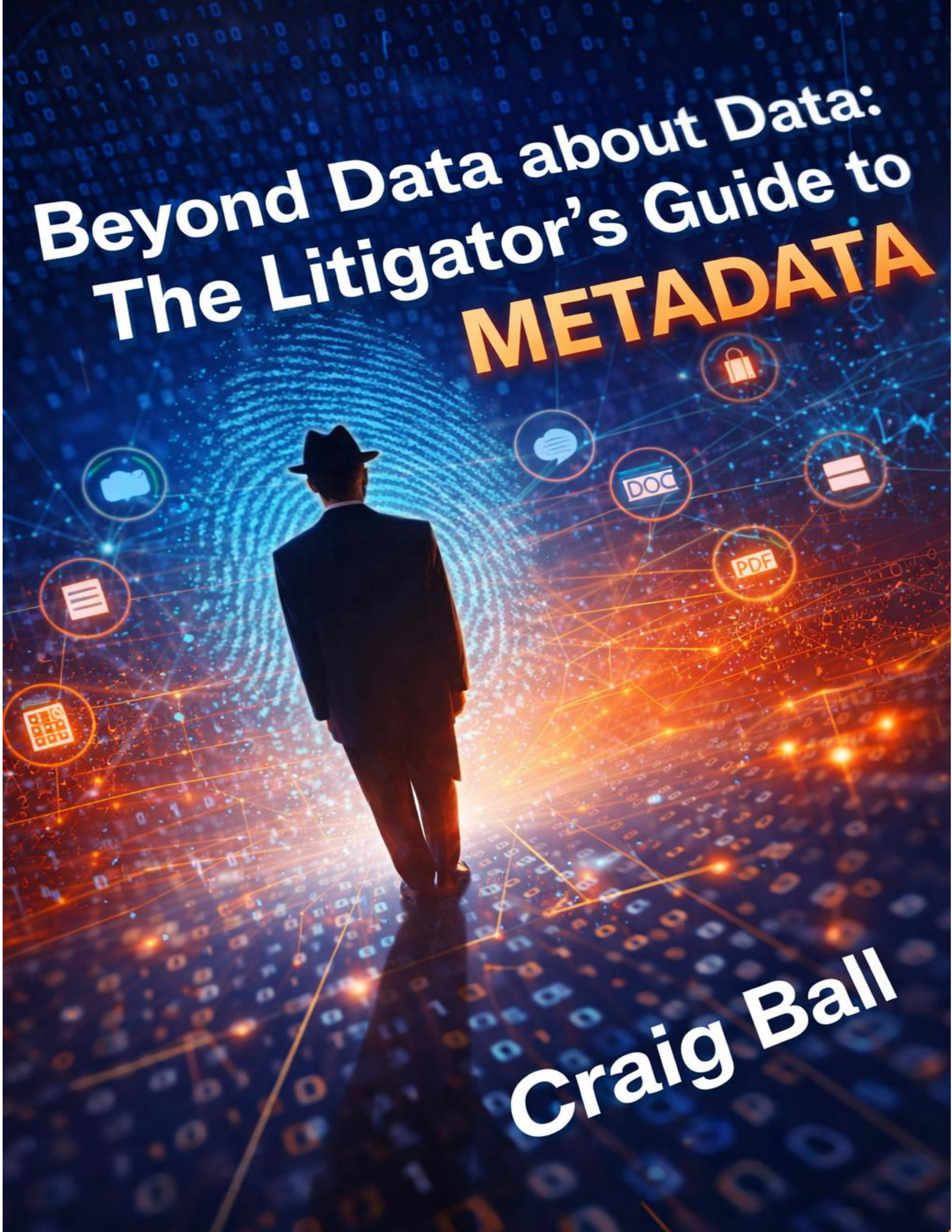


Beyond Data about Data: The Litigator's Guide to METADATA

Craig Ball

Beyond Data about Data: The Litigator’s Guide to Metadata

Craig Ball

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[1] 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 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.[2]

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.

[1] I cringe every time I refer to digital information as being stored as "ones and zeroes" because that's just a convenient way to *notate* the data, <u>not</u> how it's stored on digital media.

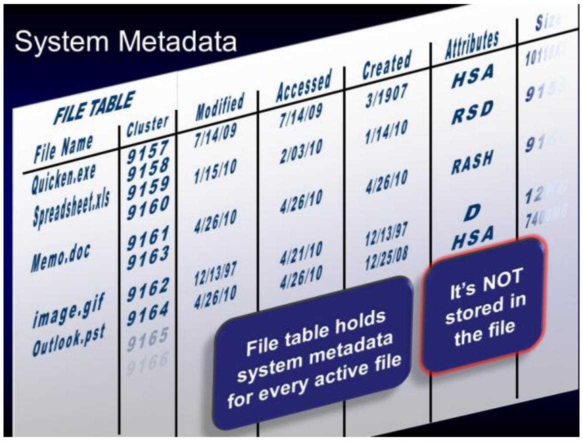
[2] On modern storage systems, including solid-state drives and virtualized storage, these locations are *logical* rather than fixed physical positions, and may change over time without the file system's awareness.

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.[3] 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 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

[3] 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.

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, **a**nd, where enabled, 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.

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

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.

Matryoshka Metadata
File System
PST
Message
DOC

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*.[4]

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

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

[4] 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.

preservation and production of metadata and act to meet changed circumstances consistent with proportionality and evolving needs.

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.[5]

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

[5] 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.

- 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.

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.[6]

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. 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,

[6] 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.

- 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, and prior metadata values may be lost. Such alteration can impair chronology, complicate review, and invite spoliation claims.[7] 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, inconsistently 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

[7] 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.

modern review tools are designed to avoid metadata alteration. Finally, containerized copies[8] 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

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 evidence processing 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.

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

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.”[9] 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,

[9] 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.

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.

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.

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

Collect Metadata

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,[10] 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.

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

[10] 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.

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 how 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

SECTOR
512bytes

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]*