts, and if your opponent won't undertake them, consider whether you should force forensic preservation by court order, even if you must bear the cost.

9. What metadata are relevant, and how will it be preserved, extracted and produced?

Metadata is evidence, typically stored electronically, that describes the characteristics, origins, usage and validity of other electronic evidence. There are all kinds of metadata found in various places in different forms. Some is supplied by the user, and some is created by the system. Some is crucial evidence, and some is just digital clutter. You will never face the question of whether a file has metadata—all active files do. Instead, the issues are what *kinds* of metadata exist, *where* it

resides and whether it's potentially *relevant* such that it must be preserved and produced. Understanding the difference—knowing what metadata exists and what evidentiary significance it holds--is an essential skill for attorneys dealing with electronic discovery.

The most important distinction is between *application metadata* and *system metadata.* The former is used by an application like Microsoft Word to embed tracked changes and commentary. Unless redacted, this data accompanies native production (that is, production in the form in which a file was created, used and stored by its associated application); but for imaged production, you'll need to ensure that application metadata is made visible before imaging or furnished in a useful form via a separate container called a "load file."

System metadata is information like a file's name, size, location, and modification date that a computer's file system uses to track and deploy stored data. Unlike application metadata, computers store system metadata outside the file. It's information essential to searching and sorting voluminous data and therefore it should be routinely preserved and produced.

Try to get your opponent to agree on the metadata fields to be preserved and produced, and be sure your opponent understands the ways in which improper examination and collection methods corrupt metadata values. Also discuss how the parties will approach the redaction of metadata holding privileged content.

10. What are the parties' data retention policies and practices?

A retention policy might fairly be called a destruction plan, and there's always a gap—sometimes a chasm—between an ESI retention policy and reality. The more onerous the policy, the greater ingenuity employees bring to its evasion to hang on to their e-mail and documents. Consequently, you can't trust a statement that ESI doesn't exist simply because a policy says it *should* be gone.

E-mail is a telling example. When a corporate e-mail system imposes an onerous purge policy, employees find ways to store messages on, *e.g.,* local hard drives, thumb drives, Cloud storage and personal accounts. Gone from the e-mail server rarely means gone for good.

Moreover, even companies that are diligent about rotating their backup tapes and that regularly overwrite old contents with new may retain complete sets of backup tapes at regular intervals. They also fail to discard obsolete tape formats when they adopt newer formats.

To meet their discovery obligations, parties may need to modify or suspend certain data retention practices. Discuss what they are doing and whether they will, as needed, agree to, *e.g.,* modify purge settings and ensure that systems of departing employees subject to a litigation hold aren't wiped or re-assigned without preservation.

11. Are there legacy systems to be addressed?

Computers and servers tend to stick around even if they've fallen off the organization's radar. That old laptop in someone's drawer can serve as a time tunnel back to evidence thought long gone. You should discuss whether potentially relevant legacy systems exist and how they will be identified and processed. Likewise, you may need to address what happens when a key custodian departs. Will the system be re-assigned, and if so, what steps will be taken to preserve potentially relevant ESI?

12. What are the current and prior e-mail applications?

E-mail systems are Grand Central Station for ESI. Understanding the current e-mail system and other systems used in the relevant past is key to understanding where evidence resides and how it can be identified and preserved. On-premises corporate e-mail systems are increasingly rare. Most companies deploy their e-mail systems online, in the cloud. Microsoft 365 predominates, though Google Workspace is also widely used. A changeover from an old system to a new system, or even from an old e-mail client to a new one, can result in a large volume of "orphaned" e-mail on media that would not otherwise be ripe for search.

13. Are personal e-mail accounts, mobile devices and computer systems involved?

Those who work from home, out on the road or from abroad may use personal e-mail accounts for business, exchange business texts on personal phones and store relevant ESI on their home or laptop machines or other portable devices. Parties should address the potential for relevant ESI to reside on personal and portable machines and devices and agree upon steps to be taken to preserve and produce that data. **With the post-pandemic prevalence of remote work, the lines between personal and business devices and accounts have blurred, making this inquiry essential.**

14. What electronic formats are common and in what anticipated volumes?

Making the right choices about how to preserve, search, produce and review ESI depends upon the forms and volume of data. Producing a Word document as a TIFF image may be acceptable where producing a native voice mail format as a TIFF is inconceivable. It's difficult to designate suitable forms for production of ESI when you don't know its native forms. Moreover, the tool you'll employ to review millions of e-mails is likely much different than the tool you'll use for thousands. If your opponent has no idea how much data they have or the forms it takes, encourage or compel them to use sampling of representative custodians to perform a "data biopsy" and gain insight into their collection.

15. How will we handle social networking, test messaging and other challenging ESI?

Producing parties routinely ignore electronic evidence like social networking posts and text messaging by acting too late to preserve it or deciding that the retention burden outweighs any benefit. *When it's relevant,* will the other side archive texts, voice mail messages, social networking

content, mobile device application content (*e.g.,* Teams, Slack, WhatsApp) or a host of other potentially relevant ESI that's often overlooked?

16. What relevant databases exist and how will their contents be discovered?

From R&D to HR and from finance to the factory floor, businesses run on databases. When they hold relevant evidence, you'll need to know the platform (e.g., SQL, Oracle, SAP, Salesforce) and how the data's structured (its "schema") before proposing sensible ways to preserve and produce it. Options include generating standard reports, running agreed queries, exporting relevant data to standard delimited formats or even (in the very rare case) mirroring the entire contents to a functional environment.

Database discovery is challenging and contentious, so know what you need and articulate why and how you need it. Be prepared to propose reasonable solutions that won't unduly disrupt operations.

17. Will paper documents be scanned, with what resolution, OCR and metadata?

Paper is still with us and ideally joins the deluge of ESI in ways that make it electronically searchable. Though parties are not obliged to convert paper to electronic forms, they commonly do so by scanning, coding and use of Optical Character Recognition (OCR). You'll want to ensure that paper documents are scanned so as to be legible and suited to OCR and are accompanied by information about their source (custodian, location, container, etc.) and logical unitization (i.e., foldering and stapled and clipped groupings).

18. Are there privilege issues unique to ESI?

Discussing privilege at meet and confer entails more than just agreeing to return items that slip through the net via so-called "claw back agreements" or a Federal Rules of Evidence Rule 502 agreement or order. It's important to surface practices that overreach. If the other side uses keywords to sidetrack potentially privileged ESI, are search terms absurdly overbroad? Simply because a document has the word "law" or "legal" in it or was copied to someone in the legal department doesn't justify its languishing in privilege purgatory. When automated mechanisms replace professional judgment concerning the privileged character of ESI, those mechanisms must be closely scrutinized and challenged when flawed.

Asserting privilege is a *privilege* that should be narrowly construed to protect either genuinely confidential communications exchanged for the purpose of seeking or receiving legal counsel or the thinking and strategy of counsel. Moreover, even documents with privileged content may

contain non-privileged material that should be parsed and produced. All the messages in a long thread aren't necessarily privileged because a lawyer got copied on the last one.[130]

Electronic evidence presents unique privilege issues for litigants, in part because of the potential for application metadata (like document comments and other collaboration features) to serve as communication tools. Comments and Tracked Changes aren't fundamentally different from e-mails discussing suggested amendments to documents, yet the former tend not to be reviewed or produced by defendants. Instead, some parties will, e.g., convert Word documents to TIFF images, suppressing the embedded communications as if they never occurred so as to avoid having to review them for privilege. If these communications exist and may be relevant, you must work to ensure this evidence is not ignored.

19. What search techniques will be used to identify responsive or privileged ESI?

Transparency of process is vitally important with respect to the mechanisms of automated search and filtering employed to identify or exclude information, yet opponents may resist sharing these details, characterizing it as work product. Privilege protects the terms and techniques facilitating an attorney's assessment of a case, but search and filtering mechanisms that effectively eliminate the exercise of attorney judgment (by excluding data as irrelevant) should be disclosed so that they may be tested and, if flawed, challenged. Likewise, if the producing party uses mechanized search to segregate data as privileged, the requesting party should be made privy to same in case it is inappropriately exclusive, though here, redaction may be appropriate to shield searches tending to reveal privileged information. Finally, use of advanced analytic techniques like technology-assisted review (TAR), including AI-powered continuous active learning (CAL) and large language model (LLM) tools, should be thoroughly explored to ensure that the processes employed are well-understood and, as feasible, the sampling and thresholds are mutually acceptable.

20. If keyword searching is contemplated, can the parties agree on keywords?

If you've been to Las Vegas, you know Keno, that game where you pick the numbers, and if enough of your picks light up on the board, you win. Keyword searching ESI is like that. The other side has you pick keywords and then goes off somewhere to run them. Later, they tell you they looked through the matches and, sorry, you didn't win. As a consolation prize, you may get the home game: a zillion jumbled images of non-searchable nonsense.

Perhaps because it performs so well in the regimented setting of online legal research, lawyers and judges invest too much confidence in keyword search. It's a seductively simple proposition: pick the words most likely to uniquely appear in responsive documents and then review for relevance and privilege just those documents containing the key words. Thanks to, *e.g.,* misspellings,

[130] See, e.g., *Muro v. Target Corporation*, 243 F.R.D. 301 (N.D. Ill. June 7, 2007) and *In re Vioxx Products Liability Litigation,* 501 F. Supp. 789 (E.D. La. Sept. 4, 2007)

acronyms, synonyms, IM-speak, noise words, OCR errors, indexing issues and peculiar industry lexicons, keyword search performs far below most lawyers' expectations, finding perhaps 20% of responsive material on first pass.[131]

Warts and all, keyword search remains a common method employed to tackle large volumes of ESI, and a method still enjoying considerable favor with courts.

Never allow your opponent to position keyword search as a single shot in the dark. You must be afforded the opportunity to use information gleaned from the first effort or subsequent efforts to narrow and target succeeding searches. The earliest searches are best used to acquaint both sides with the argot of the case. What shorthand references and acronyms did they use? Were products searched by their trade or technical names?"

Collaborating on search terms is optimum, but a requesting party must be wary of an opponent who, despite enjoying superior access to and understanding of its own business data, abdicates its obligation to identify responsive information. Beware of an invitation to "give us your search terms" if the plan is to review only documents "hit" by your terms and ignore the rest. Also, ensure that terms are tested on representative samples of ESI to ensure that search tools and queries are performing as expected. Be especially wary of stop word exclusions and documents whose textual content was not extracted and indexed.

21. How will deduplication be handled, and will data be re-populated for production?

ESI, especially e-mail, is characterized by enormous repetition. A message may appear in the mail boxes of thousands of custodians or be replicated dozens or hundreds of times through periodic backup. Deduplication is the process by which identical items are reduced to a single instance for purposes of review. Deduplication can be *vertical*, meaning the elimination of duplicates within a single custodian's collection, or *horizontal*, where identical items of multiple custodians are reduced to single instances.

Depending upon the review platform you employ, if production will be made on a custodial basis (person-by-person), it may be desirable to request re-population of content deduplicated horizontally, so each custodian's collection is complete. This will re-inject duplicates; however, each custodian's collection will be complete, witness-by-witness.

22. What forms of production are offered or sought?

[131] See, *e.g.*, The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery (2007) (describing the famous Blair and Maron study, which demonstrated the significant gap between the assumptions of lawyers that they would find 75% of the total universe of relevant documents, versus the reality that they had in fact found only 20% of the total relevant documents in a 40,000 document collection).

The Federal Rules give requesting parties the right to designate the form or forms in which ESI is to be produced. A responding party may object to producing the designated form or forms, but if the parties don't subsequently agree and the court doesn't order the use of particular forms, the responding party must produce ESI as it is ordinarily maintained or in a form that is reasonably usable. Moreover, responding parties may not simply dump undesignated forms on the requesting party, but must disclose the intended forms before making production to afford the requesting party the opportunity to ask the court to compel production in the designated form or forms.[132]

Options for forms of production include native file format, near-native forms (e.g., individual e-mail messages in MSG or EML formats), imaged production (PDF or more commonly, TIFF images accompanied by load files containing searchable text and metadata) and even paper printouts for very small collections. It is not necessary—and rarely advisable—to employ a single form of production for all items; instead, tailor the form to the data in a *hybrid* production. TIFF and load files may suffice for simple textual content like e-mail without attachments or word-processed documents, but native forms are best for spreadsheets, documents with pertinent application metadata (comments and tracked changes) and social media content. Native forms are essential for rich media, like animated PowerPoint presentations or audio and video files. Near-native forms are well-suited to e-mail and database exports.

A requesting party uncertain of what he needs plays into the other side's hands. You must be able to articulate both what you seek *and the form in which you seek it*. The native forms of ESI dictate the optimum forms for its production, but rarely is there just one option. The alternatives entail tradeoffs, typically sacrificing utility or searchability of electronic information to make it function more like paper documents. Before asking for anything, know how you'll house, review and use it. That means "know your review platform."[133] That is, know the needs and capabilities of the applications or tools you'll employ to index, sort, search and access electronic evidence.

---

[132] Fed. R. Civ. P. 34(b)

[133] If a question about your review platform gives you that deer-in-headlights look, you're probably not ready for meet and confer. Even if you're determined to look at every page of every item they produce, you'll still need a system to view, search and manage electronic information. If you wait until the data start rolling in to pick your platform, you're likely to get ESI in forms you can't use, meaning you'll have to expend time and money to convert them. Knowing your intended platform allows you to designate proper load file formats and determine if you can handle native production.

Choosing the right review platform for your practice requires understanding your work flow, your people, the way you'll search ESI and the forms in which the ESI will be produced. *You should not use native applications to review native production in e-discovery.* Instead, a platform geared to review of ESI in native formats--one able to open the various types of data received without corrupting its content or metadata--should be employed. ESI can be like Russian nesting dolls in that a compressed backup file (.BKF) may hold an encrypted Outlook e-mail container (.PST) that houses a message transmitting a compressed archive (.ZIP) attachment containing an Adobe portable document (.PDF). Clearly, a review platform needs to be able to access the textual content of compressed and proprietary formats and drill down or "recurse" through all the nested levels.

Finally, don't let your opponent confuse the medium of production with the form of production. Telling you that the data is coming on a thumb drive tells you nothing about what data you're getting.

23. How will you handle redaction of privileged or confidential content?

Producing parties often seek to redact ESI in the way they once redacted paper documents: by blacking out text. To make that possible, ESI is converted to non-searchable TIFF images in a process that destroys electronic searchability. So after redaction, electronic searchability must be restored by using OCR to extract text from the TIFF image.

A TIFF-OCR redaction method works reasonably well for text documents, but it fails miserably when applied to complex and dynamic documents like spreadsheets and databases. Unlike text, you can't spell check numbers, so the inevitable errors introduced by OCR make it impossible to have confidence in numeric content or reliably search the data. Moreover, converting a spreadsheet to a TIFF image strips away its essential functionality by jettisoning the underlying formulae that distinguishes a spreadsheet from a table.

For common productivity applications like Adobe Acrobat and Microsoft Office, it's feasible and cost-effective to redact natively so as to preserve the integrity and searchability of evidence; consequently, where it's important to preserve the integrity and searchability of redacted documents, you should determine what redaction methods are contemplated and seek to agree upon methods best suited to the task. At all events, redaction tends to implicate a relatively small population of information items in a production; so, don't let the preferred method of redaction adversely impact the form or forms of production employed for items not requiring redaction. That is, *don't let the redaction tail wag the production dog.*

24. Will load files accompany document images, and how will they be populated?

Converting ESI to TIFF images strips the evidence of its electronic searchability and metadata. Accordingly, load files accompany TIFF image productions to hold searchable text and selected metadata. Load files are constructed of delimited text, meaning that values in each row of data follow a rigid sequence and are separated by characters like commas, tabs or quotation marks. Using load files entails negotiating their organization or specifying the content and the use of a structure geared to review software.

There are many review platforms on the market, including Cloud-hosted review environments like Relativity and DISCO, and proprietary platforms marketed by e-discovery service providers touting more bells and whistles than a Mardi Gras parade.

Review platforms can be cost-prohibitive for some practitioners. If you don't currently have one in-house, your case may warrant hiring a vendor offering a hosted platform suited to the ESI. When tight budgets make even that infeasible, employ whatever productivity tools you can cobble together on a shoestring. You may have to forego the richer content of native production in favor of paper-like forms such as Tagged Image File Format (TIFF) images because you can view them in a web browser.

25. How will the parties approach file naming and Bates numbering?

It's common for file names to change to facilitate unique identification when ESI is processed for review and production. Assigned names may reflect, *e.g.*, unique values derived from a data fingerprinting process called hashing or contain sequential control numbers tied to a project management database. Native productions don't lend themselves to conventional paged formats, so aren't suited to embossed Bates numbering on a page-by-page basis; however, this is no impediment to native production in that Bates numbers can serve as filenames for native files, with page numbers embossed on the items only when converted to paged formats for use in proceedings.

26. What ESI will be claimed as not reasonably accessible, and on what bases?

Pursuant to Rule 26(b)(2)(B) of the Federal Rules of Civil Procedure, a litigant must show good cause to discover ESI that is "not reasonably accessible," but the burden of proving a claim of inaccessibility lies with the party resisting discovery. So, it's important that your opponent identify the ESI it claims is not reasonably accessible and furnish sufficient information about that claim to enable you to gauge its merit.

The meet and confer is an opportune time to resolve inaccessibility claims without court intervention—to work out sampling protocols, cost sharing and filtering strategies—or when agreements can't be reached, at least secure commitments that the disputed data will be preserved long enough to permit the court to resolve issues.

27. Can costs be minimized by shared providers, neutral experts or special masters?

Significant savings may flow from sharing costs of e-discovery service providers and online repositories, or by eliminating dueling experts in favor of a single neutral expert for thorny e-discovery issues or computer forensics. Additionally, referral of issues to a well-qualified ESI Special Master can afford the parties speedier resolution and more deliberate assessment of technical issues than a busy docket allows.

Endgame: Transparency of Process and Cooperation

Courts and commentators uniformly cite the necessity for transparency and cooperation in electronic discovery, but old habits die hard. Too many treat meet and confer as a perfunctory exercise, reluctant to offer a peek behind the curtain. Some are paying dearly for their intransigence, sanctioned for obstructive conduct or condemned to spend obscene sums chasing data that might never have been sought had there been communication and candor. Others are paying attention and have begun to understand that candor and cooperation in e-discovery isn't a sign of weakness, but a hallmark of professionalism.

The outsize cost and complexity of e-discovery will diminish as electronic records management improves and ESI procedures become standardized, but the meet and confer process is likely to endure and grow within federal and state procedure. Accordingly, learning to navigate meet and confer—to consistently ask the right questions and be ready with the right answers—is an essential advocacy skill.



Exercise 23: AI-Assisted Review

GOALS: The goals of this exercise are for the student to:

1. Explore the promise of AI and LLMs to identify relevant and responsive ESI

OUTLINE: Students will receive a handout in Canvas describing the tasks and tools to be employed in this exercise and perform hands-on tasks to critically assess AI-generated outputs.

Background

From Wikipedia 3/17/24:

> "During her tenure as United States Secretary of State, Hillary Clinton drew controversy by using a private email server for official public communications rather than using official State Department email accounts maintained on federal servers. After a years-long FBI investigation, it was determined that Clinton's server did not contain any information or emails that were clearly marked classified. Federal agencies did, however, retrospectively determine that 100 emails contained information that should have been deemed classified at the time they were sent, including 65 emails deemed "Secret" and 22 deemed "Top Secret". An additional 2,093 emails were retroactively designated confidential by the State Department."
>
> "Some experts, officials, and members of Congress contended that Clinton's use of a private email system and a private server violated federal law, specifically 18 U.S. Code § 1924, regarding the unauthorized removal and retention of classified documents or materials, as well as State Department protocols and procedures, and regulations governing recordkeeping. Clinton claimed that her use complied with federal laws and State Department regulations, and that former secretaries of state had also maintained personal email accounts (however Clinton was the only secretary of state to use a private server)…."
>
> "The controversy was a major point of discussion and contention during the 2016 presidential election, in which Clinton was the Democratic nominee. In May, the State Department's Office of the Inspector General released a report about the State Department's email practices, including Clinton's. In July, FBI director James Comey announced that the FBI investigation had concluded that Clinton had been "extremely careless" but recommended that no charges be filed because Clinton did not act with criminal intent, the historical standard for pursuing prosecution."
>
> "On October 28, 2016, eleven days before the election, Comey notified Congress that the FBI had started looking into newly discovered emails. On November 6, Comey notified Congress that the FBI had not changed its conclusion. Comey's timing was contentious,

with critics saying that he had violated Department of Justice guidelines and precedent, and prejudiced the public against Clinton. The controversy received more media coverage than any other topic during the presidential campaign. Clinton and other observers argue that the reopening of the investigation contributed to her loss in the election."…

"At the time of Senate confirmation hearings on Hillary Clinton's nomination as Secretary of State, the domain names clintonemail.com, wjcoffice.com, and presidentclinton.com were registered to Eric Hoteham, with the home of Clinton and her husband in Chappaqua, New York, as the contact address. The domains were pointed to a private email server that Clinton (who never had a state.gov email account) used to send and receive email, and which was purchased and installed in the Clintons' home for her 2008 presidential campaign."

"The email server was located in the Clintons' home in Chappaqua, New York, from January 2009 until 2013, when it was sent to a data center in New Jersey before being handed over to Platte River Networks, a Denver-based information technology firm that Clinton hired to manage her email system."

Problem: You are an associate attorney working as John Podesta's counsel. A Request for Production received in connection with civil litigation states:

Request for Production

For the period January 2009 through March 2016, produce any and all documents and communications in John Podesta's care, custody or control that discuss, describe or otherwise relate to the existence, operation or contents of a private email server used by or for Hillary Rodham Clinton (HRC) and believed to have been situated in the basement of her home in Chappaqua, New York.

Your supervisor, the senior partner working on the case, gives you the following direction:

"I want you to find all the material in Podesta's email and attachments responsive to the request so I can review them. Don't worry about excluding privileged content; I'll deal with that. Don't miss anything responsive, but what I really don't want is a lot of non-responsive stuff wasting my time! You can't review every message manually but don't rely on keyword search either; I want you to use those fancy Gen AI tools to save your time and our client's money!"

Here's the query you should first try (verbatim from the Request for Production):

1) For the period January 2009 through March 2016, produce any and all documents and communications in John Podesta's care, custody or control that discuss, describe or otherwise relate to the existence, operation or contents of a private email server used

by or for Hillary Rodham Clinton (HRC) and believed to have been situated in the basement of her home in Chappaqua, New York.

Critique AI's performance on this query and, if you find the query to be insufficient to the task, edit the query to improve its performance using Harvey (or the LLM of your choice) (including, if you see a need for same, suggesting that it be used in concert with any filters, text searches or other DISCO skills you've acquired). Supply the better query.

***In class discussion*, be prepared to address:**

- **What specific types of legal discovery tasks do AI/LLMs appear best suited for?**
- **Where do Harvey and similar AI tools show significant limitations or potential risks?**



Exercise 24: Meet and Confer

Students will form teams representing the plaintiff or defendant in a hypothetical case styled, ***Lost Creek Engineering, LLC v. Keith Austin Weird and Artemis Energy Solutions, Inc.,*** pending in a United States District Court. The nature of the case is described in Appendix A to this workbook.

You will prepare for and engage in an FRCP Rule 26(f) meet and confer process with an opposing team. Teams should be prepared to answer questions that should be anticipated in a meet and confer and assess issues and information of importance to your client(s), most particularly on those points that must be addressed and reported to the Court pursuant to FRCP Rule 16 and Rule 26(f). Each side will be privy to information not known to their opponent that will influence how to proceed and the proper level of transparency and cooperation to offer and expect. In the classroom session, the teams will appear at a scheduling conference (i.e., a hearing) before the judge where they will demonstrate their ability to present and explain the discovery plan and expertly and succinctly present unresolved issues to the court for resolution.

Special Instructions

Your team will receive a confidential plaintiff or defendant briefing. ***You are not to share the contents of this briefing with anyone other than your own team.*** You should not furnish or display same to a member of any other team nor should you look at an opponent's briefing, if available to you. Acting in the best interests of your client(s) and consistent with your ethical duties, you may disclose information gleaned from the briefing in the meet and confer process *only as legal requirements, good practice and sound strategy dictate*.

The following provisions of the Federal Rules of Civil Procedure govern the conferences with your opponents and the Court. ***You should focus on the aspects of the process that bear on electronically stored information. You should not devote significant time to the merits of the action or to procedural matters that do not bear on e-discovery.***

Rule 26. Duty to Disclose; General Provisions; Governing Discovery

…

(f) Conference of the Parties; Planning for Discovery.

(1) *Conference Timing.* Except in a proceeding exempted from initial disclosure under Rule 26(a)(1)(B) or when the court orders otherwise, the parties must confer as soon as practicable—and in any event at least 21 days before a scheduling conference is to be held or a scheduling order is due under Rule 16(b).

(2) *Conference Content; Parties' Responsibilities.* In conferring, the parties must consider the nature and basis of their claims and defenses and the possibilities for promptly settling or resolving the case; make or arrange for the disclosures required by Rule 26(a)(1); discuss any issues about preserving discoverable information; and develop a proposed discovery plan. The attorneys of record and all unrepresented parties that have appeared in the case are jointly responsible for arranging the conference, for attempting in good faith to agree on the proposed discovery plan, and for submitting to the court within 14 days after the conference a written report outlining the plan. The court may order the parties or attorneys to attend the conference in person.

(3) *Discovery Plan.* A discovery plan must state the parties' views and proposals on:

(A) what changes should be made in the timing, form, or requirement for disclosures under Rule 26(a), including a statement of when initial disclosures were made or will be made;

(B) the subjects on which discovery may be needed, when discovery should be completed, and whether discovery should be conducted in phases or be limited to or focused on particular issues;

(C) any issues about disclosure, discovery, or preservation of electronically stored information, including the form or forms in which it should be produced;

(D) any issues about claims of privilege or of protection as trial-preparation materials, including—if the parties agree on a procedure to assert these claims after production—whether to ask the court to include their agreement in an order under Federal Rule of Evidence 502;

(E) what changes should be made in the limitations on discovery imposed under these rules or by local rule, and what other limitations should be imposed; and

(F) any other orders that the court should issue under Rule 26(c) or under Rule 16(b) and (c).

Rule 16. Pretrial Conferences; Scheduling; Management

(a) Purposes of a Pretrial Conference. In any action, the court may order the attorneys and any unrepresented parties to appear for one or more pretrial conferences for such purposes as:

(1) expediting disposition of the action;

(2) establishing early and continuing control so that the case will not be protracted because of lack of management;

(3) discouraging wasteful pretrial activities;

(4) improving the quality of the trial through more thorough preparation; and

(5) facilitating settlement.

(b) Scheduling.

(1) *Scheduling Order.* Except in categories of actions exempted by local rule, the district judge—or a magistrate judge when authorized by local rule—must issue a scheduling order:

(A) after receiving the parties' report under Rule 26(f); or

(B) after consulting with the parties' attorneys and any unrepresented parties at a scheduling conference.

(2) *Time to Issue.* The judge must issue the scheduling order as soon as practicable, but unless the judge finds good cause for delay, the judge must issue it within the earlier of 90 days after any defendant has been served with the complaint or 60 days after any defendant has appeared.

(3) *Contents of the Order.*

(A) *Required Contents.* The scheduling order must limit the time to join other parties, amend the pleadings, complete discovery, and file motions.

(B) *Permitted Contents.* The scheduling order may:

(i) modify the timing of disclosures under Rules 26(a) and 26(e)(1);

(ii) modify the extent of discovery;

(iii) provide for disclosure, discovery, or preservation of electronically stored information;

(iv) include any agreements the parties reach for asserting claims of privilege or of protection as trial-preparation material after information is produced, including agreements reached under Federal Rule of Evidence 502;

(v) direct that before moving for an order relating to discovery, the movant must request a conference with the court;

(vi) set dates for pretrial conferences and for trial; and

(vii) include other appropriate matters.

(4) *Modifying a Schedule.* A schedule may be modified only for good cause and with the judge's consent.

APPENDIX A: Materials for use with Exercises 16 and 24

Lost Creek Engineering, LLC v. Keith Austin Weird and Artemis Energy Solutions, Inc.

All of the events and persons described in this hypothetical scenario are fictional. Any resemblance to persons, living or dead, or to business entities is purely coincidental.

This hypothetical case concerns the alleged misappropriation of intellectual property by a senior design engineer at an engineering company. The engineer, **Keith Austin Weird** , worked for Lost Creek Engineering, LLC for 17 years, rising to the position of Assistant Vice-President of Engineering. Weird led the design and development of Lost Creek's very profitable Arnold™ line of intelligent pipeline pigs, as well as a yet-to-be-introduced line of next generation products codenamed "When Pigs Fly."

Pigs, in the context of pipelines, are devices inserted into pipelines that travel with the flowing content for the purpose of conducting inspection, maintenance, product separation and other functions. Pipeline pigs must operate under conditions of high pressure, extreme temperatures and highly corrosive conditions. Intelligent or "smart" pigs are sophisticated robots that, until now, have been required to operate autonomously because the radio-blocking character of steel pipelines and the enormous distances traversed made it infeasible for pigs to communicate with remote operators or GPS satellites.

Lost Creek's "When Pigs Fly" innovation was the pairing of its smart pigs with an accompanying drone aircraft outside the pipeline. The innovation employs proprietary technology to enable high-bandwidth, multichannel ultrasonic communications between pig and drone, allowing a distant operator to see real time data and video from the pig, obtain precise GPS coordinates and remotely control the pig. Precise location data means that repair crews operate more efficiently and at lower cost. Real time remote control permits complex repairs to be accomplished without the risk and cost of dispatching crews and heavy equipment to distant work sites.

Weird was hired by former Lost Creek V.P. of Engineering and Development, **Montgomery Bonnell** in 2007. Weird reported directly to Bonnell for the decade that both worked together at Lost Creek. The two are close friends, and their families frequently socialize outside of work. In 2017, Bonnell left Lost Creek to found Artemis Energy Solutions, Inc. in Houston. Artemis manufactures and sells pipeline telemetry products to the energy sector. Weird sought to be considered for Bonnell's position but was told he was too valuable in his current position and encouraged to acquire some

managerial seasoning. When an outsider was brought in to replace Bonnell, Weird was assured by the CEO that his desire to advance would not be forgotten. Bonnell's replacement left Lost Creek at the start of 2025, and Weird learned that management contacted a headhunter to fill the position.

With no promotion forthcoming, Weird resigned from Lost Creek on February 14, 2025. He gave two weeks' notice and noted that, now that his kids were in college, he was heading to Houston to work for his old friend, Monty Bonnell, at Artemis Energy Solutions, Inc. Weird participated in a required exit interview, confirmed his familiarity with all Lost Creek polices impacting departing employees, and received a generous severance package to resolve unused vacation time and other benefits. Weird's last day at Lost Creek was February 28, 2025 , and he took two weeks off before starting at Artemis. Weird joined Artemis as its Executive VP of Technology.

On June 25, 2025 , Lost Creek's outside counsel, Lamar Street, sent letters to Weird and Bonnell invoking the Non-Disclosure Agreement and Covenant Not to Compete Weird signed when first hired by Lost Creek. Lost Creek demanded that Weird cease work for Artemis on anything involving pipeline pigs or telemetry. The letter to Weird also sought return of Weird's Lost Creek laptop and access to all of Weird's personal computers, digital media and e-mail accounts for the purpose of conducting an examination to assess compliance.

During May of 2025, three Lost Creek engineers, **Percy Pennybacker, Claudia Johnson and Barton Springs** , tendered their resignations. All had worked under Weird at Lost Creek in the development and testing of intelligent pipeline pigs. All joined Artemis and all once more report to Weird.

In October of 2025, Artemis' internal SharePoint newsletter announced that the company would be introducing the AirHog™ line of sophisticated intelligent pipeline drone pigs that, by the description of their capabilities, would mirror the capabilities of Lost Creek's yet-to-be-introduced When Pigs Fly technology. The article offered rosy financial projections for the new product line, prompting a blizzard of Tweets and texts between Artemis employees, Lost Creek employees and industry insiders.

On October 15, 2025, Lost Creek filed suit against Weird and Artemis in the Western District of Texas seeking injunctive relief and damages on three counts:
Count 1 – Misappropriation of Trade Secrets under the Defend Trade Secrets Act, 18 U.S.C. § 1836
Count 2 – Misappropriation of Trade Secrets under the Texas Uniform Trade Secrets Act, Tex. Civ. Prac. & Rem. Code § 134A.001 et seq. and
Count 3 – Breach of Non-Disclosure Agreement and Covenant Not to Compete
The Defendants answered, asserting various affirmative defenses.

Lost Creek has been in business for over 40 years. It is headquartered in Austin, Texas and maintains manufacturing sales and service centers in China, Australia and Europe, as well as representatives and technicians in more than 20 countries. Lost Creek is a closely held company that employs over 400 people, including 40+ persons in its Product Development and Engineering Division. Its sales and earnings figures are not made public.

Artemis Energy Solutions, Inc. was formed in 2016 and is headquartered in Houston, Texas. Artemis employed 150 people as of December 31, 2025 , and projected gross annual sales of approximately $75 million for 2025. In January of 2026 , Artemis was acquired by Prytania Oil, S.A., a conglomerate headquartered in Greece, and Artemis became a wholly-owned foreign subsidiary of Prytania Oil, S.A.



Timeline of Events

September 4, 2007: Keith Austin Weird hired by Lost Creek; executes Non-Disclosure Agreement and Covenant Not to Compete

February 10, 2025: Weird receives job offer from Artemis and copies Lost Creek data to an external hard disk drive

February 14, 2025: Weird tenders his resignation to Lost Creek

February 28, 2025: Weird's last day at Lost Creek; exit interview

March 1 – March 14, 2025: Weird on vacation
March 17, 2025: Weird's first day at Artemis
May 2025: Three Lost Creek engineering employees quit to join Artemis
June 25, 2025: Demand for return of Weird's Lost Creek laptop and to inspect his e-mail, home systems, hard drives and thumb drives
October 3, 2025: Artemis announces forthcoming AirHog™ product line
October 15, 2025: Original Complaint filed
November 19, 2025: Original Answer filed
December 15, 2025: Amended Complaint Filed
December 17, 2025: Amended Answer filed
January 15, 2026: Prytania Oil, S.A. acquires all shares in Artemis

TARRYTOWN, OLDE & RICH, L.L.P
LAWYERS
1313 Guadalupe
Suite 1900
Austin, Texas 78701

Lamar Street **(512) 555-6066**
Partner

June 25, 2025

Keith Austin Weird
200 Congress Avenue, Apt. 5701 ***Via Hand Delivery***
Austin, TX 78701

The undersigned represents the legal interests of Lost Creek Engineering, LLC ("Lost Creek" or the "Company"). As you know, in connection with your employment with Lost Creek, you were given specialized training and were provided with certain of the Company's confidential, proprietary, and trade secret information. You expressly acknowledged this in Non-Disclosure Agreement and Covenant Not to Compete (the "Agreement"). A copy of the Agreement is enclosed for your reference.

Additionally, your contract of employment includes an agreement to refrain from working for a competitive business following the termination of your employment from Lost Creek. In the Agreement you promised that, for a period of two (2) years following your termination from Lost Creek, you would not engage in or work for any business in direct competition with Lost Creek by manufacturing and/or selling intelligent pipeline pigs that resemble or imitate the pipeline pigs manufactured and sold by Lost Creek. See Agreement at 1.

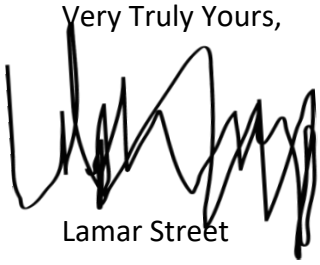
In your letter of resignation dated February 14, 2025, you indicated that you would be taking a position with Artemis Energy Solutions, Inc. as Technology Director-Pipeline Products. Although, Lost Creek does not consider Artemis to be directly competitive with its interests. Any work by you in support of the design and production of intelligent pipeline pigs is in direct competition with Lost Creek and in direct violation of the Agreement. As we now understand that your work with Artemis will be in research and development in remote sensing pipeline repair devices, a technical knowledge that you gained exclusively during your tenure at Lost Creek, the purpose of this correspondence is to notify you of your breach of the Agreement and demand that you cease your intent to continue employment with Artemis and refrain from doing so for a period of two (2) years. We also remind you that your agreement to protect confidential information that belongs to Lost Creek is not limited in any timeframe and is your obligation regardless of employment status.

On behalf of Lost Creek, we demand immediate return of all files, materials, information, technology or other property owned by Lost Creek which may be in your possession. To be assured that you have complied with this request, we request that you deliver your Lost Creek laptop, any home computer(s) and any external drives and thumb drives to our forensic examiner (see attached business card) for forensic review of the hard drives and external drives to assure that no confidential information or property of Lost Creek resides on any drive. We further request that you make the contents of any e-mail or webmail accounts you have used within the last two (2) years available to us for inspection and copying. We will also seek confirmation that you have not distributed or transferred any such information to any third-party including Artemis Energy Solutions, Inc. or any other manufacturer in the pipeline pig industry. Lost Creek will withhold the six (6) month severance pay provided in your Agreement pending compliance with this request.

In further effort to assure compliance with these post-employment requirements of you, Lost Creek has asked that you complete and sign the enclosed verification which confirms your representations that you do not have any information which could be considered confidential information belonging to Lost Creek.

Know that Lost Creek must and will protect its legal interests. Failure to immediately cease your employment with Artemis Energy Solutions, Inc. and provide the undersigned with satisfactory notice thereof will require the Company to take action to protect its legal interests. Such action will include the immediate imposition of suit against you to enforce the Agreement. In addition to the actual damages caused by your breach of the Agreement, Lost Creek will seek recovery of its attorneys' fees, costs, and interest. Please provide me with the requisite notice of termination of employment with Artemis Energy Solutions, Inc. at your earliest convenience and evidence of your compliance with the request that you deliver your computers to our forensic examiner.

Very Truly Yours,



Lamar Street

NON-DISCLOSURE AGREEMENT AND
COVENANT NOT TO COMPETE

Lost Creek Engineering, LLC ("Lost Creek") hereby promises that, upon Keith Austin Weird's acceptance of employment with Lost Creek, Lost Creek will provide Keith Austin Weird with specialized training unique to it and not otherwise available in the industry or elsewhere. Further, Lost Creek promises to provide Keith Austin Weird with information it holds as confidential, as well as certain trade secret information relating to intelligent pipeline pigs designed, manufactured and sold by Lost Creek, including access to certain privileged materials.

During the term of employment and without limitation thereafter, Keith Austin Weird hereby covenants and agrees to keep strictly confidential all knowledge to which he gains by virtue of his employment with Lost Creek. This includes all trade secrets, business practices, finances, documents, blueprints, market data, other intellectual property and other confidential information. Keith Austin Weird agrees not to disclose the above mentioned confidential information, directly or indirectly to any other person, company or corporation, or use it for his own benefit. Keith Austin Weird agrees that he will only use the confidential information as an employee of Lost Creek.

All confidential or trade secret information relating to the business of Lost Creek which Keith Austin Weird shall develop, conceive, produce, construct or observe during his employment with Lost Creek shall remain the sole property of Lost Creek.

Keith Austin Weird further agrees that upon termination of his employment, Keith Austin Weird will surrender and deliver to Lost Creek all confidential information, including but not limited to work papers, books, records, and data of every kind relating to or in connection with Lost Creek.

Keith Austin Weird agrees, upon termination of employment with Lost Creek and for a period of two (2) years thereafter, Keith Austin weird will not directly or indirectly engage in any business or work for any business which is in direct competition with Lost Creek by manufacturing and/or selling pipeline pigs that resemble or imitate the pipeline pigs manufactured and sold by Lost Creek. Keith Austin Weird agrees that this paragraph prohibits him from accepting employment on a worldwide basis with any pipeline pig manufacturer for the two (2) year period.

Executed this 4th day of September, 2007

Signature

9/4/2007

Today's Date

The State of Texas }
}
County of Travis }

BEFORE ME the undersigned authority, on this day personally appeared Keith A. Weed known to me to be the person whose name is subscribed to the foregoing instrument, and acknowledged to me that he had executed same in the capacities and for the purposes and consideration therein expressed.

GIVEN UNDER MY HAND AND SEAL OF OFFICE THIS 4th day of September, 2007

JANE DOE
Notary Public, State of Texas
Commission Expires 05-15-2007

Notary Public in the State of Texas

VERIFICATION

My name is Keith Austin Weird. I have been employed with Lost Creek Engineering, LLC ("Lost Creek") as Vice-president of Engineering and later Chief of Engineering since September 4, 2007. I have separated from employment with Lost Creek effective February 28, 2025. In connection with my separation, I have been asked to represent and warrant that I am in compliance with certain agreements related to my employment. Accordingly, I represent and warrant that:

I am aware of my obligations under that certain agreement dated September 4, 2007 entitled *Non-Disclosure Agreement and Covenant Not to Compete (*the "Agreement") and agree to comply with my obligations under the Agreement to the fullest extent possible. I understand and agree that confidential information and trade secrets includes all trade secrets, customer and vendor information, business practices, finances documents, blueprints, market data, other intellectual property relating to Lost Creek's work in the Pipeline pig industry, including remote sensing pipeline repair devices. I acknowledge that all information regarding remote sensing pipeline repair devices I have has been gained during my tenure with Lost Creek. I have not removed any confidential information or trade secrets from Lost Creek at any time during my employment. If I have any confidential information or trade secrets in my possession in written or electronic form, I will return it to Lost Creek immediately and no later than Friday, July 18, 2025.

I have not transferred any confidential information or trade secrets to any third party prior to my departure from Lost Creek. I agree to provide any computer and all external drives or devices, including jump drives, in my possession or use at home or elsewhere to Lost Creek's designated agent for forensic review on or before July 18, 2025 or at such time as Lost Creek directs for the purpose of verifying removal of all confidential information belonging to Lost Creek from such computer. I further consent to allow Lost Creek's designated agent to access and copy any personal e-mail or webmail account I have used for the last two (2) years.

Date: _____ _____

Keith Austin Weird

TARRYTOWN, OLDE & RICH, L.L.P
LAWYERS
1313 Guadalupe
Suite 1900
Austin, Texas 78701
Lamar Street **(512) 555-6066**
Partner

June 30, 2025

Montgomery Bonnell
Chief Executive Officer
Artemis Energy Solutions, Inc.
One Big Oil Boulevard ***Via Hand Delivery***
Houston, TX 77041

RE: Lost Creek Engineering, LLC.

Dear Mr. Bonnell:

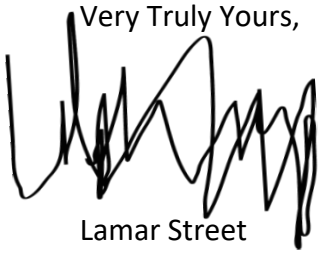
We are counsel to Lost Creek Engineering, LLC ("Lost Creek"). We have been advised of the fact that Keith Austin Weird has been offered employment with Artemis Energy Solutions, Inc. or one of its affiliates ("Artemis"). Lost Creek has recently learned that Mr. Weird's employment may involve research and development of intelligent pipeline pigs and/or remote sensing repair tools. If so, Mr. Weird would be performing the same (if not identical) services for Artemis as he performed for Lost Creek. We are writing, in part, to give you notice that Mr. Weird is subject to a prohibition from employment with a competitor of Lost Creek. A copy of Mr. Weird's agreement with Lost Creek is enclosed for your review. We are concerned that Mr. Weird's employment with Artemis may be in violation of the non-competition agreement and request your assistance in assuring his compliance with it.

Lost Creek is further concerned with Mr. Weird's compliance with his agreement to protect confidential information belonging to Lost Creek. He possesses confidential information from Lost Creek' files and may not use such information in connection with his employment with Artemis. Mr. Weird's entire knowledge regarding the intelligent pipeline pig industry has been gained during his employment with Lost Creek and we believe that, even with the best of intentions, it would be impossible for him to work in research and development regarding pipeline pigs without using Lost Creek's confidential information in violation of his agreement to protect it. We therefore suggest to you that any employment of Mr. Weird which is in violation of his non-compete agreement or

work performed by Mr. Weird and which may cause him to disclose confidential information belonging to Lost Creeks could result in legal action on behalf of Lost Creek. We trust that Artemis will work with Lost Creek to assure that there are no violations of the agreements or of other laws.

Please contact me if you have any questions.

Very Truly Yours,



Lamar Street



IN THE UNITED STATES DISTRICT COURT
WESTERN DISTRICT OF TEXAS

LOST CREEK ENGINEERING, L.L.C.	**§**	
	§	
V.	**§**	**CIVIL ACTION NO.**
	§	**5:18-CV-01234**
KEITH AUSTIN WEIRD,	**§**	
and	**§**	**JURY REQUESTED**
ARETEMIS ENERGY SOLUTIONS, INC.	**§**	
	§	

AMENDED COMPLAINT

TO THE HONORABLE JUDGE OF SAID COURT:
COMES NOW, Lost Creek Engineering, L.L.C., hereinafter referred to as Plaintiff or "Lost Creek," complaining of Keith Austin Weird and Artemis Energy Solutions, Inc. ("Artemis"), hereinafter referred to as Defendants, and for cause of action would respectfully show unto the Court and jury as follows:

I. PARTIES

1. Plaintiff is a corporation with an office in Travis County, Texas, and which has authority to do business in the State of Texas.

2. Defendant Keith Austin Weird has been served and answered.

3. Defendant, Artemis Energy Solutions, Inc., ("Artemis") is a foreign corporation doing business in Texas and has been served and answered.

II. VENUE AND JURISDICTION

4. This Court has federal question jurisdiction pursuant to 28 U.S.C. § 1331 over Plaintiff's claims under the Defend Trade Secrets Act, 18 U.S.C. § 1836 et seq., and the Computer Fraud and Abuse Act, 18 U.S.C. § 1030. This Court has supplemental jurisdiction pursuant to 28 U.S.C. § 1367 over Plaintiff's state law claims arising from the same nucleus of operative facts.
5. Venue is proper in the Western District of Texas because Weird resides in Travis County, Texas, and because a substantial part of the events or omissions giving rise to the claims plead below occurred in Travis County, Texas.

III. FACTUAL BACKGROUND

6. Lost Creek Engineering is a manufacturer of specialized tools for the inspection, maintenance and repair of petroleum and natural gas pipelines. Sophisticated and sensitive in-line inspection (ILI) tools travel through the pipe and measure and record irregularities that may represent corrosion, cracks, laminations, deformations or other defects. Lost Creek is a world leader in

the design, development and sale of pipeline smart pigs, robots designed to pass through pipelines performing specialized tasks in highly challenging environments. Lost Creek's Arnold™ line of smart pigs employ proprietary state-of-the-art magnetic flux and ultrasonic sensing devices and high-definition imagery in ways that uniquely distinguish Lost Creek's products in the marketplace.

7. Typically, smart pigs are inserted into the pipeline at a location, such as a valve or pump station, that has a special configuration of pipes and valves where the tool can be loaded into a receiver, the receiver can be closed and sealed, and the flow of the pipeline product can be directed to launch the tool into the main line of the pipeline. A similar setup is located downstream, where the tool is directed out of the main line into a receiver, the tool is removed, and the recorded data retrieved for analysis and reporting. Historically, smart pigs have been required to operate autonomously because the radio-blocking "Faraday cage" character of steel pipelines and the enormous distances traversed made it infeasible for pigs to communicate with remote operators or GPS satellites.

8. In utmost secrecy and through its investment of large sums of time and money, Lost Creek developed a unique and innovative technology to enable remote control and geolocation of the next generation of smart pig technology. Lost Creek's Project "When Pigs Fly" (WPF) innovation was the pairing of its smart pigs with an accompanying drone aircraft outside the pipeline. The innovation employs proprietary technology to enable high-bandwidth, multichannel ultrasonic communications between pig and drone, allowing a distant operator to see real time data and video from the pig, obtain precise GPS coordinates and remotely control the pig. Precise location data means that repair crews operate more efficiently and at lower cost. Real time remote control permits complex repairs to be accomplished without the risk and cost of dispatching crews and heavy equipment to distant work sites.

9. The design of Lost Creek's WPF of smart pigs has been a time-consuming and expensive process. Lost Creek continually tests, researches and improves the components, materials, designs and manufacturing processes of its products. It has taken years of field tests, experiments, research and development for Lost Creek to develop the unique technologies it is poised to market to customers. There are specific design characteristics of Lost Creek's smart pigs that are not used by other smart pig manufacturers and are not found in the open market. Such unique design characteristics include the following: (1) high-bandwidth, multichannel ultrasonic communications hardware, circuits and software; (2) Drone control and synchronization programming; (3) image and data compression algorithms; and (4) associated tools for inspection, optimization, deployment and operation of WPF drone/pig pairs.

10. These unique design characteristics were discovered and innovated by Lost Creek's engineers over the past ten years through testing, research and experience. It is these design characteristics that differentiate Lost Creek's smart pigs from other smart pigs on the market.

11. In order to design, test and (ultimately) manufacture smart pigs with WPF capabilities for its customers, Lost Creek uses specialized designs, test mechanisms, source code and algorithms

("WPF Proprietary Technology"). The information comprising Lost Creek's WPF Proprietary Technology derive from and is the product of many years of experience, the labor of dozens of Lost Creek's skilled employees, and millions of dollars invested by Lost Creek in research, testing, innovation and application. A competitor in possession of Lost Creek's WPF Proprietary Technology would have the ability to develop products and compete with Lost Creek without expending the time, energy, and resources that Lost Creek expended to develop its unique products and technology. The information comprising Lost Creek's WPF Proprietary Technology (e.g., specific formulas, designs, dimensions, safety factors, tolerances, programming source code, etc.) is not legitimately known outside of Lost Creek and provides a competitive advantage to Lost Creek in the marketplace.

12. Lost Creek has taken great care to ensure that the custom design features of its products and manufacturing processes are kept confidential and remain a trade secret. Lost Creek's designs, testing, algorithms and other details of Lost Creek's custom features cannot be found in the open market and are not available to competitors to view or reverse engineer. Lost Creek's WPF Proprietary Technology is only accessible to a limited number of Lost Creek employees and are protected from disclosure through the compulsory use of access cards, usernames and passwords required to access the information. Furthermore, each Lost Creek employee that works with the WPF Proprietary Technology is required to sign a confidentiality agreement protecting such information from disclosure. As such, Lost Creek's WPF Proprietary Technology is a trade secret of Lost Creek's business.

13. Weird executed and agreed to his Non-Disclosure Agreement and Covenant Not to Compete (NDA/CNC) on September 4, 2007. Pursuant to the NDA/CNC, Weird agreed that upon termination of his employment with Plaintiff that he would maintain the confidentiality of Plaintiff's technology, trade secrets and proprietary and confidential information. Weird also agreed not to compete against Plaintiff for two years after such termination of employment with Plaintiff and to refrain from certain activities in competition against Plaintiff, such as providing the same or similar function with a competitor as they provided to Lost Creek.

14. Defendant Keith Austin Weird was a long-time, trusted employee of Lost Creek. Weird worked for Lost Creek as an engineer for over seventeen years in its offices in Austin, Travis County, Texas. He was ultimately promoted to the position of Assistant Vice President of Engineering. Weird had duties and obligations to protect Lost Creek's trade secrets and other confidential proprietary information from disclosure.

15. When he began employment with Lost Creek, Weird signed an NDA/CNC providing:

    "During the term of employment and without limitation thereafter, Keith Austin Weird hereby covenants and agrees to keep strictly confidential all knowledge to which he gains by virtue of his employment with Lost Creek. This includes all trade secrets, business practices, finances, documents, blueprints, market data, other intellectual property and other confidential information. Keith Austin Weird agrees not to disclose the above mentioned confidential information, directly or indirectly to any other person, company

or corporation, or use it for his own benefit. Keith Austin Weird agrees that he will only use the confidential information as an employee of Lost Creek.

All confidential or trade secret information relating to the business of Lost Creek which Keith Austin Weird shall develop, conceive, produce, construct or observe during his employment with Lost Creek shall remain the sole property of Lost Creek.

Keith Austin Weird further agrees that upon termination of his employment, Keith Austin Weird will surrender and deliver to Lost Creek all confidential information, including but not limited to work papers, books, records, and data of every kind relating to or in connection with Lost Creek.
Keith Austin Weird agrees, upon termination of employment with Lost Creek and for a period of two (2) years thereafter, Keith Austin Weird will not directly or indirectly engage in any business or work for any business which is in direct competition with Lost Creek by manufacturing and/or selling pipeline pigs that resemble or imitate the pipeline pigs manufactured and sold by Lost Creek. Keith Austin Weird agrees that this paragraph prohibits him from accepting employment on a worldwide basis with any pipeline pig manufacturer for the two (2) year period."

16. During Weird's employment with Lost Creek, he worked with other Lost Creek engineers to develop the unique WPF Proprietary Technology. As a Lost Creek employee, Weird was involved in the research, development, calculations, drawings, testing and design of Lost Creek's products. Through his work for Lost Creek, Weird had knowledge of and access to research and designs, to the technical aspects of Lost Creek's products and to the applications in which Lost Creek's products function.

17. On February 10, 2025, Weird received a written offer of employment by e-mail from Montgomery Bonnell, CEO of Artemis and a former Vice-President of Lost Creek who hired and supervised Weird beginning in 2007 until Bonnell's departure in 2017.

18. On February 11, 2025, Weird connected an external Western Digital My Passport hard drive to his Lost Creek laptop computer and downloaded almost thirty gigabytes of data comprising thousands of Lost Creek's confidential business documents and trade secrets. Included among this material were the complete contents of Weird's "Documents" folder holding WPF Proprietary Technology. Also on February 11, 2025, Weird connected one or more USB thumb drives to his Lost Creek laptop.

19. On February 14, 2025, Weird submitted his resignation letter to Lost Creek, effective February 28, 2025. In his resignation letter, Weird advised Lost Creek that he would be assuming a position with Artemis Energy Solutions, Inc. ("Artemis") as Technology Director-Pipeline Products. At the time of his resignation, Weird advised Lost Creek that prior to his departure, he would "return any and all confidential material belonging to Lost Creek that is in [his] possession."

20. Upon information and belief, CEO Montgomery Bonnell and other Artemis officers or employees induced Weird to misappropriate Lost Creek's confidential information and trade secrets for use in Artemis' business operations.

21. Following Weird's departure, Lost Creek discovered that Weird transferred numerous emails containing confidential and trade secret information to his personal webmail account.

22. On February 28, 2025, Weird participated in an exit interview wherein he was instructed to return any confidential business or trade secret information. Weird claimed he did not have any such information. When asked to return his Lost Creek laptop computer, Weird stated that he had left it at his home and promised to return it at a later date. Despite repeated requests that he do so, Weird has not returned his Lost Creek laptop. Weird has further declined to permit inspection of his webmail and has failed to respond to a written demand that he make his personal and Artemis computers, phones, tablets and data storage devices available for inspection.

23. Since Weird's departure, Artemis has hired three former Lost Creek engineering employees, Percy Pennybacker, Claudia Johnson and Barton Springs, who worked on development and testing of Lost Creek's WPF smart pig.

24. It is clear that Artemis targeted Lost Creek to poach its employees to start a smart pig division and begin manufacturing smart pigs in direct competition with Lost Creek. Artemis CEO, Montgomery Bonnell, approached Weird and, on information and belief, other Lost Creek employees with offers of employment and inducements of bonuses. Since Artemis had no smart pig division nor a smart pig product, hiring Lost Creek engineers was the shortest route to market.

25. On information and belief, Artemis began aggressively pursuing development of a WPF-like smart pig product line approximately six months before Weird was hired, but encountered difficulties due to the complexity of the complex technological challenges resolved by use of Lost Creek's WPF technology. Weird was hired by Artemis to gain access to Lost Creek's WPF Proprietary Technology as it enabled Artemis to develop competing products without expending the time and resources required to develop competing products through research and testing.

26. In October 2025, Artemis distributed a newsletter announcing that it would be expanding its product offerings to feature a new line of AirHog™ drone-paired, remote-controlled pipeline smart pigs. Weird was identified as leading the effort to bring the new products to market. Prior to Weird's employment with Artemis, Artemis did not manufacture or sell any type of smart pig products that competed with Lost Creek's products, let alone any product with the innovative and sophisticated features of Lost Creek's WPF Proprietary Technology.

27. On information and belief, Artemis has contracted with existing clients of Lost Creek for the sale of AirHog™ products that imitate or resemble the WPF smart pigs developed by Lost Creek.

Weird and Artemis have further applied for a patent on features of the design of the AirHog™ Remote-Controlled Pipeline Smart Pig. It is implausible that Artemis, lacking experience in the design and manufacture of smart pig products could design, develop, manufacture, patent and sell such products in less than eight months without unauthorized use of the WPF Proprietary Technology developed by Lost Creek.

IV. APPLICATION FOR INJUNCTIVE RELIEF

28. All previous paragraphs are incorporated herein.

29. Lost Creek requests a Permanent Injunction that Defendants, and each of their agents, servants, representatives, and all other persons or entities in active concert or participation with Defendants who receive actual notice of this Order by personal service or otherwise be and hereby are enjoined as follows:

a. Defendants are restrained from violating the Non-Disclosure Agreement and Covenant Not to Compete entered into between Lost Creek and Weird or participating in the violation of said NDA/CNC;

b. Defendants are ordered to return to Lost Creek, and to cease and desist from using, any Lost Creek proprietary documents, electronic files or other property, including but not limited to Lost Creek's WPF Proprietary Technology or any Artemis document that uses Lost Creek's information;

c. Defendants are restrained from altering or deleting any electronic files on their personal or work computers, mobile devices, PDAs, smart phones, webmail accounts, online storage repositories (including social networking sites) and any other electronic storage devices or services;

d. Defendants are restrained from inducing or attempting to induce, or from causing any person or other entity to induce or attempt to induce, any person who is an employee of Lost Creek to breach a contract with Lost Creek and to leave the employ of Lost Creek;

e. Weird is restrained from the design, development, testing, manufacture, promotion, lease or sale of any products that resemble or imitate any pipeline pig manufactured, sold or developed by Lost Creek or providing the same or similar functions for Artemis that he performed for Lost Creek until February 28, 2027;

f. Defendants are ordered to cease and desist from leasing, selling, promoting, or otherwise commercially using the AirHog™ Remote-Controlled Pipeline Smart Pig or any other tool designed or derived by using Lost Creek's trade secrets or confidential information, including but not limited to the WPF Proprietary Technology.

30. Upon information and belief, Defendants used, misappropriated, and disclosed Lost Creek's trade secrets and/or proprietary confidential information and continue to do so for the purposes of furthering Artemis' business. Defendants have solicited and continue to solicit Lost Creek's customers. It is believed that Defendants may continue to solicit Lost Creek's employees to breach contracts with Lost Creek in order to work for Artemis. The evidence of Defendants' breach of contract and misappropriation of trade secrets supports this Court's granting of its request for injunction. Lost Creek would similarly be entitled to the requested relief after a trial on the merits.

31. If Lost Creek's Application is not granted, harm is imminent because upon information and belief, Defendants are presently in possession of Lost Creek's trade secrets, proprietary confidential information and/or have transmitted Lost Creek's trade secrets, proprietary confidential information to others to facilitate their use of that information for their own benefit. In addition, upon information and belief, Defendants have solicited and continue to solicit Lost Creek's former, current, and/or prospective customers and its employees. These actions violate federal and state trade secret law and Weird's contractual obligations to Lost Creek.

32. The harm that will result if the Permanent Injunction is not issued is in part irreparable. Lost Creek cannot be fully compensated for all such harm. Money cannot fully compensate Lost Creek for the loss of its trade secrets and proprietary confidential information, which Lost Creek invests substantial time, money, and human capital resources to develop, and which gives Lost Creek a competitive advantage in the marketplace and which, if used, gives to Defendants a commercial advantage. Lost Creek also cannot be fully compensated for the continued loss of its employees to Artemis. Lost Creek cannot be fully compensated by the loss of its goodwill that will result from the loss of its trade secrets, proprietary confidential information, employees, and business opportunities.

33. The injury Lost Creek faces outweighs the injury that would be sustained by the Defendants as a result of the injunctive relief. The injunctive relief sought would not adversely affect public policy or the public interest.

34. Lost Creek is willing to post the necessary reasonable bond to facilitate the above injunctive relief requested.

V. CAUSES OF ACTION

Count 1 – Misappropriation of Trade Secrets under the Defend Trade Secrets Act (18 U.S.C. § 1836)

35. The foregoing paragraphs are incorporated by reference as if fully stated herein.

36. Lost Creek's WPF Proprietary Technology, including but not limited to its specialized designs, test mechanisms, source code, algorithms, formulas, dimensions, safety factors, and tolerances, constitutes trade secrets within the meaning of the Defend Trade Secrets Act, 18 U.S.C. § 1839(3), because: (a) Lost Creek has taken reasonable measures to keep such information secret; and (b) such information derives independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable through proper means by, another person who can obtain economic value from the disclosure or use of the information.

37. Defendants misappropriated Lost Creek's trade secrets within the meaning of 18 U.S.C. § 1839(5) by acquiring such trade secrets through improper means and by disclosing or using such trade secrets without Lost Creek's consent. Weird acquired Lost Creek's trade secrets by downloading confidential files to an external hard drive and USB thumb drives and by forwarding confidential information to his personal webmail account. Upon information and belief, Weird disclosed and/or used Lost Creek's trade secrets for the benefit of Artemis in the development of Artemis' AirHog™ product line.

38. Lost Creek's trade secrets are related to products and services used in, and intended for use in, interstate and foreign commerce.
39. As a direct and proximate result of Defendants' misappropriation, Lost Creek has suffered and continues to suffer damages, including but not limited to actual losses caused by the misappropriation and unjust enrichment caused by the misappropriation that is not addressed in computing actual loss. Lost Creek seeks to recover damages pursuant to 18 U.S.C. § 1836(b)(3)(B), including actual damages, unjust enrichment, and, in lieu thereof, a reasonable royalty.

40. Defendants' misappropriation was willful and malicious. Accordingly, Lost Creek seeks exemplary damages in an amount not more than two times the amount of damages awarded pursuant to 18 U.S.C. § 1836(b)(3)(C).

41. Lost Creek further seeks injunctive relief pursuant to 18 U.S.C. § 1836(b)(3)(A) to prevent any actual or threatened misappropriation, including affirmative acts to protect the trade secret as the Court deems reasonable.

42. Lost Creek seeks an award of reasonable attorneys' fees pursuant to 18 U.S.C. § 1836(b)(3)(D).

Count 2 – Misappropriation of Trade Secrets under the Texas Uniform Trade Secrets Act (Tex. Civ. Prac. & Rem. Code § 134A.001 et seq.)
43. The foregoing paragraphs are incorporated by reference as if fully stated herein.

44. Lost Creek's WPF Proprietary Technology, including but not limited to its specialized designs, test mechanisms, source code, algorithms, formulas, dimensions, safety factors, and tolerances, constitutes trade secrets within the meaning of the Texas Uniform Trade Secrets Act, Tex. Civ. Prac. & Rem. Code § 134A.002(6), because: (a) Lost Creek has taken reasonable measures under the circumstances to keep such information secret; and (b) such information derives

independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable through proper means by, other persons who can obtain economic value from its disclosure or use.

45. Defendants misappropriated Lost Creek's trade secrets within the meaning of Tex. Civ. Prac. & Rem. Code § 134A.002(3) by acquiring such trade secrets through improper means and by disclosing or using such trade secrets without Lost Creek's consent. Weird acquired Lost Creek's trade secrets by downloading confidential files to an external hard drive and USB thumb drives and by forwarding confidential information to his personal webmail account. Upon information and belief, Weird disclosed and/or used Lost Creek's trade secrets for the benefit of Artemis in the development of Artemis' AirHog™ product line. Artemis acquired Lost Creek's trade secrets knowing or having reason to know that the trade secrets were acquired by Weird through improper means.

46. As a direct and proximate result of Defendants' misappropriation, Lost Creek has suffered and continues to suffer damages. Lost Creek seeks to recover damages pursuant to Tex. Civ. Prac. & Rem. Code § 134A.004, including both the actual loss caused by misappropriation and the unjust enrichment caused by misappropriation that is not taken into account in computing actual loss. In lieu of damages measured in this manner, Lost Creek alternatively seeks damages measured by imposition of liability for a reasonable royalty for Defendants' unauthorized disclosure or use of Lost Creek's trade secrets.

47. Defendants' misappropriation was willful and malicious. Accordingly, Lost Creek seeks exemplary damages in an amount not exceeding twice any award made under Tex. Civ. Prac. & Rem. Code § 134A.004(a), pursuant to Tex. Civ. Prac. & Rem. Code § 134A.004(b).

48. Lost Creek further seeks injunctive relief pursuant to Tex. Civ. Prac. & Rem. Code § 134A.003 to prevent any actual or threatened misappropriation.

49. Lost Creek seeks an award of reasonable attorneys' fees pursuant to Tex. Civ. Prac. & Rem. Code § 134A.005.

Count 3 – Breach of Non-Disclosure Agreement and Covenant Not to Compete

50. The foregoing paragraphs are incorporated by reference as if fully stated herein.

51. The Non-Disclosure Agreement and Covenant Not to Compete executed and agreed to by Weird precludes Weird from competing against Lost Creek for a period of two (2) years. The Non-Disclosure Agreement and Covenant Not to Compete executed by Weird also includes Weird's promises not to disclose or use Lost Creek's confidential information and trade secrets.

52. Weird's Non-Disclosure Agreement and Covenant Not to Compete agreement is enforceable under Texas law. Weird's promises in the agreement were each made in exchange for Lost Creek's promises to provide Weird with specialized knowledge and training, Lost Creek's trade secrets, Lost Creek's proprietary confidential information and Lost Creek's goodwill. Lost Creek

fulfilled each of these promises with respect to Weird. Each of the covenants arise out of the trade secret agreement because the covenant is: (1) designed to protect Lost Creek's trade secrets, Lost Creek's confidential and proprietary information, Lost Creek's goodwill, and the specialized training and knowledge Lost Creek provided to Weird; and (2) to enforce Weird's promises regarding the same.

53. Weird's covenants not to compete have reasonable time, territory, and activity limitations. The covenants' limitations do not impose greater restraint than necessary to protect Lost Creek's business interests; and Lost Creek does not seek to enforce the covenants in any unreasonable manner or to any unreasonable extent.

54. Upon information and belief, Weird violated his Non-Disclosure Agreement and Covenant Not to Compete by divulging, disclosing, and using trade secrets and/or proprietary confidential information as discussed above.

55. The above breaches are material. As a natural, probable, and foreseeable consequence and proximate cause of Weird's actions, Lost Creek has suffered and continues to suffer damages for which Weird and Artemis are liable. Lost Creek seeks to recover all special, general, consequential, actual, and exemplary damages allowed by law as well as attorney fees, court costs, prejudgment, and post-judgment interest. Lost Creek has or will suffer damages to its business in the form of lost profits, loss of customers, loss of future business opportunities, loss of the exclusive right to use Lost Creek's trade secrets, and loss of goodwill. Lost Creek seeks to recover lost profits from contracts that were awarded to Artemis as a result of Weird's breaches of contract. In order to fully develop its lost profit claims, Lost Creek must examine Artemis' documents to determine the value of the jobs Artemis obtained. In the alternative, and in the event that Lost Creek's lost profits are unascertainable, Lost Creek seeks unjust enrichment damages.

Count 4 – The Computer Fraud and Abuse Act (18 U.S.C. § 1030)

56. The foregoing paragraphs are incorporated by reference as if fully stated herein.

57. Lost Creek's computers are used in interstate commerce; thus, Lost Creek's computers are protected computers pursuant to 18 U.S.C. § 1030(e)(2)(B).

58. Weird knowingly and with intent to defraud, accessed and used the computer(s) assigned by Lost Creek, without authorization or in a manner exceeding any authorization he may claim that he had. By means of such conduct, Weird furthered the intended fraud.

59. Lost Creek believes that, in February 2025 and on other occasions, Weird used Lost Creek's computer(s) to misappropriate, use, and share Lost Creek's trade secrets and proprietary confidential information without authorization.

60. Because of Weird's actions, Lost Creek suffered losses in excess of $75,000, including costs related to a computer forensic preservation and analysis of Weird's Lost Creek issued laptop and iPhone.

V. ATTORNEY FEES AND INTEREST

61. Pursuant to statute, common law, and the contracts with Defendants, Plaintiff is entitled to an award of its reasonable and necessary attorney fees with respect to Defendants for this cause and any appeals.

VII. EXEMPLARY DAMAGES

62. The conduct of Defendants, as alleged above, including misappropriation of trade secrets and breach of contract, was aggravated by the kind of willfulness, wantonness and malice for which the law allows for the imposition of exemplary damages. Moreover, Defendants' wrongdoing was committed knowingly and with a conscious indifference to Lost Creek's rights. Defendants acted with intent to harm Lost Creek and their misconduct was intentional, willful, wanton and without justification or excuse. Therefore, Lost Creek seeks to recover exemplary damages from Defendants in an amount to be determined by the Court, including exemplary damages as provided under 18 U.S.C. § 1836(b)(3)(C) and Tex. Civ. Prac. & Rem. Code § 134A.004(b).

VIII. CONDITIONS PRECEDENT

63. All conditions precedent to an outcome favorable to the party represented by the undersigned in this action have been performed, have occurred or have been waived.

IX. PRAYER

WHEREFORE, PREMISES CONSIDERED, Plaintiff prays for the following relief:

a) A permanent injunction for the relief requested above, including injunctive relief authorized by 18 U.S.C. § 1836(b)(3)(A) and Tex. Civ. Prac. & Rem. Code § 134A.003;

b) Upon final trial, judgment against Defendants, jointly and severally, for full permanent injunctive relief as requested herein, and, for the full amount of the Plaintiff's damages, special, general, consequential, actual, and exemplary;

c) Assignment of Defendants' Provisional Patent Application and/or Patent on the AirHog™ Remote-Controlled Pipeline Smart Pig and/or related technologies;

d) Prejudgment interest;

e) Post judgment interest;

f) Plaintiff's reasonable and necessary attorney fees in prosecuting its claims through trial and, if necessary appeal, including attorneys' fees as provided under 18 U.S.C. § 1836(b)(3)(D) and Tex. Civ. Prac. & Rem. Code § 134A.005;

g) All costs of suit; and

h) Such other and further relief, at law or in equity, to which Plaintiff may show itself justly entitled.
Respectfully submitted,

Respectfully submitted,
Massive International Law, LLP

/S/
By: _____
William E. Nelson
TSB No. 00003723
1 Congress Ave., Suite 20000
Austin, Texas Austin 78701
Tel: (512) 555-1234
willenelson@milaw.com
LEAD ATTORNEYS FOR PLAINTIFF

Of Counsel
Lamar Street
Tarrytown, Olde & Rich, Attorneys
1313 Guadalupe, Suite 1900
Austin, Texas 78701

CERTIFICATE OF SERVICE
I certify that a true copy of the foregoing has been served on all attorneys of record and persons pro se in this cause, by electronic service, electronic mail, facsimile and/or certified mail, return receipt requested, by depositing same, postpaid, in an official depository under the care and custody of the United States Postal Service on December 15, 2025.

//s// William E. Nelson



IN THE UNITED STATES DISTRICT COURT
WESTERN DISTRICT OF TEXAS

LOST CREEK ENGINEERING, L.L.C.	§	
	§	
V.	§	**CIVIL ACTION NO.**
	§	**5:25-CV-01234**
KEITH AUSTIN WEIRD,	§	
and	§	**JURY REQUESTED**
ARETEMIS ENERGY SOLUTIONS, INC.	§	
	§	

FIRST AMENDED ANSWER OF KEITH AUSTIN WEIRD AND ARTEMIS ENERGY SOLUTIONS, INC.

Defendants Keith Austin Weird and Artemis Energy Solutions, Inc., file this Amended Answer in response to the Amended Complaint and Application for Injunctive Relief filed by Plaintiff, Lost Creek Engineering, Inc. ("Lost Creek").

I. FIRST AMENDED ANSWER

1. Defendants are not required to admit or deny the allegations of Paragraph 1.
2. Defendants admit the allegations of paragraph 2.
3. Defendants admit the allegations of paragraph 3
4. Paragraph 4 is a statement of jurisdiction which Defendants are not required to admit or deny.
5. Paragraph 5 is a statement of venue which Defendants are not required to admit or deny.
6. Upon information and belief, Defendants admit the allegations of Paragraph 6.
7. Upon information and belief, Defendants admit the allegations of Paragraph 7.
8. With regard to Paragraph 8, Defendants deny the allegation that Lost Creek's When Pigs Fly (WPF) smart pig technology, if any, represent unique or innovative technology. Defendants admit all other allegations in Paragraph 8.
9. With regard to paragraph 9, Defendants deny that Lost Creek's WPF technologies (if any) are not found in the open market and are not used by other smart pig manufacturers. Defendants

contend that all or part of these allegedly proprietary and confidential WPF technologies (if any) derive from open sources and/or were not developed by Lost Creek. Defendants admit all other allegations in Paragraph 9.

10. Defendants deny all allegations in Paragraph 10.

11. Defendants deny that the information referred to as "e WPF Proprietary Technology "is not known outside of the Lost Creek and provides a competitive advantage to Lost Creek in the marketplace. Defendants do not have sufficient information to either admit or deny the other allegations in Paragraph 11.

12. Defendants deny that Lost Creek has taken great care to ensure that the custom design features of its products and manufacturing processes are kept confidential and remain a trade secret. Defendants deny that Lost Creek's WPF Proprietary Technology and other details of Lost Creek's custom features cannot be found in the open market and are not available to competitors to view or reverse engineer. Defendants admit all other allegations in Paragraph 12.

13. With regard to Paragraph 13, Defendants admit that Weird executed a Non-Disclosure Agreement and Covenant Not to Compete on September 1, 2001 after being ordered to do so by Lost Creek. Defendants admit that Non-Disclosure Agreement and Covenant Not to Compete is an industry-wide, unenforceable restraint of trade that purports to forbid Weird from competing directly or indirectly with Lost Creek for a period of two years, without territorial restriction. Defendants do not have sufficient information to either admit or deny any other allegations in Paragraph 13.

14. Defendants admit the allegations of paragraph 14.

15. Defendants admit the allegations of paragraph 15.

16. Defendants deny that the information referenced as "unique WPF Proprietary Technology' is unique, proprietary or the property of Plaintiff Lost Creek. Defendants admit the other allegations of paragraph 16.

17. Defendants admit the allegations of paragraph 17.

18. Defendants deny that Weird downloaded almost thirty gigabytes of data comprising thousands of Lost Creek's confidential business documents and trade secrets. Defendant admits that he may have sought to back up certain iTunes music he personally purchased as well as family photographs. Defendants contend that any business documents copied by Weird were either copied inadvertently or were copied for the purpose of completing work for Lost Creek's sole and

exclusive benefit. Defendants do not have sufficient information to either admit or deny any other allegations in Paragraph 18.

19. Defendants admit the allegations of paragraph 19.

20. Defendants deny all allegations of paragraph 20.

21. Defendants do not have sufficient information to either admit or deny the allegations in Paragraph 21. Defendants admit that over the course of 15 years of employment, Weird may have used his personal e-mail for his former employer's benefit.

22. Defendants deny that there have been repeated requests made for the return of Weird's Lost Creek laptop or that Weird has declined (or failed to respond to) requests for inspection. Many of the devices and sources described hold confidential personal and privileged information and communications. Defendants admit that Weird participated in an exit interview.

23. Defendants admit the allegations of paragraph 23.

24. Defendants deny all allegations of paragraph 24.

25. Defendants deny all allegations of paragraph 25.

26. Defendants deny that prior to Weird's employment with Artemis, Artemis did not manufacture or sell any type of smart pig products that competed with Lost Creek's products. Defendants admit all other allegations of Paragraph 26.

27. Defendants admit Artemis has applied for a patent on unique and innovative design features of certain of its intelligent pipeline pig products. Defendants deny all other allegations of Paragraph 27

28. Defendants incorporate their prior responses to Paragraphs 1-27.

29. Defendants deny all allegations of paragraph 29.

30. Defendants deny all allegations of paragraph 30.

31. Defendants deny all allegations of paragraph 31.

32. Defendants deny all allegations of paragraph 32.

33. Defendants deny all allegations of paragraph 33.

34. Defendants do not have sufficient information to either admit or deny any allegations in Paragraph 34.

35. Defendants incorporate their prior responses to Paragraphs 1-34.

36. Defendants admit the allegations of paragraph 36.

37. Defendants deny all allegations of paragraph 37.

38. Defendants deny all allegations of paragraph 38.

39. Defendants deny all allegations of paragraph 39.

40. Defendants deny all allegations of paragraph 40.

41. Defendants incorporate their prior responses to Paragraphs 1-40.

42. Defendants do not have sufficient information to either admit or deny any allegations in Paragraph 42.

43. Defendants deny all allegations of paragraph 43.

44. Defendants incorporate their prior responses to Paragraphs 1-43.

45. Defendants deny all allegations of paragraph 45.

46. Defendants incorporate their prior responses to Paragraphs 1-45.

47. Defendants deny all allegations of paragraph 47.

48. Defendants deny all allegations of paragraph 48.

49. Defendants incorporate their prior responses to Paragraphs 1-48.

50. Defendants deny all allegations of paragraph 50.

51. Defendants incorporate their prior responses to Paragraphs 1-50.

52. Defendants deny all allegations of paragraph 52.

53. Defendants deny all allegations of paragraph 53.

54. Defendants incorporate their prior responses to Paragraphs 1-53.

55. Defendants deny all allegations of paragraph 55.

56. Defendants deny all allegations of paragraph 56.

57. Defendants incorporate their prior responses to Paragraphs 1-56.

58. Defendants deny all allegations of paragraph 58.

59. Defendants deny all allegations of paragraph 59.

60. Defendants deny all allegations of paragraph 60.

61. Defendants deny all allegations of paragraph 61.

62. Defendants deny all allegations of paragraph 62.

63. Defendants deny all allegations of paragraph 63.

II. AFFIRMATIVE DEFENSES

FIRST AFFIRMATIVE DEFENSE: FAILURE TO STATE A CLAIM.

64. Defendants affirmatively assert that Lost Creek's claims are barred, in whole or in

part, because Lost Creek has failed to state a claim upon which relief may be granted.

SECOND AFFIRMATIVE DEFENSE: WAIVER.

65. Defendants affirmatively assert that Lost Creek's claims are barred by the doctrine of waiver.

THIRD AFFIRMATIVE DEFENSE: ESTOPPEL.

66. Defendants affirmatively assert that Lost Creek's claims are barred by the doctrine of estoppel.

FOURTH AFFIRMATIVE DEFENSE: INJUNCTIVE RELIEF IS UNECESSARY.

67. Defendants affirmatively assert that Lost Creek's claims are barred, in whole or in part, because injunctive relief is unnecessary as pled.

FIFTH AFFIRMATIVE DEFENSE: JUSTIFICATION.

68. Defendants affirmatively asserts that Lost Creeks claims are barred because of the doctrine of justification.

SIXTH AFFIRMATIVE DEFENSE: PRIVILEGE.

69. Defendants affirmatively assert that Lost Creek's claims are barred, in whole or in part, because of the doctrine of privilege.

SEVENTH AFFIRMATIVE DEFENSE:
PREEMPTION OF ATTORNEY'S FEES AWARD.

70. Defendants affirmatively assert that Lost Creek's claims for attorney's fees are barred, in whole or in part, because such claims are preempted by the Texas Covenant not to Compete Act. TEX. Bus. & COM. CODE ANN. § 15.51 and § 15.52.

III. ATTORNEY'S FEES

71. The primary purpose of the "agreement" to which Lost Creek claims the Non-Disclosure Agreement and Covenant Not to Compete was ancillary to, was to obligate Weird to render personal services. Plaintiff knew that the Non-Disclosure Agreement and Covenant Not to Compete did not contain limitations as to time, geographical area, and scope of activity to be restrained that were reasonable and the limitations imposed a greater restraint than necessary to protect the goodwill or other business interest of Plaintiff.

72. Plaintiff is also seeking to enforce the covenant to a greater extent than is necessary to protect Plaintiffs goodwill or other business interest. Therefore, Pursuant to Section 15.51 of the Texas Business and Commerce Code, Defendants seek to recover reasonable attorney's fees and costs as are equitable and just.

73. Additionally, Defendants seek to recover reasonable attorney's fees and costs pursuant to Section 134.005 of the Texas Civil Practice & Remedies Code.

Respectfully Submitted,

Bevo ★ Orange ★ Tower, P.C.

By: Tex S Tower

"Tex" S. Tower
Federal ID No. 123456
State Bar No. 010101010
2300 Inner Campus Drive
Austin, Texas 78713
TEL: (512) 555-3377

ATTORNEY IN CHARGE FOR DEFENDANTS KEITH AUSTIN WEIRD AND ARTEMIS ENERGY SOLUTIONS, INC.

CERTIFICATE OF SERVICE

The undersigned hereby certifies that a true and correct copy of the above and foregoing was served pursuant to the Federal Rules of Civil Procedure on this the 17th day of December, 2025, to:

William E. Nelson
1 Congress Ave., Suite 20000
Austin, Texas Austin 78701

Tex S Tower

The "E-Discovery Rules" (1,16,26,34 & 45) of the Federal Rules of Civil Procedure

with Committee Notes accompanying 2006 and 2015 Amendments

[Note: Some provisions highlighted to emphasize their importance]

Rule 1. Scope and Purpose

These rules govern the procedure in all civil actions and proceedings in the United States district courts, except as stated in Rule 81. They should be construed, administered, and employed by the court and the parties to secure the just, speedy, and inexpensive determination of every action and proceeding.

Notes

(As amended Dec. 29, 1948, eff. Oct. 20, 1949; Feb. 28, 1966, eff. July 1, 1966; Apr. 22, 1993, eff. Dec. 1, 1993; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 29, 2015, eff. Dec. 1, 2015.)

Committee Notes on Rules—2015 Amendment

Rule 1 is amended to emphasize that just as the court should construe and administer these rules to secure the just, speedy, and inexpensive determination of every action, so the parties share the responsibility to employ the rules in the same way. Most lawyers and parties cooperate to achieve these ends. But discussions of ways to improve the administration of civil justice regularly include pleas to discourage over-use, misuse, and abuse of procedural tools that increase cost and result in delay. Effective advocacy is consistent with — and indeed depends upon — cooperative and proportional use of procedure.

This amendment does not create a new or independent source of sanctions. Neither does it abridge the scope of any other of these rules.

***

Rule 16. Pretrial Conferences; Scheduling; Management

(a) Purposes of a Pretrial Conference. In any action, the court may order the attorneys and any unrepresented parties to appear for one or more pretrial conferences for such purposes as:

(1) expediting disposition of the action;

(2) establishing early and continuing control so that the case will not be protracted because of lack of management;

(3) discouraging wasteful pretrial activities;

(4) improving the quality of the trial through more thorough preparation; and

(5) facilitating settlement.

(b) Scheduling.

(1) *Scheduling Order.* Except in categories of actions exempted by local rule, the district judge—or a magistrate judge when authorized by local rule—must issue a scheduling order:

(A) after receiving the parties' report under Rule 26(f); or

(B) after consulting with the parties' attorneys and any unrepresented parties at a scheduling conference.

(2) *Time to Issue.* The judge must issue the scheduling order as soon as practicable, but unless the judge finds good cause for delay, the judge must issue it within the earlier of 90 days after any defendant has been served with the complaint or 60 days after any defendant has appeared.

(3) *Contents of the Order.*

(A) *Required Contents.* The scheduling order must limit the time to join other parties, amend the pleadings, complete discovery, and file motions.

(B) *Permitted Contents.* The scheduling order may:

(i) modify the timing of disclosures under Rules 26(a) and 26(e)(1);

(ii) modify the extent of discovery;

(iii) provide for disclosure, discovery, or preservation of electronically stored information;

(iv) include any agreements the parties reach for asserting claims of privilege or of protection as trial-preparation material after information is produced, including agreements reached under Federal Rule of Evidence 502;

(v) direct that before moving for an order relating to discovery, the movant must request a conference with the court;

(vi) set dates for pretrial conferences and for trial; and

(vii) include other appropriate matters.

(4) *Modifying a Schedule.* A schedule may be modified only for good cause and with the judge's consent.

(c) Attendance and Matters for Consideration at a Pretrial Conference.

(1) *Attendance.* A represented party must authorize at least one of its attorneys to make stipulations and admissions about all matters that can reasonably be anticipated for discussion at a pretrial conference. If appropriate, the court may require that a party or its representative be present or reasonably available by other means to consider possible settlement.

(2) *Matters for Consideration.* At any pretrial conference, the court may consider and take appropriate action on the following matters:

(A) formulating and simplifying the issues, and eliminating frivolous claims or defenses;

(B) amending the pleadings if necessary or desirable;

(C) obtaining admissions and stipulations about facts and documents to avoid unnecessary proof, and ruling in advance on the admissibility of evidence;

(D) avoiding unnecessary proof and cumulative evidence, and limiting the use of testimony under Federal Rule of Evidence 702;

(E) determining the appropriateness and timing of summary adjudication under Rule 56;

(F) controlling and scheduling discovery, including orders affecting disclosures and discovery under Rule 26 and Rules 29 through 37;

(G) identifying witnesses and documents, scheduling the filing and exchange of any pretrial briefs, and setting dates for further conferences and for trial;

(H) referring matters to a magistrate judge or a master;

(I) settling the case and using special procedures to assist in resolving the dispute when authorized by statute or local rule;

(J) determining the form and content of the pretrial order;

(K) disposing of pending motions;

(L) adopting special procedures for managing potentially difficult or protracted actions that may involve complex issues, multiple parties, difficult legal questions, or unusual proof problems;

(M) ordering a separate trial under Rule 42(b) of a claim, counterclaim, crossclaim, third-party claim, or particular issue;

(N) ordering the presentation of evidence early in the trial on a manageable issue that might, on the evidence, be the basis for a judgment as a matter of law under Rule 50(a) or a judgment on partial findings under Rule 52(c);

(O) establishing a reasonable limit on the time allowed to present evidence; and

(P) facilitating in other ways the just, speedy, and inexpensive disposition of the action.

(d) Pretrial Orders. After any conference under this rule, the court should issue an order reciting the action taken. This order controls the course of the action unless the court modifies it.

(e) Final Pretrial Conference and Orders. The court may hold a final pretrial conference to formulate a trial plan, including a plan to facilitate the admission of evidence. The conference must be held as close to the start of trial as is reasonable, and must be attended by at least one attorney who will conduct the trial for each party and by any unrepresented party. The court may modify the order issued after a final pretrial conference only to prevent manifest injustice.

(f) Sanctions.

(1) *In General.* On motion or on its own, the court may issue any just orders, including those authorized by Rule 37(b)(2)(A)(ii)–(vii), if a party or its attorney:

(A) fails to appear at a scheduling or other pretrial conference;

(B) is substantially unprepared to participate—or does not participate in good faith—in the conference; or

(C) fails to obey a scheduling or other pretrial order.

(2) *Imposing Fees and Costs.* Instead of or in addition to any other sanction, the court must order the party, its attorney, or both to pay the reasonable expenses—including attorney's fees—incurred because of any noncompliance with this rule, unless the noncompliance was substantially justified or other circumstances make an award of expenses unjust.

Notes

(As amended Apr. 28, 1983, eff. Aug. 1, 1983; Mar. 2, 1987, eff. Aug. 1, 1987; Apr. 22, 1993, eff. Dec. 1, 1993; Apr. 12, 2006, eff. Dec. 1, 2006; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 29, 2015, eff. Dec. 1, 2015.)

Committee Notes on Rules—2006 Amendment

The amendment to Rule 16(b) is designed to alert the court to the possible need to address the handling of discovery of electronically stored information early in the litigation if such discovery is

expected to occur. Rule 26(f) is amended to direct the parties to discuss discovery of electronically stored information if such discovery is contemplated in the action. Form 35 is amended to call for a report to the court about the results of this discussion. In many instances, the court's involvement early in the litigation will help avoid difficulties that might otherwise arise.

Rule 16(b) is also amended to include among the topics that may be addressed in the scheduling order any agreements that the parties reach to facilitate discovery by minimizing the risk of waiver of privilege or work-product protection. Rule 26(f) is amended to add to the discovery plan the parties' proposal for the court to enter a case-management or other order adopting such an agreement. The parties may agree to various arrangements. For example, they may agree to initial provision of requested materials without waiver of privilege or protection to enable the party seeking production to designate the materials desired or protection for actual production, with the privilege review of only those materials to follow. Alternatively, they may agree that if privileged or protected information is inadvertently produced, the producing party may by timely notice assert the privilege or protection and obtain return of the materials without waiver. Other arrangements are possible. In most circumstances, a party who receives information under such an arrangement cannot assert that production of the information waived a claim of privilege or of protection as trial-preparation material.

An order that includes the parties' agreement may be helpful in avoiding delay and excessive cost in discovery. *See Manual for Complex Litigation*(4th) §11.446. Rule 16(b)(6) recognizes the propriety of including such agreements in the court's order. The rule does not provide the court with authority to enter such a case-management or other order without party agreement, or limit the court's authority to act on motion.

Committee Notes on Rules—2015 Amendment

The provision for consulting at a scheduling conference by "telephone, mail, or other means" is deleted. A scheduling conference is more effective if the court and parties engage in direct simultaneous communication. The conference may be held in person, by telephone, or by more sophisticated electronic means.

The time to issue the scheduling order is reduced to the earlier of 90 days (not 120 days) after any defendant has been served, or 60 days (not 90 days) after any defendant has appeared. This change, together with the shortened time for making service under Rule 4(m), will reduce delay at the beginning of litigation. At the same time, a new provision recognizes that the court may find good cause to extend the time to issue the scheduling order. In some cases it may be that the parties cannot prepare adequately for a meaningful Rule 26(f) conference and then a scheduling conference in the time allowed. Litigation involving complex issues, multiple parties, and large organizations, public or private, may be more likely to need extra time to establish meaningful

collaboration between counsel and the people who can supply the information needed to participate in a useful way. Because the time for the Rule 26(f) conference is geared to the time for the scheduling conference or order, an order extending the time for the scheduling conference will also extend the time for the Rule 26(f) conference. But in most cases it will be desirable to hold at least a first scheduling conference in the time set by the rule.

Three items are added to the list of permitted contents in Rule 16(b)(3)(B).

The order may provide for preservation of electronically stored information, a topic also added to the provisions of a discovery plan under Rule 26(f)(3)(C). Parallel amendments of Rule 37(e) recognize that a duty to preserve discoverable information may arise before an action is filed.

The order also may include agreements incorporated in a court order under Evidence Rule 502 controlling the effects of disclosure of information covered by attorney-client privilege or work-product protection, a topic also added to the provisions of a discovery plan under Rule 26(f)(3)(D).

Finally, the order may direct that before filing a motion for an order relating to discovery the movant must request a conference with the court. Many judges who hold such conferences find them an efficient way to resolve most discovery disputes without the delay and burdens attending a formal motion, but the decision whether to require such conferences is left to the discretion of the judge in each case.

***

Rule 26. Duty to Disclose; General Provisions Governing Discovery

(a) Required Disclosures.

(1) *Initial Disclosure.*

(A) *In General.* Except as exempted by Rule 26(a)(1)(B) or as otherwise stipulated or ordered by the court, a party must, without awaiting a discovery request, provide to the other parties:

(i) the name and, if known, the address and telephone number of each individual likely to have discoverable information—along with the subjects of that information—that the disclosing party may use to support its claims or defenses, unless the use would be solely for impeachment;

(ii) a copy—or a description by category and location—of all documents, electronically stored information, and tangible things that the disclosing party has in its possession, custody, or control and may use to support its claims or defenses, unless the use would be solely for impeachment;

(iii) a computation of each category of damages claimed by the disclosing party—who must also make available for inspection and copying as under Rule 34 the documents or other evidentiary material, unless privileged or protected from disclosure, on which each computation is based, including materials bearing on the nature and extent of injuries suffered; and

(iv) for inspection and copying as under Rule 34, any insurance agreement under which an insurance business may be liable to satisfy all or part of a possible judgment in the action or to indemnify or reimburse for payments made to satisfy the judgment.

(B) *Proceedings Exempt from Initial Disclosure.* The following proceedings are exempt from initial disclosure:

(i) an action for review on an administrative record;

(ii) a forfeiture action in rem arising from a federal statute;

(iii) a petition for habeas corpus or any other proceeding to challenge a criminal conviction or sentence;

(iv) an action brought without an attorney by a person in the custody of the United States, a state, or a state subdivision;

(v) an action to enforce or quash an administrative summons or subpoena;

(vi) an action by the United States to recover benefit payments;

(vii) an action by the United States to collect on a student loan guaranteed by the United States;

(viii) a proceeding ancillary to a proceeding in another court; and

(ix) an action to enforce an arbitration award.

(C) *Time for Initial Disclosures—In General.* A party must make the initial disclosures at or within 14 days after the parties' Rule 26(f)conference unless a different time is set by stipulation or court order, or unless a party objects during the conference that initial disclosures are not appropriate in this action and states the objection in the proposed discovery plan. In ruling on the objection, the court must determine what disclosures, if any, are to be made and must set the time for disclosure.

(D) *Time for Initial Disclosures—For Parties Served or Joined Later.* A party that is first served or otherwise joined after the Rule 26(f)conference must make the initial disclosures within 30 days after being served or joined, unless a different time is set by stipulation or court order.

(E) *Basis for Initial Disclosure; Unacceptable Excuses.* A party must make its initial disclosures based on the information then reasonably available to it. A party is not excused from making its

disclosures because it has not fully investigated the case or because it challenges the sufficiency of another party's disclosures or because another party has not made its disclosures.

(2) *Disclosure of Expert Testimony.*

(A) *In General.* In addition to the disclosures required by Rule 26(a)(1), a party must disclose to the other parties the identity of any witness it may use at trial to present evidence under Federal Rule of Evidence 702, 703, or 705.

(B) *Witnesses Who Must Provide a Written Report.* Unless otherwise stipulated or ordered by the court, this disclosure must be accompanied by a written report—prepared and signed by the witness—if the witness is one retained or specially employed to provide expert testimony in the case or one whose duties as the party's employee regularly involve giving expert testimony. The report must contain:

(i) a complete statement of all opinions the witness will express and the basis and reasons for them;

(ii) the facts or data considered by the witness in forming them;

(iii) any exhibits that will be used to summarize or support them;

(iv) the witness's qualifications, including a list of all publications authored in the previous 10 years;

(v) a list of all other cases in which, during the previous 4 years, the witness testified as an expert at trial or by deposition; and

(vi) a statement of the compensation to be paid for the study and testimony in the case.

(C) *Witnesses Who Do Not Provide a Written Report.* Unless otherwise stipulated or ordered by the court, if the witness is not required to provide a written report, this disclosure must state:

(i) the subject matter on which the witness is expected to present evidence under Federal Rule of Evidence 702, 703, or 705; and

(ii) a summary of the facts and opinions to which the witness is expected to testify.

(D) Time to Disclose Expert Testimony. A party must make these disclosures at the times and in the sequence that the court orders. Absent a stipulation or a court order, the disclosures must be made:

(i) at least 90 days before the date set for trial or for the case to be ready for trial; or

(ii) if the evidence is intended solely to contradict or rebut evidence on the same subject matter identified by another party under Rule 26(a)(2)(B) or (C), within 30 days after the other party's disclosure.

(E) *Supplementing the Disclosure.* The parties must supplement these disclosures when required under Rule 26(e).

(3) *Pretrial Disclosures.*

(A) *In General.* In addition to the disclosures required by Rule 26(a)(1) and (2), a party must provide to the other parties and promptly file the following information about the evidence that it may present at trial other than solely for impeachment:

(i) the name and, if not previously provided, the address and telephone number of each witness—separately identifying those the party expects to present and those it may call if the need arises;

(ii) the designation of those witnesses whose testimony the party expects to present by deposition and, if not taken stenographically, a transcript of the pertinent parts of the deposition; and

(iii) an identification of each document or other exhibit, including summaries of other evidence—separately identifying those items the party expects to offer and those it may offer if the need arises.

(B) *Time for Pretrial Disclosures; Objections.* Unless the court orders otherwise, these disclosures must be made at least 30 days before trial. Within 14 days after they are made, unless the court sets a different time, a party may serve and promptly file a list of the following objections: any objections to the use under Rule 32(a) of a deposition designated by another party under Rule 26(a)(3)(A)(ii); and any objection, together with the grounds for it, that may be made to the admissibility of materials identified under Rule 26(a)(3)(A)(iii). An objection not so made—except for one under Federal Rule of Evidence 402 or 403—is waived unless excused by the court for good cause.

(4) *Form of Disclosures.* Unless the court orders otherwise, all disclosures under Rule 26(a) must be in writing, signed, and served.

(b) Discovery Scope and Limits.

(1) *Scope in General.* Unless otherwise limited by court order, the scope of discovery is as follows: Parties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the case, considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit. Information within this scope of discovery need not be admissible in evidence to be discoverable.

(2) *Limitations on Frequency and Extent.*

(A) *When Permitted.* By order, the court may alter the limits in these rules on the number of depositions and interrogatories or on the length of depositions under Rule 30. By order or local rule, the court may also limit the number of requests under Rule 36.

(B) *Specific Limitations on Electronically Stored Information.* A party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost. On motion to compel discovery or for a protective order, the party from whom discovery is sought must show that the information is not reasonably accessible because of undue burden or cost. If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)(2)(C). The court may specify conditions for the discovery.

(C) *When Required.* On motion or on its own, the court must limit the frequency or extent of discovery otherwise allowed by these rules or by local rule if it determines that:

(i) the discovery sought is unreasonably cumulative or duplicative, or can be obtained from some other source that is more convenient, less burdensome, or less expensive;

(ii) the party seeking discovery has had ample opportunity to obtain the information by discovery in the action; or

(iii) the proposed discovery is outside the scope permitted by Rule 26(b)(1).

(3) *Trial Preparation: Materials.*

(A) *Documents and Tangible Things.* Ordinarily, a party may not discover documents and tangible things that are prepared in anticipation of litigation or for trial by or for another party or its representative (including the other party's attorney, consultant, surety, indemnitor, insurer, or agent). But, subject to Rule 26(b)(4), those materials may be discovered if:

(i) they are otherwise discoverable under Rule 26(b)(1); and

(ii) the party shows that it has substantial need for the materials to prepare its case and cannot, without undue hardship, obtain their substantial equivalent by other means.

(B) *Protection Against Disclosure.* If the court orders discovery of those materials, it must protect against disclosure of the mental impressions, conclusions, opinions, or legal theories of a party's attorney or other representative concerning the litigation.

(C) *Previous Statement.* Any party or other person may, on request and without the required showing, obtain the person's own previous statement about the action or its subject matter. If the

request is refused, the person may move for a court order, and Rule 37(a)(5) applies to the award of expenses. A previous statement is either:

(i) a written statement that the person has signed or otherwise adopted or approved; or

(ii) a contemporaneous stenographic, mechanical, electrical, or other recording—or a transcription of it—that recites substantially verbatim the person's oral statement.

(4) *Trial Preparation: Experts.*

(A) *Deposition of an Expert Who May Testify.* A party may depose any person who has been identified as an expert whose opinions may be presented at trial. If Rule 26(a)(2)(B) requires a report from the expert, the deposition may be conducted only after the report is provided.

(B) *Trial-Preparation Protection for Draft Reports or Disclosures.* Rules 26(b)(3)(A) and (B) protect drafts of any report or disclosure required under Rule 26(a)(2), regardless of the form in which the draft is recorded.

(C) *Trial-Preparation Protection for Communications Between a Party's Attorney and Expert Witnesses.* Rules 26(b)(3)(A) and (B) protect communications between the party's attorney and any witness required to provide a report under Rule 26(a)(2)(B), regardless of the form of the communications, except to the extent that the communications:

(i) relate to compensation for the expert's study or testimony;

(ii) identify facts or data that the party's attorney provided and that the expert considered in forming the opinions to be expressed; or

(iii) identify assumptions that the party's attorney provided and that the expert relied on in forming the opinions to be expressed.

(D) *Expert Employed Only for Trial Preparation.* Ordinarily, a party may not, by interrogatories or deposition, discover facts known or opinions held by an expert who has been retained or specially employed by another party in anticipation of litigation or to prepare for trial and who is not expected to be called as a witness at trial. But a party may do so only:

(i) as provided in Rule 35(b); or

(ii) on showing exceptional circumstances under which it is impracticable for the party to obtain facts or opinions on the same subject by other means.

(E) *Payment.* Unless manifest injustice would result, the court must require that the party seeking discovery:

(i) pay the expert a reasonable fee for time spent in responding to discovery under Rule 26(b)(4)(A) or (D); and

(ii) for discovery under (D), also pay the other party a fair portion of the fees and expenses it reasonably incurred in obtaining the expert's facts and opinions.

(5) *Claiming Privilege or Protecting Trial-Preparation Materials.*

(A) *Information Withheld.* When a party withholds information otherwise discoverable by claiming that the information is privileged or subject to protection as trial-preparation material, the party must:

(i) expressly make the claim; and

(ii) describe the nature of the documents, communications, or tangible things not produced or disclosed—and do so in a manner that, without revealing information itself privileged or protected, will enable other parties to assess the claim.

(B) *Information Produced.* If information produced in discovery is subject to a claim of privilege or of protection as trial-preparation material, the party making the claim may notify any party that received the information of the claim and the basis for it. After being notified, a party must promptly return, sequester, or destroy the specified information and any copies it has; must not use or disclose the information until the claim is resolved; must take reasonable steps to retrieve the information if the party disclosed it before being notified; and may promptly present the information to the court under seal for a determination of the claim. The producing party must preserve the information until the claim is resolved.

(c) Protective Orders.

(1) *In General.* A party or any person from whom discovery is sought may move for a protective order in the court where the action is pending—or as an alternative on matters relating to a deposition, in the court for the district where the deposition will be taken. The motion must include a certification that the movant has in good faith conferred or attempted to confer with other affected parties in an effort to resolve the dispute without court action. The court may, for good cause, issue an order to protect a party or person from annoyance, embarrassment, oppression, or undue burden or expense, including one or more of the following:

(A) forbidding the disclosure or discovery;

(B) specifying terms, including time and place or the allocation of expenses, for the disclosure or discovery;

(C) prescribing a discovery method other than the one selected by the party seeking discovery;

(D) forbidding inquiry into certain matters, or limiting the scope of disclosure or discovery to certain matters;

(E) designating the persons who may be present while the discovery is conducted;

(F) requiring that a deposition be sealed and opened only on court order;

(G) requiring that a trade secret or other confidential research, development, or commercial information not be revealed or be revealed only in a specified way; and

(H) requiring that the parties simultaneously file specified documents or information in sealed envelopes, to be opened as the court directs.

(2) *Ordering Discovery.* If a motion for a protective order is wholly or partly denied, the court may, on just terms, order that any party or person provide or permit discovery.

(3) *Awarding Expenses.* Rule 37(a)(5) applies to the award of expenses.

(d) Timing and Sequence of Discovery.

(1) *Timing.* A party may not seek discovery from any source before the parties have conferred as required by Rule 26(f), except in a proceeding exempted from initial disclosure under Rule 26(a)(1)(B), or when authorized by these rules, by stipulation, or by court order.

(2) Early Rule 34 Requests.

Time to Deliver. More than 21 days after the summons and complaint are served on a party, a request under Rule 34 may be delivered:

(i) to that party by any other party, and

(ii) by that party to any plaintiff or to any other party that has been served.

(B) When Considered Served. The request is considered to have been served at the first Rule 26(f) conference.

(3) *Sequence.* Unless the parties stipulate or the court orders otherwise for the parties' and witnesses' convenience and in the interests of justice:

(A) methods of discovery may be used in any sequence; and

(B) discovery by one party does not require any other party to delay its discovery.

(e) Supplementing Disclosures and Responses.

(1) *In General.* A party who has made a disclosure under Rule 26(a)—or who has responded to an interrogatory, request for production, or request for admission—must supplement or correct its disclosure or response:

(A) in a timely manner if the party learns that in some material respect the disclosure or response is incomplete or incorrect, and if the additional or corrective information has not otherwise been made known to the other parties during the discovery process or in writing; or

(B) as ordered by the court.

(2) *Expert Witness.* For an expert whose report must be disclosed under Rule 26(a)(2)(B), the party's duty to supplement extends both to information included in the report and to information given during the expert's deposition. Any additions or changes to this information must be disclosed by the time the party's pretrial disclosures under Rule 26(a)(3) are due.

(f) Conference of the Parties; Planning for Discovery.

(1) *Conference Timing.* Except in a proceeding exempted from initial disclosure under Rule 26(a)(1)(B) or when the court orders otherwise, the parties must confer as soon as practicable—and in any event at least 21 days before a scheduling conference is to be held or a scheduling order is due under Rule 16(b).

(2) *Conference Content; Parties' Responsibilities.* In conferring, the parties must consider the nature and basis of their claims and defenses and the possibilities for promptly settling or resolving the case; make or arrange for the disclosures required by Rule 26(a)(1); discuss any issues about preserving discoverable information; and develop a proposed discovery plan. The attorneys of record and all unrepresented parties that have appeared in the case are jointly responsible for arranging the conference, for attempting in good faith to agree on the proposed discovery plan, and for submitting to the court within 14 days after the conference a written report outlining the plan. The court may order the parties or attorneys to attend the conference in person.

(3) *Discovery Plan.* A discovery plan must state the parties' views and proposals on:

(A) what changes should be made in the timing, form, or requirement for disclosures under Rule 26(a), including a statement of when initial disclosures were made or will be made;

(B) the subjects on which discovery may be needed, when discovery should be completed, and whether discovery should be conducted in phases or be limited to or focused on particular issues;

(C) any issues about disclosure, discovery, or preservation of electronically stored information, including the form or forms in which it should be produced;

(D) any issues about claims of privilege or of protection as trial-preparation materials, including—if the parties agree on a procedure to assert these claims after production—whether to ask the court to include their agreement in an order under Federal Rule of Evidence 502;

(E) what changes should be made in the limitations on discovery imposed under these rules or by local rule, and what other limitations should be imposed; and

(F) any other orders that the court should issue under Rule 26(c) or under Rule 16(b) and (c).

(4) *Expedited Schedule.* If necessary to comply with its expedited schedule for Rule 16(b) conferences, a court may by local rule:

(A) require the parties' conference to occur less than 21 days before the scheduling conference is held or a scheduling order is due under Rule 16(b); and

(B) require the written report outlining the discovery plan to be filed less than 14 days after the parties' conference, or excuse the parties from submitting a written report and permit them to report orally on their discovery plan at the Rule 16(b) conference.

(g) Signing Disclosures and Discovery Requests, Responses, and Objections.

(1) *Signature Required; Effect of Signature.* Every disclosure under Rule 26(a)(1) or (a)(3) and every discovery request, response, or objection must be signed by at least one attorney of record in the attorney's own name—or by the party personally, if unrepresented—and must state the signer's address, e-mail address, and telephone number. By signing, an attorney or party certifies that to the best of the person's knowledge, information, and belief formed after a reasonable inquiry:

(A) with respect to a disclosure, it is complete and correct as of the time it is made; and

(B) with respect to a discovery request, response, or objection, it is:

(i) consistent with these rules and warranted by existing law or by a nonfrivolous argument for extending, modifying, or reversing existing law, or for establishing new law;

(ii) not interposed for any improper purpose, such as to harass, cause unnecessary delay, or needlessly increase the cost of litigation; and

(iii) neither unreasonable nor unduly burdensome or expensive, considering the needs of the case, prior discovery in the case, the amount in controversy, and the importance of the issues at stake in the action.

(2) *Failure to Sign.* Other parties have no duty to act on an unsigned disclosure, request, response, or objection until it is signed, and the court must strike it unless a signature is promptly supplied after the omission is called to the attorney's or party's attention.

(3) *Sanction for Improper Certification.* If a certification violates this rule without substantial justification, the court, on motion or on its own, must impose an appropriate sanction on the signer, the party on whose behalf the signer was acting, or both. The sanction may include an order to pay the reasonable expenses, including attorney's fees, caused by the violation.

Notes

(As amended Dec. 27, 1946, eff. Mar. 19, 1948; Jan. 21, 1963, eff. July 1, 1963; Feb. 28, 1966, eff. July 1, 1966; Mar. 30, 1970, eff. July 1, 1970; Apr. 29, 1980, eff. Aug. 1, 1980; Apr. 28, 1983, eff. Aug. 1, 1983; Mar. 2, 1987, eff. Aug. 1, 1987; Apr. 22, 1993, eff. Dec. 1, 1993; Apr. 17, 2000, eff. Dec. 1, 2000; Apr. 12, 2006, eff. Dec. 1, 2006; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 28, 2010, eff. Dec. 1, 2010; Apr. 29, 2015, eff. Dec. 1, 2015.)

Committee Notes on Rules—2006 Amendment

Subdivision (a). Rule 26(a)(1)(B) is amended to parallel Rule 34(a) by recognizing that a party must disclose electronically stored information as well as documents that it may use to support its claims or defenses. The term "electronically stored information" has the same broad meaning in Rule 26(a)(1) as in Rule 34(a). This amendment is consistent with the 1993 addition of Rule 26(a)(1)(B). The term "data compilations" is deleted as unnecessary because it is a subset of both documents and electronically stored information.

Changes Made After Publication and Comment. As noted in the introduction [omitted], this provision was not included in the published rule. It is included as a conforming amendment, to make Rule 26(a)(1) consistent with the changes that were included in the published proposals.

[*Subdivision (a)(1)(E)*.] Civil forfeiture actions are added to the list of exemptions from Rule 26(a)(1) disclosure requirements. These actions are governed by new Supplemental Rule G. Disclosure is not likely to be useful.

Subdivision (b)(2). The amendment to Rule 26(b)(2) is designed to address issues raised by difficulties in locating, retrieving, and providing discovery of some electronically stored information. Electronic storage systems often make it easier to locate and retrieve information. These advantages are properly taken into account in determining the reasonable scope of discovery in a particular case. But some sources of electronically stored information can be accessed only with substantial burden and cost. In a particular case, these burdens and costs may make the information on such sources not reasonably accessible.

It is not possible to define in a rule the different types of technological features that may affect the burdens and costs of accessing electronically stored information. Information systems are designed to provide ready access to information used in regular ongoing activities. They also may be designed so as to provide ready access to information that is not regularly used. But a system may retain information on sources that are accessible only by incurring substantial burdens or costs. Subparagraph (B) is added to regulate discovery from such sources.

Under this rule, a responding party should produce electronically stored information that is relevant, not privileged, and reasonably accessible, subject to the (b)(2)(C) limitations that apply to all discovery. The responding party must also identify, by category or type, the sources containing potentially responsive information that it is neither searching nor producing. The identification should, to the extent possible, provide enough detail to enable the requesting party to evaluate the burdens and costs of providing the discovery and the likelihood of finding responsive information on the identified sources.

A party's identification of sources of electronically stored information as not reasonably accessible does not relieve the party of its common-law or statutory duties to preserve evidence. Whether a responding party is required to preserve unsearched sources of potentially responsive information that it believes are not reasonably accessible depends on the circumstances of each case. It is often useful for the parties to discuss this issue early in discovery.

The volume of—and the ability to search—much electronically stored information means that in many cases the responding party will be able to produce information from reasonably accessible sources that will fully satisfy the parties' discovery needs. In many circumstances the requesting party should obtain and evaluate the information from such sources before insisting that the responding party search and produce information contained on sources that are not reasonably accessible. If the requesting party continues to seek discovery of information from sources identified as not reasonably accessible, the parties should discuss the burdens and costs of accessing and retrieving the information, the needs that may establish good cause for requiring all or part of the requested discovery even if the information sought is not reasonably accessible, and conditions on obtaining and producing the information that may be appropriate.

If the parties cannot agree whether, or on what terms, sources identified as not reasonably accessible should be searched and discoverable information produced, the issue may be raised either by a motion to compel discovery or by a motion for a protective order. The parties must confer before bringing either motion. If the parties do not resolve the issue and the court must decide, the responding party must show that the identified sources of information are not reasonably accessible because of undue burden or cost. The requesting party may need discovery to test this assertion. Such discovery might take the form of requiring the responding party to conduct a sampling of information contained on the sources identified as not reasonably accessible;

allowing some form of inspection of such sources; or taking depositions of witnesses knowledgeable about the responding party's information systems.

Once it is shown that a source of electronically stored information is not reasonably accessible, the requesting party may still obtain discovery by showing good cause, considering the limitations of Rule 26(b)(2)(C) that balance the costs and potential benefits of discovery. The decision whether to require a responding party to search for and produce information that is not reasonably accessible depends not only on the burdens and costs of doing so, but also on whether those burdens and costs can be justified in the circumstances of the case. Appropriate considerations may include: (1) the specificity of the discovery request; (2) the quantity of information available from other and more easily accessed sources; (3) the failure to produce relevant information that seems likely to have existed but is no longer available on more easily accessed sources; (4) the likelihood of finding relevant, responsive information that cannot be obtained from other, more easily accessed sources; (5) predictions as to the importance and usefulness of the further information; (6) the importance of the issues at stake in the litigation; and (7) the parties' resources.

The responding party has the burden as to one aspect of the inquiry—whether the identified sources are not reasonably accessible in light of the burdens and costs required to search for, retrieve, and produce whatever responsive information may be found. The requesting party has the burden of showing that its need for the discovery outweighs the burdens and costs of locating, retrieving, and producing the information. In some cases, the court will be able to determine whether the identified sources are not reasonably accessible and whether the requesting party has shown good cause for some or all of the discovery, consistent with the limitations of Rule 26(b)(2)(C), through a single proceeding or presentation. The good-cause determination, however, may be complicated because the court and parties may know little about what information the sources identified as not reasonably accessible might contain, whether it is relevant, or how valuable it may be to the litigation. In such cases, the parties may need some focused discovery, which may include sampling of the sources, to learn more about what burdens and costs are involved in accessing the information, what the information consists of, and how valuable it is for the litigation in light of information that can be obtained by exhausting other opportunities for discovery.

The good-cause inquiry and consideration of the Rule 26(b)(2)(C) limitations are coupled with the authority to set conditions for discovery. The conditions may take the form of limits on the amount, type, or sources of information required to be accessed and produced. The conditions may also include payment by the requesting party of part or all of the reasonable costs of obtaining information from sources that are not reasonably accessible. A requesting party's willingness to share or bear the access costs may be weighed by the court in determining whether there is good

cause. But the producing party's burdens in reviewing the information for relevance and privilege may weigh against permitting the requested discovery.

The limitations of Rule 26(b)(2)(C) continue to apply to all discovery of electronically stored information, including that stored on reasonably accessible electronic sources.

Changes Made after Publication and Comment. This recommendation modifies the version of the proposed rule amendment as published. Responding to comments that the published proposal seemed to require identification of information that cannot be identified because it is not reasonably accessible, the rule text was clarified by requiring identification of sources that are not reasonably accessible. The test of reasonable accessibility was clarified by adding "because of undue burden or cost."

The published proposal referred only to a motion by the requesting party to compel discovery. The rule text has been changed to recognize that the responding party may wish to determine its search and potential preservation obligations by moving for a protective order.

The provision that the court may for good cause order discovery from sources that are not reasonably accessible is expanded in two ways. It now states specifically that the requesting party is the one who must show good cause, and it refers to consideration of the limitations on discovery set out in present Rule 26(b)(2)(i), (ii), and (iii).

The published proposal was added at the end of present Rule 26(b)(2). It has been relocated to become a new subparagraph (B), allocating present Rule 26(b)(2) to new subparagraphs (A) and (C). The Committee Note was changed to reflect the rule text revisions. It also was shortened. The shortening was accomplished in part by deleting references to problems that are likely to become antique as technology continues to evolve, and in part by deleting passages that were at a level of detail better suited for a practice manual than a Committee Note.

The changes from the published proposed amendment to Rule 26(b)(2) are set out below. [Omitted]

Subdivision (b)(5). The Committee has repeatedly been advised that the risk of privilege waiver, and the work necessary to avoid it, add to the costs and delay of discovery. When the review is of electronically stored information, the risk of waiver, and the time and effort required to avoid it, can increase substantially because of the volume of electronically stored information and the difficulty in ensuring that all information to be produced has in fact been reviewed. Rule 26(b)(5)(A) provides a procedure for a party that has withheld information on the basis of privilege or protection as trial-preparation material to make the claim so that the requesting party can decide whether to contest the claim and the court can resolve the dispute. Rule 26(b)(5)(B) is added to provide a procedure for a party to assert a claim of privilege or trial-preparation material protection

after information is produced in discovery in the action and, if the claim is contested, permit any party that received the information to present the matter to the court for resolution.

Rule 26(b)(5)(B) does not address whether the privilege or protection that is asserted after production was waived by the production. The courts have developed principles to determine whether, and under what circumstances, waiver results from inadvertent production of privileged or protected information. Rule 26(b)(5)(B) provides a procedure for presenting and addressing these issues. Rule 26(b)(5)(B) works in tandem with Rule 26(f), which is amended to direct the parties to discuss privilege issues in preparing their discovery plan, and which, with amended Rule 16(b), allows the parties to ask the court to include in an order any agreements the parties reach regarding issues of privilege or trial-preparation material protection. Agreements reached under Rule 26(f)(4) and orders including such agreements entered under Rule 16(b)(6) may be considered when a court determines whether a waiver has occurred. Such agreements and orders ordinarily control if they adopt procedures different from those in Rule 26(b)(5)(B).

A party asserting a claim of privilege or protection after production must give notice to the receiving party. That notice should be in writing unless the circumstances preclude it. Such circumstances could include the assertion of the claim during a deposition. The notice should be as specific as possible in identifying the information and stating the basis for the claim. Because the receiving party must decide whether to challenge the claim and may sequester the information and submit it to the court for a ruling on whether the claimed privilege or protection applies and whether it has been waived, the notice should be sufficiently detailed so as to enable the receiving party and the court to understand the basis for the claim and to determine whether waiver has occurred. Courts will continue to examine whether a claim of privilege or protection was made at a reasonable time when delay is part of the waiver determination under the governing law.

After receiving notice, each party that received the information must promptly return, sequester, or destroy the information and any copies it has. The option of sequestering or destroying the information is included in part because the receiving party may have incorporated the information in protected trial-preparation materials. No receiving party may use or disclose the information pending resolution of the privilege claim. The receiving party may present to the court the questions whether the information is privileged or protected as trial-preparation material, and whether the privilege or protection has been waived. If it does so, it must provide the court with the grounds for the privilege or protection specified in the producing party's notice, and serve all parties. In presenting the question, the party may use the content of the information only to the extent permitted by the applicable law of privilege, protection for trial-preparation material, and professional responsibility.

If a party disclosed the information to nonparties before receiving notice of a claim of privilege or protection as trial-preparation material, it must take reasonable steps to retrieve the information and to return it, sequester it until the claim is resolved, or destroy it.

Whether the information is returned or not, the producing party must preserve the information pending the court's ruling on whether the claim of privilege or of protection is properly asserted and whether it was waived. As with claims made under Rule 26(b)(5)(A), there may be no ruling if the other parties do not contest the claim.

Changes Made After Publication and Comment. The rule recommended for approval is modified from the published proposal. The rule is expanded to include trial-preparation protection claims in addition to privilege claims.

The published proposal referred to production "without intending to waive a claim of privilege." This reference to intent was deleted because many courts include intent in the factors that determine whether production waives privilege.

The published proposal required that the producing party give notice "within a reasonable time." The time requirement was deleted because it seemed to implicate the question whether production effected a waiver, a question not addressed by the rule, and also because a receiving party cannot practicably ignore a notice that it believes was unreasonably delayed. The notice procedure was further changed to require that the producing party state the basis for the claim.

Two statements in the published Note have been brought into the rule text. The first provides that the receiving party may not use or disclose the information until the claim is resolved. The second provides that if the receiving party disclosed the information before being notified, it must take reasonable steps to retrieve it. [1]

The rule text was expanded by adding a provision that the receiving party may promptly present the information to the court under seal for a determination of the claim.

The published proposal provided that the producing party must comply with Rule 26(b)(5)(A) after making the claim. This provision was deleted as unnecessary.

Changes are made in the Committee Note to reflect the changes in the rule text.

The changes from the published rule are shown below. [Omitted]

Subdivision (f). Rule 26(f) is amended to direct the parties to discuss discovery of electronically stored information during their discovery-planning conference. The rule focuses on "issues relating to disclosure or discovery of electronically stored information"; the discussion is not required in cases not involving electronic discovery, and the amendment imposes no additional requirements

in those cases. When the parties do anticipate disclosure or discovery of electronically stored information, discussion at the outset may avoid later difficulties or ease their resolution.

When a case involves discovery of electronically stored information, the issues to be addressed during the Rule 26(f) conference depend on the nature and extent of the contemplated discovery and of the parties' information systems. It may be important for the parties to discuss those systems, and accordingly important for counsel to become familiar with those systems before the conference. With that information, the parties can develop a discovery plan that takes into account the capabilities of their computer systems. In appropriate cases identification of, and early discovery from, individuals with special knowledge of a party's computer systems may be helpful.

The particular issues regarding electronically stored information that deserve attention during the discovery planning stage depend on the specifics of the given case. *See Manual for Complex Litigation* (4th) §40.25(2) (listing topics for discussion in a proposed order regarding meet-and-confer sessions). For example, the parties may specify the topics for such discovery and the time period for which discovery will be sought. They may identify the various sources of such information within a party's control that should be searched for electronically stored information. They may discuss whether the information is reasonably accessible to the party that has it, including the burden or cost of retrieving and reviewing the information. *See* Rule 26(b)(2)(B). Rule 26(f)(3) explicitly directs the parties to discuss the form or forms in which electronically stored information might be produced. The parties may be able to reach agreement on the forms of production, making discovery more efficient. Rule 34(b) is amended to permit a requesting party to specify the form or forms in which it wants electronically stored information produced. If the requesting party does not specify a form, Rule 34(b) directs the responding party to state the forms it intends to use in the production. Early discussion of the forms of production may facilitate the application of Rule 34(b) by allowing the parties to determine what forms of production will meet both parties' needs. Early identification of disputes over the forms of production may help avoid the expense and delay of searches or productions using inappropriate forms.

Rule 26(f) is also amended to direct the parties to discuss any issues regarding preservation of discoverable information during their conference as they develop a discovery plan. This provision applies to all sorts of discoverable information, but can be particularly important with regard to electronically stored information. The volume and dynamic nature of electronically stored information may complicate preservation obligations. The ordinary operation of computers involves both the automatic creation and the automatic deletion or overwriting of certain information. Failure to address preservation issues early in the litigation increases uncertainty and raises a risk of disputes.

The parties' discussion should pay particular attention to the balance between the competing needs to preserve relevant evidence and to continue routine operations critical to ongoing

activities. Complete or broad cessation of a party's routine computer operations could paralyze the party's activities. *Cf. Manual for Complex Litigation* (4th) §11.422 ("A blanket preservation order may be prohibitively expensive and unduly burdensome for parties dependent on computer systems for their day-to-day operations.") The parties should take account of these considerations in their discussions, with the goal of agreeing on reasonable preservation steps.

The requirement that the parties discuss preservation does not imply that courts should routinely enter preservation orders. A preservation order entered over objections should be narrowly tailored. Ex parte preservation orders should issue only in exceptional circumstances.

Rule 26(f) is also amended to provide that the parties should discuss any issues relating to assertions of privilege or of protection as trial-preparation materials, including whether the parties can facilitate discovery by agreeing on procedures for asserting claims of privilege or protection after production and whether to ask the court to enter an order that includes any agreement the parties reach. The Committee has repeatedly been advised about the discovery difficulties that can result from efforts to guard against waiver of privilege and work-product protection. Frequently parties find it necessary to spend large amounts of time reviewing materials requested through discovery to avoid waiving privilege. These efforts are necessary because materials subject to a claim of privilege or protection are often difficult to identify. A failure to withhold even one such item may result in an argument that there has been a waiver of privilege as to all other privileged materials on that subject matter. Efforts to avoid the risk of waiver can impose substantial costs on the party producing the material and the time required for the privilege review can substantially delay access for the party seeking discovery.

These problems often become more acute when discovery of electronically stored information is sought. The volume of such data, and the informality that attends use of e-mail and some other types of electronically stored information, may make privilege determinations more difficult, and privilege review correspondingly more expensive and time consuming. Other aspects of electronically stored information pose particular difficulties for privilege review. For example, production may be sought of information automatically included in electronic files but not apparent to the creator or to readers. Computer programs may retain draft language, editorial comments, and other deleted matter (sometimes referred to as "embedded data" or "embedded edits") in an electronic file but not make them apparent to the reader. Information describing the history, tracking, or management of an electronic file (sometimes called "metadata") is usually not apparent to the reader viewing a hard copy or a screen image. Whether this information should be produced may be among the topics discussed in the Rule 26(f) conference. If it is, it may need to be reviewed to ensure that no privileged information is included, further complicating the task of privilege review.

Parties may attempt to minimize these costs and delays by agreeing to protocols that minimize the risk of waiver. They may agree that the responding party will provide certain requested materials for initial examination without waiving any privilege or protection—sometimes known as a "quick peek." The requesting party then designates the documents it wishes to have actually produced. This designation is the Rule 34 request. The responding party then responds in the usual course, screening only those documents actually requested for formal production and asserting privilege claims as provided in Rule 26(b)(5)(A). On other occasions, parties enter agreements—sometimes called "clawback agreements"—that production without intent to waive privilege or protection should not be a waiver so long as the responding party identifies the documents mistakenly produced, and that the documents should be returned under those circumstances. Other voluntary arrangements may be appropriate depending on the circumstances of each litigation. In most circumstances, a party who receives information under such an arrangement cannot assert that production of the information waived a claim of privilege or of protection as trial-preparation material.

Although these agreements may not be appropriate for all cases, in certain cases they can facilitate prompt and economical discovery by reducing delay before the discovering party obtains access to documents, and by reducing the cost and burden of review by the producing party. A case-management or other order including such agreements may further facilitate the discovery process. Form 35 is amended to include a report to the court about any agreement regarding protections against inadvertent forfeiture or waiver of privilege or protection that the parties have reached, and Rule 16(b) is amended to recognize that the court may include such an agreement in a case- management or other order. If the parties agree to entry of such an order, their proposal should be included in the report to the court.

Rule 26(b)(5)(B) is added to establish a parallel procedure to assert privilege or protection as trial-preparation material after production, leaving the question of waiver to later determination by the court.

Changes Made After Publication and Comment. The Committee recommends a modified version of what was published. Rule 26(f)(3) was expanded to refer to the form "or forms" of production, in parallel with the like change in Rule 34. Different forms may be suitable for different sources of electronically stored information.

The published Rule 26(f)(4) proposal described the parties' views and proposals concerning whether, on their agreement, the court should enter an order protecting the right to assert privilege after production. This has been revised to refer to the parties' views and proposals concerning any issues relating to claims of privilege, including—if the parties agree on a procedure to assert such claims after production—whether to ask the court to include their agreement in an order. As with

Rule 16(b)(6), this change was made to avoid any implications as to the scope of the protection that may be afforded by court adoption of the parties' agreement.

Rule 26(f)(4) also was expanded to include trial-preparation materials.

The Committee Note was revised to reflect the changes in the rule text.

The changes from the published rule are shown below. [Omitted]

Committee Notes on Rules—2015 Amendment

Rule 26(b)(1) is changed in several ways.

Information is discoverable under revised Rule 26(b)(1) if it is relevant to any party's claim or defense and is proportional to the needs of the case. The considerations that bear on proportionality are moved from present Rule 26(b)(2)(C)(iii), slightly rearranged and with one addition.

Most of what now appears in Rule 26(b)(2)(C)(iii) was first adopted in 1983. The 1983 provision was explicitly adopted as part of the scope of discovery defined by Rule 26(b)(1). Rule 26(b)(1) directed the court to limit the frequency or extent of use of discovery if it determined that "the discovery is unduly burdensome or expensive, taking into account the needs of the case, the amount in controversy, limitations on the parties' resources, and the importance of the issues at stake in the litigation." At the same time, Rule 26(g) was added. Rule 26(g) provided that signing a discovery request, response, or objection certified that the request, response, or objection was "not unreasonable or unduly burdensome or expensive, given the needs of the case, the discovery already had in the case, the amount in controversy, and the importance of the issues at stake in the litigation." The parties thus shared the responsibility to honor these limits on the scope of discovery.

The 1983 Committee Note stated that the new provisions were added "to deal with the problem of overdiscovery. The objective is to guard against redundant or disproportionate discovery by giving the court authority to reduce the amount of discovery that may be directed to matters that are otherwise proper subjects of inquiry. The new sentence is intended to encourage judges to be more aggressive in identifying and discouraging discovery overuse. The grounds mentioned in the amended rule for limiting discovery reflect the existing practice of many courts in issuing protective orders under Rule 26(c).... On the whole, however, district judges have been reluctant to limit the use of the discovery devices."

The clear focus of the 1983 provisions may have been softened, although inadvertently, by the amendments made in 1993. The 1993 Committee Note explained: "[F]ormer paragraph (b)(1) [was] subdivided into two paragraphs for ease of reference and to avoid renumbering of paragraphs (3)

and (4).” Subdividing the paragraphs, however, was done in a way that could be read to separate the proportionality provisions as “limitations,” no longer an integral part of the (b)(1) scope provisions. That appearance was immediately offset by the next statement in the Note: “Textual changes are then made in new paragraph (2) to enable the court to keep tighter rein on the extent of discovery.”

The 1993 amendments added two factors to the considerations that bear on limiting discovery: whether “the burden or expense of the proposed discovery outweighs its likely benefit,” and “the importance of the proposed discovery in resolving the issues.” Addressing these and other limitations added by the 1993 discovery amendments, the Committee Note stated that “[t]he revisions in Rule 26(b)(2) are intended to provide the court with broader discretion to impose additional restrictions on the scope and extent of discovery . . . .”

The relationship between Rule 26(b)(1) and (2) was further addressed by an amendment made in 2000 that added a new sentence at the end of (b)(1): “All discovery is subject to the limitations imposed by Rule 26(b)(2)(i), (ii), and (iii)[now Rule 26(b)(2)(C)].” The Committee Note recognized that “[t]hese limitations apply to discovery that is otherwise within the scope of subdivision (b)(1).” It explained that the Committee had been told repeatedly that courts were not using these limitations as originally intended. “This otherwise redundant cross-reference has been added to emphasize the need for active judicial use of subdivision (b)(2) to control excessive discovery.”

The present amendment restores the proportionality factors to their original place in defining the scope of discovery. This change reinforces the Rule 26(g) obligation of the parties to consider these factors in making discovery requests, responses, or objections.

Restoring the proportionality calculation to Rule 26(b)(1) does not change the existing responsibilities of the court and the parties to consider proportionality, and the change does not place on the party seeking discovery the burden of addressing all proportionality considerations.

Nor is the change intended to permit the opposing party to refuse discovery simply by making a boilerplate objection that it is not proportional. The parties and the court have a collective responsibility to consider the proportionality of all discovery and consider it in resolving discovery disputes.

The parties may begin discovery without a full appreciation of the factors that bear on proportionality. A party requesting discovery, for example, may have little information about the burden or expense of responding. A party requested to provide discovery may have little information about the importance of the discovery in resolving the issues as understood by the requesting party. Many of these uncertainties should be addressed and reduced in the parties’ Rule 26(f) conference and in scheduling and pretrial conferences with the court. But if the parties

continue to disagree, the discovery dispute could be brought before the court and the parties' responsibilities would remain as they have been since 1983. A party claiming undue burden or expense ordinarily has far better information — perhaps the only information — with respect to that part of the determination. A party claiming that a request is important to resolve the issues should be able to explain the ways in which the underlying information bears on the issues as that party understands them. The court's responsibility, using all the information provided by the parties, is to consider these and all the other factors in reaching a case-specific determination of the appropriate scope of discovery.

The direction to consider the parties' relative access to relevant information adds new text to provide explicit focus on considerations already implicit in present Rule 26(b)(2)(C)(iii). Some cases involve what often is called "information asymmetry." One party — often an individual plaintiff — may have very little discoverable information. The other party may have vast amounts of information, including information that can be readily retrieved and information that is more difficult to retrieve. In practice these circumstances often mean that the burden of responding to discovery lies heavier on the party who has more information, and properly so.

Restoring proportionality as an express component of the scope of discovery warrants repetition of parts of the 1983 and 1993 Committee Notes that must not be lost from sight. The 1983 Committee Note explained that "[t]he rule contemplates greater judicial involvement in the discovery process and thus acknowledges the reality that it cannot always operate on a self-regulating basis." The 1993 Committee Note further observed that "[t]he information explosion of recent decades has greatly increased both the potential cost of wide-ranging discovery and the potential for discovery to be used as an instrument for delay or oppression." What seemed an explosion in 1993 has been exacerbated by the advent of e-discovery. The present amendment again reflects the need for continuing and close judicial involvement in the cases that do not yield readily to the ideal of effective party management. It is expected that discovery will be effectively managed by the parties in many cases. But there will be important occasions for judicial management, both when the parties are legitimately unable to resolve important differences and when the parties fall short of effective, cooperative management on their own.

It also is important to repeat the caution that the monetary stakes are only one factor, to be balanced against other factors. The 1983 Committee Note recognized "the significance of the substantive issues, as measured in philosophic, social, or institutional terms. Thus the rule recognizes that many cases in public policy spheres, such as employment practices, free speech, and other matters, may have importance far beyond the monetary amount involved." Many other substantive areas also may involve litigation that seeks relatively small amounts of money, or no money at all, but that seeks to vindicate vitally important personal or public values.

So too, consideration of the parties' resources does not foreclose discovery requests addressed to an impecunious party, nor justify unlimited discovery requests addressed to a wealthy party. The 1983 Committee Note cautioned that "[t]he court must apply the standards in an even-handed manner that will prevent use of discovery to wage a war of attrition or as a device to coerce a party, whether financially weak or affluent."

The burden or expense of proposed discovery should be determined in a realistic way. This includes the burden or expense of producing electronically stored information. Computer-based methods of searching such information continue to develop, particularly for cases involving large volumes of electronically stored information. Courts and parties should be willing to consider the opportunities for reducing the burden or expense of discovery as reliable means of searching electronically stored information become available.

A portion of present Rule 26(b)(1) is omitted from the proposed revision. After allowing discovery of any matter relevant to any party's claim or defense, the present rule adds: "including the existence, description, nature, custody, condition, and location of any documents or other tangible things and the identity and location of persons who know of any discoverable matter." Discovery of such matters is so deeply entrenched in practice that it is no longer necessary to clutter the long text of Rule 26 with these examples. The discovery identified in these examples should still be permitted under the revised rule when relevant and proportional to the needs of the case. Framing intelligent requests for electronically stored information, for example, may require detailed information about another party's information systems and other information resources.

The amendment deletes the former provision authorizing the court, for good cause, to order discovery of any matter relevant to the subject matter involved in the action. The Committee has been informed that this language is rarely invoked. Proportional discovery relevant to any party's claim or defense suffices, given a proper understanding of what is relevant to a claim or defense. The distinction between matter relevant to a claim or defense and matter relevant to the subject matter was introduced in 2000. The 2000 Note offered three examples of information that, suitably focused, would be relevant to the parties' claims or defenses. The examples were "other incidents of the same type, or involving the same product"; "information about organizational arrangements or filing systems"; and "information that could be used to impeach a likely witness." Such discovery is not foreclosed by the amendments. Discovery that is relevant to the parties' claims or defenses may also support amendment of the pleadings to add a new claim or defense that affects the scope of discovery.

The former provision for discovery of relevant but inadmissible information that appears "reasonably calculated to lead to the discovery of admissible evidence" is also deleted. The phrase has been used by some, incorrectly, to define the scope of discovery. As the Committee Note to the 2000 amendments observed, use of the "reasonably calculated" phrase to define the scope of

discovery "might swallow any other limitation on the scope of discovery." The 2000 amendments sought to prevent such misuse by adding the word "Relevant" at the beginning of the sentence, making clear that "'relevant' means within the scope of discovery as defined in this subdivision . . . ." The "reasonably calculated" phrase has continued to create problems, however, and is removed by these amendments. It is replaced by the direct statement that "Information within this scope of discovery need not be admissible in evidence to be discoverable." Discovery of nonprivileged information not admissible in evidence remains available so long as it is otherwise within the scope of discovery.

Rule 26(b)(2)(C)(iii) is amended to reflect the transfer of the considerations that bear on proportionality to Rule 26(b)(1). The court still must limit the frequency or extent of proposed discovery, on motion or on its own, if it is outside the scope permitted by Rule 26(b)(1).

Rule 26(c)(1)(B) is amended to include an express recognition of protective orders that allocate expenses for disclosure or discovery. Authority to enter such orders is included in the present rule, and courts already exercise this authority. Explicit recognition will forestall the temptation some parties may feel to contest this authority. Recognizing the authority does not imply that cost-shifting should become a common practice. Courts and parties should continue to assume that a responding party ordinarily bears the costs of responding.

Rule 26(d)(2) is added to allow a party to deliver Rule 34 requests to another party more than 21 days after that party has been served even though the parties have not yet had a required Rule 26(f) conference. Delivery may be made by any party to the party that has been served, and by that party to any plaintiff and any other party that has been served. Delivery does not count as service; the requests are considered to be served at the first Rule 26(f) conference. Under Rule 34(b)(2)(A) the time to respond runs from service. This relaxation of the discovery moratorium is designed to facilitate focused discussion during the Rule 26(f) conference. Discussion at the conference may produce changes in the requests. The opportunity for advance scrutiny of requests delivered before the Rule 26(f) conference should not affect a decision whether to allow additional time to respond.

Rule 26(d)(3) is renumbered and amended to recognize that the parties may stipulate to case-specific sequences of discovery.

Rule 26(f)(3) is amended in parallel with Rule 16(b)(3) to add two items to the discovery plan — issues about preserving electronically stored information and court orders under Evidence Rule 502.

***

Rule 34. Producing Documents, Electronically Stored Information, and Tangible Things, or Entering onto Land, for Inspection and Other Purposes

(a) In General. A party may serve on any other party a request within the scope of Rule 26(b):

(1) to produce and permit the requesting party or its representative to inspect, copy, test, or sample the following items in the responding party's possession, custody, or control:

(A) any designated documents or electronically stored information—including writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations—stored in any medium from which information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form; or

(B) any designated tangible things; or

(2) to permit entry onto designated land or other property possessed or controlled by the responding party, so that the requesting party may inspect, measure, survey, photograph, test, or sample the property or any designated object or operation on it.

(b) Procedure.

(1) *Contents of the Request.* The request:

(A) must describe with reasonable particularity each item or category of items to be inspected;

(B) must specify a reasonable time, place, and manner for the inspection and for performing the related acts; and

(C) may specify the form or forms in which electronically stored information is to be produced.

(2) *Responses and Objections.*

(A) *Time to Respond.* The party to whom the request is directed must respond in writing within 30 days after being served or — if the request was delivered under Rule 26(d)(2) — within 30 days after the parties' first Rule 26(f) conference. A shorter or longer time may be stipulated to under Rule 29 or be ordered by the court.

(B) *Responding to Each Item.* For each item or category, the response must either state that inspection and related activities will be permitted as requested or state with specificity the grounds for objecting to the request, including the reasons. The responding party may state that it will produce copies of documents or of electronically stored information instead of permitting inspection. The production must then be completed no later than the time for inspection specified in the request or another reasonable time specified in the response.

(C) *Objections.* An objection must state whether any responsive materials are being withheld on the basis of that objection. An objection to part of a request must specify the part and permit inspection of the rest.

(D) *Responding to a Request for Production of Electronically Stored Information.* The response may state an objection to a requested form for producing electronically stored information. If the responding party objects to a requested form—or if no form was specified in the request—the party must state the form or forms it intends to use.

(E) *Producing the Documents or Electronically Stored Information.* Unless otherwise stipulated or ordered by the court, these procedures apply to producing documents or electronically stored information:

(i) A party must produce documents as they are kept in the usual course of business or must organize and label them to correspond to the categories in the request;

(ii) If a request does not specify a form for producing electronically stored information, a party must produce it in a form or forms in which it is ordinarily maintained or in a reasonably usable form or forms; and

(iii) A party need not produce the same electronically stored information in more than one form.

(c) Nonparties. As provided in Rule 45, a nonparty may be compelled to produce documents and tangible things or to permit an inspection.

Notes

(As amended Dec. 27, 1946, eff. Mar. 19, 1948; Mar. 30, 1970, eff. July 1, 1970; Apr. 29, 1980, eff. Aug. 1, 1980; Mar. 2, 1987, eff. Aug. 1, 1987; Apr. 30, 1991, eff. Dec. 1, 1991; Apr. 22, 1993, eff. Dec. 1, 1993; Apr. 12, 2006, eff. Dec. 1, 2006; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 29, 2015, eff. Dec. 1, 2015.)

Committee Notes on Rules—2006 Amendment

Subdivision (a). As originally adopted, Rule 34 focused on discovery of "documents" and "things." In 1970, Rule 34(a) was amended to include discovery of data compilations, anticipating that the use of computerized information would increase. Since then, the growth in electronically stored information and in the variety of systems for creating and storing such information has been dramatic. Lawyers and judges interpreted the term "documents" to include electronically stored information because it was obviously improper to allow a party to evade discovery obligations on the basis that the label had not kept pace with changes in information technology. But it has become increasingly difficult to say that all forms of electronically stored information, many

dynamic in nature, fit within the traditional concept of a "document." Electronically stored information may exist in dynamic databases and other forms far different from fixed expression on paper. Rule 34(a) is amended to confirm that discovery of electronically stored information stands on equal footing with discovery of paper documents. The change clarifies that Rule 34 applies to information that is fixed in a tangible form and to information that is stored in a medium from which it can be retrieved and examined. At the same time, a Rule 34 request for production of "documents" should be understood to encompass, and the response should include, electronically stored information unless discovery in the action has clearly distinguished between electronically stored information and "documents."

Discoverable information often exists in both paper and electronic form, and the same or similar information might exist in both. The items listed in Rule 34(a) show different ways in which information may be recorded or stored. Images, for example, might be hard-copy documents or electronically stored information. The wide variety of computer systems currently in use, and the rapidity of technological change, counsel against a limiting or precise definition of electronically stored information. Rule 34(a)(1) is expansive and includes any type of information that is stored electronically. A common example often sought in discovery is electronic communications, such as e-mail. The rule covers—either as documents or as electronically stored information—information "stored in any medium," to encompass future developments in computer technology. Rule 34(a)(1) is intended to be broad enough to cover all current types of computer-based information, and flexible enough to encompass future changes and developments.

References elsewhere in the rules to "electronically stored information" should be understood to invoke this expansive approach. A companion change is made to Rule 33(d), making it explicit that parties choosing to respond to an interrogatory by permitting access to responsive records may do so by providing access to electronically stored information. More generally, the term used in Rule 34(a)(1) appears in a number of other amendments, such as those to Rules 26(a)(1), 26(b)(2), 26(b)(5)(B), 26(f), 34(b), 37(f), and 45. In each of these rules, electronically stored information has the same broad meaning it has under Rule 34(a)(1). References to "documents" appear in discovery rules that are not amended, including Rules 30(f), 36(a), and 37(c)(2). These references should be interpreted to include electronically stored information as circumstances warrant.

The term "electronically stored information" is broad, but whether material that falls within this term should be produced, and in what form, are separate questions that must be addressed under Rules 26(b), 26(c), and 34(b).

The Rule 34(a) requirement that, if necessary, a party producing electronically stored information translate it into reasonably usable form does not address the issue of translating from one human language to another. *See In re Puerto Rico Elect. Power Auth.*, 687 F.2d 501, 504–510 (1st Cir. 1989).

Rule 34(a)(1) is also amended to make clear that parties may request an opportunity to test or sample materials sought under the rule in addition to inspecting and copying them. That opportunity may be important for both electronically stored information and hard-copy materials. The current rule is not clear that such testing or sampling is authorized; the amendment expressly permits it. As with any other form of discovery, issues of burden and intrusiveness raised by requests to test or sample can be addressed under Rules 26(b)(2) and 26(c). Inspection or testing of certain types of electronically stored information or of a responding party's electronic information system may raise issues of confidentiality or privacy. The addition of testing and sampling to Rule 34(a) with regard to documents and electronically stored information is not meant to create a routine right of direct access to a party's electronic information system, although such access might be justified in some circumstances. Courts should guard against undue intrusiveness resulting from inspecting or testing such systems.

Rule 34(a)(1) is further amended to make clear that tangible things must—like documents and land sought to be examined—be designated in the request.

Subdivision (b). Rule 34(b) provides that a party must produce documents as they are kept in the usual course of business or must organize and label them to correspond with the categories in the discovery request. The production of electronically stored information should be subject to comparable requirements to protect against deliberate or inadvertent production in ways that raise unnecessary obstacles for the requesting party. Rule 34(b) is amended to ensure similar protection for electronically stored information.

The amendment to Rule 34(b) permits the requesting party to designate the form or forms in which it wants electronically stored information produced. The form of production is more important to the exchange of electronically stored information than of hard-copy materials, although a party might specify hard copy as the requested form. Specification of the desired form or forms may facilitate the orderly, efficient, and cost-effective discovery of electronically stored information. The rule recognizes that different forms of production may be appropriate for different types of electronically stored information. Using current technology, for example, a party might be called upon to produce word processing documents, e-mail messages, electronic spreadsheets, different image or sound files, and material from databases. Requiring that such diverse types of electronically stored information all be produced in the same form could prove impossible, and even if possible could increase the cost and burdens of producing and using the information. The rule therefore provides that the requesting party may ask for different forms of production for different types of electronically stored information.

The rule does not require that the requesting party choose a form or forms of production. The requesting party may not have a preference. In some cases, the requesting party may not know what form the producing party uses to maintain its electronically stored information, although Rule

26(f)(3) is amended to call for discussion of the form of production in the parties' prediscovery conference.

The responding party also is involved in determining the form of production. In the written response to the production request that Rule 34 requires, the responding party must state the form it intends to use for producing electronically stored information if the requesting party does not specify a form or if the responding party objects to a form that the requesting party specifies. Stating the intended form before the production occurs may permit the parties to identify and seek to resolve disputes before the expense and work of the production occurs. A party that responds to a discovery request by simply producing electronically stored information in a form of its choice, without identifying that form in advance of the production in the response required by Rule 34(b), runs a risk that the requesting party can show that the produced form is not reasonably usable and that it is entitled to production of some or all of the information in an additional form. Additional time might be required to permit a responding party to assess the appropriate form or forms of production.

If the requesting party is not satisfied with the form stated by the responding party, or if the responding party has objected to the form specified by the requesting party, the parties must meet and confer under Rule 37(a)(2)(B) in an effort to resolve the matter before the requesting party can file a motion to compel. If they cannot agree and the court resolves the dispute, the court is not limited to the forms initially chosen by the requesting party, stated by the responding party, or specified in this rule for situations in which there is no court order or party agreement.

If the form of production is not specified by party agreement or court order, the responding party must produce electronically stored information either in a form or forms in which it is ordinarily maintained or in a form or forms that are reasonably usable. Rule 34(a) requires that, if necessary, a responding party "translate" information it produces into a "reasonably usable" form. Under some circumstances, the responding party may need to provide some reasonable amount of technical support, information on application software, or other reasonable assistance to enable the requesting party to use the information. The rule does not require a party to produce electronically stored information in the form it [sic] which it is ordinarily maintained, as long as it is produced in a reasonably usable form. But the option to produce in a reasonably usable form does not mean that a responding party is free to convert electronically stored information from the form in which it is ordinarily maintained to a different form that makes it more difficult or burdensome for the requesting party to use the information efficiently in the litigation. If the responding party ordinarily maintains the information it is producing in a way that makes it searchable by electronic means, the information should not be produced in a form that removes or significantly degrades this feature.

Some electronically stored information may be ordinarily maintained in a form that is not reasonably usable by any party. One example is "legacy" data that can be used only by superseded systems. The questions whether a producing party should be required to convert such information to a more usable form, or should be required to produce it at all, should be addressed under Rule 26(b)(2)(B).

Whether or not the requesting party specified the form of production, Rule 34(b) provides that the same electronically stored information ordinarily be produced in only one form.

Changes Made after Publication and Comment. The proposed amendment recommended for approval has been modified from the published version. The sequence of "documents or electronically stored information" is changed to emphasize that the parenthetical exemplifications apply equally to illustrate "documents" and "electronically stored information." The reference to "detection devices" is deleted as redundant with "translated" and as archaic.

The references to the form of production are changed in the rule and Committee Note to refer also to "forms." Different forms may be appropriate or necessary for different sources of information.

The published proposal allowed the requesting party to specify a form for production and recognized that the responding party could object to the requested form. This procedure is now amplified by directing that the responding party state the form or forms it intends to use for production if the request does not specify a form or if the responding party objects to the requested form.

The default forms of production to be used when the parties do not agree on a form and there is no court order are changed in part. As in the published proposal, one default form is "a form or forms in which [electronically stored information] is ordinarily maintained." The alternative default form, however, is changed from "an electronically searchable form" to "a form or forms that are reasonably usable." "[A]n electronically searchable form" proved to have several defects. Some electronically stored information cannot be searched electronically. In addition, there often are many different levels of electronic searchability—the published default would authorize production in a minimally searchable form even though more easily searched forms might be available at equal or less cost to the responding party.

The provision that absent court order a party need not produce the same electronically stored information in more than one form was moved to become a separate item for the sake of emphasis.

The Committee Note was changed to reflect these changes in rule text, and also to clarify many aspects of the published Note. In addition, the Note was expanded to add a caveat to the published amendment that establishes the rule that documents—and now electronically stored information—may be tested and sampled as well as inspected and copied. Fears were expressed

that testing and sampling might imply routine direct access to a party's information system. The Note states that direct access is not a routine right, "although such access might be justified in some circumstances."

The changes in the rule text since publication are set out below. [Omitted]

Committee Notes on Rules—2015 Amendment

Several amendments are made in Rule 34, aimed at reducing the potential to impose unreasonable burdens by objections to requests to produce.

Rule 34(b)(2)(A) is amended to fit with new Rule 26(d)(2). The time to respond to a Rule 34 request delivered before the parties' Rule 26(f) conference is 30 days after the first Rule 26(f) conference.

Rule 34(b)(2)(B) is amended to require that objections to Rule 34 requests be stated with specificity. This provision adopts the language of Rule 33(b)(4), eliminating any doubt that less specific objections might be suitable under Rule 34. The specificity of the objection ties to the new provision in Rule 34(b)(2)(C) directing that an objection must state whether any responsive materials are being withheld on the basis of that objection. An objection may state that a request is overbroad, but if the objection recognizes that some part of the request is appropriate the objection should state the scope that is not overbroad. Examples would be a statement that the responding party will limit the search to documents or electronically stored information created within a given period of time prior to the events in suit, or to specified sources. When there is such an objection, the statement of what has been withheld can properly identify as matters "withheld" anything beyond the scope of the search specified in the objection.

Rule 34(b)(2)(B) is further amended to reflect the common practice of producing copies of documents or electronically stored information rather than simply permitting inspection. The response to the request must state that copies will be produced. The production must be completed either by the time for inspection specified in the request or by another reasonable time specifically identified in the response. When it is necessary to make the production in stages the response should specify the beginning and end dates of the production.

Rule 34(b)(2)(C) is amended to provide that an objection to a Rule 34 request must state whether anything is being withheld on the basis of the objection. This amendment should end the confusion that frequently arises when a producing party states several objections and still produces information, leaving the requesting party uncertain whether any relevant and responsive information has been withheld on the basis of the objections. The producing party does not need to provide a detailed description or log of all documents withheld, but does need to alert other parties to the fact that documents have been withheld and thereby facilitate an informed

discussion of the objection. An objection that states the limits that have controlled the search for responsive and relevant materials qualifies as a statement that the materials have been "withheld."

***

Rule 45. Subpoena

(a) In General.

(1) *Form and Contents.*

(A) *Requirements—In General.* Every subpoena must:

(i) state the court from which it issued;

(ii) state the title of the action and its civil-action number;

(iii) command each person to whom it is directed to do the following at a specified time and place: attend and testify; produce designated documents, electronically stored information, or tangible things in that person's possession, custody, or control; or permit the inspection of premises; and

(iv) set out the text of Rule 45(d) and (e).

(B) *Command to Attend a Deposition—Notice of the Recording Method.* A subpoena commanding attendance at a deposition must state the method for recording the testimony.

(C) *Combining or Separating a Command to Produce or to Permit Inspection; Specifying the Form for Electronically Stored Information.* A command to produce documents, electronically stored information, or tangible things or to permit the inspection of premises may be included in a subpoena commanding attendance at a deposition, hearing, or trial, or may be set out in a separate subpoena. A subpoena may specify the form or forms in which electronically stored information is to be produced.

(D) *Command to Produce; Included Obligations.* A command in a subpoena to produce documents, electronically stored information, or tangible things requires the responding person to permit inspection, copying, testing, or sampling of the materials.

(2) *Issuing Court.* A subpoena must issue from the court where the action is pending.

(3) *Issued by Whom.* The clerk must issue a subpoena, signed but otherwise in blank, to a party who requests it. That party must complete it before service. An attorney also may issue and sign a subpoena if the attorney is authorized to practice in the issuing court.

(4) *Notice to Other Parties Before Service.* If the subpoena commands the production of documents, electronically stored information, or tangible things or the inspection of premises before trial, then before it is served on the person to whom it is directed, a notice and a copy of the subpoena must be served on each party.

(b) Service.

(1) *By Whom and How; Tendering Fees.* Any person who is at least 18 years old and not a party may serve a subpoena. Serving a subpoena requires delivering a copy to the named person and, if the subpoena requires that person's attendance, tendering the fees for 1 day's attendance and the mileage allowed by law. Fees and mileage need not be tendered when the subpoena issues on behalf of the United States or any of its officers or agencies.

(2) *Service in the United States.* A subpoena may be served at any place within the United States.

(3) *Service in a Foreign Country.* 28 U.S.C. §1783 governs issuing and serving a subpoena directed to a United States national or resident who is in a foreign country.

(4) *Proof of Service.* Proving service, when necessary, requires filing with the issuing court a statement showing the date and manner of service and the names of the persons served. The statement must be certified by the server.

(c) Place of Compliance.

(1) *For a Trial, Hearing, or Deposition.* A subpoena may command a person to attend a trial, hearing, or deposition only as follows:

(A) within 100 miles of where the person resides, is employed, or regularly transacts business in person; or

(B) within the state where the person resides, is employed, or regularly transacts business in person, if the person

(i) is a party or a party's officer; or

(ii) is commanded to attend a trial and would not incur substantial expense.

(2) *For Other Discovery.* A subpoena may command:

(A) production of documents, electronically stored information, or tangible things at a place within 100 miles of where the person resides, is employed, or regularly transacts business in person; and

(B) inspection of premises at the premises to be inspected.

(d) Protecting a Person Subject to a Subpoena; Enforcement.

(1) *Avoiding Undue Burden or Expense; Sanctions.* A party or attorney responsible for issuing and serving a subpoena must take reasonable steps to avoid imposing undue burden or expense on a person subject to the subpoena. The court for the district where compliance is required must enforce this duty and impose an appropriate sanction—which may include lost earnings and reasonable attorney's fees—on a party or attorney who fails to comply.

(2) *Command to Produce Materials or Permit Inspection.*

(A) *Appearance Not Required.* A person commanded to produce documents, electronically stored information, or tangible things, or to permit the inspection of premises, need not appear in person at the place of production or inspection unless also commanded to appear for a deposition, hearing, or trial.

(B) *Objections.* A person commanded to produce documents or tangible things or to permit inspection may serve on the party or attorney designated in the subpoena a written objection to inspecting, copying, testing or sampling any or all of the materials or to inspecting the premises—or to producing electronically stored information in the form or forms requested. The objection must be served before the earlier of the time specified for compliance or 14 days after the subpoena is served. If an objection is made, the following rules apply:

(i) At any time, on notice to the commanded person, the serving party may move the court for the district where compliance is required for an order compelling production or inspection.

(ii) These acts may be required only as directed in the order, and the order must protect a person who is neither a party nor a party's officer from significant expense resulting from compliance.

(3) *Quashing or Modifying a Subpoena.*

(A) *When Required.* On timely motion, the court for the district where compliance is required must quash or modify a subpoena that:

(i) fails to allow a reasonable time to comply;

(ii) requires a person to comply beyond the geographical limits specified in Rule 45(c);

(iii) requires disclosure of privileged or other protected matter, if no exception or waiver applies; or

(iv) subjects a person to undue burden.

(B) *When Permitted.* To protect a person subject to or affected by a subpoena, the court for the district where compliance is required may, on motion, quash or modify the subpoena if it requires:

(i) disclosing a trade secret or other confidential research, development, or commercial information; or

(ii) disclosing an unretained expert's opinion or information that does not describe specific occurrences in dispute and results from the expert's study that was not requested by a party.

(C) *Specifying Conditions as an Alternative.* In the circumstances described in Rule 45(d)(3)(B), the court may, instead of quashing or modifying a subpoena, order appearance or production under specified conditions if the serving party:

(i) shows a substantial need for the testimony or material that cannot be otherwise met without undue hardship; and

(ii) ensures that the subpoenaed person will be reasonably compensated.

(e) Duties in Responding to a Subpoena.

(1) *Producing Documents or Electronically Stored Information.* These procedures apply to producing documents or electronically stored information:

(A) *Documents.* A person responding to a subpoena to produce documents must produce them as they are kept in the ordinary course of business or must organize and label them to correspond to the categories in the demand.

(B) *Form for Producing Electronically Stored Information Not Specified.* If a subpoena does not specify a form for producing electronically stored information, the person responding must produce it in a form or forms in which it is ordinarily maintained or in a reasonably usable form or forms.

(C) *Electronically Stored Information Produced in Only One Form.* The person responding need not produce the same electronically stored information in more than one form.

(D) *Inaccessible Electronically Stored Information.* The person responding need not provide discovery of electronically stored information from sources that the person identifies as not reasonably accessible because of undue burden or cost. On motion to compel discovery or for a protective order, the person responding must show that the information is not reasonably accessible because of undue burden or cost. If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)(2)(C). The court may specify conditions for the discovery.

(2) *Claiming Privilege or Protection.*

(A) *Information Withheld.* A person withholding subpoenaed information under a claim that it is privileged or subject to protection as trial-preparation material must:

(i) expressly make the claim; and

(ii) describe the nature of the withheld documents, communications, or tangible things in a manner that, without revealing information itself privileged or protected, will enable the parties to assess the claim.

(B) *Information Produced.* If information produced in response to a subpoena is subject to a claim of privilege or of protection as trial-preparation material, the person making the claim may notify any party that received the information of the claim and the basis for it. After being notified, a party must promptly return, sequester, or destroy the specified information and any copies it has; must not use or disclose the information until the claim is resolved; must take reasonable steps to retrieve the information if the party disclosed it before being notified; and may promptly present the information under seal to the court for the district where compliance is required for a determination of the claim. The person who produced the information must preserve the information until the claim is resolved.

(f) Transferring a Subpoena-Related Motion. When the court where compliance is required did not issue the subpoena, it may transfer a motion under this rule to the issuing court if the person subject to the subpoena consents or if the court finds exceptional circumstances. Then, if the attorney for a person subject to a subpoena is authorized to practice in the court where the motion was made, the attorney may file papers and appear on the motion as an officer of the issuing court. To enforce its order, the issuing court may transfer the order to the court where the motion was made.

(g) Contempt. The court for the district where compliance is required — and also, after a motion is transferred, the issuing court — may hold in contempt a person who, having been served, fails without adequate excuse to obey the subpoena or an order related to it.

Notes

(As amended Dec. 27, 1946, eff. Mar. 19, 1948; Dec. 29, 1948, eff. Oct. 20, 1949; Mar. 30, 1970, eff. July 1, 1970; Apr. 29, 1980, eff. Aug. 1, 1980; Apr. 29, 1985, eff. Aug. 1, 1985; Mar. 2, 1987, eff. Aug. 1, 1987; Apr. 30, 1991, eff. Dec. 1, 1991; Apr. 25, 2005, eff. Dec. 1, 2005; Apr. 12, 2006, eff. Dec. 1, 2006; Apr. 30, 2007, eff. Dec. 1, 2007; Apr. 16, 2013, eff. Dec. 1, 2013.)

Committee Notes on Rules—2006 Amendment

Rule 45 is amended to conform the provisions for subpoenas to changes in other discovery rules, largely related to discovery of electronically stored information. Rule 34 is amended to provide in

greater detail for the production of electronically stored information. Rule 45(a)(1)(C) is amended to recognize that electronically stored information, as defined in Rule 34(a), can also be sought by subpoena. Like Rule 34(b), Rule 45(a)(1) is amended to provide that the subpoena can designate a form or forms for production of electronic data. Rule 45(c)(2) is amended, like Rule 34(b), to authorize the person served with a subpoena to object to the requested form or forms. In addition, as under Rule 34(b), Rule 45(d)(1)(B) is amended to provide that if the subpoena does not specify the form or forms for electronically stored information, the person served with the subpoena must produce electronically stored information in a form or forms in which it is usually maintained or in a form or forms that are reasonably usable. Rule 45(d)(1)(C) is added to provide that the person producing electronically stored information should not have to produce the same information in more than one form unless so ordered by the court for good cause.

As with discovery of electronically stored information from parties, complying with a subpoena for such information may impose burdens on the responding person. Rule 45(c) provides protection against undue impositions on nonparties. For example, Rule 45(c)(1) directs that a party serving a subpoena "shall take reasonable steps to avoid imposing undue burden or expense on a person subject to the subpoena," and Rule 45(c)(2)(B) permits the person served with the subpoena to object to it and directs that an order requiring compliance "shall protect a person who is neither a party nor a party's officer from significant expense resulting from" compliance. Rule 45(d)(1)(D) is added to provide that the responding person need not provide discovery of electronically stored information from sources the party identifies as not reasonably accessible, unless the court orders such discovery for good cause, considering the limitations of Rule 26(b)(2)(C), on terms that protect a nonparty against significant expense. A parallel provision is added to Rule 26(b)(2).

Rule 45(a)(1)(B) is also amended, as is Rule 34(a), to provide that a subpoena is available to permit testing and sampling as well as inspection and copying. As in Rule 34, this change recognizes that on occasion the opportunity to perform testing or sampling may be important, both for documents and for electronically stored information. Because testing or sampling may present particular issues of burden or intrusion for the person served with the subpoena, however, the protective provisions of Rule 45(c) should be enforced with vigilance when such demands are made. Inspection or testing of certain types of electronically stored information or of a person's electronic information system may raise issues of confidentiality or privacy. The addition of sampling and testing to Rule 45(a) with regard to documents and electronically stored information is not meant to create a routine right of direct access to a person's electronic information system, although such access might be justified in some circumstances. Courts should guard against undue intrusiveness resulting from inspecting or testing such systems.

Rule 45(d)(2) is amended, as is Rule 26(b)(5), to add a procedure for assertion of privilege or of protection as trial-preparation materials after production. The receiving party may submit the information to the court for resolution of the privilege claim, as under Rule 26(b)(5)(B).

Federal Rule of Evidence 502

Attorney-Client Privilege and Work Product; Limitations on Waiver

The following provisions apply, in the circumstances set out, to disclosure of a communication or information covered by the attorney-client privilege or work-product protection.

(a) Disclosure Made in a Federal Proceeding or to a Federal Office or Agency; Scope of a Waiver. When the disclosure is made in a federal proceeding or to a federal office or agency and waives the attorney-client privilege or work-product protection, the waiver extends to an undisclosed communication or information in a federal or state proceeding only if:

(1) the waiver is intentional;
(2) the disclosed and undisclosed communications or information concern the same subject matter; and
(3) they ought in fairness to be considered together.

(b) Inadvertent Disclosure. When made in a federal proceeding or to a federal office or agency, the disclosure does not operate as a waiver in a federal or state proceeding if:

(1) the disclosure is inadvertent; (2) the holder of the privilege or protection took reasonable steps to prevent disclosure; and (3) the holder promptly took reasonable steps to rectify the error, including (if applicable) following Federal Rule of Civil Procedure 26 (b)(5)(B).

(c) Disclosure Made in a State Proceeding. When the disclosure is made in a state proceeding and is not the subject of a state-court order concerning waiver, the disclosure does not operate as a waiver in a federal proceeding if the disclosure:

(1) would not be a waiver under this rule if it had been made in a federal proceeding; or (2) is not a waiver under the law of the state where the disclosure occurred.

(d) Controlling Effect of a Court Order. A federal court may order that the privilege or protection is not waived by disclosure connected with the litigation pending before the court — in which event the disclosure is also not a waiver in any other federal or state proceeding.

(e) Controlling Effect of a Party Agreement. An agreement on the effect of disclosure in a federal proceeding is binding only on the parties to the agreement, unless it is incorporated into a court order.

(f) Controlling Effect of this Rule. Notwithstanding Rules 101 and 1101, this rule applies to state proceedings and to federal court-annexed and federal court-mandated arbitration proceedings, in the circumstances set out in the rule. And notwithstanding Rule 501, this rule applies even if state law provides the rule of decision.

(g) Definitions. In this rule:

(1) “attorney-client privilege” means the protection that applicable law provides for confidential attorney-client communications; and

(2) “work-product protection” means the protection that applicable law provides for tangible material (or its intangible equivalent) prepared in anticipation of litigation or for trial.

FRE Rule 902. Evidence That Is Self-Authenticating *[as amended effective 2017, footnote added]*

The following items of evidence are self-authenticating; they require no extrinsic evidence of authenticity in order to be admitted:

…

(13) Certified Records Generated by an Electronic Process or System. A record generated by an electronic process or system that produces an accurate result, as shown by a certification of a qualified person that complies with the certification requirements of Rule 902(11) or (12).[134] The proponent must also meet the notice requirements of Rule 902(11).

(14) Certified Data Copied from an Electronic Device, Storage Medium, or File. Data copied from an electronic device, storage medium, or file, if authenticated by a process of digital identification, as shown by a certification of a qualified person that complies with the certification requirements of Rule (902(11) or (12). The proponent also must meet the notice requirements of Rule 902 (11).

Committee Notes on Rules—2017 Amendment

Paragraph (14). The amendment sets forth a procedure by which parties can authenticate data copied from an electronic device, storage medium, or an electronic file, other than through the testimony of a foundation witness. As with the provisions on business records in Rules 902(11) and (12), the Committee has found that the expense and inconvenience of producing an authenticating witness for this evidence is often unnecessary. It is often the case that a party goes to the expense of producing an authentication witness, and then the adversary either stipulates authenticity before the witness is called or fails to challenge the authentication testimony once it is presented. The amendment provides a procedure in which the parties can determine in advance of trial whether a real challenge to authenticity will be made, and can then plan accordingly.

Today, data copied from electronic devices, storage media, and electronic files are ordinarily authenticated by "hash value". A hash value is a number that is often represented as a sequence

---

[134] **(11) *Certified Domestic Records of a Regularly Conducted Activity*.** The original or a copy of a domestic record that meets the requirements of Rule 803(6)(A)-(C), as shown by a certification of the custodian or another qualified person that complies with a federal statute or a rule prescribed by the Supreme Court. Before the trial or hearing, the proponent must give an adverse party reasonable written notice of the intent to offer the record — and must make the record and certification available for inspection — so that the party has a fair opportunity to challenge them.
(12) *Certified Foreign Records of a Regularly Conducted Activity*. In a civil case, the original or a copy of a foreign record that meets the requirements of Rule 902(11), modified as follows: the certification, rather than complying with a federal statute or Supreme Court rule, must be signed in a manner that, if falsely made, would subject the maker to a criminal penalty in the country where the certification is signed. The proponent must also meet the notice requirements of Rule 902(11).

of characters and is produced by an algorithm based upon the digital contents of a drive, medium, or file. If the hash values for the original and copy are different, then the copy is not identical to the original. If the hash values for the original and copy are the same, it is highly improbable that the original and copy are not identical. Thus, identical hash values for the original and copy reliably attest to the fact that they are exact duplicates. This amendment allows self-authentication by a certification of a qualified person that she checked the hash value of the proffered item and that it was identical to the original. The rule is flexible enough to allow certifications through processes other than comparison of hash value, including by other reliable means of identification provided by future technology.

Nothing in the amendment is intended to limit a party from establishing authenticity of electronic evidence on any ground provided in these Rules, including through judicial notice where appropriate.

A proponent establishing authenticity under this Rule must present a certification containing information that would be sufficient to establish authenticity were that information provided by a witness at trial. If the certification provides information that would be insufficient to authenticate the record of the certifying person testified, then authenticity is not established under this Rule.

The reference to the "certification requirements of Rule 902(11) or (12)" is only to the procedural requirements for a valid certification. There is no intent to require, or permit, a certification under this Rule to prove the requirements of Rule 803(6). Rule 902(14) is solely limited to authentication, and any attempt to satisfy a hearsay exception must be made independently.

A certification under this Rule can only establish that the proffered item is authentic. The opponent remains free to object to admissibility of the proffered item on other grounds—including hearsay, relevance, or in criminal cases the right to confrontation. For example, in a criminal case in which data copied from a hard drive is proffered, the defendant can still challenge hearsay found in the hard drive, and can still challenge whether the information on the hard drive was placed there by the defendant.

A challenge to the authenticity of electronic evidence may require technical information about the system or process at issue, including possibly retaining a forensic technical expert; such factors will affect whether the opponent has a fair opportunity to challenge the evidence given the notice provided.

The reference to Rule 902(12) is intended to cover certifications that are made in a foreign country.

Exemplar FRE 902(14) Certification

The Rules don't supply the language required to appear in an FRE 902 (14) certification; however, a declaration like the following would seem to suffice to prove up a duplicate, amended to fit the particular data and methods used and the declarant. *Be sure to make clear that the baseline hash value used for comparison and authentication was calculated contemporaneously with the acquisition of the copy or image being authenticated*. A hash value only describes data at the time of hashing, so may only authenticate the contents of a dataset as they existed at the time the contents were hashed. *Note: this only serves to prove identicality, not cure hearsay objections.*

My name is_____. I am a domiciliary of_____, over the age of eighteen years and competent and qualified to make this declaration.

The supplied data was transferred to me on a [*DESCRIBE TRANSFER MEDIA*], serial number _____, and consisted of:

[*IDENTIFY AND DESCRIBE DATASET*]

Based on information supplied to me, [*PERSON MAKING DUPLICATE*] copied the source data using [*IDENTIFY TOOL*]. I am familiar with this tool and have used it routinely and successfully over many years to create forensically sound images of digital media. I know it to be a capable tool for forensically-sound duplication of digital media and one widely accepted as standard and reliable by digital forensic experts. Based on the reporting supplied and routinely generated by the tool, the acquisition completed successfully, imaging ____sectors (___GB) without identification of bad sectors.

The reported [*MD5/SHA1/SHA256*] hash value of the data supplied when acquired was [*HASH VALUE*]. Using [*IDENTIFY TOOL*], I verified that the data supplied to me matched the hash value obtained when the data was acquired. I am trained and experienced in the generation and use of cryptographic hash values for testing and authenticating the integrity of digital evidence.

On [*DATE*], I re-verified the integrity of the drive image used in my work in this case. The [*MD5/SHA1/SHA256*] hash value is unchanged from the hash value calculated at acquisition. Accordingly, I can attest that the data received by me is authentic in that it was and remains an exact duplicate of the contents of the [SOURCE MEDIA], serial number: _____, attributed to [IDENTIFY SOURCE DEVICE AND CUSTODIAN] and originally acquired on [*ACQUISITION DATE*]. All my processing, extraction, production and interpretation of the contents of the drive image were made using the hash-authenticated data. <SIGNED>

Texas Rules of Civil Procedure

Part II – Rules of Practice in District and County Court

TRCP Rule 196.4 Electronic or Magnetic Data (enacted 1999)

To obtain discovery of data or information that exists in electronic or magnetic form, the requesting party must specifically request production of electronic or magnetic data and specify the form in which the requesting party wants it produced. The responding party must produce the electronic or magnetic data that is responsive to the request and is reasonably available to the responding party in its ordinary course of business. If the responding party cannot – through reasonable efforts – retrieve the data or information requested or produce it in the form requested, the responding party must state an objection complying with these rules. If the court orders the responding party to comply with the request, the court must also order that the requesting party pay the reasonable expenses of any extraordinary steps required to retrieve and produce the information.

Questions to Consider:

1. What is entailed in specifically requesting electronic or magnetic data?
2. How much specificity is needed to specify the form of production sought?
3. Are the native forms of data as created, used and stored by the responding party's software and systems synonymous with the forms "reasonably available to the responding party in its ordinary course of business?"
4. Is a responding party obliged by this rule to undertake reasonable efforts to retrieve requested data? Is it ever appropriate to object without making any effort?
5. What sorts of circumstances would give rise to "extraordinary" steps required to retrieve and produce information so as to trigger mandatory cost shifting? If litigation is "extraordinary," are any steps to respond to e-discovery also extraordinary?
6. What is the role of "proportionality" in limiting the scope of discovery and appropriate forms of production?