



# **E-Discovery: Right...from the Start**

**Craig Ball**

## **Employment Law Collection**

**Right from the Start: 8 Ways to Hit the Ground Running**  
**Surefire Steps to Splendid Search**

**Geek Speak: Lawyer's Guide to the Language of Data Storage and Networking**  
**Meeting the Challenge of E-Mail in Civil Discovery**

**Preservation of ESI after Layoffs Discovery**

**Selecting, Engaging and Working with E-Discovery Service Providers**

**Piecing Together the E-Discovery Plan: a Plaintiff's Guide to Meet and Confer**  
**Selected BIYC Columns on Electronic Discovery for Employment Lawyers**

**E-Discovery:  
Right...from the Start**  
Employment Law Collection  
Craig Ball

<b>About this Collection .....</b>	<b>2</b>
<b>Right from the Start: 8 Ways to Hit the Ground Running.....</b>	<b>5</b>
<b>Surefire Steps to Splendid Search.....</b>	<b>8</b>
<b>Geek Speak: Lawyer’s Guide to the Language of Data Storage and Networking..</b>	<b>17</b>
<b>Meeting the Challenge of E-Mail in Civil Discovery .....</b>	<b>37</b>
<b>Preservation of ESI after Layoffs .....</b>	<b>75</b>
<b>Selecting, Engaging and Working with E-Discovery Service Providers.....</b>	<b>80</b>
<b>Piecing Together the E-Discovery Plan: a Plaintiff’s Guide to Meet and Confer ..</b>	<b>88</b>
<b>Selected BIYC Columns on Electronic Discovery for Employment Lawyers .....</b>	<b>103</b>
<b>About the Author .....</b>	<b>163</b>

**About this Collection**

I'm not an employment law specialist, but the majority of cases in which I serve as an expert are disputes between employers and employees, particularly departed employees who, through guile or error, left with proprietary data. Through computer forensics, I've exposed many a data thief, but I'm equally proud of the cases where my work proved that proprietary data either didn't run off or wasn't exploited.

In *The Common Law*, Justice Oliver Wendell Holmes, Jr. wrote, "even a dog distinguishes between being stumbled over and being kicked." That distinction so often lies at the heart of what we seek to divine from the electronic evidence. Was the act complained of mean or careless? Did the employee spirit their e-mail to a thumb drive to abscond with client lists or pricing, or did he or she mean only to retrieve personal messages? Are there genuine issues of, e.g., harassment, sexism, racism, ageism or religious discrimination? If so, are they confined to an aberrant corporate cul-de-sac or entrenched in the company culture?

When a data theft case involves a veteran employee or senior executive, the central question isn't *whether* they hold proprietary data--if they worked from home or on the road, company information likely resides on their home machines, thumb drives and personal e-mail accounts.

Instead, the issues are *what* do they have, *where* is it and *where* did it come from, *when* did they acquire it, *who* are they sharing it with and *how* are they using it.

Whether damages should flow, whether injunctive relief should issue and what steps are needed to rectify abuse all hinge on whether the dog was kicked or stumbled over. On the face of things, one looks like the other. Often, it's the electronic evidence--perhaps only the digital trail visible to a skilled analyst--that allows us to tell the difference.

America's halcyon days of hammer and harness are behind us. We are knowledge workers now; yet, even those who drive trucks or empty bedpans are tasked by pixels and tracked by bytes. The evidence of what we do and say, of when and where and how we go, of what we own and earn and spend is digital. *More than 99% of it will never exist as anything but electronically stored information*, and most takes forms that require special tools or expertise to see and interpret. This irritates and intimidates old school lawyers. At great cost to unwitting clients, the old school cling to what they know and disregard the rest. They print documents or convert them to paper-like formats like TIFF. They unleash armies of reviewers against hordes of irrelevant documents. They thunder that e-discovery is "out-of-control," extolling the merits of raw meat rather than learning to make fire.

Sure it's scary to face the fact that something you can't see and don't understand can hurt you; but, we've been down that road before. Take germs. It's barely 150 years since John Snow challenged the miasma theory of disease and stemmed a cholera epidemic by proving a contaminated well was the common source of infection. He posited that invisible beasties in the water were responsible. What a nutcase! Then, Dr. Oliver Wendell Holmes, Sr., the famous jurist's father, made himself a pariah in the medical community by advocating antiseptic techniques. "More invisible bogeybugs," the eminent doctors scoffed, "and now he wants us to wash our hands, gowns and instruments!" They also thought it too costly, complicated and "out of control."

Just as we demand clean hands from surgeons, clients and courts expect "clean hands" from lawyers in the sense of meeting a task ethically and in good faith. A lawyer without the skills needed to properly preserve, collect, analyze and present electronic evidence is all-but-incompetent to manage litigation today, and visiting the cost to compensate for those shortcomings upon the client is its own ethical minefield.

That's why you must make it your mission to master electronic discovery.

Now, let me tell you why you'll be *glad* you did.

Occasionally you'll win a case on charm, a good or bad judge, an appealing client, a hateful opponent or just dumb luck. But, without any of these things, you'll win most of the time *if you have the evidence proving your case*. Much of that evidence is digital. It's there. It's waiting for you--eager to tell its compelling story, ready to show your client was right and the other side should pay big or go hence without day. The employment lawyer who can get to the digital evidence--find it, understand it and use it--enjoys an enormous competitive advantage. It's bad to want to win at any cost. It's worse to accept defeat without a whimper.

The selected articles and columns that follow were chosen with the interests of employment lawyers in mind, but they are but a sampling of the articles I've written about electronic discovery and computer forensics and make available at **[www.craigball.com](http://www.craigball.com)**. I hope you find them, along with my blog posts, webcasts and other resources, to be an valuable, accessible introduction to the technology and best practices of electronic discovery.

Craig Ball, July 2009

# Right from the Start Smart First Steps in Electronic Discovery

Craig Ball

© 2008

Certainly it's smart to prepare for e-discovery—to be “proactive” about electronically stored information (ESI) and implement early case assessment systems and strategies. But sometimes, the lawsuit's the first sign of trouble, and you have to choose which fires to fight...and fast.

Don't be paralyzed by fear of failure or confusion about where to begin. There are no perfect e-discovery efforts. Before the ESI experts come aboard, there are things you can and must do. Here's a quick compendium of eight ways to hit the ground running:

1. **Apply the five Ws of good journalism—who, what, when, where and why**—to get a handle on your core preservation duties. Immediately make a list of the people, events, time intervals, business units, records and communications central to the case.
  - a. List the apparent key players (don't forget assistants who, *e.g.*, handle the boss' email and significant third parties over whom your client has a right of direction or control).
  - b. Hone in on what happened—both from your perspective and theirs—and posit what ESI sheds light either way or tends to explain or challenge the key players' actions and attitudes.
  - c. Decide what dates and time periods are relevant for preservation. Is there a continuing preservation obligation going forward?
  - d. Determine which business units, facilities, systems and devices most likely hold relevant ESI.

Your lists will change over time, but a focused, thoughtful and well-documented effort, diligently implemented, is more defensible, less costly and invariably more effective than a scattershot approach. Don't delay. It needn't be flawless right now; reasonable will do.

2. **Focus on the fragile first.** What potentially relevant ESI has the shortest shelf life and requires quickest action to preserve while it's still reasonably accessible? Voice mail, web mail and text messaging, computers requiring forensic examination, web content and surveillance video are examples of ESI that tend to be rapidly discarded or overwritten. Grabbing e-mail of key custodians *before* it migrates to backup media can save a bundle and accelerate search and processing.

3. **Protect employees from themselves.** People who wouldn't dream of shredding a paper record will purge ESI with nary a thought. In the blink of an eye, history will be reinvented as employees delete overly candid e-mail and commingled personal and business communications. The results are often catastrophic and always costly. Assess whether those entrusted with preservation can be trusted to perform, and don't rely on custodial preservation *alone* when its failure is reasonably foreseeable.
4. **Holds should be instructional, not merely aspirational.** Too many lawyers draft legal hold instructions designed to protect lawyers. Broadly disseminating a form hold directive saying "keep everything" isn't helpful and will come back to haunt you at deposition. "I *got* that memo," they say, "but I didn't *do* anything."

Custodians need to know where to start. Tell them what to do and how to do it. Give examples that inform and deadlines that demand action. Get management buy in for the time needed to comply. Better a handful of key players take the hold directive seriously than dozens or hundreds of minor players wink at it.

5. **Boots on the ground.** Good doctors don't diagnose over the phone. Likewise, good lawyers meet key players and get a firsthand sense of how they operate. Seek out the people who manage the systems that hold the evidence, and learn the "who, what, when, where and why" of your client's ESI face-to-face. It's not just enormously helpful—it's what courts demand.
6. **Build the data map, including local collections and databases.** Federal practice requires identification of potentially relevant ESI, but it's a best practice everywhere. That goes for the less-accessible stuff, too. Courts won't accept, "We don't know what we have or where it is," so be ready to identify potentially relevant ESI that you will and won't explore or produce. Data stored off the servers or on databases pose special challenges made harder by turning a blind eye to its existence. Don't fall prey to, "If we don't tell them we have it, they won't ask for it."
7. **Consider how you'll collect, store, search, review and produce ESI.** All ESI is just a bunch of ones and zeros. Making sense of it, controlling costs and minimizing frustrating "do-overs," rides on how you choose to process and produce information. So add an "H"—*How*—to those five Ws, and ponder your options for how the data gets from here to there.
8. **Engage the other side.** Even warring nations cease fire to carry off fallen comrades. You don't have to like or trust the opposition, but you have to be straight with them if you want to stay out of trouble in e-discovery. Tell the other side what you're doing and

what you're unwilling to do. Collaborate anywhere you can. Lawyers over-discover cases more from ignorance and mistrust than guile or greed; but, even when you face someone gaming the system, your documented candor and good faith effort to cooperate will serve you well in court.







## Surefire Steps to Splendid Search

Craig Ball

© 2009

Hear that rumble? It's the bench's mounting frustration with the senseless, slipshod way lawyers approach keyword search.

It started with Federal Magistrate Judge John Facciola's observation that keyword search entails a complicated interplay of sciences beyond a lawyer's ken. He said lawyers selecting search terms without expert guidance were truly going "where angels fear to tread."

Federal Magistrate Judge Paul Grimm called for "careful advance planning by persons qualified to design effective search methodology" and testing search methods for quality assurance. He added that, "the party selecting the methodology must be prepared to explain the rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented."

Most recently, Federal Magistrate Judge Andrew Peck issued a "wake up call to the Bar," excoriating counsel for proposing *thousands* of artless search terms.

Electronic discovery requires cooperation between opposing counsel and transparency in all aspects of preservation and production of ESI. Moreover, where counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI's custodians as to the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of 'false positives.' It is time that the Bar—even those lawyers who did not come of age in the computer era—understand this.

### No Help

Despite the insights of Facciola, Grimm and Peck, lawyers still don't know what to do when it comes to effective, defensible keyword search. Attorneys aren't *trained* to craft keyword searches of ESI or implement quality control testing for same. And their experience using Westlaw, Lexis or Google serves only to inspire false confidence in search prowess.

Even saying "hire an expert" is scant guidance. Who's an expert in ESI search for your case? A linguistics professor or litigation support vendor? Perhaps the misbegotten offspring of William Safire and Sergey Brin?

The most admired figure in e-discovery search today—*the Sultan of Search*—is Jason R. Baron at the National Archives and Records Administration, and Jason would be the first to admit he has no training in search. The persons most qualified to design effective search in e-discovery

earned their stripes by spending thousands of hours running searches in real cases--making mistakes, starting over and tweaking the results to balance efficiency and accuracy.

### **The Step-by-Step of Smart Search**

So, until the courts connect the dots or better guidance emerges, here's my step-by-step guide to craftsmanlike keyword search. I promise these ten steps will help you fashion more effective, efficient and defensible queries.

- 1. Start with the Request for Production**
- 2. Seek Input from Key Players**
- 3. Look at what You've Got and the Tools you'll Use**
- 4. Communicate and Collaborate**
- 5. Incorporate Misspellings, Variants and Synonyms**
- 6. Filter and Deduplicate First**
- 7. Test, Test, Test!**
- 8. Review the hits**
- 9. Tweak the Queries and Retest**
- 10. Check the Discards**

#### **1. Start with the Request for Production**

Your pursuit of ESI should begin at the first anticipation of litigation in support of the obligation to identify and preserve potentially relevant data. Starting on receipt of a request for production (RFP) is starting late. Still, it's against the backdrop of the 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 demands in everyday English to get closer to the terms most likely to appear in the ESI. Add 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 varying 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 some 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."

## **2. Seek Input from Key Players**

Judge Peck was particularly exercised by the parties' failure to elicit search assistance from the custodians of the data being searched. 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. At the latest TREC Legal Track challenge, higher precision and recall seemed to closely correlate with the amount of time devoted to questioning persons who understood the documents and why they were relevant. The need to do so seems obvious, but lawyers routinely dive into search before dipping a toe into the pool of subject matter experts.

## **3. Look at what You've Got and the Tools You'll Use**

Analyze the pertinent documentary and e-mail evidence you have. Unique phrases will turn up threads. Look for words and short phrases that tend to distinguish the communication as being about the topic at issue. What content, context, sender or recipients would prompt you to file the message or attachment in a responsive folder had it occurred in a paper document?

Knowing what you've got also means understanding the forms of ESI you must search. Textual content stored in TIFF images or facsimiles demands a different search technique than that used for e-mail container files or word processed documents.

You can't implement a sound search if you don't know the capabilities and limitations of your search tool. Don't rely on what a vendor tells you their tool can do, test it against actual data and evidence. Does it find the responsive data you already know to be there? If not, why not? Any search tool must be able to handle the most common productivity formats, e.g., .doc, docx, .ppt, .pptx, .xls, .xlsx, and .pdf, thoroughly process the contents of common container files, e.g., .pst, .ost, .zip, and recurse through nested content and e-mail attachments.

As importantly, search tools need to clearly identify any "exceptional" files unable to be searched, such as non-standard file types or encrypted ESI. If you've done a good job collecting and preserving ESI, you should have a sense of the file types comprising the ESI under scrutiny. Be sure that you or your service providers analyze the complement of file types

and flags any that can't be searched. Unless you make it clear that certain file types won't be searched, the natural assumption will be that you thoroughly searched all types of ESI.

#### **4. Communicate and Collaborate**

Engaging in genuine, good faith collaboration is the most important step you can take to insure successful, defensible search. Cooperation with the other side is not a sign of weakness, and courts expect to see 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 make Draconian demands.

#### **5. 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 high 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. Until a tool emerges that lists common variants and predicts the likelihood of false hits, try a site like <http://www.dumbtionalary.com> that checks keywords against over 10,000 common misspellings and consult Wikipedia's list of more than 4,000 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 the 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

## **6. 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 have to know *how* your search tool handles encoded, embedded, multibyte and compressed content.

Filtering irrelevant file types can be accomplished various ways, including culling by binary signatures, file extensions, paths, dates or sizes and by de-NISTing for known hash values. 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*.

## **7. Test, Test, Test!**

The single 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 likely save a lot of wasted time and money.

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 a large number of hits in wholly irrelevant collections is compelling justification for limiting or eliminating unsuitable keywords.

Similarly, 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.

## **8. Review the Hits**

My practice when testing keywords is to generate spreadsheet-style views letting me preview search hits in context, that is, flanked by 20 to 30 words on each side of the hit. It's efficient and illuminating to scan a column of hits, pinpoint searches gone awry and select particular documents for further scrutiny. Not all search tools support this ability, so check with your service provider to see what options they offer.

Armed with the results of your test runs, determine whether the keywords employed are hitting on a reasonably high incidence of potentially responsive documents. If not, what usages are throwing the search off? What file types are appearing on exceptions lists as unsearchable due to, e.g., obscure encoding, password protection or encryption?

As responsive documents are identified, review them for additional keywords, acronyms and misspellings. Are terms that should be finding known responsive documents failing to achieve hits? Are there any consistent features in the documents with noise hits that would allow them to be excluded by modifying the query?

Effective search is an *iterative* process, and success depends on new insight from each pass. So expect to spend considerable time assessing the results of your sample search. It's time wisely invested.



## 9. Tweak the Queries and Retest

As you review the sample searches, look for ways you can tweak the queries to achieve better precision without adversely affecting recall. Do keyword pairs tend to cluster in responsive documents such that using a Boolean *and* connector will reduce noise hits? Can you approximate the precise context you seek by controlling for proximity between terms?

If very short (e.g., three letter) acronyms or words are generating too many noise hits, you may improve performance by controlling for case (e.g., all caps) or searching for discrete occurrences (i.e., the term is flanked only by spaces or punctuation).

## 10. Check the Discards

Keyword search must be judged both by what it *finds* and what it *misses*. That's the "quality assurance" courts demand. A defensible search protocol includes limited examination of the items not generating hits to assess whether relevant documents are being passed over.

Examination of the discards will be more exacting for your representative sample searches as you seek to refine and gain confidence in your queries. Thereafter, random sampling should suffice.

No court has proposed a benchmark or rule-of-thumb for random sampling, but there's more science to sampling than simply checking every hundredth document. If your budget doesn't allow for expert statistical advice, and you can't reach a consensus with the other side, be prepared to articulate why your sampling method was chosen and why it strikes a fair balance between quality assurance and economy. The sampling method you employ needn't be foolproof, but it must be rational.

Remember that the purpose of sampling the discards is to promptly *identify and resolve* ineffective searches. If quality assurance examinations reveal that responsive documents are turning up in the discards, those failures must receive prompt attention.

## 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.

Effective searches are tailored to the data under scrutiny. For example, it's silly to run a custodian's name or e-mail address against his or her own e-mail, but sensible for other collections. It's often smart to *tier* your ESI and employ keywords suited to each tier or, when feasible, to limit searches to just those file types or segments of documents (i.e., message body

and subject) likely to be responsive. This requires understanding what you're searching and how it's structured.

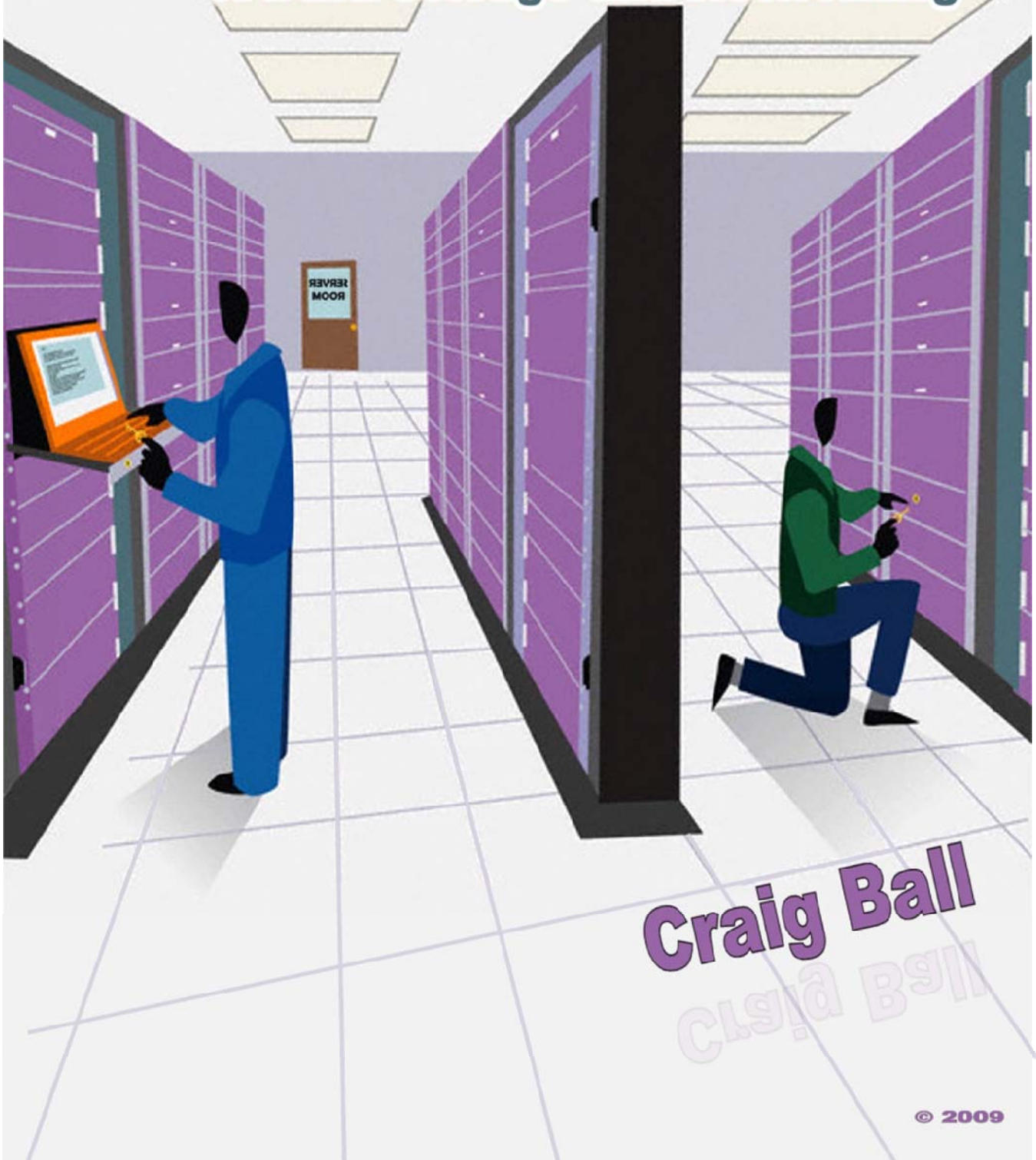
When searching e-mail for recipients, it's almost always better to search by e-mail address than by name. In a company with dozens of Bob Browns, each must have a unique e-mail address. Be sure to check whether users employ e-mail aliasing (assigning idiosyncratic "nicknames" to addressees) or distribution lists, as these can thwart search by e-mail address or name.

**Search is a Science...**

...but one lawyers *can* master. I guarantee these steps will wring more quality and trim the fat from text retrieval. *It's worth the trouble*, because the lowest cost e-discovery effort is the one done right from the start.

# *Geek Speak*

**A Lawyer's Guide to the Language  
of Data Storage and Networking**



**Craig Ball**

Craig Ball

© 2009

# Geek Speak

## A Lawyer's Guide to the Language of Data Storage and Networking

Craig Ball

© 2009

In 1624, when John Donne mused, “No man is an island,” he could scarcely have imagined how connected we’ve become. The bell not only tolls for thee, it beeps and vibrates, too. No iPhone is an iLand.

Networks are the ties that bind our global village and make the world flat. Without networks, our laptops, iPods and Blackberries are just pricey pocket calculators. Networks also transit and store much of the electronic evidence sought in electronic discovery. This article looks at network architecture and data storage devices in the form of an occasionally irreverent glossary offered to help lawyers be at ease discussing the technology of electronic discovery.<sup>1</sup>

Dealing with electronically stored information (ESI) is like living with a teenager—always running in, changing its clothes and heading out again, tracking metadata all over the carpet! But litigants and lawyers aren’t relieved of the duty to find and collect potentially relevant ESI just because it’s flitting about and messy. They’re still obliged to track down the data and make sure it’s safe from harm and will stay put (or come home) until needed in discovery. Rooting out responsive data begins with knowing where to look and the right questions to ask, so it helps to have a working knowledge of the terminology of data storage and networking.

### Storage and Network and Memory, Oh My!

Though the terms “storage” and “network” are surely familiar, the technologies they describe take many forms, prompting some confusion. Many mistakenly refer to data *storage* devices like hard drives as “memory.” Hard drives are *storage*; that is, any non-volatile and semi-permanent electronic, optical, mechanical or magnetic device into which data can be entered and subsequently retrieved on demand. Storage is also a location on a network that enables access to storage devices. *Memory* is a term that should be reserved to devices, particularly *Random Access Memory* or *RAM*, where data resides temporarily during processing but is typically lost or overwritten when an application closes or power is interrupted.<sup>2</sup>

---

<sup>1</sup> For a more comprehensive (and sober) glossary of e-discovery terms, download The Sedona Conference Glossary for E-Discovery and Digital Information Management (2nd Ed.) from [http://www.thesedonaconference.org/dltForm?did=TSCGlossary\\_12\\_07.pdf](http://www.thesedonaconference.org/dltForm?did=TSCGlossary_12_07.pdf)

<sup>2</sup> The line between storage and memory is getting harder to find. Non-volatile flash *memory* is widely used as a means of data *storage* in cameras, thumb drives and solid state drives. Flash memory has almost entirely supplanted photographic film, and solid state drives will soon replace hard drives in laptops and MP3 player. Moreover, it’s unclear how *long* information must be “stored” to be called electronically *stored* information. One case has lawyers worried that the interval may be measured in mere nanoseconds. *Columbia Pictures, Inc. v. Bunnell*, 245 F.R.D. 443 (C.D. Cal. 2007) (defendants ordered to produce contents of RAM).

A “network” can be any number of computers or devices connected for the purpose of sharing information or capabilities. The largest and most widely used network is, of course, the Internet; but, businesses and homes deploy Wide or Local Area Networks (WANs or LANs) to share databases, mail systems, applications, printers and Internet service. There can be a lot of overlap. WANs may be composed of multiple LANs and connect to the Internet.

## B

### Backup

Although sharing information and resources is the *raison d'être* for networking generally, an imperative for business networking is the ability to backup many user's data from a single location. Without networking and the mapping of users' storage areas to networked storage devices, users must periodically backup their own data—a responsibility consuming many hours and fostering tragic outcomes.

With networking, each user can be allotted space on a common storage server and the network configured to route that user's activities to the assigned storage location when the user logs on. The user's machine may be configured to assign a specified drive letter (e.g., M:) or folder name to the user's networked storage location. Because the network storage *device* is shared among many users, its allotments are called **network shares**. But these user-assigned storage areas are typically not “shared” with (i.e., *accessible* to) multiple users. Still other allocations may be open to all or just particular users granted access privileges.

With many users' critical data consolidated in a single locale, albeit in discrete “shares,” it falls to the **information technology (IT)** staff to insure that all that data gets thoroughly and reliably duplicated at regular intervals to protect against its loss as a consequence of system failure or other disaster. Ideally, the duplicate data is physically or electronically transported to a distant secure location unlikely to be affected by the disaster and is then used to get the downed machines back up again; hence the duplicates are called **backups** and their use is termed **disaster recovery**.

Because it's cheap, durable and portable, magnetic tape is the most common medium used for backup, although remote duplication (**mirroring**) to other network storage devices is fast becoming a viable alternative as hard drive costs plummet. To save time and space, backup regimens seldom copy commercial software programs that can be reinstalled from other media. More time and space is saved—along with network bandwidth--by only occasionally making **full backups** of all user created data, opting instead to create more frequent **differential backups** of files created or changed since the last full backup and **incremental backups** of just what's been created or changed since the last incremental backup. When disaster strikes, the full,

differential and/or incremental sets are pieced together like Humpty-Dumpty, a process called **tape restoration**.

Businesses only need disaster recovery data for a brief interval because no business wants to restore its systems with stale data. Accordingly, the only backup tapes essential for recovery are the last complete, uncorrupted set before the river rose. As a cost savings practice, older tapes may be reused by overwriting them with the latest data, a practice called **tape rotation**.

In practice, companies may keep backup tapes well beyond their utility for disaster recovery--often years longer and occasionally past the companies' ability to access tapes created with obsolete software or hardware. These **legacy tapes** are business records—sometimes the last surviving copy—but afforded little in the way of records management. Even businesses that overwrite tapes every two weeks replace their tape sets from time to time as faster, bigger options hit the market. Consequently, old tapes get set aside and forgotten in offsite storage or a box in the corner until their existence is uncovered in discovery.

Backup tapes store data in significantly different ways than the computer systems they protect. Further, large complex enterprises demand large, complex backup systems protecting hundreds of servers. Such backup systems may occupy room-sized silos where robotic arms ceaselessly cycle through thousands of tapes, and databases are required just to track their convoluted contents. This is an arena where broad brush e-discovery efforts go horribly awry and where transparency, close analysis and well-honed choices are vital. Cooperation between opposing sides is essential, and Judges should tread carefully before issuing orders with untoward costs and consequences.<sup>3</sup>

## C

### Cache

Downloading data over a network is slower than accessing data on a local hard drive, so networked computers sometimes store or “cache” data obtained from the network to avoid the need to download the same data when later needed. Used as a noun, a cache is an area where oft-used information is stored to facilitate its faster access. Devices like hard drives and processors use caching to improve performance, as do certain software programs. For example, Windows computers running the Internet Explorer web browser use a file cache on the local hard drive called Temporary Internet Files which (with some exceptions) holds the HTML code and images of each web page viewed on the machine until the cache is full or emptied by the user. Users revisiting a cached website experience faster page loads because the browser can pull identical data from the cache instead of downloading it from the Web.

---

<sup>3</sup> Judges and counsel may find value in Ball, *What Judges Should Know about Discovery from Backup Tapes* (2008); Available at [http://www.craigball.com/What\\_Judges\\_Backup\\_Tapes-200806.pdf](http://www.craigball.com/What_Judges_Backup_Tapes-200806.pdf)



Though this requires the system to compare the network and cached data to determine if the network data has changed, caching is still faster than needlessly downloading the data a second time.

From the standpoint of electronic discovery, information in the Temporary Internet Files cache may be relevant, especially where Internet usage is at issue or where data (like web mail) may not be available from more accessible locations.

## Client

A client, as in **client-server model**, is a program, computer or other device that connects via a network to another computer or device called the **server**. Internet browsers are client applications that obtain web pages from web servers. Microsoft Outlook is an e-mail client that connects to e-mail servers like Microsoft's Exchange server. When the client is a personal computer and performs much of the processing of the data, it's ungraciously called a **fat client**. When the client device or application cedes most processing to the server, it's called a **thin client** (or even a **dumb terminal** when it has no processing or local storage capabilities at all).

## Cloud Computing

Cloud Computing refers to reliance on web-based tools and resources to supplant local applications and storage. It encompasses **Software as a Service (SaaS)**, where users "lease" programs via the Internet (Google Apps is a prime example), as well as the much-touted, yet elusive **Web 2.0**--a catchall for all manner of web-enabled phenomena: **social networking, blogs, wikis, Twitter, YouTube, Google mashups** and arguably any web-centric venture that survived the great dot-com meltdown.

Gen Xers and Millennials embrace "cloud computing" as if they invented it, but Boomers knew cloud computing when it was called client-server or thin client. Then as now, it was screens and keyboards talking to Big Iron elsewhere, the latter doing the heavy lifting. With SaaS and Web 2.0, we've come full circle and are richer for the journey. As cloud computing takes hold, the bits and bytes of our lives will again move out and get their own places, this time in the ether, but we'll have their cell numbers and can call when we need them.

Cloud computing creates new opportunities in e-discovery because the candid, probative revelations once the exclusive province of e-mail now flood **MySpace** and **Facebook**. But cloud computing creates new challenges for e-discovery because it's harder for employers to isolate and search custodial collections without physical dominion of the storage devices and their users' log in credentials. Additionally, repatriation of cloud content depends on the compatibility of cloud formats with local storage formats, including the ability to preserve and produce relevant metadata. Consider **Gmail**. Though it's feasible to download Gmail messages into a local mail client application like Microsoft Outlook using Gmail's POP3 support

feature, the functionality, searchability and some associated metadata will vary between cloud and local counterparts.

## Collection

As a noun in e-discovery, collection refers to any discrete set of electronically stored information, particularly the set amassed after targeted retrieval and culling efforts have occurred. However, it's not uncommon to hear parties speak of their entire universe of ESI as the "collection." For this reason, it's important to define the parameters of any ESI collection to insure common expectations.

## Container Files

Sometimes called **compound files**, container files hold other files, often in compressed, encrypted or proprietary formats or nested—container-within-container--like Russian matryoshka dolls. Container files commonly encountered in e-discovery include compressed Zip and RAR archives, Outlook PST and OST mail files and Lotus Notes NSF mail files. Container files can severely distort document volume estimations as a function of data volume, e.g., a one gigabyte mail container can easily hold tens of thousands of messages and attachments.

## Custodian

A custodian is a caretaker, and in the context of e-discovery, the term refers to a person who holds or is charged with overseeing and maintaining potentially relevant information, whether stored electronically, on paper or by other means. For litigation purposes, one is the custodian of his own e-mail, locally and server-stored documents, voice and electronic messaging, smart phone data and any other information to which he has a right of ownership, access or control, including information in the hands of third parties over whom he may exercise direction or control. Custodian also refers to the persons to whom legal hold notices are directed.

Identifying custodians becomes particularly important when ESI resides in shared network repositories and no one person bears the duty to preserve, search or produce the data. When *everyone* is responsible, often *no one* steps up. Accordingly, efforts to identify potentially responsive ESI should always inquire into the existence of, or rights of access to, shared repositories.

# D

## Database

A database is a structured collection of records or information organized according to a framework called a **data model** or **schema** that typically facilitates search and recall of the records using **query language**. Massive, costly and enormously complex, databases play vital roles in most large enterprises. For companies like Google, Amazon.com and e-Bay,

databases serve as the nexus of virtually all operations. Yet, databases come in all sizes and forms, for tasks as varied as balancing checkbooks, organizing family photos and tracking stock portfolios. Even many common file formats are structured as databases, including Microsoft Outlook mails containers and Adobe Acrobat PDF files.

Databases are the most important resources shared across networks, and they also serve as repositories for much information of importance in e-discovery. Many transactions and documents that would once have been memorialized on paper now exist solely as disparate records stored within databases. Because databases assemble documents on-the-fly and are constantly being updated and purged, they can be particularly challenging sources from which to preserve, isolate and produce responsive data. E-discovery from databases requires detailed assessment of the contents, users, capabilities, applications and schema. Responsive contents may need to be extracted using queries constructed expressly for the purpose of isolating evidence and protecting privileged or confidential content, and the form of production is a key consideration, as many requesting parties lack the hardware and software to assimilate database contents in its native format.

### **Distributed Data**

Distributed data might also be called “willy-nilly data,” in that it describes all the potentially responsive ESI that’s not on the server, but is strewn across laptops, handheld devices, external hard drives, flash drives, CDs, DVDs, home machines, online storage and webmail. Distributed data is costly to collect and sometimes difficult to process because it tends to be the most idiosyncratic ESI and that most prone to obstructive intervention by custodians. A common mistake in e-discovery is assuming that the responsive ESI is on the server without taking reasonable steps to preserve and assess (even by sampling) the contents of distributed data sources.

### **Domain**

A domain is a group of networked computers (typically in the same physical facility) that share common peripherals, directories and storage areas. E-mail systems are customarily organized and backed up by domain.

### **Domino Server**

A Domino server is a network-accessible computer holding users’ centralized e-mail stores and employing the IBM Lotus Notes e-mail application. If an IT person mentions the company’s Domino server (and you aren’t discussing pizza delivery), be prepared for Lotus Notes e-mail and the unique e-discovery challenges and opportunities it entails.

# E

## ECM

Enterprise Content Management is an umbrella term describing a range of technologies designed to help companies identify, access and use the information stored in their documents, photographs, video, web content, databases and e-mail, especially siloed repositories and unstructured content that tends to be unavailable or difficult to access companywide. ECM applications tend to encompass document management and version control, integration of paper records, records management and retention, web content management and collaboration tools. The most familiar implementation of ECM is probably Microsoft's SharePoint Services (MOSS and WSS).

From an e-discovery perspective, the consequences of a substantial ECM implementation are manifold. ECM may operate at cross-purposes with—or at least complicate—legal hold obligations. Further, collaborative environments are heavily dependent on metadata to support functionality, making preservation and production of a broad range of metadata essential to meet the obligation to produce ESI in reasonably usable forms. Within some ECM environments, documents exist in untraditional and proprietary formats necessitating new and creative approaches to selecting forms of production that preserve look, feel and function of multimedia and informational content. On the positive side, a successful ECM system should facilitate cost-effective identification and search of responsive ESI (though cynics might suggest that savings will be offset by having to deal with all the potentially responsive ESI that ECM makes impossible to ignore).

## Enterprise

Enterprise is variously the flagship Federation starship commanded by Captain James T. Kirk, a low cost rental car company favored by skinflint insurance carriers or, in e-discovery, the term of choice when “company” or “business” are insufficiently pretentious.

## Ethernet

A set of network cabling and communication protocols for bus topology<sup>4</sup> local area networks. That is, an agreed-upon set of instructions, akin to a language, that permits devices to exchange information. If that's not helpful, think of it as the *other* way computers talk to each other when they're not speaking Internet (TCP/IP).

## Exchange Server

An Exchange server is a network accessible computer holding users' centralized e-mail stores and running the Microsoft Exchange e-mail and calendaring application. Typically, users access Exchange servers with Microsoft Outlook mail clients. Microsoft Exchange accounts for

---

<sup>4</sup> See “Topology,” *infra*, for further discussion of network topologies.

65% of market share among all organizations, with significantly larger shares among businesses with fewer than 49 employees and those in the health care and telecommunications sectors. Consequently, Exchange Server e-mail crops up in the overwhelming majority of cases and understanding its architecture is an essential e-discovery skill.<sup>5</sup> **See also** the discussion of Microsoft Outlook, *infra*.

### Extensible Markup Language (XML)

Extensible Markup Language or XML provides a basic syntax that can be used to share information between different kinds of computers, applications and organizations without first converting it. XML employs coded identifiers paired with text and other information. These identifiers can define the appearance of content (much like the Reveal Codes screen of WordPerfect documents) or serve to tag content to distinguish whether 09011957 is a birth date (09/01/1957), a phone number (0-901-1957) or a Bates number. Plus, markup languages allow machines to talk to each other in ways humans understand.

Like multilingual speakers agreeing to converse in a common language, as long as two systems employ the same XML tags and structure (typically shared as an XML Schema Definition or .XSD file), they can quickly and intelligibly share information. Parties and vendors exchanging data can fashion a common schema tailored to their data or employ a published schema suited to the task, such as that under development by the Electronic Discovery Reference Model.<sup>6</sup>

### Extranet

An extranet is a private network made available via the Internet to a select group of users, typically customers or suppliers. When used to support transactions, extranets are often called **virtual deal rooms**. Extranets are increasingly used as a collaborative tool in e-discovery and as a host repository for ESI. Access may be secured by use of a VPN connection or by a conventional link employing user ID and password alone.

## F

### File Server

File servers, the heart of any client-server network, are computers typically equipped with fast, redundant storage devices that store and deliver each user's files and other data. Very small networks may not use dedicated file servers but instead allow workstations to share data amongst themselves in a peer-to-peer configuration.

### FTP

File Transfer Protocol or FTP is a set of standards and instructions that permit transfer of files

---

<sup>5</sup> For a more detailed discussion of Exchange Servers and e-discovery, *see* Ball, *Meeting the Challenge of E-Mail in Civil Discovery* (2009) at p.25 et seq., *infra* and available at <http://www.craigball.com/em2008.pdf>

<sup>6</sup> <http://edrm.net>

between networked computers, most often via the Internet. You'll encounter FTP in e-discovery both as a potential repository to be explored for "orphaned" responsive data not available from other accessible sources and as a mechanism to transfer large volumes of data to and from clients and e-discover service providers.

## G

### Gateway

A gateway is a combination of hardware and software that allows two networks to communicate. A gateway is essentially a protocol translator that enables, e.g., the wireless network in your home to communicate with the Internet. In this role, the gateway is also called a *router*.

## H

### Hub

A hub allows multiple computers to share a network connection, not unlike a power strip allows multiple electrical devices to share AC power from an outlet. Hubs support simple peer-to-peer networking between computers.

## I

### IM

Instant Messaging or IM is a form of real-time textual communication between two or more persons where such messages are carried by the Internet or a cell phone network. It is the instantaneous receipt and response of IM and its evanescence that distinguishes IM from e-mail. Though relevant, non-privileged IM messages are as subject to preservation and production duties as any other evidence, IM messages typically reside only on the local device sending or receiving the message, not on network servers, and not in active data unless the user has enabled message logging. Accordingly, litigants obliged to preserve IM traffic must either compel message logging and periodic collection of the logs or implement a packet capture mechanism to scan for IM traffic and snare and copy messages as they enter and leave the company's Internet gateway. Neither method is wholly satisfactory.

When a company obliged to preserve IM traffic fails to do so, the data loss may be mitigated by collection from other parties to the dialog or by forensic examination of the machines or devices employed, although recovery of message traffic is by no means assured.

### Internet

You're not *really* going to make me define Internet, are you? Where have you been the last 15 years?! Okay, if you insist.



Turning to none other than the august personage of former (convicted but charges dropped) Alaska Senator Ted Stevens in a speech delivered on June 28, 2006 as chairman of the Senate Committee on Commerce, Science and Transportation:

[T]he Internet is not something that you just dump something on. It's not a big truck. It's a series of tubes. And if you don't understand, those tubes can be filled and if they are filled, when you put your message in, it gets in line and it's going to be delayed by anyone that puts into that tube enormous amounts of material, enormous amounts of material.

So, the Internet is a series of tubes, not a big truck, and it's best to keep a plumber's helper at hand while Web surfing.

### **Intranet**

An intranet is a private web site, typically reserved to the exclusive use of an organization's employees or members. Intranets tend to be hosted internally on a local access network, but may be Internet-enabled so as to permit secure connections by authorized users via the Internet.

### **IP Address**

An Internet Protocol or IP address is a unique series of four numbers joined by periods and sometimes called a Dotted Quad. It is the numerical designation of the host system that connects you to the Internet and is cross-referenced to the domain name such that either the name or the number can be employed to correctly designate your host system. An IP address can also serve as a unique identifier for computers and other Web-enabled devices on a network employing the standard TCP/IP protocol that serves as the basic computer-to-computer language of the Internet. For example, the IP address of the computer used to write this article is 192.168.0.189.

IP addresses can be useful in e-discovery when constructing a company's data map. Using IP addresses, machines claimed to exist can be correlated against those actually connected to a network. An IP address can also tie ESI to a particular device and, thus, a particular user.

### **ISP**

An Internet Service Provider or ISP is a business or other entity that supplies Internet access via dial-up, cable modem, DSL or ISDN lines or dedicated high speed connections. ISPs routinely host their customers' e-mail accounts and thus may be a source of ESI by subpoena or constitute a third party custodian who should be put on notice of legal hold obligations.

## J

### Journaling

Journaling is a means of archiving electronic messages, principally e-mail, but potentially IM and VM, too. A journaling mail server copies all messages or, per established rules, certain incoming and outgoing messages to a mailbox or storage location serving as the journaling repository. Journaling serves to preempt ultimate reliance on individual users for litigation preservation and regulatory compliance. Properly implemented, it should be entirely transparent to users and secured in a manner that eliminates the ability to alter the journaled collection.

Accordingly, journaling is a valuable safety net for companies obliged to preserve e-mail because of litigation or regulatory obligations, and counsel should inquire to determine if journaling was enabled, as journaled e-mail traffic can mitigate custodial preservation errors and misconduct. Journaling also helps protect the company against rogue employees seeking to conceal wrongdoing by destroying their e-mail stores before leaving.

Exchange Server supports three types of journaling:

- Message-only journaling, which does not account for blind carbon copy recipients, recipients from transport forwarding rules, or recipients from distribution group expansions;
- Bcc journaling, which is identical to Message-only journaling except that it captures Bcc addressee data; and
- Envelope Journaling which captures all data about the message, including information about those who received it.

Envelope journaling is the mechanism best suited to e-discovery preservation and regulatory compliance. Unlike messages preserved after delivery, journaled messages won't include metadata reflecting the addressee's handling of the message, such as foldering or indications that the message was read.

Journaling should be distinguished from e-mail archiving, which may implement only selective, rules-based retention and customarily entails removal of archived items from the server for offline or near-line storage to minimize strain on IT resources and/or implement electronic records management. However, Exchange journaling also has the ability to implement rules-based storage, so each can conceivably be implemented to play the role of the other.

## L

### LAN

A Local Area Network or LAN is an interconnected group of computers typically situated in a

single location and connected by cable or wirelessly. LANs tend to be used in offices and homes to share Internet connections, files and printers, though they may also be configured to exchange e-mail internally.

## Lotus Notes

Lotus Notes is an IBM client application supporting e-mail, calendaring, web browsing and a host of collaborative features. Notes works in conjunction with an IBM Lotus Domino server, although it can also be configured to retrieve e-mail from Microsoft Exchange servers. Though Lotus Notes reportedly has just a 10% overall market share, it enjoys a much higher percentage base among manufacturers with at least 5,000 employees, and IBM claims it has sold 140 million Notes licenses worldwide. Still, the relative infrequency with which E-discovery service providers encounter Lotus Notes means that not all providers are equipped or experienced to process Notes content.

Unlike Microsoft Exchange, which is a purpose-built application designed for messaging and calendaring, Lotus Notes is more like a toolkit for building whatever capabilities you need to deal with documents—mail documents, calendaring documents and any other type of document used in business. Notes wasn't designed for e-mail—e-mail just happened to be one of the things it was tasked to do.

Notes is database-driven and distinguished by its replication and security. Lotus Notes is all about copies. Notes content, stored in **Notes Storage facility** or **NSF** files, is constantly being replicated (synchronized) here and there across the network. This guards against data loss and enables data access when the network is unavailable, but it also means there can be many versions of Notes data stashed in various places within an enterprise. Thus, discoverable Notes mail may not be gone, but lurks within a laptop that hasn't connected to the network since the last business trip.

## M

### Mail Client

A mail client is any software application used to prepare, send, receive and read e-mail. E-mail clients can be rudimentary or, more common today, feature-laden productivity tools like Microsoft Outlook or Lotus Notes, which offer a sophisticated and highly-customizable interface. The configuration of a user's mail client may determine whether messages are stored locally, on the mail server or in both places. Additionally, the mail client records and manages key metadata detailing a user's handling of e-mail, including the user's folder structure and various flags indicating whether, *inter alia*, the user opened a particular message, tied it to a calendar entry or flagged it for action.

## Microsoft Outlook

Microsoft Outlook is an e-mail client and calendaring tool coupled with several other productivity features to comprise a personal information manager (PIM) toolset. Outlook serves as both a standalone mail client compatible with all mail protocols in common use, but in business, it's usually deployed in conjunction with **Microsoft Exchange Server** or, lately, **Microsoft Office SharePoint Server** (MOSS).

Despite the confusing similarity of their names, Outlook is a much different and substantially more sophisticated application than Outlook Express (now called Windows Mail). One of many important differences is that where Outlook Express stores messages in plain text, Outlook encrypts and compresses messages. The most significant challenge Outlook poses in discovery is the fact that all of its message data and folder structure, along with all other information managed by the program (except the user's Contact data), is stored within a single, often massive, database file with the file extension .pst. The Outlook PST file format is proprietary and its structure is poorly documented, limiting your options when trying to view or process its contents to Outlook itself or one of a handful of PST file reader programs available for purchase and download via the Internet.

While awareness of the Outlook PST file has grown, even many lawyers steeped in e-discovery fail to consider a user's Outlook .ost file. The OST or offline synchronization file is commonly encountered on laptops configured for Exchange Server environments. Designed to afford access to cached messages when the user has no active network connection., e.g., while on airplanes, local OST files often hold messages purged from the server—at least until re-synchronization. It's not unusual for an OST file to hold e-mail unavailable from any other comparably-accessible source.

By default, when a user opens an attachment to a message from within Outlook (as opposed to saving the attachment to disk and then opening it), Outlook stores a copy of the attachment in a "temporary" folder. But don't be misled by the word "temporary." In fact, the folder isn't going anywhere, and its contents—sometimes voluminous--tend to long outlast the messages that transported the attachments. Thus, litigants should be cautious about representing that Outlook e-mail is "gone" if the attachments are not.

The Outlook "viewed attachment folder" will have a varying name for every user and on every machine, but it will always begin with the letters "OLK" followed by several randomly generated numbers and uppercase letters (e.g., OLK943B, OLK7AE, OLK167, etc.).

## Mirroring

Mirroring refers to the creation of an exact copy of a dataset. Mirroring may be used locally for data integrity and protection or across a network as a form of backup, duplicating the entire

contents of a server to some distant, identical system. Disk mirroring, also called RAID 1, entails simultaneously writing identical data to two different hard drives, affording redundancy should either drive fail.

## N

### Nearline Storage

Nearline storage refers to voluminous data that, while not in such demand as to require instantaneous access via the network, must nonetheless be available from time-to-time without human intervention. Nearline data tends to be stored on high capacity media (like magnetic tape) that can be robotically loaded on demand, occasioning only a brief delay between a request and delivery of data.

### NAS

Networked Attached Storage or NAS is a dedicated file server designed expressly for data storage. Because a NAS isn't called upon to do general computing tasks, it can employ a file system built exclusively for its limited role. When inquiring about devices, be careful not to reference only computers and servers, as a too-literal interpretation might allow someone to overlook a NAS.

### Node

Anything connected to a network can be termed a "node;" however, anyone who uses the word node in this way must be termed a "nerd."

## O

### Offline Data

Offline data denotes ESI housed on media that is not connected to the network and requires human intervention, e.g., mounting or restoration, to access the contents. Backup tapes sent offsite for storage, legacy systems in the warehouse and even a CD-R in your desk drawer are examples.

The e-discovery challenge of offline data is that it must be proven not reasonably accessible to be excluded from search and production. Even then, producing parties must identify offline data with sufficient specificity to allow the requesting party to determine if the producing party is right about the data's inaccessibility. But there's the catch: how does a producing party do that without examining the contents?

To economically manage offline data, insure that its contents are indexed and the media clearly labeled *when the data goes offline* so as to obviate the costly and time-consuming need to bring it online, albeit briefly, to identify its contents. This isn't going to help with legacy data, but it's a no-brainer going forward.

# P

## Partition

A partition is a division of the storage area of a hard drive such that a single physical drive can be seen by the computer as multiple drives. If you think of an unpartitioned hard drive as a big metal cabinet, a partition is the division of that cabinet into file drawers. Though it's most common to encounter drives created with a single partition encompassing the entire storage area of the drive, in Windows, a hard drive can currently have up to four primary partitions or three **primary partitions** and one so-called **extended partition** that can be subdivided into as many as 24 extended partitions. Only one of the four partitions can be designated as an active partition, signaling the partition that holds the operating system the machine should boot on start up.

Partitioned hard drives can hold multiple operating systems such that a snippet of code called a **boot loader** can point the system to a partition other than the active partition to initiate a different operating system. Thus, a machine with a single drive can be configured to boot in Windows Vista, Linux or Windows XP via a start up menu. From the standpoint of e-discovery, a thorough search for ESI should include accounting for the full storage capacity of a hard disk, in case responsive data lurks on another partition. If you think this sounds farfetched, take a look at *Phoenix Four, Inc. v. Strategic Res. Corp.*<sup>7</sup>

## Path

The complete local or network address to a particular folder, file or device, expressed hierarchically from a root location of a server or disk volume. If I were a file, the path to me might be expressed as **Earth:\Worth AmericaUSA\Texas\Austin\78735\3723 Lost Creek Blvd\Lab\Craig Ball**. Traversing a path to a file is sometimes called "drilling down."

## Peer-to-Peer Network

In a **peer-to-peer** or **P2P** network, each connected computer serves as both client and server for the purpose of sharing resources, but most often for sharing files (notably copyrighted music and video, as well as adult content and pirated software).

## Peripheral

Just about any device you connect to a computer by cabling or networking (other than another computer or server) is called a peripheral. It most commonly refers to printers and scanners.

## Protocol

An agreed-upon set of instructions, akin to a language, that permit devices to exchange information. Networks notably employ Ethernet or TCP/IP protocols to intelligibly transmit and receive data. As language can be thought of as a "protocol" for written or oral communications,

---

<sup>7</sup> No. 05 Civ. 4837, 2006 WL 1409413 (S.D.N.Y. May 23, 2006).

a network protocol is a framework to sensibly interpret the ones and zeroes of digital communications.

## R

### RAID

A **Redundant Array of Independent (or Inexpensive) Disks** or **RAID** is a way of combining multiple hard drives to achieve greater performance, greater reliability or a mix of the two. The various types of RAID configurations are numbered. The three most commonly used configurations are RAID 0, RAID 1 and RAID 5.

A RAID 0 divides (or *stripes*, in storage parlance) data between two hard drives to combine the capacity into a single large volume and to increase the speed at which data is read and written. But because the data zigzags across two drives, a failure of either drive means the loss of all data.

A RAID 1 opts for complete redundancy, mirroring all contents between two drives such that a failure of either drive results in no loss of data--the trade off being that you can use only half of the combined capacity of the two drives and get no performance boost.

A RAID 5 uses three or more disks, garnering some of the speed boost seen in RAID 0 and the ability to fully recover all data should any one drive fail.

Because any one drive in a RAID 5 array can fail without data loss, RAID storage allows for the removal and replacement of drives from the array without the need to down the server. Thus, RAID storage—particularly RAID 5 configurations with more than 3 disks—are ubiquitous in mission critical servers. RAID 5 arrays are typically seen by the server as a single logical disk with a capacity of about two-thirds of the combined capacity of all disks in the array.

Despite its reliability, a RAID is not a substitute for a backup. A fire, flood or disgruntled employee won't destroy just one or two drives in the array, and all data will be unrecoverable absent a backup.

### Root

Root refers to top level of a file system's directory structure, typically C:\ in a Windows system. In hacking, it also refers to a level of unrestricted access to a system, where "getting root" means taking unauthorized control of the system, often using hacker tools called **root kits**.

### Router

A router (sometimes called a **switch**) is a device that directs the flow of the data packets by which information is transferred across a network. Unlike a hub, which merely relays all



packets to all connections, a router actually assigns unique addresses to connections and steers packets to and from those addresses.

## S

### SaaS

**Software as a Service** or **SaaS** is software distribution mechanism where, instead of purchasing applications and installing them, programs are accessed on the Internet or downloaded on-the-fly as needed. The advantage of SaaS is that there is no need to purchase upgrades or install patches because the software's always up-to-date. The down side is that you do not own the software and must continue to pay for its use, as well as security concerns. In e-discovery, complications derive from the loss of physical dominion of the devices storing the data, as discussed previously under Cloud Computing. A notable example of SaaS is the Google Apps package of applications, which virtualizes a user's e-mail, contacts and calendar, along with document, spreadsheet and presentation authoring tools. The provider of SaaS is called an **Application Service Provider** or **ASP**.

### SAN

A **Storage Area Network** or **SAN** is a mass storage configuration that allows *network*-attached devices to be shared among servers at very high speeds yet appear as if they are *physically* attached to each server. SANs are tied to two important trends in networking: **storage replication** (where data is remotely mirrored for disaster recovery) and **virtualization** (where physical devices are subdivided into multiple virtual devices that appear to be distinct, physical machines like servers but actually exist as emulations using software). SANs allow large aggregations of physical storages devices to be logically re-allocated to various servers and tasks. Instead of adding a 120GB hard drive to a server, a 120GB "slice" of a multi-terabyte array can be assigned to appear and function as a physically-connected 120GB drive.

### Server

A server is a device or application that delivers information to networked devices. When applied to hardware, server usually denotes a computer optimized and tasked to perform certain functions for other machines on the network. Servers tend to be isolated in locked and refrigerated server rooms, protected by backup systems and equipped with fail-safe or redundant components mounted in accessible racks, all to minimize downtime and increase security. Though a single server can perform a variety of tasks, businesses tend to dedicate servers to particular functions, such as storing user data, running applications like databases, delivering web content, managing printing, routing Internet traffic, handling e-mail stores, etc.

### Share

Also called a **Network Share**, see the discussion of shares in **Backup**, above.

### Single Instance Storage

Networks and e-mail systems are replete with multiple iterations of identical documents. When an entire department receives an e-mail with the same attachment, or when thousands of



employees keep a copy of the same memo, storage is wasted. Single instance storage performs **de-duplication** and replaces the individual copies with a *pointer* to an identical master copy. SIS aids backup by facilitating the use of fewer tapes and reducing the time required to complete the task. When dealing with a SIS volume in e-discovery, be careful to collect the de-duplicated document and not just its SIS pointer.

## T

### TCP/IP

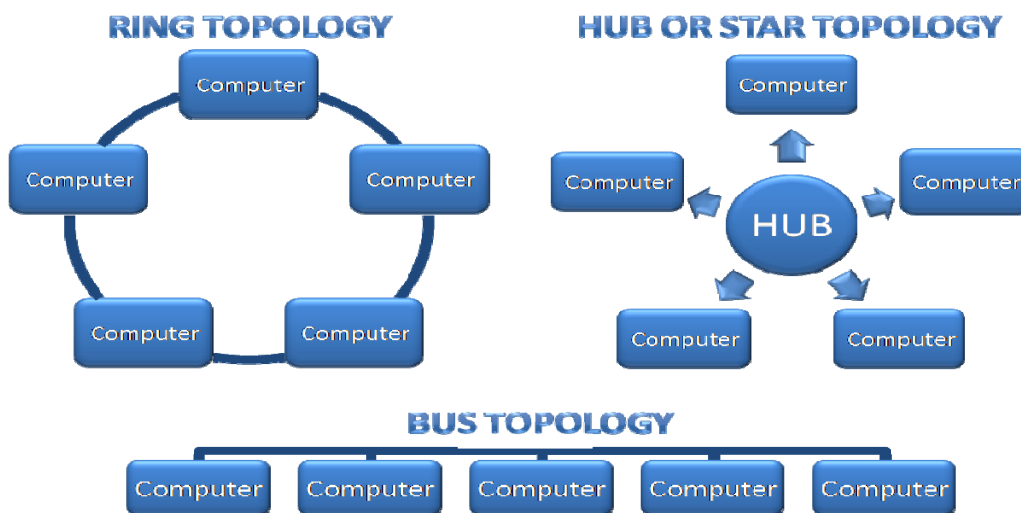
**Transmission Control Protocol/Internet Protocol** or **TCP/IP** is the universal computer-to-computer language of the Internet, but can also be implemented to support an intranet.

### Thin Client

See *Client*

### Topology

A geometric description of a network's structure based upon the way devices interconnect. Compare communication routes of the Ring, Hub or Star and Bus topologies depicted below.



## V

### Virtual Machine

**Virtual machine** or **VM** refers to the use of software to emulate or mimic the presence and function of hardware. Using VM software, a complete hardware and software computing environment, including operating systems, applications, data and emulated peripherals, can be stored in a single file. When that file is loaded to a VM player, it looks and works just like a real machine, but runs in a window, like any other piece of software.

Virtual machines have found enthusiastic acceptance in the IT world as a means to deploy, protect and backup virtualized servers, as well as a method to extract more value from hardware because one “real” machine can run many virtual machines without a notable drop in performance.

Because VMs can replicate almost any computing platform or environment, it promises to be a viable form of production for complex ESI. Virtualization enables opposing sides to enjoy comparable levels of functionality in native production even when one side lacks the hardware and software resources of the other. Not only does the evidence look the same for both sides, but it *works* the same way and can be easily shielded from inadvertent alteration and intentional manipulation.

### Volume

A volume is a logical division of a hard drive that can hold a single operating system. Where a partition was akin to the physical drawer in a file cabinet, a volume speaks to the division of that drawer into compartments to hold file systems and files.

### VPN

A **Virtual Private Network** or **VPN** is a private (i.e., secure) network that employs public pathways (i.e., the Internet). By employing authentication protocols and encryption of data as it traverses public pathways, the network traffic over a VPN is protected from interception and thus said to “tunnel” through public areas.

## W

### Workgroup

A workgroup is a subset of users in a local area network environment who are assigned privileges enabling them to collaborate by sharing files and peripherals. Microsoft Windows uses the term workgroup to identify the participants in a peer-to-peer network.



# Meeting the Challenge: E-Mail in Civil Discovery

Craig Ball

**Meeting the Challenge: E-Mail in Civil Discovery**  
**Craig Ball**  
©2009

**Table of Contents**

<b>Introduction.....</b>	<b>39</b>
<b>Not Enough Eyeballs.....</b>	<b>40</b>
<b>Test Your E.Q.....</b>	<b>41</b>
<b>Staying Out of Trouble.....</b>	<b>42</b>
<b>...And You Could Make Spitballs with It, Too.....</b>	<b>42</b>
<b>Did You Say <i>Billion</i>?.....</b>	<b>42</b>
<b>Net Full of Holes.....</b>	<b>43</b>
<b>New Tools.....</b>	<b>43</b>
<b>E-Mail Systems and Files.....</b>	<b>43</b>
<b>A Snippet about Protocols.....</b>	<b>44</b>
<b>Incoming Mail: POP, IMAP, MAPI and HTTP E-Mail.....</b>	<b>44</b>
<b>Outgoing Mail: SMTP and MTA.....</b>	<b>46</b>
<b>Anatomy of an E-Mail Header.....</b>	<b>47</b>
<b>E-Mail Autopsy: Tracing a Message’s Incredible Journey.....</b>	<b>49</b>
<b>Hashing and Deduplication.....</b>	<b>53</b>
<b>Local E-Mail Storage Formats and Locations.....</b>	<b>54</b>
<b>Looking for E-Mail 101.....</b>	<b>56</b>
<b>Finding Outlook E-Mail.....</b>	<b>57</b>
<b>Outlook Mail Stores Paths.....</b>	<b>58</b>
<b>“Temporary” OLK Folders.....</b>	<b>58</b>
<b>Finding Outlook Express E-Mail.....</b>	<b>59</b>
<b>Finding Windows Mail and Windows Live Mail E-Mail Stores.....</b>	<b>60</b>
<b>Finding Netscape E-Mail.....</b>	<b>61</b>
<b>Microsoft Exchange Server.....</b>	<b>61</b>
<b>The ABCs of Exchange.....</b>	<b>63</b>
<b>Recovery Storage Groups and ExMerge.....</b>	<b>63</b>
<b>Journaling, Archiving and Transport Rules.....</b>	<b>64</b>

<b>Lotus Domino Server and Notes Client .....</b>	<b>65</b>
<b>Novell GroupWise.....</b>	<b>66</b>
<b>Webmail.....</b>	<b>67</b>
<b>Computer Forensics.....</b>	<b>68</b>
<b>Why Deleted Doesn't Mean Gone.....</b>	<b>70</b>
<b>Forms of Production .....</b>	<b>72</b>
<b>Conclusion.....</b>	<b>74</b>

## Introduction

This paper looks at e-mail from the standpoint of what lawyers should know about the nuts-and-bolts of these all-important communications systems. It's technical; sometimes, *very* technical.

When you finish the paper, you'll know *a lot* more about e-mail, and along the way, you may realize that discoverable e-mail can be found in far more places than your client probably checked before the last time you said, "Yes, your Honor, we've given them the e-mail."

So, if you know what's good for you, you should probably stop reading right now.

....

Still here? Okay, you asked for it.

*Get the e-mail!* It's the war cry in discovery today. More than simply a feeding frenzy, it's an inevitable recognition of e-mail's importance and ubiquity. We go after e-mail because it accounts for the majority of business communications and because e-mail users tend to let their guard down and reveal plainspoken truths they'd never dare put in a memo. Or do they? A 2008 study<sup>29</sup> demonstrated that employees are significantly more likely to lie in e-mail messages than in traditional pen-and-paper communications. Whether replete with ugly truths or ugly lies, e-mail is telling and compelling evidence.

If you're on the producing end of a discovery request, you not only worry about what the messages say, but also whether you and your client can find, preserve and produce all responsive items. Questions like these *should* keep you up nights:

- Will the client simply conceal damning messages, leaving counsel at the mercy of an angry judge or disciplinary board?
- Will employees seek to rewrite history by deleting "their" e-mail from company systems?
- Will the searches employed prove reliable and be directed to the right digital venues?

<sup>29</sup> [http://www3.lehigh.edu/News/V2news\\_story.asp?iNewsID=2892](http://www3.lehigh.edu/News/V2news_story.asp?iNewsID=2892) (visited 11/1/08)

- Will review processes unwittingly betray privileged or confidential communications?

Meeting these challenge begins with understanding e-mail technology well enough to formulate a sound, defensible strategy. For requesting parties, it means grasping the technology well enough to assess the completeness and effectiveness of your opponent's e-discovery efforts.

This paper seeks to equip the corporate counsel or trial lawyer with some of what's needed to meet the challenge of e-mail discovery in civil litigation. It's intended to be technical because technical knowledge is what's most needed and most lacking in continuing legal education today. Even if you went to law school because you had no affinity for matters technical, it's time to dig in and learn enough to stay in the fray.

### **Not Enough Eyeballs**

Futurist Arthur C. Clarke said, "Any sufficiently advanced technology is indistinguishable from magic." E-mail, like electricity or refrigeration, is one of those magical technologies we use every day without knowing quite how it works. But, "It's magic to me, your Honor," won't help you when the e-mail pulls a disappearing act. Judges expect you to pull that e-mail rabbit out of your hat.

A lawyer managing electronic discovery is obliged to do more than just tell their clients to "produce the e-mail." You've got to make an effort to understand their systems and procedures and ask the right questions. Plus, you have to know when you aren't getting the right answers. Perhaps that's asking a lot, but well over 95% of all business documents are born digitally and only a tiny fraction are ever printed.<sup>30</sup> Hundreds of billions of e-mails traverse the Internet *daily*, far more than telephone and postal traffic combined,<sup>31</sup> and the average business person sends and receives between 50 and 150 e-mails *every business day*. E-mail contributes *500 times greater volume* to the Internet than web page content.

Think that's a lot? Then best not think about the fact that the volume is expected to nearly double by 2012,<sup>32</sup> and none of these numbers take into account the explosive growth in instant messaging, unified messaging or the next insanely great communication or collaboration technology that—starting next year and every year—we can hardly live without. The volume keeps increasing, and there's no end in sight. It's simply too easy, too quick and too cheap to expect anything else.

Neither should we anticipate a significant decline in users' propensity to retain their e-mail. Here again, it's too easy and, at first blush, too cheap to expect users to selectively dispose of e-mail

---

<sup>30</sup> Extrapolating from a 2003 updated study compiled by faculty and students at the School of Information Management and Systems at the University of California at Berkeley.

<http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>

<sup>31</sup> <http://www.radicati.com/?p=638> (visited 11/1/08)

<sup>32</sup> Id.



and still meet business, litigation hold and regulatory obligations. Our e-mail is so twisted up with our lives that to abandon it is to part with our personal history.

Another difficulty is that this startling growth isn't happening in just one locale. E-mail lodges on servers, cell phones, laptops, home systems, thumb drives and in "the cloud," a term ethereally denoting all the places we store information online, little knowing or caring about its physical location. Within the systems, applications and devices we use to store and access e-mail, most users and even many IT professionals don't know where messages lodge or how long they hang around.

In discovery, we overlook so much that we're obliged to consider, and with respect to what we do collect, it's increasingly infeasible to put enough pairs of trained eyes in front of enough computers to review every potentially responsive electronic document. Instead, we must employ shortcuts that serve as proxies for lawyer judgment. Here, too, our success hinges upon our understanding of the technologies we use to extend and defend our reach.

### **Test Your E.Q.**

Suppose opposing counsel serves a preservation demand or secures an order compelling your client to preserve electronic messaging. Are you assured that your client can and will faithfully back up and preserve responsive data? Even if it's practicable to capture and set aside the current server e-mail stores of key custodians—and even if you hold onto backup tapes for a few significant points in time—are you *really* capturing all or even most of the discoverable communications? How much is falling outside your net, and how do you assess its importance?

Here are a dozen questions you should be able to confidently answer about your client's communication systems:

1. What messaging environment(s) does your client employ? Microsoft Exchange, Lotus Domino, Novell GroupWise or something else?
2. Do *all* discoverable electronic communications come in and leave via the company's e-mail server?
3. Is the e-mail system configured to support synchronization with local e-mail stores on laptops and desktops?
4. How long have the current e-mail client and server applications been used?
5. What are the message purge, dumpster, journaling and archival settings for each key custodian?
6. Can your client disable a specific custodian's ability to delete messages?
7. Does your client's backup or archival system capture e-mail stored on individual user's hard drives, including company-owned laptops?
8. Where are e-mail container files stored on laptops and desktops?
9. How should your client collect and preserve relevant web mail?

10. Do your clients' employees use home machines, personal e-mail addresses or browser-based e-mail services (like Gmail or Yahoo! Mail) for discoverable business communications?
11. Do your clients' employees use Instant Messaging on company computers or over company-owned networks?
12. How do your clients' voice messaging systems store messages, and how long are they retained?

If you are troubled that you can't answer some of these questions, you should be; but know you're not alone. Many other lawyers can't either. And don't delude yourself that these are exclusively someone else's issues, e.g., your litigation support services vendor or IT expert. These are the inquiries that will soon be coming at *you* in court and when conferring with the other side. You do confer on ESI, right?

### **Staying Out of Trouble**

Fortunately, the rules of discovery don't require you to do the impossible. All they require is diligence, reasonableness and good faith. To that end, you must be able to establish that you and your client acted swiftly, followed a sound plan, and took such action as reasonable minds would judge adequate to the task. It's also important to keep the lines of communication open with the opposing party and the court, seeking agreement with the former or the protection of the latter where fruitful. I'm fond of quoting Oliver Wendell Holmes' homily, "Even a dog knows the difference between being stumbled over and being kicked." Judges, too, have a keen ability to distinguish error from arrogance. There's no traction for sanctions when it is clear that the failure to produce electronic evidence occurred despite good faith and due diligence.

### **...And You Could Make Spitballs with It, Too**

Paper discovery enjoyed a self-limiting aspect because businesses tended to allocate paper records into files, folders and cabinets according to persons, topics, transactions or periods of time. The space occupied by paper and the high cost to create, manage and store paper records served as a constant impetus to cull and discard them, or even to avoid creating them in the first place. By contrast, the ephemeral character of electronic communications, the ease of and perceived lack of cost to create, duplicate and distribute them and the very low direct cost of data storage have facilitated a staggering and unprecedented growth in the creation and retention of electronic evidence. At fifty e-mails per day, a company employing 100,000 people could find itself storing well over *1.5 billion* e-mails annually.

### **Did You Say *Billion*?**

But volume is only part of the challenge. Unlike paper records, e-mail tends to be stored in massive data blobs. The single file containing my Outlook e-mail is over four gigabytes in size and contains tens of thousands of messages, many with multiple attachments covering virtually every aspect of my life and many other people's lives, too. In thousands of those e-mails, the subject line bears only a passing connection to the contents as "Reply to" threads strayed



further and further from the original topic. E-mails meander through disparate topics or, by absent-minded clicks of the “Forward” button, lodge in my inbox dragging with them, like toilet paper on a wet shoe, the unsolicited detritus of other people’s business.

To respond to a discovery request for e-mail on a particular topic, I’d either need to skim/read countless messages or I’d have to naively rely on keyword search to flush out all responsive material. If the request for production implicated material I no longer kept on my current computer or web mail collections, I’d be forced to root around through a motley array of archival folders, old systems, obsolete disks, outgrown hard drives, ancient backup tapes (for which I currently have no tape reader) and unlabeled CDs. Ugh!

### **Net Full of Holes**

I’m just one guy. What’s a company to do when served with a request for “all e-mail” on a particular matter in litigation? Surely, I mused, someone must have found a better solution than repeating, over and over again, the tedious and time-consuming process of accessing individual e-mail servers at far-flung locations along with the local drives of all key players’ computers?

For this article, I contacted colleagues in both large and small electronic discovery consulting groups, inquiring about “the better way” for enterprises, and was struck by the revelation that, if there was a better mousetrap, they hadn’t discovered it either. Uniformly, we recognized such enterprise-wide efforts were gargantuan undertakings fraught with uncertainty and concluded that counsel must somehow seek to narrow the scope of the inquiry—either by data sampling or through limiting discovery according to offices, regions, time span, business sectors or key players. Trying to capture *everything*, enterprise-wide, is trawling with a net full of holes.

### **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 de-duplication,” rules-based journaling and other customizable features.

These tools, especially enterprise archival, 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.

### **E-Mail Systems and Files**

The corporate and government e-mail environment is dominated by two well-known, competitive product pairs: Microsoft Exchange Server and its Outlook e-mail client and IBM Lotus Domino

server and its Lotus Notes client. A legacy environment called Novell GroupWise occupies a distant third place, largely among government users.

Per a 2008 study by Ferris Research,<sup>33</sup> Microsoft Exchange accounts for 65% of market share among all organizations, with significantly larger shares among businesses with fewer than 49 employees and those in the health care and telecommunications sectors. Lotus Notes was found to have just 10% of overall market share, but a much higher percentage base among manufacturers with at least 5,000 employees. GroupWise's share was termed "negligible," except in niches—notably organizations in the financial services and government sectors with 100 to 999 employees—where its share reached as high as 10-15%. Blackberry servers transmit a large percentage of e-mail as well, but these messages typically find their way to or through an Exchange or Lotus mail server.

Of course, when one looks at personal and small office/home office business e-mail, it's rare to encounter server-based Exchange or Domino systems. Here, the market belongs to Internet service providers (e.g., AOL, the major cable and telephone companies and hundreds of smaller, local players) and web mail providers (e.g., Gmail, Yahoo! Mail or Hot Mail). Users employ a variety of e-mail client applications, including Microsoft Outlook, Windows Mail (formerly Outlook Express), Eudora, Entourage (on Apple machines) and, of course, their web browser and webmail. This motley crew and the enterprise behemoths are united by common e-mail *protocols* that allow messages and attachments to be seamlessly handed off between applications, providers, servers and devices.

### **A Snippet about Protocols**

Computer network specialists are always talking about this "protocol" and that "protocol." Don't let the geek-speak get in the way. An *application protocol* 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 adjacent to the addressee information. 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—or just fail to drop it in the mail—and Grandma gets no Hallmark this year! 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.

### **Incoming Mail: POP, IMAP, MAPI and HTTP E-Mail**

Although Microsoft Exchange Server rules the roost in enterprise e-mail, it's by no means the most common e-mail system for the individual and small business user. When you access your

---

<sup>33</sup> <http://www.ferris.com/2008/01/31/email-products-market-shares-versions-deployed-migrations-and-software-cost/> visited 11/10/08.

personal e-mail from your own Internet Service Provider (ISP), chances are your e-mail comes to you from your ISP's e-mail server in one of three ways: POP3, IMAP or HTTP, the last commonly called web- or browser-based e-mail. Understanding how these three protocols work—and differ—helps in identifying where e-mail can be found.

**POP3** (for Post Office Protocol, version 3) is the oldest and most common of the three approaches and the one most familiar (by function, if not by name) to users of the Windows Mail, Outlook Express and Eudora e-mail clients. Using POP3, you connect to a mail server, download copies of all messages and, unless you have configured your e-mail client to leave copies on the server, the e-mail is deleted on the server and now resides on the hard drive of the computer you used to pick up mail. Leaving copies of your e-mail on the server seems like a great idea as it allows you to have a back up if disaster strikes and facilitates easy access of your e-mail, again and again, from different computers. However, few ISPs afford unlimited storage space on their servers for users' e-mail, so mailboxes quickly become “clogged” with old e-mails, and the servers start bouncing new messages. As a result, POP3 e-mail typically resides only on the local hard drive of the computer used to read the mail and on the back up system for the servers which transmitted, transported and delivered the messages. In short, POP is locally-stored e-mail that supports some server storage.

**IMAP** (Internet Mail Access Protocol) functions in much the same fashion as most Microsoft Exchange Server installations in that, when you check your messages, your e-mail client downloads just the headers of e-mail it finds on the server and only retrieves the body of a message when you open it for reading. Else, the entire message stays in your account on the server. Unlike POP3, where e-mail is searched and organized into folders locally, IMAP e-mail is organized and searched on the server. Consequently, the server (and its back up tapes) retains not only the messages but also the way the user *structured* those messages for archival.

Since IMAP e-mail “lives” on the server, how does a user read and answer it without staying connected all the time? The answer is that IMAP e-mail clients afford users the ability to synchronize the server files with a local copy of the e-mail and folders. When an IMAP user reconnects to the server, local e-mail stores are updated (synchronized) and messages drafted offline are transmitted. So, to summarize, IMAP is server-stored e-mail, with support for synchronized local storage.

A notable distinction between POP3 and IMAP e-mail centers on where the “authoritative” collection resides. Because each protocol allows for messages to reside both locally (“downloaded”) and on the server, it's common for there to be a difference between the local and server collections. Under POP3, the *local* collection is deemed authoritative whereas in IMAP the *server* collection is authoritative. But for e-discovery, the important point is that the contents of the local and server e-mail stores can and do *differ*.

**MAPI** (Messaging Application Programming Interface) is the e-mail protocol at the heart of Windows and Microsoft's Exchange Server applications. Simple MAPI comes preinstalled on Windows machines to provide basic messaging services for Windows Mail/Outlook Express. A substantially more sophisticated version of MAPI (Extended MAPI) is installed with Microsoft Outlook and Exchange. Like IMAP, MAPI e-mail is typically stored on the server and not necessarily on the client machine. The local machine may be configured to synchronize with the server mail stores and keep a copy of mail on the local hard drive (typically in an Offline Synchronization file with the extension .OST), but this is user- and client application-dependent. Though it's exceedingly rare (especially for laptops) for there to be no local e-mail stores for a MAPI machine, it's nonetheless possible, and e-mail won't be found on the local hard drive except to the extent fragments may turn up through computer forensic examination.

**HTTP** (Hyper Text Transfer Protocol) mail, or web-based/browser-based e-mail, dispenses with the local e-mail client and handles all activities on the server, with users managing their e-mail using their Internet browser to view an interactive web page. Although most browser-based e-mail services support local POP3 or IMAP synchronization with an e-mail client, users may have no local record of their browser-based e-mail transactions except for messages they've affirmatively saved to disk or portions of e-mail web pages which happen to reside in the browser's cache (e.g., Internet Explorer's Temporary Internet Files folder). Gmail, AOL, Hotmail and Yahoo! Mail are popular examples of browser-based e-mail services, although many ISPs (including all the national providers) offer browser-based e-mail access in addition to POP and IMAP connections.

The protocol used to carry e-mail is not especially important in electronic discovery except to the extent that it signals the most likely place where archived and orphaned e-mail can be found. Companies choose server-based e-mail systems (e.g., IMAP and MAPI) for two principal reasons. First, such systems make it easier to access e-mail from different locations and machines. Second, it's easier to back up e-mail from a central location. Because IMAP and MAPI systems store e-mail on the server, the back up system used to protect server data can yield a mother lode of server e-mail.

Depending upon the back up procedures used, access to archived e-mail can prove a costly and time-consuming task or a relatively easy one. The enormous volume of e-mail residing on back up tapes and the potentially high cost to locate and restore that e-mail makes discovery of archived e-mail from backup tapes a major bone of contention between litigants. In fact, most reported cases addressing cost-allocation in e-discovery seem to have been spawned by disputes over e-mail on server back up tapes.

### **Outgoing Mail: SMTP and MTA**

Just as the system that brings water into your home works in conjunction with a completely different system that carries wastewater away, the protocol that delivers e-mail to you is

completely different from the one that transmits your e-mail. Everything discussed in the preceding paragraph concerned the protocols used to *retrieve* e-mail from a mail server.

Yet another system altogether, called **SMTP** for *Simple Mail Transfer Protocol*, takes care of outgoing e-mail. SMTP is indeed a very simple protocol and doesn't even require authentication, in much the same way as anyone can anonymously drop a letter into a mailbox. A server that uses SMTP to route e-mail over a network to its destination is called an **MTA** for *Message Transfer Agent*. Examples of MTAs you might hear mentioned by IT professionals include Sendmail, Exim, Qmail and Postfix. Microsoft Exchange Server is an MTA, too. In simplest terms, an MTA is the system that carries e-mail between e-mail servers and sees to it that the message gets to its destination. Each MTA reads the code of a message and determines if it is addressed to a user in its domain and, if not, passes the message on to the next MTA after adding a line of text to the message identifying the route to later recipients. If you've ever set up an e-mail client, you've probably had to type in the name of the servers handling your outgoing e-mail (perhaps *SMTP.yourISP.com*) and your incoming messages (perhaps *mail.yourISP.com* or *POP.yourISP.com*).

### **Anatomy of an E-Mail Header**

Now that we've waded through the alphabet soup of protocols managing the movement of an e-mail message, let's take a look inside the message itself. Considering the complex systems on which it lives, an e-mail is astonishingly simple in structure. The Internet protocols governing e-mail transmission require electronic messages to adhere to rigid formatting, making individual e-mails fairly easy to dissect and understand. The complexities and headaches associated with e-mail don't really attach until the e-mails are stored and assembled into databases and local stores.

An e-mail is just a plain text file. Though e-mail can be "tricked" into carrying non-text binary data like application files (*i.e.*, a Word document) or image attachments (*e.g.*, GIF or JPEG files), this piggybacking requires binary data be *encoded into text* for transmission. Consequently, even when transmitting files created in the densest computer code, *everything in an e-mail is plain text*.

Figure 1 is an e-mail I sent from one of my e-mail addresses to another with a small image attached. Transmitted and received in seconds using the same machine, the message was sliced-and-diced into two versions (plain text and HTML), and its image attachment was encoded into Base 64, restructured to comply with rigid Internet protocols. It then winged its way across several time zones and servers, each server prepending its own peculiar imprimatur.

Figure 1 is just one of a variety of different ways in which an e-mail client application (in this instance the webmail application, Gmail) may display a message. When you view e-mail onscreen or print it out, you're seeing just part of the data contained in the message and attachment. Moreover, the e-mail client may be interpreting the message data according to, e.g., the time zone and daylight savings time settings of your machine or its ability to read embedded formatting information. What you don't see—or see accurately—may be of little import, or it may be critical evidence. You've got to know what lies beneath to gauge its relevance.

**Figure 1: Exemplar E-Mail as Displayed in Client Application**

**An E-Mail's Incredible Journey** Inbox | X

☆ **Craig Ball** to craig show details 5:51 PM (2 hours ago) Reply

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 Technologist  
Certified Computer Forensic Examiner  
3723 Lost Creek Blvd.  
Austin, Texas 78746  
TEL: 512-514-0182  
E-MAIL: [craig@ball.net](mailto:craig@ball.net)

---


 **Ball-photo\_76x50 pixels\_B&W.jpg**  
2K [View](#) [Download](#)

Figure 2 (opposite) shows the source code of the Figure 1 e-mail, sent using a browser-based Gmail account. The e-mail came from the account [computerforensics@gmail.com](mailto:computerforensics@gmail.com) and was addressed to [craig@ball.net](mailto:craig@ball.net). A small photograph in JPEG format was attached.

Before we dissect the e-mail message in Figure 2, note that any e-mail can be divided into two parts, the header and body of the message. By design, the header details the journey taken by the e-mail from origin to destination; but be cautioned that it's a fairly simple matter for a hacker to spoof (falsify) the identification of all but the final delivery server. Accordingly, where the origin or origination date of an e-mail is suspect, the actual route of the message may need to be validated at each server along its path.

In an e-mail header, each line which begins with the word "Received:" represents the transfer of the message between or within systems. The transfer sequence is reversed chronologically such that those closest to the top of the header were inserted after those that follow, and the topmost line reflects delivery to the recipient's e-mail server. As the message passes through intervening hosts, each adds its own identifying information along with the date and time of transit.

### **E-Mail Autopsy: Tracing a Message's Incredible Journey**

In this header, section **(A)** indicates the parts of the message designating the sender, addressee, recipient, date, time and subject line of the message. Importantly, the header also identifies the message as being formatted in MIME (MIME-Version: 1.0).<sup>34</sup> The **Content-Type: multipart/mixed** reference that follows indicates that the message holds both text and one or more attachments.

Though a message may be assigned various identification codes by the servers it transits in its journey (each enabling the administrator of the transiting e-mail server to track the message in the server logs), the message will contain one unique identifier assigned by the originating Message Transfer Agent. The unique identifier assigned to this message at **(B)** labeled "Message-ID:" is:

**1023f46e0811102015gd55453fpec00af81eb38dfaa@mail.gmail.com.**

In the line labeled "Date," both the date and time of transmittal are indicated. The time indicated is 22:15:33, and the "-0600" which follows denotes the time *difference* between the sender's local time (the system time on my computer in Austin, Texas in standard time) and Coordinated Universal Time (UTC), roughly equivalent to Greenwich Mean Time. As the offset from UTC is minus six hours on November 10, 2008, we deduce that the message was sent from a machine set to Central Standard Time, giving some insight into the sender's location. Knowing the originating computer's time and time zone can occasionally prove useful in demonstrating fraud or fabrication.

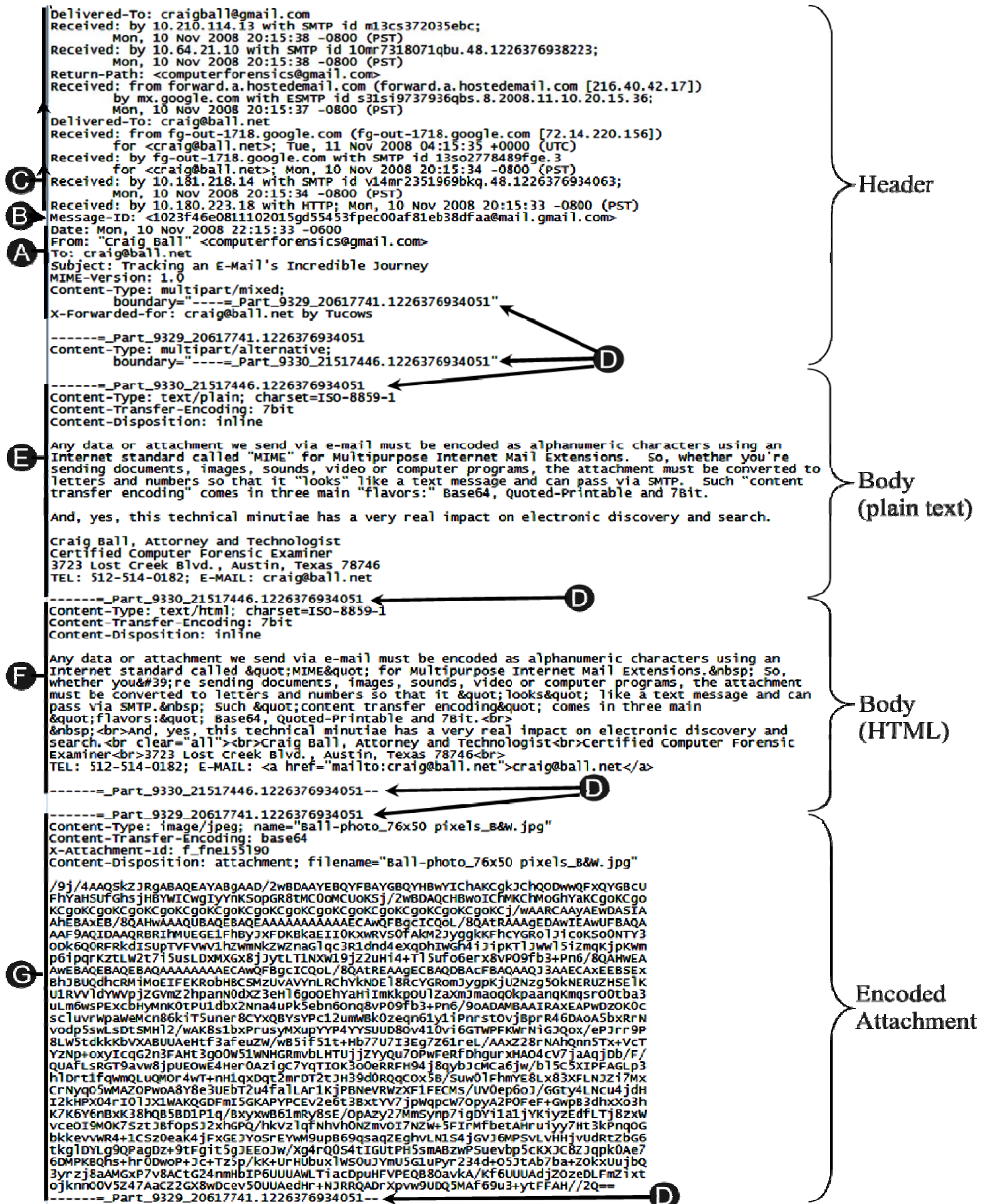
At **(A)**, we see that the message was addressed to [craig@ball.net](mailto:craig@ball.net) from [computerforensics@gmail.com](mailto:computerforensics@gmail.com); yet, the ultimate recipient of the message is (as seen at

---

<sup>34</sup> 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.



Figure 2: Anatomy of an E-Mail



the very top of the page) [craigball@gmail.com](mailto:craigball@gmail.com). How this transpired can be deciphered from the header data, read from the bottom up.

The message was created and sent using Gmail web interface; consequently the first hop **(C)** indicates that the message was transmitted using HTTP and first received by IP (Internet Protocol) address 10.180.223.18 at 20:15:33 -0800 (PST). Note that the server marks time in Pacific Standard Time, suggesting it may be located on the West Coast. The message is immediately handed off to another IP address 10.181.218.14 using Simple Mail Transfer Protocol, denoted by the initials SMTP. Next, we see another SMTP hand off to Google's server named "fg-out-1718.google.com" (IP address 72.14.220.156), which immediately transmits the message to a server with the IP address 216.40.42.17 and keeping time in UTC. A check of that IP address reveals that it's registered to Tucows International in Toronto, Canada.

Tucows is the host of my [craig@ball.net](mailto:craig@ball.net) address, which is configured to forward incoming messages to my other Gmail address, [craigball@gmail.com](mailto:craigball@gmail.com). The forwarding is handled by a server called *forward.a.hostedemail.com*, and we then see the message received by server *MX.google.com*, transferred via SMTP to a server at IP address 10.64.21.10, then finally come to rest, delivered via SMTP to my [craigball@gmail.com](mailto:craigball@gmail.com) address via a server at 10.210.114.

As we examine the structure of the e-mail, we see that it contains content boundaries separating its constituent parts **(D)**. These content boundary designators serve as delimiters; that is, sequences of one or more characters used to specify the boundary between text or data streams.<sup>35</sup> In order to avoid confusion of the boundary designator with message text, a complex sequence of characters is generated to serve as the two boundary designators used in this message. The first, called "\_Part\_9329\_20617741.1226376934051," serves to separate the message header from the message body and signal the end of the message. The second delimiter, called "----=\_Part\_9330\_21517446.1226376934051," denotes the boundaries between the segments of the message body: here, plain text content **(E)**, HTML content **(F)** and the encoded attachment **(G)**.

I didn't draft the message in *both* plain text and HTML formats, but my e-mail client thoughtfully did so to insure that my message won't confuse recipients using e-mail clients unable to display the richer formatting supported by HTML. For these recipients, there is a plain text version, too (albeit without the bolding, italics, hyperlinks and other embellishments of HTML). That the message carries alternative versions of the text is flagged by the designation at the break between header and message body stating: "**Content-Type: multipart/alternative.**"

Looking more closely at the message boundaries, we see that each boundary delimiter is followed by Content-Type and Content-Transfer-Encoding designations. The plain text version

---

<sup>35</sup> The use of delimiters should be a familiar concept to those accustomed to specifying load file formats to accompany document image productions employed in e-discovery, where commas typically serve as field delimiters. Hence, these load files are sometimes referred to as CSV files (for comma-separated values).

of the message (E) begins: **“Content-Type: text/plain; charset=ISO-8859-1,”** followed by **“Content-Transfer-Encoding: 7bit.”** The first obviously denotes plain text content using the very common ISO-8859-1 character encoding more commonly called “Latin 1.”<sup>36</sup> The second signals that the content that follows consists of standard ASCII characters which historically employ 7 bits to encode 128 characters.

Not surprisingly, the boundary for the HTML version uses the Content-Type designator “text/html.”

The most interesting and complex part of the message (F) starts after the second to last boundary delimiter with the specifications:

**Content-Type: image/jpeg; name="Ball-photo\_76x50 pixels\_B&W.jpg"**  
**Content-Transfer-Encoding: base64**

The content type is self explanatory: an image in the JPEG format common to digital photography. The “name” segment obviously carries the name to be re-assigned to the attached photograph when decoded at its destination. But where, exactly, is the photograph?

Recall that to travel as an e-mail attachment, binary content (like photos, sound files, video or machine codes) must first be converted to plain text characters. Thus, the photograph has been encoded to a format 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.<sup>37</sup>

Accordingly, the attached JPEG photograph with the filename “Ball-photo\_76x50 pixels\_B&W.jpg,” has been encoded from non-printable binary code into those 26 lines of gibberish comprising nearly 2,000 plain text characters (G) and **Figure 3**. It’s now able to traverse the network as an e-mail, yet easily be converted back to binary data when the message reaches its destination.



<sup>36</sup> In simplest terms, a character set or encoding pairs a sequence of characters (like the Latin alphabet) with numbers, byte values or other signals in much the same way as Morse code substitutes particular sequences of dots and dashes for letters. It’s the digital equivalent of the Magic Decoder Rings once found in boxes of Cracker Jacks.

<sup>37</sup> A third common transfer encoding is called “quoted-printable” or “QP encoding.” It facilitates transfer of non-ASCII 8-bit data as 7-bit ASCII characters using three ASCII characters (the “equals” sign followed by two hexadecimal characters: 0-9 and A-F) to stand in for a byte of data. Quoted-printable is employed where the content to be encoded is predominantly ASCII text coupled with some non-ASCII items. Its principal advantage is that it allows the encoded data to remain largely intelligible to readers.

Clearly, e-mail clients don't display all the information contained in a message's source but instead parse the contents into the elements we most want to see: To, From, Subject, body, and attachment. If you decide to try a little digital detective work on your own e-mail, you'll find that some e-mail client software doesn't make it easy to see complete header information. Microsoft's Outlook mail client makes it difficult to see the complete message source; however, you can see message headers for individual e-mails by opening the e-mail, then selecting "View" followed by "Options" until you see the "Internet headers" window on the Message Option menu. In Microsoft Outlook Express (now Windows Mail), highlight the e-mail item you want to analyze and then select "File" from the Menu bar, then "Properties," then click the "Details" tab followed by the "Message Source" button. For Gmail, select "Show Original" from the Reply button pull-down menu.

The lesson from this is that what you see displayed in your e-mail client application isn't really the e-mail. It's an *arrangement* of selected *parts* of the message, frequently modified in some respects from the native message source that traversed the network and Internet and, as often, supplemented by metadata (like message flags, contact data and other feature-specific embellishments) unique to your software and setup. What you see handily displayed as a discrete attachment is, in reality, encoded into the message body. The time assigned to message is calculated relative to your machine's time and DST settings. Even the sender's name may be altered based upon the way your machine and contact's database is configured. What you see is not always what you get (or got).

### **Hashing and Deduplication**

Hashing is the use of mathematical algorithms to calculate a unique sequence of letters and numbers to serve as a "fingerprint" for digital data. These fingerprint sequences are called "message digests" or, more commonly, "hash values."

The ability to "fingerprint" data makes it possible to identify identical files without the necessity of examining their content. If the hash values of two files are identical, the files are identical. This file-matching ability allows hashing to be used to de-duplicate collections of electronic files before review, saving money and minimizing the potential for inconsistent decisions about privilege and responsiveness for identical files.

Although hashing is a useful and versatile technology, it has a few shortcomings. Because the tiniest change in a file will alter that file's hash value, hashing is of little value in comparing files that have any differences, even if those differences have no bearing on the substance of the file. Applied to e-mail, we understand from our e-mail "autopsy" that messages contain unique identifiers, time stamps and routing data that would frustrate efforts to compare one complete message to another using hash values. Looking at the message as a whole, multiple recipients of the same message have different versions insofar as their hash values.

Consequently, deduplication of e-mail messages is accomplished by calculating hash values for selected segments of the messages and comparing those segment values. Thus, hashing e-mails for deduplication will omit the parts of the header data reflecting, e.g., the message identifier and the transit data. Instead, it will hash just the data seen in, e.g., the To, From, Subject and Date lines, message body and encoded attachment. If these match, the message can be said to be *practically* identical.

For example, a deduplication application might hash only segments **(A)**, **(E)** and **(G)** of Figure 2. If the hash values of these segments match the hash values of the same segments of another message, can we say they are the same message? Probably, but it could also be important to evaluate emphasis added by HTML formatting (e.g., text in red or underlined) or information about blind carbon copy recipients. The time values or routing information in the headers may also be important to reliably establishing authenticity, reliability or sequence.

By hashing particular segments of messages and selectively comparing the hash values, it's possible to gauge the *relative* similarity of e-mails and perhaps eliminate the cost to review messages that are *inconsequentialy* different. This concept is called "near deduplication." It works, but it's important to be aware of exactly what it's excluding and why. It's also important to advise your opponents when employing near deduplication and ascertain whether you're mechanically excluding evidence the other side deems relevant and material.

Hash deduplication of e-mail is tricky. Time values may vary, along with the apparent order of attachments. These variations, along with minor formatting discrepancies, may serve to prevent the exclusion of items defined as duplicates. When this occurs, be certain to delve into the reasons *why* apparent duplicates aren't deduplicating, as such errors may be harbingers of a broader processing problem.

### **Local E-Mail Storage Formats and Locations**

Suppose you're faced with a discovery request for a client's e-mail and there's no budget or time to engage an e-discovery service provider or ESI expert?

*Where are you going to look to find stored e-mail, and what form will it take?*

"Where's the e-mail?" It's a simple question, and one answered too simply and often wrongly by, "It's on the server" or "The last 60 days of mail is on the server and the rest is purged." Certainly, much e-mail will reside on the server, but most e-mail is elsewhere; and it's never all gone in practice, notwithstanding retention policies. The true location and extent of e-mail depends on systems configuration, user habits, backup procedures and other hardware, software and behavioral factors. This is true for mom-and-pop shops, for large enterprises and for everything in-between.



Going to the server isn't the wrong answer. It's just not the whole answer. In a matter where I was tasked to review e-mails of an employee believed to have stolen proprietary information, I went first to the company's Microsoft Exchange e-mail server and gathered a lot of unenlightening e-mail. Had I stopped there, I would've missed the Hotmail traffic in the Temporary Internet Files folder and the Short Message Service (SMS) exchanges in the PDA synchronization files. I'd have overlooked the Microsoft Outlook archive file (archive.pst) and offline synchronization file (Outlook.ost) on the employee's laptop, collectively holding thousands more e-mails, including some "smoking guns" absent from the server. These are just some of the many places e-mails without counterparts on the server may be found. Though an exhaustive search of every nook and cranny may not be required, you need to know your options in order to assess feasibility, burden and cost.

E-mail resides in some or all of the following venues, grouped according to relative accessibility:

**Easily Accessible:**

- **E-Mail Server:** Online e-mail residing in active files on enterprise servers: MS Exchange e.g., (.edb, .stm, .log files), Lotus Notes (.nsf files), Novell GroupWise (.db files)
- **File Server:** E-mail saved as individual messages or in container files on a user's network file storage area ("network share").
- **Desktops and Laptops:** E-mail stored in active files on local or external hard drives of user workstation hard drives (e.g., .pst, .ost files for Outlook and .nsf for Lotus Notes), laptops (.ost, .pst, .nsf), mobile devices, and home systems, particularly those with remote access to networks.
- OLK system subfolders holding viewed attachments to Microsoft Outlook messages, *including deleted messages*.
- Nearline e-mail: Optical "juke box" devices, backups of user e-mail folders.
- Archived or journaled e-mail: e.g., Autonomy Zantaz Enterprise Archive Solution, EMC EmailXtender, Mimosa NearPoint, Symantec Enterprise Vault.

**Accessible, but Often Overlooked:**

- E-mail residing on non-party servers: ISPs (IMAP, POP, HTTP servers), Gmail, Yahoo! Mail, Hotmail, etc.
- E-mail forwarded and cc'd to external systems: Employee forwards e-mail to self at personal e-mail account.
- E-mail threaded as text behind subsequent exchanges.
- Offline local e-mail stored on removable media: External hard drives, thumb drives and memory cards, optical media: CD-R/RW, DVD-R/RW, floppy drives, zip drives.
- Archived e-mail: Auto-archived or saved under user-selected filename.
- Common user "flubs": Users experimenting with export features unwittingly create e-mail archives.
- Legacy e-mail: Users migrate from e-mail clients "abandoning" former e-mail stores. Also, e-mail on mothballed or re-tasked machines and devices.

- E-mail saved to other formats: PDF, .tiff, .txt, .eml, .msg, etc.
- E-mail contained in review sets assembled for other litigation/compliance purposes.
- E-mail retained by vendors or third- parties (e.g., former service provider or attorneys)
- Paper print outs.

#### **Less Accessible:**

- Offline e-mail on server backup tapes and other media.
- E-mail in forensically accessible areas of local hard drives and re-tasked/reimaged legacy machines: deleted e-mail, internet cache, unallocated clusters.

The levels of accessibility above speak to practical challenges to ease of access, not to the burden or cost of review. The burden continuum isn't a straight line. That is, it may be less burdensome or costly to turn to a small number of less accessible sources holding relevant data than to broadly search and review the contents of many accessible sources. Ironically, it typically costs much more to process and review the contents of a mail server than to undertake forensic examination of a key player's computer; yet, the former is routinely termed "reasonably accessible" and the latter not.

The issues in the case, key players, relevant time periods, agreements between the parties, applicable statutes, decisions and orders of the court determine the extent to which locations must be examined; however, the failure to diligently identify relevant e-mail carries such peril that caution should be the watchword. Isn't it wiser to invest more effort to know exactly what the client has—even if it's not reasonably accessible and will not be searched or produced—than concede at the sanctions hearing the client failed to preserve and produce evidence it didn't know it because no one looked?

#### **Looking for E-Mail 101**

Because an e-mail is just a text file, individual e-mails could be stored as discrete text files. But that's not a very efficient or speedy way to manage a large number of messages, so you'll find that most e-mail client software doesn't do that. Instead, e-mail clients employ proprietary database files housing e-mail messages, and each of the major e-mail clients uses its own unique format for its database. Some programs encrypt the message stores. Some applications merely display e-mail housed on a remote server and do not store messages locally (or only in fragmentary way). The only way to know with certainty if e-mail is stored on a local hard drive is to look for it.

Merely checking the e-mail client's settings is insufficient because settings can be changed. Someone not storing server e-mail today might have been storing it a month ago. Additionally, users may create new identities on their systems, install different client software, migrate from other hardware or take various actions resulting in a cache of e-mail residing on their systems without their knowledge. *If they don't know it's there, they can't tell you it's not.* On local hard



drives, you've simply got to know what to look for and where to look...*and then you've got to look for it.*

For many, computer use is something of an unfolding adventure. One may have first dipped her toes in the online ocean using browser-based e-mail or an AOL account. Gaining computer-savvy, she may have signed up for broadband access or with a local ISP, downloading e-mail with Netscape Messenger or Microsoft Outlook Express. With growing sophistication, a job change or new technology at work, the user may have migrated to Microsoft Outlook or Lotus Notes as an e-mail client. Each of these steps can orphan a large cache of e-mail, possibly unbeknownst to the user but still fair game for discovery. Again, you've simply got to know what to look for and where to look.

One challenge you'll face when seeking stored e-mail is that every user's storage path is different. This difference is not so much the result of a user's ability to specify the place to store e-mail—which few do, but which can make an investigator's job more difficult when it occurs—but more from the fact that operating systems are designed to support multiple users and so must assign unique identities and set aside separate storage areas for different users. Even if only one person has used a Windows computer, the operating system will be structured at the time of installation so as to make way for others. Thus, finding e-mail stores will hinge on your knowledge of the User's Account Name or Globally Unique Identifier (GUID) string assigned by the operating system. This may be as simple as the user's name or as obscure as the 128-bit hexadecimal value {721A17DA-B7DD-4191-BA79-42CF68763786}. Customarily, it's both.

***Caveat:*** *Before you or anyone on your behalf "poke around" on a computer system seeking a file or folder, recognize that absent the skilled use of specialized tools and techniques, such activity will result in changing data on the drive. Some of the changed data may be forensically significant (such as file access dates) and could constitute spoliation of evidence. If, under the circumstances of the case or matter, your legal or ethical obligation is to preserve the integrity of electronic evidence, then you and your client may be obliged to entrust the search only to qualified persons*

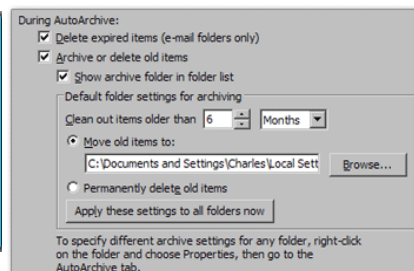
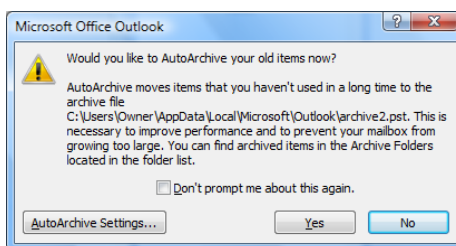
### **Finding Outlook E-Mail**

**PST:** Microsoft Outlook is by far the most widely used e-mail client in the business environment. Despite the confusing similarity of their names, Outlook is a much different and substantially more sophisticated application than Outlook Express (now called Windows Mail). One of many important differences is that where Outlook Express stores messages in plain text, Outlook encrypts and compresses messages. But the most significant challenge Outlook poses in discovery is the fact that all of its message data and folder structure, along with all other information managed by the program (except the user's Contact data), is stored within a single, often massive, database file with the file extension .pst. The Outlook PST file format is proprietary and its structure poorly documented, limiting your options when trying to view or

process its contents to Outlook itself or one of a handful of PST file reader programs available for purchase and download via the Internet.

**OST:** While awareness of the Outlook PST file has grown, even many lawyers steeped in e-discovery fail to consider a user's Outlook .ost file. The OST or offline synchronization file is commonly encountered on laptops configured for Exchange Server environments. It exists for the purpose of affording access to messages when the user has no active network connection. Designed to allow work to continue on, e.g., airplane flights, local OST files often hold messages purged from the server—at least until re-synchronization. It's not unusual for an OST file to hold e-mail unavailable from any other comparably-accessible source.

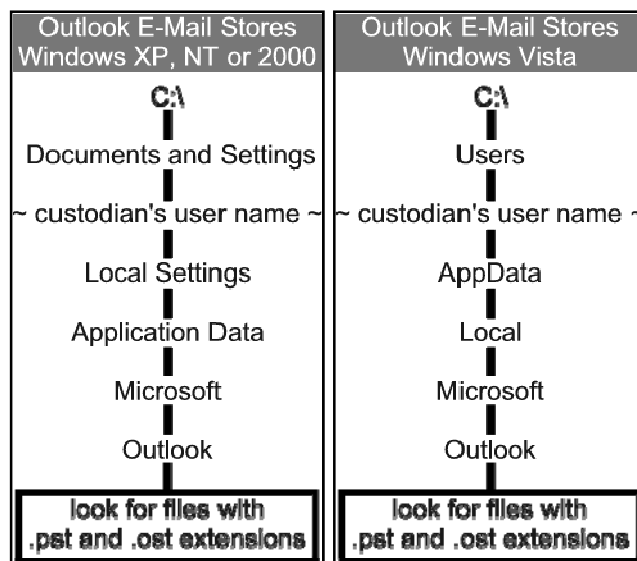
**Archive.pst:** Another file to consider is one customarily called, "archive.pst." As its name suggests, the archive.pst file holds older messages, either stored automatically or by user-initiated action. If you've used Outlook without manually



configuring its archive settings, chances are the system periodically asks whether you'd like to auto archive older items. Every other week (by default), Outlook 2003 seeks to auto archive any Outlook items older than six months (or for Deleted and Sent items older than two months for Outlook 2007). Users can customize these intervals, turn archiving off or instruct the application to permanently delete old items.

### Outlook Mail Stores Paths

To find the Outlook message stores on machines running Windows XP/NT/2000 or Vista, drill down from the root directory (C:\ for most users) according to the path diagram on the right for the applicable operating system. The default filename of Outlook.pst/ost may vary if a user has opted to select a different designation or maintains multiple e-mail stores; however, it's rare to see users depart from the default settings. Since the location of the PST and OST files can be changed by the user, it's a good idea to do a search of all files and folders to identify any files ending with the .pst and .ost extensions.

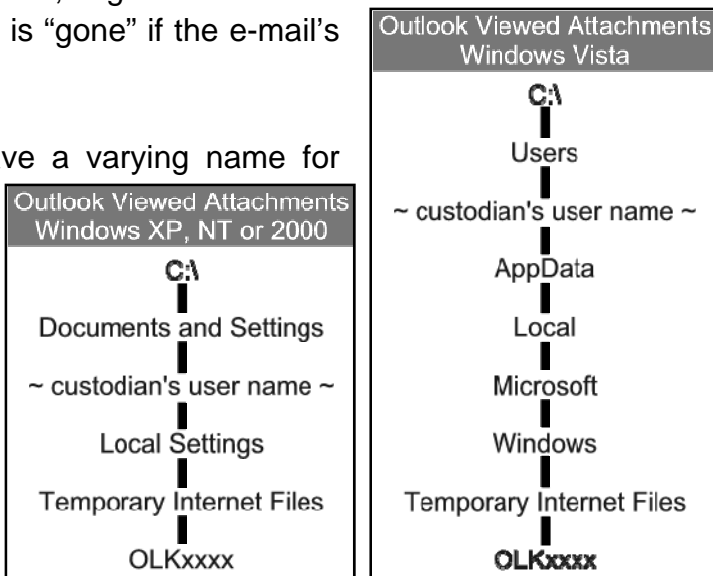


### “Temporary” OLK Folders

Note that by default, when a user opens an attachment to a message from within Outlook (as opposed to saving the attachment to disk and then opening it), Outlook stores a copy of the

attachment in a “temporary” folder. But don’t be misled by the word “temporary.” In fact, the folder isn’t going anywhere and its contents—sometimes voluminous--tend to long outlast the messages that transported the attachments. Thus, litigants should be cautious about representing that Outlook e-mail is “gone” if the e-mail’s attachments are not.

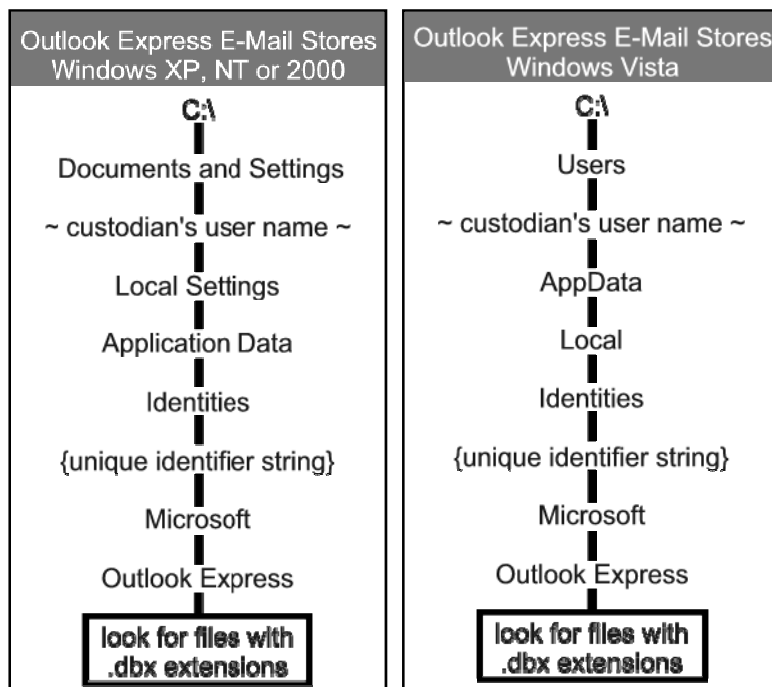
The Outlook viewed attachment folder will have a varying name for every user and on every machine, but it will always begin with the letters “OLK” followed by several randomly generated numbers and uppercase letters (e.g., OLK943B, OLK7AE, OLK167, etc.). To find the OLKxxxx viewed attachments folder on machines running Windows XP/NT/2000 or Vista, drill down from the root directory according to the path diagrams on the right for the applicable operating system.<sup>38</sup>



### Finding Outlook Express E-Mail

Outlook Express has been bundled with every Windows operating system for about fifteen years, so you are sure to find at least the framework of an e-mail cache created by the program. Beginning with the release of Microsoft Vista, the Outlook Express application was renamed Windows Mail and the method of message storage was changed from a database format to storage as individual messages. More recently, Microsoft has sought to replace both Outlook Express on Windows XP and Windows Mail on Windows Vista with a freeware application called Windows Live Mail.

Outlook Express places e-mail in database files with the extension .dbx.



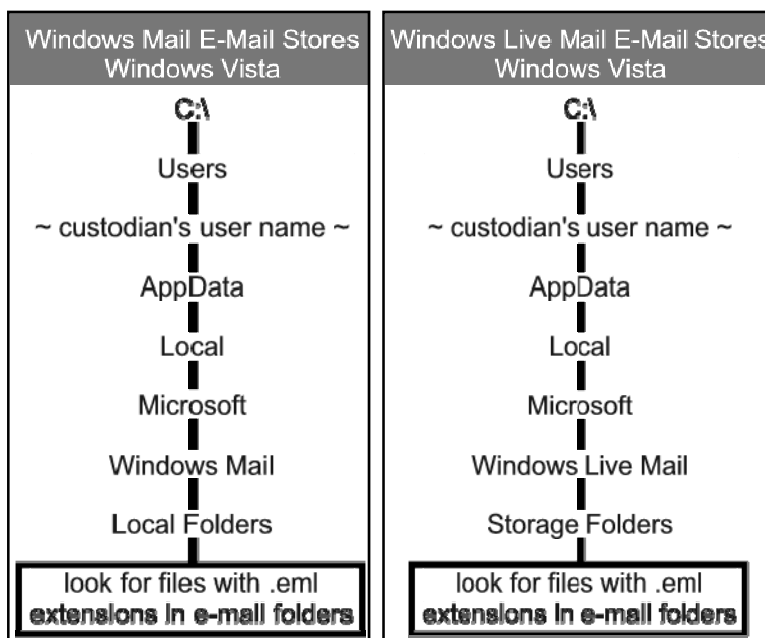
<sup>38</sup> By default, Windows hides system folders from users, so you may have to first make them visible. This is accomplished by starting Windows Explorer, then selecting 'Folder Options' from the Tools menu in Windows XP or 'Organize>Folder and Search Options' in Vista. Under the 'View' tab, scroll to 'Files and Folders' and check 'Show hidden files and folders' and uncheck 'Hide extensions for known file types' and 'Hide protected operating system files. Finally, click 'OK.'

The program creates a storage file for each e-mail storage folder that it displays, so expect to find at least Inbox.dbx, Outbox.dbx, Sent Items.dbx and Deleted Items.dbx. If the user has created other folders to hold e-mail, the contents of those folders will reside in a file with the structure *foldername.dbx*. Typically on a Windows XP/NT/2K system, you will find Outlook Express .dbx files in the path shown in the diagram at near right on the preceding page. Though less frequently encountered on a Windows Vista machine, the .dbx files would be found in the default location path shown at far right on preceding page. Multiple identifier strings (Globally Unique Identifiers) string listed in the Identities subfolder may be an indication of multiple e-mail stores and/or multiple users of the computer. You will need to check each Identity's path. Another approach is to use the Windows Search function (if under windows XP) to find all files ending .dbx, but be very careful to enable all three of the following Advanced Search options before running a search: Search System Folders, Search Hidden Files and Folders, and Search Subfolders. If you don't, you won't find any—or at least not all—Outlook Express e-mail stores. Be certain to check the paths of the files turned up by your search as it can be revealing to know whether those files turned up under a particular user identity, in Recent Files or even in the Recycle Bin.

### Finding Windows Mail and Windows Live Mail E-Mail Stores

You'll encounter Windows Mail on a machine running Windows Vista. By default, Windows Mail messages will be stored in oddly named individual files with the extension .eml and these housed in standard (*i.e.*, Inbox, Outbox, Sent Items, deleted Items, etc.) and user-created folders under the path diagrammed at near right.

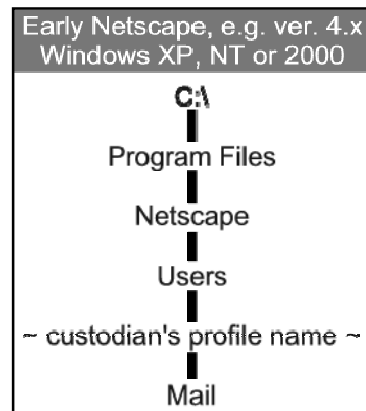
Similarly, Windows Live Mail running on Vista will store messages as oddly named individual files with the extension .eml, within standard and user-created folders under the path seen at far right.



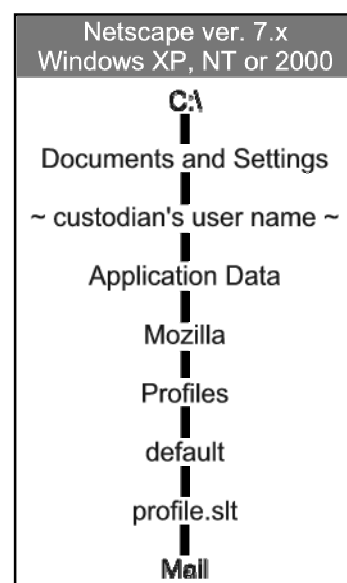
When collecting mail from these mail stores, it's important to capture both the message and the folder structure because, unlike the structured container seen in, *e.g.*, Outlook PST or OST files, the user's folder structure is not an integral part of the message storage scheme in Windows Mail or Live Mail.

## Finding Netscape E-Mail

Though infrequently seen today, Netscape and its Mozilla e-mail client ruled the Internet before the browser wars left it crippled and largely forgotten. If you come across a Netscape e-mail client installation, keep in mind that the location of its e-mail stores will vary depending upon the version of the program installed. If it is an older version of the program, such as Netscape 4.x and a default installation, you will find the e-mail stores by drilling down the path depicted at right. Expect to find two files for each mailbox folder, one containing the message text with no extension (e.g., Inbox) and another which serves as an index file with a .snm extension (e.g., Inbox.snm).



In the last version of Netscape to include an e-mail client (Netscape 7.x), both the location and the file structures/names were changed. Drill down using the default path shown at right and locate the folder for the e-mail account of interest, usually the name of the e-mail server from which messages are retrieved. If you don't see the Application Data folder, go to the Tools Menu, pull down to Folder Options, click on the View tab, and select "Show Hidden Files and Folders," then click "OK." You should find two files for each mailbox folder, one containing the message text with no extension (e.g., Sent) and another which serves as an index file with a .msf extension (e.g., Sent.msf). If you can't seem to find the e-mail stores, you can either launch a Windows search for files with the .snm and .msf extensions (e.g. \*.msf) or, if you have access to the e-mail client program, you can check its configuration settings to identify the path and name of the folder in which e-mail is stored.



## Microsoft Exchange Server

About 200 million people get their work e-mail via a Microsoft product called Exchange Server. It's been sold for about a dozen years and its latest version was introduced in 2007; although, most users continue to rely on the 2003 version of the product.

The key fact to understand about an e-mail server is that it's a *database* holding the messages (and calendars, contacts, to-do lists, journals and other datasets) of multiple users. E-mail servers are configured to maximize performance, stability and disaster recovery, with little consideration given to compliance and discovery obligations. If anyone anticipated the role e-mail would play in virtually every aspect of business today, their prescience never influenced the design of e-mail systems. E-mail evolved largely by accident, absent the characteristics of competent records management, and only lately are tools emerging that are designed to catch up to legal and compliance duties.

The other key thing to understand about enterprise e-mail systems is that, unless you administer the system, it probably doesn't work the way you imagine. The exception to that rule is if you can distinguish between Local Continuous Replication (LCR), Clustered Continuous Replication (CCR), Single Copy Cluster (SCC) and Standby Continuous Replication (SCR). In that event, I should be reading *your* paper!

But to underscore the potential for staggering complexity, appreciate that the latest Enterprise release of Exchange Server 2007 supports up to 50 storage groups per server of up to 50 message stores per group, for a database size limit of 16 terabytes. If there is an upper limit on how many users can share a single message store, I couldn't ascertain what it might be!

Though the preceding pages dealt with finding e-mail stores on local hard drives, in disputes involving medium- to large-sized enterprises, the e-mail server is likely to be the initial nexus of electronic discovery efforts. The server is a productive venue in electronic discovery for many reasons, among them:

- The periodic backup procedures which are a routine part of prudent server management tend to shield e-mail stores from those who, by error or guile, might delete or falsify data on local hard drives.
- The ability to recover deleted mail from archival server backups may obviate the need for costly and unpredictable forensic efforts to restore deleted messages.
- Data stored on a server is often less prone to tampering by virtue of the additional physical and system security measures typically dedicated to centralized computer facilities as well as the inability of the uninitiated to manipulate data in the more-complex server environment.
- The centralized nature of an e-mail server affords access to many users' e-mail and may lessen the need for access to workstations at multiple business locations or to laptops and home computers.
- Unlike e-mail client applications, which store e-mail in varying formats and folders, e-mail stored on a server can usually be located with relative ease and adhere to common file formats.
- The server is the crossroads of corporate electronic communications and the most effective chokepoint to grab the biggest "slice" of relevant information in the shortest time, for the least cost.

Of course, the big advantage of focusing discovery efforts on the mail server (*i.e.*, it affords access to thousands or millions of messages) is also its biggest disadvantage (someone has to *collect and review* thousands or millions of messages). Absent a carefully-crafted and, ideally, agreed-upon plan for discovery of server e-mail, both requesting and responding parties run the risk of runaway costs, missed data and wasted time.

E-mail originating on servers is generally going to fall into two realms, being online “live” data, which is deemed reasonably accessible, and offline “archival” data, routinely deemed inaccessible based on considerations of cost and burden.<sup>39</sup> Absent a change in procedure, “chunks” of data routinely migrate from accessible storage to less accessible realms—on a daily, weekly or monthly basis—as selected information on the server is replicated to backup media and deleted from the server’s hard drives.

### **The ABCs of Exchange**

Because it’s unlikely most readers will be personally responsible for collecting e-mail from an Exchange Server and mail server configurations can vary widely, the descriptions of system architecture here are offered only to convey a rudimentary understanding of common Exchange architecture.

The 2003 version of Exchange Server stores data in a Storage Group containing a Mailbox Store and a Public Folder Store, each composed of two files: an .edb file and a .stm file. Mailbox Store, Priv1.edb, is a rich-text database file containing user’s email messages, text attachments and headers. Priv1.stm is a streaming file holding SMTP messages and containing multimedia data formatted as MIME data. Public Folder Store, Pub1.edb, is a rich-text database file containing messages, text attachments and headers for files stored in the Public Folder tree. Pub1.stm is a streaming file holding SMTP messages and containing multimedia data formatted as MIME data. Exchange Server 2007 did away with STM files altogether, shifting their content into the EDB database files.

Storage Groups also contain system files and transaction logs. Transaction logs serve as a disaster recovery mechanism that helps restore an Exchange after a crash. Before data is written to an EDB file, it is first written to a transaction log. The data in the logs can thus be used to reconcile transactions after a crash.

By default, Exchange data files are located in the path **X:\Program files\Exchsrvr\MDBDATA**, where X: is the server’s volume root. But, it’s common for Exchange administrators to move the mail stores to other file paths.

### **Recovery Storage Groups and ExMerge**

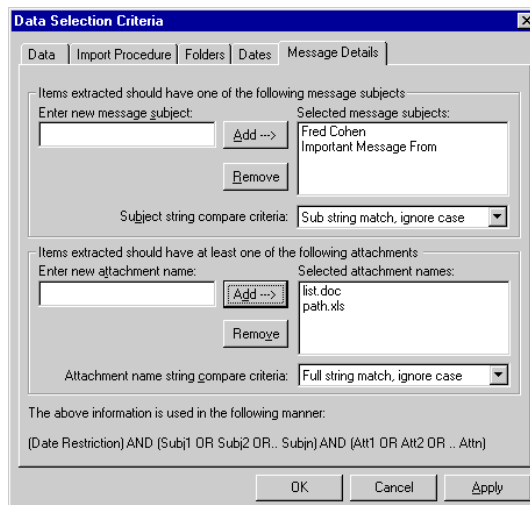
Two key things to understand about Microsoft Exchange are that, since 2003, an Exchange feature called **Recovery Storage Group** supports collection of e-mail from the server without

---

<sup>39</sup> Lawyers and judges intent on distilling the complexity of electronic discovery to rules of thumb are prone to pigeonhole particular ESI as “accessible” or “inaccessible” based on the media on which it resides. In fact, ESI’s storage medium is just one of several considerations that bear on the cost and burden to access, search and produce same. Increasingly, backup tapes are less troublesome to search and access while active data on servers or strewn across many “accessible” systems and devices is a growing challenge.



any need to interrupt its operation or restore data to a separate recovery computer. The second key thing is that Exchange includes a simple utility for exporting the server-stored e-mail of individual custodians to separate PST container files. This utility, officially the Exchange Server Mailbox Merge Wizard but universally called **ExMerge** allows for rudimentary filtering of messages for export, including (right) by message dates, folders, attachments and subject line content.



ExMerge also plays a crucial role in recovering e-mails “double deleted” by users if the Exchange server has been configured to support a “dumpster retention period.” When a user deletes an e-mail, it’s automatically relegated to a “dumpster” on the Exchange Server. The dumpster holds the message for 30 days by default or until a full backup of your Exchange database is run, whichever comes first. The retention interval can be customized for a longer or shorter interval.

### Journaling, Archiving and Transport Rules

Journaling is the practice of copying all e-mail to and from all users or particular users to one or more repositories inaccessible to most users. Journaling serves to preempt ultimate reliance on individual users for litigation preservation and regulatory compliance. Properly implemented, it should be entirely transparent to users and secured in a manner that eliminates the ability to alter the journaled collection.

Exchange Server supports three types of journaling: **Message-only journaling** which does not account for blind carbon copy recipients, recipients from transport forwarding rules, or recipients from distribution group expansions; **Bcc journaling**, which is identical to Message-only journaling except that it captures Bcc addressee data; and **Envelope Journaling** which captures all data about the message, including information about those who received it. Envelope journaling is the mechanism best suited to e-discovery preservation and regulatory compliance.

Journaling should be distinguished from **e-mail archiving**, which may implement only selective, rules-based retention and customarily entails removal of archived items from the server for offline or near-line storage, to minimize strain on IT resources and/or implement electronic records management. However, Exchange journaling also has the ability to implement rules-based storage, so each can conceivably be implemented to play the role of the other.

A related concept is the use of **Transport Rules** in Exchange, which serve, *inter alia*, to implement “Chinese Walls” between users or departments within an enterprise who are ethically or legally obligated not to share information, as well as to guard against dissemination of

confidential information. In simplest terms, software called *transport rules agents* “listen” to e-mail traffic, compare the content or distribution to a set of rules (conditions, exceptions and actions) and if particular characteristics are present, intercedes to block, route, flag or alter suspect communications.

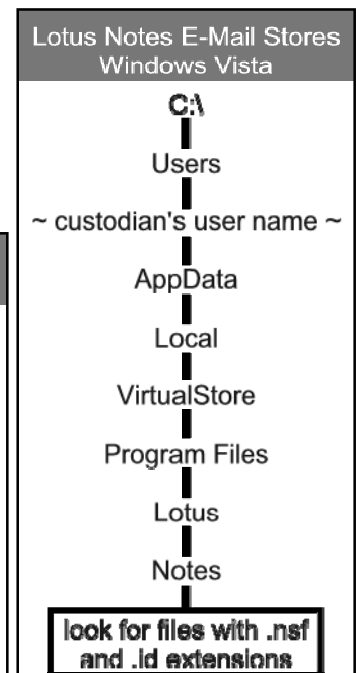
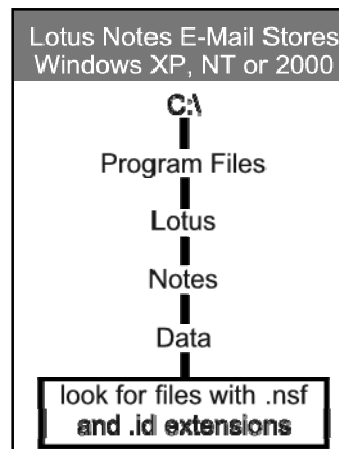
### Lotus Domino Server and Notes Client

Though Microsoft’s Exchange and Outlook e-mail products have a greater overall market share, IBM’s Lotus Domino and Notes products hold powerful sway within the world’s largest corporations, especially giant manufacturing concerns and multinationals. IBM boasts of 140 million Notes licenses sold to date worldwide.

Lotus Notes can be unhelpfully described as a “cross-platform, secure, distributed document-oriented database and messaging framework and rapid application development environment.” The main takeaway with Notes is that, unlike Microsoft Exchange, which is a purpose-built application designed for messaging and calendaring, Lotus Notes is more like a toolkit for *building* whatever capabilities you need to deal with documents—mail documents, calendaring documents and any other type of document used in business. Notes wasn’t *designed* for e-mail—e-mail just happened to be one of the things it was tasked to do.<sup>40</sup> Notes is database driven and distinguished by its replication and security.

Lotus Notes is all about copies. Notes content, stored in Notes Storage facility or **NSF** files, are constantly being replicated (synchronized) here and there across the network. This guards against data loss and enables data access when the network is unavailable, but it also means that there can be many versions of Notes data stashed in various places within an enterprise. Thus, discoverable Notes mail may not be gone, but lurks within a laptop that hasn’t connected to the network since the last business trip.

By default, local iterations of users’ NSF and ID files will be found on desktops and laptops in the paths shown in the diagrams at right. It’s imperative to collect the user’s .id file along with the .nsf message container or you may find yourself locked out of encrypted content. It’s also important to secure each custodian’s Note’s password. It’s common for Notes to be installed in ways other than the default configuration, so search by



<sup>40</sup> Self-anointed “Technical Evangelist,” Jeff Atwood describes Lotus Notes this way: “It is death by a thousand tiny annoyances -- the digital equivalent of being kicked in the groin upon arrival at work every day.” <http://blogs.vertigosoftware.com/jatwood/archive/2005/08/11/1366.aspx>. In fairness, Lotus Notes has been extensively overhauled since he made that observation.

extension to insure that .nsf and .id files are not also found elsewhere. Also, check the files' last modified date to assess whether the date is consistent with expected last usage. If there is a notable disparity, look carefully for alternate file paths housing later replications.

Local replications play a significant role in e-discovery of Lotus Notes mail because, built on a database and geared to synchronization of data stores, deletion of an e-mail within Lotus "broadcasts" the deletion of the same message system wide. Thus, it's less common to find undeleted iterations of messages in a Lotus environment unless you resort to backup media or find a local iteration that hasn't been synchronized after deletion.

### **Novell GroupWise**

Experienced lawyers—that sound better than "older"—probably remember GroupWise. It originated as a WordPerfect virtual desktop product for messaging and calendaring called "WordPerfect Library," then became "WordPerfect Office." It changed to GroupWise when WordPerfect was acquired in 1993 by another deposed tech titan, Novell. GroupWise is alive (some might say "alive and well") in a handful of niche sectors, particularly government; but GroupWise's market share been so utterly eclipsed by its rivals as to make it seem almost extinct.

GroupWise is another tool thought of as "just an e-mail application" when it's really a Swiss army knife of data management features that happens to do e-mail, too. Because it's not a standalone e-mail server and client and because few vendors and experts have much recent experience with GroupWise, it's presents greater challenges and costs in e-discovery.

GroupWise is built on a family of databases which collectively present data comprising messages to users. That's an important distinction. Messages are less like discrete communications than reports *about* the communication, queried from a series of databases and presented *in the form of* an e-mail. User information is pulled from one database (ofuser), message content emerges from a second (ofmsg) and attachments are managed by a third database (offiles). When a user sends a GroupWise e-mail, the message is created in the user's message database and pointers to that message go to the user's Sent Items folder and the Recipients' Inboxes. Attachments go to the offiles database and pointers to attachments go out. Naturally, a more traditional method must be employed when message are sent beyond the GroupWise environment.

The prevailing practice in dealing with GroupWise e-mail is to convert messages to Outlook PST formats. The sole rationale for this seems to be that most e-discovery service providers are equipped to deal with PSTs and not native GroupWise data. Thus, the decision is driven by ignorance not evidence. Accordingly, a cottage industry has emerged dedicated to converting GroupWise ESI to other formats, but a few vendors tout their ability to work natively with GroupWise data. As often as not, conversion is a costly but harmless hurdle; but recognize that some data won't survive the leap between formats and, in choosing whether to deal with

GroupWise data by conversion, you must assess whether the data sacrificed to the conversion process may be relevant and material.

## Webmail

An estimated 1.2 billion people use webmail worldwide.<sup>41</sup> Ferris Research puts the number of business e-mail users in 2007 at around 780 million, accounting for some 6 *trillion* non-spam e-mails in sent in 2006. In April 2008, *USA Today*<sup>42</sup> reported the leading webmail providers' market share as:

Microsoft webmail properties:	256.2 million users
Yahoo:	254.6 million users
Google:	91.6 million users
AOL webmail properties:	48.9 million users

Any way you slice it, webmail can't be ignored in e-discovery. Webmail holding discoverable ESI presents legal, technical and practical challenges, but the literature is nearly silent about how to address them.

The first hurdle posed by webmail is the fact that it's stored "in the cloud" and off the company grid. Short of a subpoena or court order, the only legitimate way to access and search employee web mail is with the employee's cooperation, and that's not always forthcoming. Courts nonetheless expect employers to exercise control over employees and insure that relevant, non-privileged webmail isn't lost or forgotten.

One way to assess the potential relevance of webmail is to search server e-mail for webmail traffic. If a custodian's Exchange e-mail reveals that it was the custodian's practice to e-mail business documents to or from personal webmail accounts, the webmail accounts may need to be addressed in legal hold directives and vetted for responsive material.

A second hurdle stems from the difficulty in collecting responsive webmail. How do you integrate webmail content into your review and production system? Where a few pages might be "printed" to searchable Adobe Acrobat PDF formats or paper, larger volumes require a means to dovetail online content and local collections. The most common approach is to employ a POP3 client application to download messages from the webmail account. All of the leading webmail providers support POP3 transfer, and with the user's cooperation, it's simple to configure a clean installation of any of the client applications already discussed to capture online message stores. Before proceeding, the process should be tested against accounts that don't evidence to determine what metadata values may be changed, lost or introduced by POP3 collection.

---

<sup>41</sup> October 2007 report by technology market research firm The Radicati Group, expected to rise to 1.6 billion by 2011.

<sup>42</sup> [http://www.usatoday.com/tech/products/2008-04-15-google-gmail-webmail\\_N.htm](http://www.usatoday.com/tech/products/2008-04-15-google-gmail-webmail_N.htm)

Keep in mind that webmail content can be fragile compared to server content. Users rarely employ a mechanism to back up webmail messages (other than the POP3 retrieval just discussed) and webmail accounts may purge content automatically after periods of inactivity or when storage limits are exceeded. Further, users tend to delete embarrassing or incriminating content more aggressively on webmail, perhaps because they regard webmail content as personal property or the evanescent nature of account emboldens them to believe spoliation will be harder to detect and prove.

## Computer Forensics

Virtually any information that traverses a personal computer or other device has the potential to leave behind content that can be recovered in an examination of the machine or device by a skilled computer forensic examiner. Even container files like Outlook PST or OST files have a propensity to hold a considerable volume of recoverable information long after the user believes such data has been deleted.

Though the scope and methodology of a thorough computer forensic examination for hidden or deleted e-mail is beyond the scope of this paper,<sup>43</sup> readers should be mindful that a computer's operating system or **OS** (e.g., Windows or Vista, Mac or Linux) and installed software (**applications**) generate and store much more information than users realize. Some of this unseen information is **active data** readily accessible to users, but requiring skilled interpretation to be of value in illuminating human behavior. Examples include the data *about* data or **metadata** tracked by the OS and applications, but not displayed onscreen. For example, Microsoft Outlook records the date a Contact is created, but few of us customize the program to display that "date created" information.

Other active data reside in obscure locations or in coded formats less readily accessible to users, but enlightening when interpreted and correlated. Log files, hidden system files and information recorded in non-text formats are examples of **encoded data** that may reveal information about user behavior. As discussed, e-mail attachments and the contents of OST, PST and NSF files are all encoded data.

Finally, there are vast regions of hard drives and other data storage devices that hold **forensic data** even the operating systems and applications can't access. These "data landfills," called **unallocated clusters** and **slack space**, contain much of what a user, application or OS discards over the life of a machine. Accessing and making sense of these vast, unstructured troves demands specialized tools, techniques and skill.

---

<sup>43</sup> For further reading on computer forensics, see Ball, *Five on Forensics*, <http://www.craigball.com/cf.pdf> and Ball, *What Judges Should Know About Computer Forensics*, published by the Federal Judicial Center and available at [http://www.craigball.com/What\\_Judges\\_Computer\\_Forensics-200807.pdf](http://www.craigball.com/What_Judges_Computer_Forensics-200807.pdf)

**Computer forensics** is the expert acquisition, interpretation and presentation of the data within these three categories (**Active**, **Encoded** and **Forensic** data), along with its juxtaposition against other available information (e.g., e-mail, phone records and voice mail, credit card transactions, keycard access data, documents and instant message communications).

Most cases require no forensic-level computer examination, so courts and litigants should closely probe whether a request for access to an opponent's machines to recover e-mail is grounded on a genuine need or is simply a fishing expedition. Except in cases involving, e.g., data theft, forgery or spoliation, computer forensics will usually be an effort of last resort for identification and production of e-mail.

The Internet has so broken down barriers between business and personal communications that workplace computers are routinely peppered with personal, privileged and confidential communications, even intimate and sexual content, and home computers normally contain some business content. Further, a hard drive is more like one's office than a file drawer. It may hold data about the full range of a user's daily activity, including private or confidential information about others.

Accordingly, computer forensic examination should be governed by an agreed or court-ordered protocol to protect unwarranted disclosure of privileged and confidential information. Increasingly, courts appoint neutral forensic examiners to serve as Rule 53 Special Masters for the purpose of performing the forensic examination *in camera*. To address privilege concerns, the information developed by the neutral is first tendered to counsel for the party proffering the machines for examination, which party generates a privilege log and produces non-privileged, responsive data.<sup>44</sup>

Whether an expert or court-appointed neutral conducts the examination, the order or agreed protocol granting forensic examination of ESI should provide for handling of confidential and privileged data and narrow the scope of examination by targeting specific objectives. The examiner needs clear direction in terms of relevant keywords and documents, as well as pertinent events, topics, persons and time intervals. A common mistake is for parties to agree upon a search protocol or secure an agreed order without consulting an expert to determine feasibility, complexity or cost.

There is no more a "standard" protocol for forensic examination than there is a "standard" set of deposition questions. In either case, a good examiner tailors the inquiry to the case, follows the evidence as it develops and remains flexible enough to adapt to unanticipated discoveries. Consequently, it is desirable for a court-ordered or agreed protocol to afford the examiner discretion to adapt to the evidence and apply their expertise.

---

<sup>44</sup> For further discussion of forensic examination protocols, see Ball in Your Court, *Problematic Protocols*, November 2008, Law Technology News; [http://www.lawtechnews.com/r5/showkiosk.asp?listing\\_id=2756144&pub\\_id=5173&category\\_id=27902](http://www.lawtechnews.com/r5/showkiosk.asp?listing_id=2756144&pub_id=5173&category_id=27902)

## **Why Deleted Doesn't Mean Gone**

A computer manages its hard drive in much the same way that a librarian manages a library. The files are the “books” and their location is tracked by an index. But there are two key differentiators between libraries and computer file systems. Computers employ no Dewey decimal system, so electronic “books” can be on any shelf. Further, electronic “books” may be split into chapters, and those chapters stored in multiple locations across the drive. This is called “**fragmentation.**” Historically, libraries tracked books by noting their locations on index card in a card catalog. Computers similarly employ directories (often called “**file tables**”) to track files and fragmented portions of files.

When a user hits “Delete,” nothing happens to the actual file targeted for deletion. Instead, a change is made to the file table that keeps track of the file’s location. Thus, akin to tearing up a card in the card catalogue, the file, like its literary counterpart, is still on the “shelf,” but now...without a locator in the file table...our file is a needle in a haystack, lost among millions of other unallocated clusters.

To recover the deleted file, a computer forensic examiner employs three principal techniques:

### **File Carving by Binary Signature**

Because most files begin with a unique digital signature identifying the file type, examiners run software that scans each of the millions of unallocated clusters for particular signatures, hoping to find matches. If a matching file signature is found and the original size of the deleted file can be ascertained, the software copies or “carves” out the deleted file. If the size of the deleted file is unknown, the examiner designates how much data to carve out. The carved data is then assigned a new name and the process continues.

Unfortunately, deleted files may be stored in pieces as discussed above, so simply carving out contiguous blocks of fragmented data grabs intervening data having no connection to the deleted file and fails to collect segments for which the directory pointers have been lost. Likewise, when the size of the deleted file isn’t known, the size designated for carving may prove too small or large, leaving portions of the original file behind or grabbing unrelated data. Incomplete files and those commingled with unrelated data are generally corrupt and non-functional. Their evidentiary value is also compromised.

File signature carving is frustrated when the first few bytes of a deleted file are overwritten by new data. Much of the deleted file may survive, but the data indicating what type of file it was, and thus enabling its recovery, is gone.



File signature carving requires that each unallocated cluster be searched for each of the file types sought to be recovered. When the parties or a court direct that an examiner “recover all deleted files,” that’s an exercise that could take weeks, followed by countless hours spent culling corrupted files. Instead, the protocol should, as feasible, specify the *particular* file types of interest (i.e., e-mail and attachments) based upon how the machine’s was used and the facts and issues in the case.

### **File Carving by Remnant Directory Data**

In some file systems, residual file directory information revealing the location of deleted files may be strewn across the drive. Forensic software scans the unallocated clusters in search of these lost directories and uses this data to restore deleted files.

### **Search by Keyword**

Where it’s known that a deleted file contained certain words or phrases, the remnant data may be found using keyword searching of the unallocated clusters and slack space. Keyword search is a laborious and notoriously inaccurate way to find deleted files, but its use is necessitated in most cases by the enormous volume of ESI. When keywords are not unique or less than about 6 letters long, many false positives (“**noise hits**”) are encountered. Examiners must painstakingly look at each hit to assess relevance and then manually carve out responsive data. This process can take days or weeks for a single machine.

Keyword searching for e-mail generally involves looking for strings invariably associated with messages (e.g., e-mail addresses) or words or phrases known or expected to be seen in deleted messages (e.g., subject lines, signatures or header data).

Because e-mail is commonly encoded, encrypted and/or compressed, and because it customarily resides in container files structured more like databases than discrete messages, computer forensic analysis for e-mail recovery is particularly challenging. On the other hand, e-mail tends to lodge in so many places and formats; it’s the rare case where at least some responsive e-mail cannot be found.

As relevant, a forensic protocol geared to e-mail should include a thorough search for orphaned message collections, looking for any of the varied formats in which e-mail is stored (e.g., PST, OST, NSF, MSG, EML, MHT, DBX, IDX) and of unallocated clusters for binary signatures of deleted container files. Container files themselves should be subjected to processes that allow for recovery of double deleted messages that remain lodged within uncompact containers.<sup>45</sup>

---

<sup>45</sup> A common technique used on PST containers is to corrupt the file header on a copy of the container file and use Microsoft’s free Scanpst utility to repair it. This process sometimes recovers double deleted messages as these remain in the container until periodically compacted by Outlook. Scanpst can also be run against chunks of the unallocated clusters to ferret out deleted PSTs.

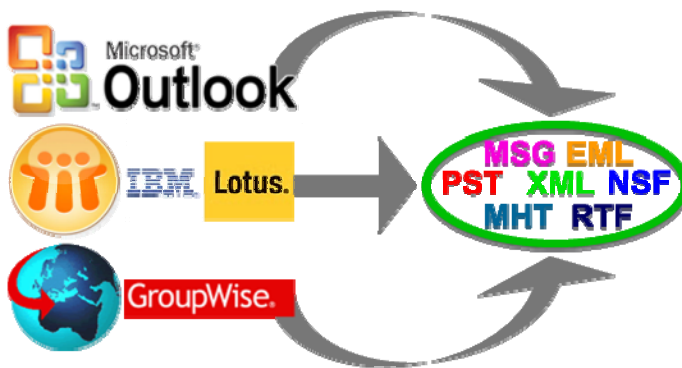
Webmail can often be found in the Internet cache (Temporary Internet Files) as well as within unallocated clusters and swap files. Desktop search and indexing programs (like Google Desktop) may also hold the full text of deleted e-mail. Moreover, devices like smart phones and PDAs employ synchronization files to store and transfer e-mail. Finally, e-mail clients like Outlook can themselves hold messages (e.g., corrupted drafts and failed transmissions) along with metadata unseen by users.

### Forms of Production

As discussed above, what users see presented onscreen as e-mail is a selective presentation of information from the header, body and attachments of the source message, determined by the capabilities and configuration of their e-mail client and engrafted with metadata supplied by that client. Meeting the obligation to produce comparable data of similar utility to the other side in discovery is no mean feat, and one that hinges on choosing suitable forms of production.

Requesting parties often demand “native production” of e-mail; but, electronic mail is rarely produced natively in the sense of supplying a duplicate of the source container file. That is, few litigants produce the entire Exchange database EDB file to the other side. Even those that produce mail in the format employed natively by the application (e.g., as a PST file) aren’t likely to produce the source file but will fashion a reconstituted PST file composed of selected messages deemed responsive and non-privileged.

As applied to e-mail, “native production” instead signifies production in a form or forms that most closely approximate the contents and usability of the source. Often, this will be an form of production identical to the original (e.g., PST or NSF) or a form (like MSG or EML) that shares many of the characteristics of the source and can deliver comparable usability when paired with additional information (e.g., information about folder structures).<sup>46</sup>



Similarly, producing parties employ imaged production and supply TIFF image files of messages, but in order to approximate the usability of the source must also create and produce accompanying load files carrying the metadata and full text of the source message keyed to its images. Collectively, the load files and image data permit recipients with compatible software (e.g., Summation, Concordance) to view and search the messages. Selection of Adobe PDF

<sup>46</sup> When e-mail is produced as individual messages, the folder structure may be lost and with it, important context. Additionally, different container formats support different complements of metadata applicable to the message. For example, a PST container may carry information about whether a message was opened, flagged or linked to a calendar entry.

documents as the form of production allows producing parties to dispense with the load files because much of the same data can be embedded in the PDF. PDF also has the added benefit of not requiring the purchase of review software.

Some producing parties favor imaged production formats in a mistaken belief that they are more secure than native production and out of a desire to emboss Bates numbers or other text (i.e., protective order language) to the face of each image. Imaged productions are more expensive than native or quasi-native productions, but, as they hew closest to the document review mechanisms long employed by law firms, they require little adaptation. It remains to be seen if clients will continue to absorb higher costs solely to insulate their counsel from embracing more modern and efficient tools and techniques.

Other possible format choices include XML<sup>47</sup> and MHT,<sup>48</sup> as well as Rich Text Format (RTF)—essentially plain text with improved formatting—and, for small collections, paper printouts.

There is no single, “perfect” form of production for e-mail, though the “best” format to use is the one on which the parties agree. Note also that there’s likely not a single production format that lends itself to *all* forms of ESI. Instead, *hybrid productions* match the form of production to the characteristics of the data being produced. In a hybrid production, images are used where they are most utile or cost-effective and native formats are employed when they offer the best fit or value.

As a rule of thumb to maximize usability of data, hew closest to the format of the source data (i.e., PST for Outlook mail and NSF for Lotus Notes), but keep in mind that whatever form is chosen should be one that the requesting party has the tools and expertise to use.

Though there is no ideal form of production, we can be guided by certain ideals in selecting the forms to employ. Absent agreement between the parties or an order of the Court, the forms of production employed for electronic mail should be either the mail’s native format or a form that will:

1. Enable the complete and faithful reproduction of all information available to the sender and recipients of the message, including layout, bulleting, tabular formats, colors, italics, bolding, underlining, hyperlinks, highlighting, embedded images and other non-textual ways we communicate and accentuate information in e-mail messages.
2. Support accurate electronic searchability of the message text and header data;

---

<sup>47</sup> XML is eXtensible Markup Language, an unfamiliar name for a familiar technology. Markup languages are coded identifiers paired with text and other information. They can define the appearance of content, like the Reveal Codes screen of Corel Inc.’s WordPerfect documents. They also serve to tag content to distinguish whether 09011957 is a birth date (09/01/1957), a phone number (0-901-1957) or a Bates number. Plus, markup languages allow machines to talk to each other in ways humans understand. For further information about the prospects for XML in e-discovery, see Ball in Your Court, *Trying to Love XML*, March 2008, Law Technology News; [http://www.lawtechnews.com/r5/showkiosk.asp?listing\\_id=1929884](http://www.lawtechnews.com/r5/showkiosk.asp?listing_id=1929884)

<sup>48</sup> MHT is a shorthand reference for MHTML or MIME Hypertext markup Language. HTML is the markup language used to create web pages and rich text e-mails. MHT formats mix HTML and encoded MIME data(see prior discussion of MIME at page to represent the header, message body and attachments of an e-mail.

3. Maintain the integrity of the header data (To, From, Cc, Bcc, Subject and Date/Time) as discrete fields to support sorting and searching by these data;
4. Preserve family relationships between messages and attachments;
5. Convey the folder structure/path of the source message;
6. Include message metadata responsive to the requester's legitimate needs;
7. Facilitate redaction of privileged and confidential content and, as feasible, identification and sequencing akin to Bates numbering; and
8. Enable reliable date and time normalization across the messages produced.<sup>49</sup>

## Conclusion

By now, you're wishing you'd taken my advice on page one and not begun. It's too late. You know too much about e-mail to ever again trot out the "I dunno" defense.

As I look back over the preceding discussion of the nerdy things that lawyers need to know about e-mail, I'm struck by how much *more* there is to cover. We've barely touched on e-mail backup systems, review platforms, visual analytics, e-mail archival, cloud computing, search and sampling, message conversion tools, unified messaging and a host of other exciting topics.

I hope you've gleaned something useful from this paper. I invite and appreciate your suggestions for corrections and improvements. Please e-mail them to [craig@ball.net](mailto:craig@ball.net).

---

<sup>49</sup> E-mails carry multiple time values depending upon, e.g., whether the message was obtained from the sender or recipient. Moreover, the times seen in an e-mail may be offset according to the time zone settings of the originating or receiving machine as well as for daylight savings time. When e-mail is produced as TIFF images or as text embedded in threads, these offsets may produce hopelessly confusing sequences. For further discussion of date/time normalization to UTC, see Ball in Your Court, *SNAFU*, September 2008, Law Technology News; [http://www.lawtechnews.com/r5/showkiosk.asp?listing\\_id=2217760](http://www.lawtechnews.com/r5/showkiosk.asp?listing_id=2217760)

# Preservation of ESI after Layoffs

Craig Ball



## Preservation of ESI after Layoffs

Craig Ball

© 2009

In a March 27, 2009 article in The National Law Journal called, "Protecting Corporate Data in Economic Downturn" authors Regina A. Jytyla and R. Jason Straight pose the question, "As work force cutbacks become commonplace, many organizations face the daunting task of locating, securing and imaging hard drives left behind by departing employees. For example, what should a corporation do with 30, 100 or even 1,000 idle computer terminals?" It's a great question. Unfortunately, the article ventured no answer.

Indeed, there's no pat response; but scant resources aren't a free pass to spoliation. Reductions-in-force are Alzheimer's to institutional memory. Suddenly, the people who know where responsive ESI lives and the ones caching stuff for litigation hold are gone. All that remains are their machines, file shares and a drawer full of little ketchup packages.

In-house and outside counsel must know how to react *molto pronto*. So, I write to put forward one low cost approach tailored to companies in crisis.

In performing preservation triage on dozens of idle machines, first undertake some mundane tasks, then make a couple of threshold decisions:

1. Label each machine and external hard drive or other storage media with the name of its former user, title, department and physical location and include relevant dates (e.g., terminated 3/31/09). If it's not a security issue, consider adding the custodian's username on that machine, along with his or her log in password and company e-mail address. Right now it's "Susan's machine," but soon it'll be just a chassis amidst a hundred others that look just like it. Be certain the labels are firmly affixed and applied in a consistent way so you can see them when the machines are stacked.

Print two more copies of the label: one for the hard drive (see below) and the other to stick on the log sheet that will serve as a preliminary inventory record and ultimately input for your discovery database. Better still, generate the label data from a database holding the same information. Sure, you can use bar codes or RFID tags, if you want to go high tech. For my money, a prosaic paper label gets the job done for the lowest cost.

2. If a former user was subject to a litigation hold, create and prominently affix a warning label to that effect. You need to insure these machines don't get wiped, re-tasked or auctioned off until their contents have been harvested in all pending and anticipated matters. After that, you can put on a new label that says "okay to wipe" with someone's signature right on it for the sake of accountability. I keep a Brother P-Touch label printer attached to my machine because **mistakes are costly, and labels are cheap.**



3. While you're printing labels, affix some identical to those described above on the hard drives *inside* the machine. I like to include the serial number or service tag for the chassis on these labels. This small effort can save you big headaches down the line.
4. Secure the machines and external media. Even before a departing employee is out the door, co-workers start circling their office stuff like vultures over carrion. In no time, that stuff grows legs and walks off. Lock the machines up where they aren't likely to get wet, hot, frozen, stolen, borrowed, played with or cannibalized for parts.
5. Cell phones and other handheld devices pose unique challenges. Years ago, I was part of a group inspecting the Chicago ESI preservation facility used by Arthur Andersen in the Enron litigation. With an overweening air of "we know exactly what we're doing," Andersen's legal team pointed to shelves laden with carefully packaged desktops, laptops, hard drives and handhelds. I was duly impressed, but since handhelds of that era lost their data soon after batteries died, I asked what provision had been made to supply *power* to the trussed-up PDAs and phones during their weeks *en plastique*. The wild-eyed looks exchanged on the other side were answer enough. No deer ever faced headlights with greater dismay. A charging protocol was quickly implemented.

Most handhelds spread data across three storage areas: the device's (sometimes volatile) memory, a removable media card and an online repository (e.g., a mobile service provider or a Blackberry Enterprise Server). If the device is synched with a computer, you have a fourth storage area to consider. Moreover, wireless devices keep interacting with the world--searching for signals, talking to towers and storing new data--if you don't intercede.

Labels are your friend here, too. The user's name, title, department, messaging ID, account number, password and relevant dates. Consider adding a piece of tape to secure removable media too small to label. It's not rocket science, but it works.

6. Now, decide if any of the machines or devices are candidates for computer forensic preservation and examination. Unless the FBI is at the door or the New York Times has lately used the company's name in the same sentence with "Ponzi scheme," the need for wholesale forensic imaging is remote. However, firings often lead inexorably to litigation and, from fear or spite, departing employees engage in delete-o-thons before they go.

On a shoestring budget, bring in forensics experts only if you have a reasonable basis to anticipate the need for forensic preservation and examination. For a list of situations where you *should* see the need, check out the *Ball in Your Court* column in the April 2009 issue of Law Technology News.



7. Sitting on dozens or hundreds of idle machines is a waste of resources, but so is collecting contents in anticipation of litigation that may never arise. Bankruptcy trustees, government overseers, even once-somnolent directors may insist that idle assets be put to work or converted to cash.

***How do you re-task or sell machines while meeting preservation duties?***

This is where it's helpful to consider computers separately from hard drives and ESI. Most people grasp the distinction between a DVD player, a DVD and a movie stored on DVD. A computer is like a DVD player, the hard drive is the DVD and the ESI on the drive is the movie: ***viewer, container and content***. With one notable exception discussed below, it's possible to swap hard drives between machines and move ESI to new media.

***Strategy: Pull the drives, then re-task or sell the CPUs.***

"But wait, "you protest, "Doesn't much of the value of a computer flow from its operating system and all the software on the system?" Sure, but the software site licenses that companies buy from Microsoft and other software publishers don't allow them to sell the operating system or the software along with the hardware. Possessing an installed copy of a program is not the same as having a license to use it. Moreover, drives bound for the auction block or charity must be cleansed of confidential company data--something that's hard to achieve without devoting hours to wiping everything. Even re-tasking a system within a company involves some reconfiguration and/or re-installation of software.

The bottom line is that new drives are cheap, and *when an employee leaves the company and preservation is required*, trying to copy their drive or keep it in service ends up costing more than simply pulling, labeling and sequestering the drive. It's certainly a lot faster than duplication.

8. What kind of lawyer would I be if I didn't add, "but it depends?" The bonds between hard drive and machine are pretty weak to begin with (e.g., configuring the right drivers for the hardware) and of little consequence in e-discovery. In EDD, we rarely collect the system and application files that boot and run the operating system and programs. In fact, we de-NIST data sets to get rid of that stuff.

But machine and drive are hopelessly conjoined in computers implementing the features of a **Trusted Platform Module** or **TPM**. Many newer laptops and some enterprise desktops sold today incorporate a TPM, though very few civilian users activate the full disk encryption and other features that prevent accessing data on a hard drive absent the TPM module used to encrypt the data. How few? Well, in the last 100 business laptops I've forensically imaged, not a single machine implemented TPM features.

If the company deployed encrypted laptops keyed to the TPM, you're probably stuck keeping the whole machine **and its password or USB activation key** or collecting potentially responsive ESI from the drive while the user is logged in.

9. When you pull hard drives for preservation, don't forget those drive labels discussed above. Be sure the machine is powered off, and then pull the plug to be sure. It's a low voltage device, but why tempt fate?

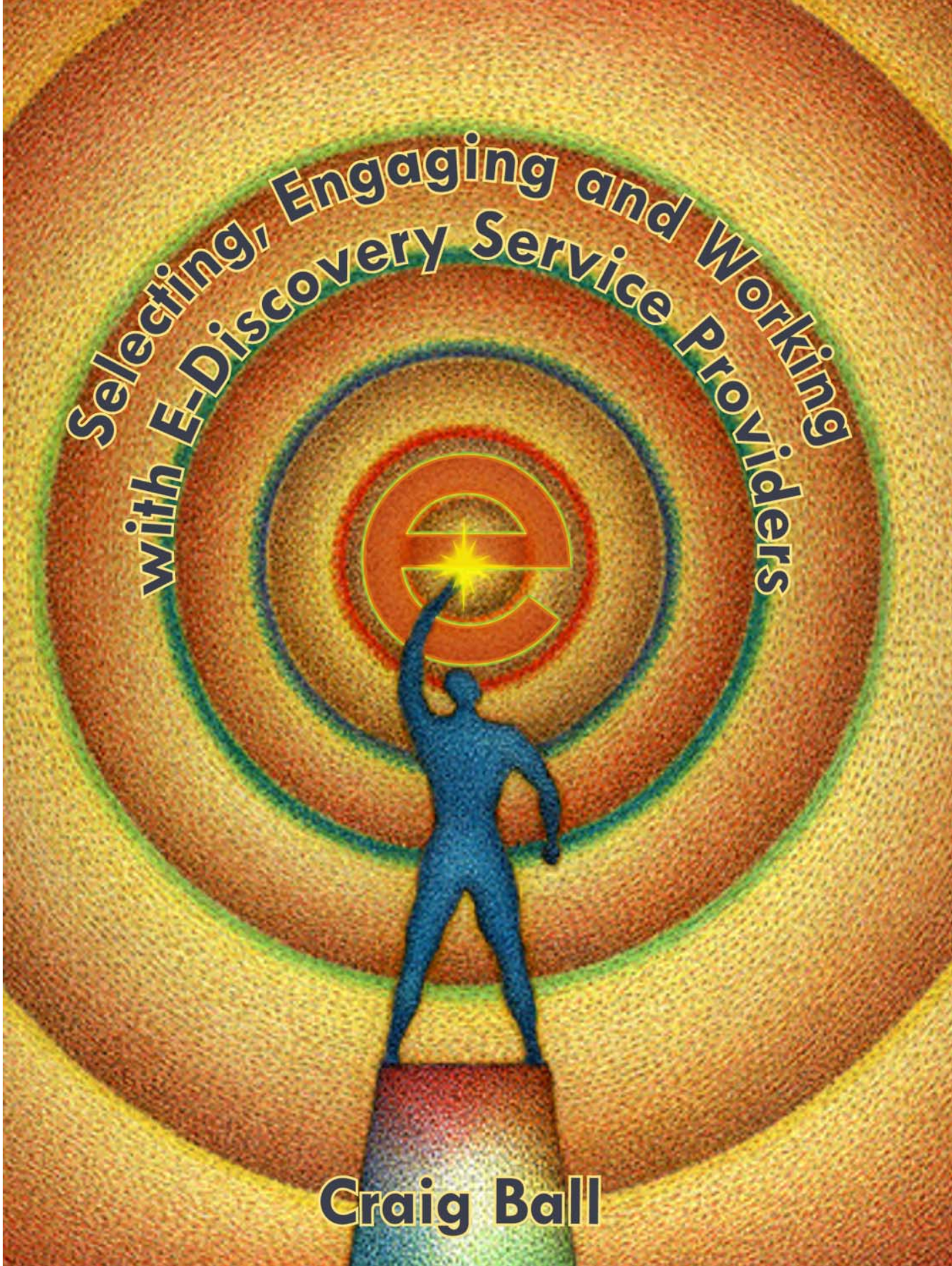
Some drives can come out in seconds without any tools. Others will require a Phillips screwdriver and some finesse. Don't cut yourself on sharp chassis edges. This downturn will pass, and your family needs you more than the employee death benefit.

The mortal enemies of a hard drive are electrostatic discharge and rough handling. Touch a grounded surface (cold water pipe, grounded system chassis or freshly unearthed vampire) right before handling the drive. Handle the drive with care; which is to say hold it by its sides and don't drop it or squeeze hard on the drive enclosure. Affix the label somewhere that it won't just peel off, and try not to cover the parts of the drive's factory label detailing the size and serial number of the drive.

Package drives to keep them clean and dry. Those specially-made, pink, antistatic bubble wrap bags are ideal; but a little low-static padding and a quart-size Ziploc freezer bag from the grocery store will do in a pinch. If the drives are individually well packaged, I find ordinary vinyl storage bins work well to store and protect as many as 20-30 drives in a secure room or closet.

10. Remember: **The trick is not losing track of what you have.** Good labeling, careful inventories and making sure that multiple people--particularly the legal team--know what's where is key to this strategy.





**Selecting, Engaging and Working  
with E-Discovery Service Providers**

**Craig Ball**



## Selecting, Engaging and Working with E-Discovery Service Providers

By Craig Ball<sup>1</sup>

The transmittal letter in the overnight package reads simply, “Enclosed please find Defendant’s production.” Peering in, you see a power supply, USB cable and external hard drive. “Finally,” you [sigh] [exult], “the documents we’ve been fighting to get.”

Eagerly connecting the drive to your computer, you browse its contents to find tens...wait...*hundreds* of thousands of cryptically named files in dozens of folders. Clicking at random prompts the computer to repeatedly ask what program you would like to use to open them. “*How should I know?*” you grumble. On the next click, the computer stops asking and launches your e-mail program. To your horror, you see that the defendant’s messages now fill your mailbox. “Uh oh,” you realize, “maybe I should have gotten some help with this.”

Increasingly, plaintiffs’ counsel succeed in discovering electronically stored information (ESI), but find they can’t handle what’s produced. Once, we might have dealt with a “document dump” by rolling up our sleeves and putting in extra hours; today, more time and effort are unavailing if you can’t search or even read the ESI.

For most plaintiffs’ counsel, the challenges of understanding and managing what’s been produced can be met only by enlisting the aid of an electronic discovery service provider.

Both sides must identify, preserve, cull and produce relevant ESI, and there’s no shortage of guidance published and sold to aid producing parties. The legal technology marketplace fairly teems with companies promising to help major law firms and corporate America get a grip on e-discovery. But, where do *you* turn for e-discovery services, and how do you protect yourself from paying too much or buying what you don’t need or won’t work? This article suggests strategies for the lawyer on the receiving end of an ESI production.

### **Know Your Needs**

Lawyers often head for the e-discovery marketplace lacking a clear picture of what they need. They buy the wrong services at the wrong price from unreliable providers and end up with little or nothing they can use.

The most effective steps you can take to insure success in your dealings with EDD service providers should occur *before* you start looking for help. You need to assess your case and do some preliminary research. Based on what you know of your opponent and the events at the heart of the claim, what types of electronic evidence are likely to exist, where and in whose custody does it reside, and what forms might it take? That’s not as hard as it sounds.

If your client is a current or former employee of the defendant, he or she can clue you into how the company managed information. If the defendant is a large corporation, there’s a good chance the Internet will yield insight into their information technology (IT) systems. A colleague who’s litigated against the defendant can help, as can a friend or associate with IT expertise, optimally--but not necessarily--in the same industry. IT systems have more similarities across

---

<sup>1</sup> The author gratefully acknowledges the contributions of colleagues Ann Marie Gibbs, Sharon Nelson, John Simek and Linda Richenderfer in generously sharing their time and insights.

companies than differences, so even someone without specific knowledge of the defendant's systems can help you master essential concepts and terminology.

### **Know your Current Capabilities**

Ideally, you specified the form of production for any ESI produced and received forms you know how to handle.<sup>2</sup> If not, you'll probably need help selecting and setting up ways to store, search, sort, review and annotate the production.

The right approach fits your practice and your budget. Help the service provider achieve that fit by being prepared to answer these questions about your capabilities:

- Where are we now in terms of computer and network hardware, applications owned and used, network bandwidth, storage capacity and in-house expertise and support?
- How do we currently use computers to manage voluminous production?
- How comfortable are we working with ESI?
- How much time can we devote to learning to use new tools?
- Considering the anticipated value of the suit, what is our budget for services, hardware and software?
- Should we invest in tools, hardware and training that we can use in more than one case?
- Can we share the cost or the work with other firms or parties in the case?

### **Know Your Goals**

Decide what capabilities you're seeking in a review platform<sup>3</sup> for particular tasks, For example:

While reviewing the other side's e-mail, I want to be able to:

- Use software and hardware I already own;

---

<sup>2</sup> *FRCP Rule 34(b) contemplates that requesting parties will specify the form or forms of ESI production sought and that, absent objection or court order, production will be made in your specified form or forms. If a requesting party fails to specify a form or if a producing party objects to the requested form or forms, the producing party is obliged to state--in a written response filed within 30 days or at such other time as the parties agree to in writing or the Court directs--the form or forms in which it intends to make production. Absent a specification of form by the requesting party, the information must be produced in the form or forms in which it is ordinarily maintained or in a reasonably usable form or forms. FRCP Rule 34(b)(ii).*

<sup>3</sup> "Review Platform" is the buzzword for the software, hardware and services used to store, display, sort, search, tag, code, annotate, redact and/or produce ESI. There are many review platforms on the market, including the familiar LexisNexis Concordance (<http://law.lexisnexis.com/concordance>) and CT Summation (<http://www.ctsummation.com/>) applications, Internet-accessible hosted review environments and proprietary "solutions" marketed by e-discovery service providers.

If your firm doesn't already own tools like Concordance or Summation that are well-suited to ESI review after some slicing and dicing of the data, consider whether powerful and low-cost desktop search tools like DT Search (<http://www.dtsearch.com/>) meet your search and retrieval needs. The tradeoff to using tools not designed for e-discovery is that they lack key work flow features like document annotation, deduplication, issue coding and metadata management. Unfortunately, the solo and small firm e-discovery market hasn't stirred sufficient interest among developers for the emergence of a right-sized and right-priced desktop review suite adapted to day-to-day litigation. However, there's a fast-growing market for an e-discovery application akin to what QuickBooks offers for small business accounting. Either an affordable EDD tool kit will emerge or the cost of online storage and review services will drop to fill the void.

- Search the messages and attachments for words or phrases;
- Use proximity, fuzzy, Boolean and/or wild card search capabilities and stemming;
- View routine attachments on the fly;
- Deduplicate more than a single instance of identical or nearly identical items;
- Collaborate with other reviewers on my team;
- Attach notes or categorize the messages in my own way; and/or
- Expose relationships among senders and recipients or automatically find similar items.

Small efficiencies gained for each item reviewed pay handsome dividends applied to thousands of items. Also, stay mindful of how you plan to authenticate, use and present ESI at deposition or in trial.

### **The E-Discovery Marketplace**

Have you ever gone to the grocery store without a list and bought things you really didn't need? If so, you'll appreciate the economy that comes from deciding what you want before approaching vendors. Certainly, you need to listen, but don't let a salesperson talk you into products or services you neither want nor fully understand. When you cut through the marketing hype, you'll find that vendors provide similar services. Differentiation arises from experience, price, support and the ability to scale to projects of varying size and complexity.

Are you seeking a consultant, processor, application service provider or software vendor? Consultants plan the work, processors do it, application service providers rent workspace and tools and software vendors sell them.

More specifically, **consultants** advise you on proper, cost-effective ways to preserve, collect, search, produce, seek and manage ESI. They help you target your discovery requests, distinguish meritorious objections from obstructionist tactics and specify forms of production. They'll evaluate vendors and bids and support you at meetings and conferences concerning ESI issues. Depending on expertise, they may also conduct computer forensic examinations, assess the accuracy and completeness of production or testify in court to defend or challenge the handling or production of the electronic evidence.

**Processors** collect ESI or convert it to preferred forms, often filtering specific content according to the needs of the case or creating indices and databases to help manage the information. You'll turn to processors when you need native e-mail or electronic documents converted to page images, text extracted from those images, paper scanned to searchable electronic formats and backup tapes restored. Processors also generate the load files holding metadata and text that must accompany TIFF page images in order for them to be searchable.

**Application service providers** (ASPs) host large volumes of ESI, making it accessible to you via secure network or Internet access, typically paired with online tools to aid searching, viewing, analyzing, annotating and categorizing the data. ASPs facilitate sharing information and allocating review tasks among litigation team members in different firms and locations. ASPs are sometimes called SaaS providers, for *Software as a Service*.

**Software vendors** sell programs that collect, archive, search, display, index, analyze, annotate, categorize and convert ESI. Software tools may be highly specialized, like those that track litigation hold efforts, or require extensive training, like those used for computer forensic

analysis. Additionally, they may necessitate investment in dedicated hardware and data storage.

**E-discovery companies** fall into one or more of these categories, and full service providers typically consult on projects as well as process and host data. The boundaries aren't always clear cut; but as a rule of thumb, you'll pay the most for full service providers and the least for processors specializing in particular tasks, like restoration of backup tapes.<sup>4</sup> Of course, a processor is no bargain if you've bought the wrong service or don't know how to use what they deliver.

Hosting data with an application service provider is a good choice when collaborating with lawyers in different firms or locations. Because you pay every month, hosting with an ASP can cost more than alternatives when a case takes longer to resolve than anticipated. Though expensive, hosting eliminates much of the investment required to purchase software and develop in-house systems and, unlike capital investments, hosting costs can generally be passed on as cases expenses.

### **The Top Tier**

Legal magazines and websites are filled with ads for national e-discovery service providers promoting their expertise and "solutions." These vendors vie for attention at CLE conferences and trade expos, sponsor webcasts, fill mailboxes with glossy solicitations and compete aggressively for their slice of the multibillion dollar e-discovery pie.

You'd think they'd welcome your business; but for the most part, they don't want anything to do with you.

The top tier e-discovery service providers are geared to serve Fortune 500 corporations and the clients of large law firms. Their bread is buttered by enterprise-wide collection efforts, processing of huge volumes of data and the sale of sophisticated review and "early assessment" tools used by platoons of reviewers. Their business model hinges on developing relationships with customers for whom being sued or responding to government document requests is recurrent or routine. Even those who want your business may decline for fear of creating a future conflict in their target market.

Notable exceptions are plaintiffs' firms and litigation groups handling class action suits where the frequency and complexity of discovery efforts, as well as available funding, render them as attractive as large corporate clients. But for the plaintiffs' attorney looking for a hand with a few disks received in production or wielding the sling in David vs. Goliath discovery, the options are more limited and local.

---

<sup>4</sup> Restoration of backup tape is a task where it particularly pays to work with a specialist. Only a handful of vendors are equipped to process tape in volumes that allow for economies of scale and fast turnaround. While the author doesn't endorse any vendor, the leading national specialists in backup tape restoration are eMag Solutions, LLC, based in Atlanta, GA; National Data Conversion, Inc., based in New York, NY; and Renew Data Corp., based in Austin, TX.



## **Going Local**

The best way to find a good e-discovery vendor is to seek recommendations from other lawyers. Be sure to ask if they have firsthand knowledge of the vendor, what type of work was done and when. Check with colleagues who've spoken or published about electronic discovery at CLE conferences. Those top tier companies who don't want your business can help in directing you to providers in or near your community who will be glad to have you as a customer.

Even communities too small to support a local e-discovery service provider may be home to a computer forensic examiner. It may be economically feasible to use a forensicist to process modest data volumes (e.g., collections from fewer than ten machines or under 500 gigabytes in aggregate post-processed production); but, the higher hourly rates of computer forensic examiners (\$150-500/hour) compared to those of e-discovery data processors may price them out of the job. Still, a local examiner should be sufficiently plugged into the e-discovery market in your region to suggest well-qualified vendors.

Searching the Internet for local leads can be frustrating. National providers tend to dominate responses, in part because search engines sell search terms and enhanced placement. Accordingly, you can't rely on responses reflecting the names of local service providers, even when you include the name of your community or a nearby city in your search.

## **Do Your Homework**

Anyone can have a flashy web presence, so be wary of companies with web sites thin on content or that shed little light on the experience and qualifications of the principals.

The lure of e-discovery dollars has prompted everyone from copy and scanning services to computers and network service technicians to offer e-discovery services. While some comelately vendors do fine work, it's difficult to know which local providers are reliable and which are merely trying to jump on the bandwagon. Don't hesitate to ask how long the vendor has been in business and offered e-discovery services. Ask what software packages they employ. Behind-the-scenes, many local providers employ the same off-the-shelf tools for e-discovery, and it pays to know if they are limited by their tools or possess the technical expertise to adapt to your needs.

Visit a processor's workplace. Is it orderly and up-to-date? Is it secure? Is it dedicated to e-discovery work or is e-discovery just a sideline? Especially in smaller communities, computer resellers or copy services may be trying to make ends meet by adding e-discovery to their repertoire; but if they don't do enough of it, they may not have the experience needed to perform.

Salespeople make plenty of promises, but they're rarely the ones who must deliver the goods. The quality of an e-discovery effort is closely tied to the skills and experience of the project manager. So find out who will be responsible for your project and get a handle on their accessibility and resourcefulness. Since you probably can't gauge those qualities in a brief meeting, ask for the names of recent customer references *and call them*. True, you won't get the horror stories, but even happy customers can open your eyes to problems. Always ask if the company met deadlines and budget projections.

Especially for local service providers, a call to the Better Business Bureau, a quick check for lawsuits and even purchasing a financial report can head off headaches. Any industry has its fly-by-nights and fools, ask around.

Don't forget to explore any conflicts of interest that may exist. You need to know if the vendor works for your opposing counsel or the defendant. You should also assess the vendor's ability to secure the premises and the data, particularly if produced subject to protective order or paired with your confidential work product.

You may also want a frank discussion about who pays for rectifying processing errors. Such errors occur with enough regularity that it's not a question of *if* they will happen but *when*. Companies that value their reputations won't hesitate to step up and fix their mistakes. Any hemming or hawing about this should give pause.

Finally, if the vendor will be conducting searches of the data on your behalf, it's imperative to establish their search capabilities and limitations. Do they support Boolean searches? Can they effectively search foreign language data employing multibyte encoding or "Unicode?" Do they offer enhanced search technologies like concept searching or visual analytics? Will they charge for "machine time" while searches are made?

### **Getting Started**

Whether a vendor charges by the page, the gigabyte or the hour, costs tend to rise with the volume of information processed and its complexity. Uncertainty leads to price padding, so one of the first steps a processor should undertake is to inventory the data. They need to know how many discrete files have been produced, their format and size. Nested files that contain other files (e.g., compressed archives and e-mail containers) can seriously skew volume and cost estimates, so you want to flag them and inventory their contents early. This is an opportune time to ensure that the vendor's systems are capable of recursing through nested data as deeply as required to extract all content and of identifying and interpreting all of the various file types produced.

Your next goal should be filtering to cull irrelevant data and deduplication to insure that you don't waste time looking at the same material over and over again. These efforts should incorporate ways to uniquely identify production items (if not already Bates stamped) and insure that the review doesn't impact the integrity of the electronic evidence. A technique called file hashing<sup>5</sup> is especially useful for this.

At this point, the vendor should be able to offer very reliable projections of the total cost to convert the data to more accessible forms. Be wary of cost projections based on "presumed" or "rule of thumb" page equivalencies. These are rarely accurate and tend to inflate the cost significantly. It's a good time to inquire about your exposure to "exception handling" charges--sums assessed when vendors must e.g., deal with oddball file types or decrypt password protected data.

---

<sup>5</sup> Hashing is the use of an algorithm to calculate a distinctive alphanumeric value called a "hash" that ably serves as an electronic fingerprint for any digital data. Hashing supports unique identification and easy authentication of digital evidence. Common hash algorithms are MD5 and SHA-1. MD5 hash fingerprints are so unique that the likelihood of two differing files having the same hash value is estimated to be one in 340 trillion trillion trillion.

## **V as in Victory...and Vendor**

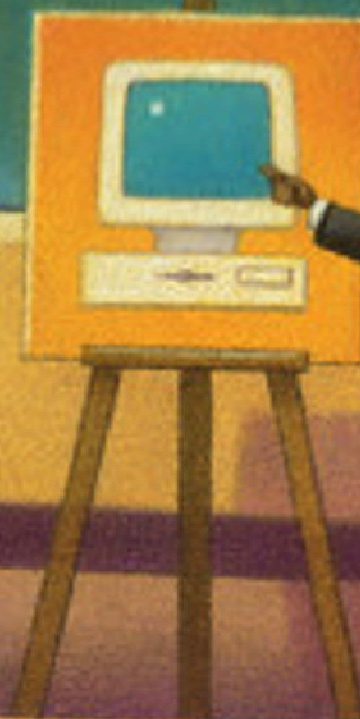
The too-candid e-mail. The *res gestae* voice message. The purloined PowerPoint. More-and-more, the most compelling documents in your cases are digital. Lawyers able to discover, navigate and present electronic evidence hold the winning hand against those who cannot, and the strength of that hand owes a lot to good working relationships with skilled, reliable e-discovery service providers.

Good help is hard to find—and some vendors you'll want to avoid—but with the right ESI expertise on your trial team, you'll be better able to unearth the buried bits and bytes, debunk the excuses and—just maybe—bring about that amazing alchemy of turning leaden production into golden justice.

## **21 Tips for Working with E-Discovery Service Providers**

1. Bone up on e-discovery before you wade in.
2. Find out all you can about the forms and volumes of ESI before approaching vendors.
3. Know what systems and software you already own capable of reviewing ESI.
4. Don't purchase products or services you or your staff don't know how to use unless you budget the time and money needed for training.
5. Visit the processing facility. Does it look like the photos on the website? Is it orderly and up-to-date? Is it secure?
6. Explore conflicts. You may not mind that the vendor is also working for the defendant or opposing counsel in other matters, but you need to know about it.
7. Establish the vendor's ability and willingness to deliver both the technical and testimonial support you'll need.
8. Insist on meeting the project manager or consultant who'll be assigned to your project. Is it a good fit? Can she communicate in non-technical language? Does she look shell shocked?
9. Get the cell phone numbers of your project manager and at least one other person working with your data.
10. Get customer references and talk with them. Did the vendor meet deadlines? How much work had to be reprocessed due to errors? Did actual charges significantly exceed projections?
11. It's a vendor's job to be certain you understand the scope of work, the reasons behind actions and what it will all cost. Don't be baffled by jargon or concerned about looking foolish. The only dumb questions are the one you don't ask.
12. Clearly establish the scope of work in writing, and be sure you understand all price components and how they are calculated.
13. For pricing based on data volumes, be certain you understand how volumes are calculated. Are bills based on raw volumes or determined after filtering? Beware of "page equivalency" calculations.
14. Guard against sticker shock by setting a threshold above which you must expressly authorize further work.
15. Even the sharpest attorneys can't find loopholes in the laws of physics. Large ESI volumes take time to duplicate, index, search and process, so be sure you allow sufficient time; else, be prepared to pay expedited rates and seek extensions.
16. Clearly communicate deadlines, and get written commitments to meet them.
17. Protect your ability to efficiently and economically recover your data if serious problems occur. Insure your data can't be "held hostage" in billing disputes.
18. Be sure you know what tasks the vendor handles in-house and what they outsource. Outsourcing may be smart, but you need to know why and what markups apply.
19. Arrange for your side's EDD technician to talk directly to your opponent's EDD technician. Their ability to speak the same language streamlines data transfer and establishes appropriate expectations.
20. The demand for talented e-discovery product managers far outstrips the supply, so there's a lot of turnover. Get acquainted with others working on your project before your project manager moves on. When crunch time comes, you don't want to hear, "She's gone, let me find out who's handling that now."
21. Don't wait until you need an e-discovery service provider to start lining up prospects. The need will be there. Having someone on deck helps you hit the ground running.

# Piecing Together the E-Discovery Plan: A Plaintiff's Guide to Meet and Confer



**Craig Ball**

# Piecing Together the E-Discovery Plan: *a Plaintiff's Guide to Meet and Confer*

By Craig Ball

© 2008

**E-discovery is challenging, but it needn't be complicated by a battle-zone mentality. Take advantage of the meet-and-confer process to ensure that your opponents know what electronically stored information they have and how they should produce it.**

Everyone wants e-discovery to be simple. The defendant's tech guru wants it to be simple because he's got too much to do. The defendant's in-house counsel wants it to be simple because she's got budget issues and thinks most claims are frivolous. Outside counsel wants it to be simple because he likes doing things the way he's always done them and doesn't like looking clueless about electronic information.

And you want it to be simple because you need it to be simple. Hiring experts and e-discovery vendors raises the stakes, and a misstep may result in significant cost-shifting to your client. Moreover, if you don't ask the right questions, you're not going to get the right information--and aren't courts starting to sanction lawyers for e-discovery foul-ups?

The problem is e-discovery is not simple. It's complex, technical and tricky. There are no shortcuts--no form, checklist, or script that's going to get the defendant to find the relevant information and turn it over in a reasonably usable way.

Face it: you've got to *fight* to get electronic evidence. You have to know what they've got, what you need and how to ask for it. You must understand the capabilities and limitations of electronic search and the forms of production best suited to the evidence.

## **Meet and Confer**

In 2006, Federal Rule of Civil Procedure 26(f) was amended to require parties to confer about preserving discoverable information and to develop a proposed discovery plan addressing discovery of electronically stored information (ESI) and the form or forms in which it should be produced. The amended rule requires parties to confer about preserving discoverable information and to develop a proposed discovery plan addressing, *inter alia*, discovery of electronically stored information (ESI) and the form or forms in which it should be produced.

This conference<sup>1</sup>, and the overall exchange of information about electronic discovery, is called “meet and confer.”<sup>2</sup>

The states are rapidly adopting rules of procedure and local practice like the Rule 26(f) meet and confer;<sup>3</sup> but even where no such rule exists, state judges often find the federal e-discovery model instructive and grant motions to compel parties to confer on ESI issues.<sup>4</sup>

---

<sup>1</sup> *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)....”*

<sup>2</sup> *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.”

<sup>3</sup> Noted e-discovery commentator Thomas Allman, a founding member of the Sedona Conference and co-chair the E-Discovery Committee of the Lawyers for Civil Justice, reports that seven states that have adopted e-discovery rules hewing closely to the Fed. R. Civ. P. (Louisiana, Minnesota, Montana, New Jersey, Utah, Arizona and Indiana). Allman notes another 14 states are considering changes to their court rules to address e-discovery (Alaska, Connecticut, Florida, Illinois, Iowa, Kansas, Maryland, Nebraska, New Mexico, North Dakota, Ohio, Tennessee, Virginia and Washington). *See* Brett Burney, *Mining E-Discovery Stateside*, Law Technology News (January 18, 2008).

<sup>4</sup> *See, e.g., Conference of Chief Justices, Guidelines For State Trial Courts Regarding Discovery Of Electronically-Stored Information, Section 3 (2006), stating that a judge should “encourage” counsel to meet and confer in an effort to agree on e-discovery issues and to exchange information, inter alia:*

(1) A list of the person(s) most knowledgeable about the relevant computer system(s) or network(s), the storage and retrieval of electronically-stored information, and the backup, archiving, retention, and routine destruction of electronically-stored information, together with pertinent contact information and a brief description of each person's responsibilities;

(2) A list of the most likely custodian(s), other than the party, of relevant electronic data, together with pertinent contact information, a brief description of each custodian's responsibilities, and a description of the electronically-stored information in each custodian's possession, custody, or control;

(3) A list of each electronic system that may contain relevant electronically-stored information and each potentially relevant electronic system that was operating during the time periods relevant to the matters in dispute, together with a general description of each system;

(4) An indication whether relevant electronically-stored information may be of limited accessibility or duration of existence (e.g., because they are stored on media, systems, or formats no longer in use, because it is subject to destruction in the routine course of business, or because retrieval may be very costly);

(5) A list of relevant electronically-stored information that has been stored offsite or off-system;



## Sizing Up The Opposition

Opponents weaned on a scorched earth, “take no prisoners” approach to litigation aren’t adapting well to the requisite openness and collaboration of meet and confer. They won’t tell you how they identified and collected responsive data. They’ll refuse to share custodial questionnaires or disclose keywords and filtering mechanisms. Deal with them by making your record and seeking the court’s intervention. Angry judges, sanctions and unhappy clients are the Darwinian factors bringing about the extinction of Obstructasaurus Lex.

Obstructive opponents aren’t your only obstacle. Well-intentioned producing parties present challenges, too, and tend to split into three camps:

**Those who accept the duty to preserve and produce ESI, want to do it right, but don’t know how:** These opponents are ill-equipped to guide preservation or ask the right questions. Here, be prepared to fill the knowledge gap in a non-threatening manner—a daunting challenge in a profession where few are willing to admit weakness—or find ways to convince your opponent to get expert help. Bringing your own expert to conferences or hearings helps the other side see they are in over their heads.

**Those who accept the duty, but know only one way to deal with ESI:** These opponents have settled on an approach that worked for them in another case and are determined to employ it in *every* case. Their method might entail, e.g., over-reliance on custodial collection or a blind devotion to TIFF image production, even when it destroys the integrity of the evidence. Here, you need to understand their approach and determine if it’s going to work. If not, be prepared to demonstrate where it falls short and offer suitable alternatives. The right solution may be a *hybrid* production integrating alternative techniques for categories of ESI that don’t lend themselves to the other side’s approach.

**Those who accept the duty, want to do it right and know how:** Here, the onus is on you to meet them on the level playing field, so know what you need and be prepared

---

*(6) A description of any efforts undertaken, to date, to preserve relevant electronically-stored information, including any suspension of regular document destruction, removal of computer media with relevant information from its operational environment and placing it in secure storage for access during litigation, or the making of forensic image back-ups of such computer media;*

*(7) The form of production preferred by the party; and*

*(8) Notice of any known problems reasonably anticipated to arise in connection with compliance with e-discovery requests, including any limitations on search efforts considered to be burdensome or oppressive or unreasonably expensive, the need for any shifting or allocation of costs, the identification of potentially relevant data that is likely to be destroyed or altered in the normal course of operations or pursuant to the party’s document retention policy.*



to settle on reasonable and effective methods to identify, preserve, select and produce the information without undue burden or cost. Be ready, and be reasonable.

### **Preparing for Meet and Confer**

E-discovery duties are reciprocal. Just because your client has little electronic evidence, you must nonetheless act to preserve and produce it. At meet and confer, be prepared to answer many of the same questions you'll pose.

**A cardinal rule for electronic discovery is to tell your opponents what you seek, plainly and clearly. They may show up empty-handed, but not because you failed to set the agenda.**

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. 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 the load file specification you want used or keywords you want searched, 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, 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 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 address the types and varieties of matters discussed at meet and confer. They're neither exhaustive nor tailored to the unique issues in your case. They're offered as talking points to stimulate discussion, not as a rigid agenda and certainly not as a form for discovery.

#### **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 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. 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, modification, and access dates of electronic documents tend to be tracked, focusing on time periods and particular events helps identify relevant ESI, but only if you understand what the dates signify and when you can or can't rely on them. When a document was created doesn't necessarily equate to when it was written, nor does "accessed" always mean "used." 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 preservation duty going forward. The preservation obligation generally begins with an expectation of litigation, but the facts and issues dictate if there is a going forward obligation. 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 backup tape slated for reuse or e-mail subject to automatic deletion, and insure its preservation. Address back up 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 dinosaurs extant 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 opponents see that reasonable disclosures that don't prompt objections 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, application service providers, family members and other nonparties. Some may retain copies of information discarded by your opponent. Absent your reminder, the other side may focus on their own data stores and fail to take steps to preserve data held by others over whom that have some right of direction or control.

## **9. 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 are required. Though skilled forensic examination is expensive, off-site, forensically-sound preservation can cost less than \$500 per system. So talk about the need for such efforts, and if your opponent won’t undertake them, consider whether you should force forensic preservation, even if you bear the cost.

## **10. 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 insure that application metadata is made visible before imaging.

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.

### **11. What are the defendant's 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.

Telling examples are e-mail and backup tapes. 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 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, the defendant may need to modify or suspend certain data retention practices. Discuss what they are doing and whether they will, as needed, agree to pull tapes from rotation or modify purge settings.

### **12. Are there legacy systems to be addressed?**

Like legacy backup tapes, old computers and servers tend to stick around even if they've fallen off the defendant's radar. 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?

### **13. What are the current and prior e-mail applications?**

E-mail systems are Grand Central Station for ESI. Understanding an opponent's 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. Corporate e-mail systems tend to split between the predominant Microsoft Exchange Server software tied to the Microsoft Outlook e-mail client on user's machines and Lotus' Domino mail server accessed by the Lotus Notes e-mail client application. 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 an opponent may fail to search.

### **14. Are personal e-mail accounts and computer systems involved?**

Those who work from home, out on the road or from abroad may use personal e-mail accounts for business or store relevant ESI on their home or laptop machines. Parties should address the

potential for relevant ESI to reside on personal and portable machines and agree upon steps to be taken to preserve and produce that data.

#### **15. 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.

#### **16. How will we handle voice mail, instant messaging and other challenging ESI?**

Producing parties routinely ignore short-lived electronic evidence like voice mail and instant messaging by acting too late to preserve it or deciding that the retention burden outweighs any benefit. Though it's not especially challenging to preserve voice mail or IM logs if one acts swiftly, defendants tend to demand a particularized need before they'll do so. *When it's relevant*, will the other side archive voice mail messages or activate local logging or packet capture of IM traffic?

#### **17. 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, Documentum) and how the data's structured before proposing sensible ways to preserve and produce it. Options include running agreed queries, exporting relevant data to standard formats like Access databases or XML or even mirroring the entire contents to a review environment.

Because databases are always changing, Michael Arkfeld, author of the respected treatise "Arkfeld on Electronic Discovery and Evidence"<sup>5</sup> cautions that both sides need to be working from the same database, asking, "Does the database ESI have a concrete beginning or ending date or is it a "rolling" database where data's added and deleted on a continuous basis?"

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.

---

<sup>5</sup> Michael R Arkfeld, Arkfeld on Electronic Discovery and Evidence (2nd Ed. 2007); <http://www.lexisnexis.com/arkfeld/>

### **18. 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 insure 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.).

### **19. 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. 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.<sup>6</sup>

### **20. 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. The terms and techniques facilitating an attorney's assessment of a case are protected, 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 defendant uses mechanized search to segregate data as privileged, plaintiffs 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.

### **21. 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

---

<sup>6</sup> 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)

they looked through the matches and, sorry, you didn't win. As a consolation prize, you may get the home game: a million 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. But according to Jason Baron, Director of Litigation at the National Archives and Records Administration, "Lawyers are waking up to the fact that keyword searching is subject to profound limitations in terms of accuracy and results."<sup>7</sup> Thanks to, e.g., misspellings, acronyms, synonyms, IM-speak, noise words, OCR errors and the peculiar "insider" lingo of colleagues, companies and industries, keyword search performs far below most lawyers' expectations, finding perhaps 20% of responsive material on first pass.<sup>8</sup>

Under the rubric of "concept search," technologies employing Google-like analysis are improving both the precision and recall of electronic search, but Baron cautions, "Despite the hype, artificial intelligence, data mining, and content analytics is just not sufficiently advanced to ensure that substantially all relevant documents in a large collection of ESI will be found."<sup>9</sup>

**Never allow opposing counsel to position keyword search as a single shot in the dark. You must be afforded the opportunity to use information gleaned from the first or subsequent efforts to narrow and target succeeding searches.**

Warts and all, keyword search remains the most common method employed to tackle large volumes of ESI, and a method still enjoying considerable favor with courts.

Baron notes that "keyword searches--indeed, any form of searching--is more effective when employed in an iterative way, as part of a cooperative and informed process."<sup>10</sup> In other words,

---

<sup>7</sup> Mr. Baron should know, as he is Editor in Chief of The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery (2007) and also responsible for searching through 20 million White House presidential emails in response to massive discovery in the U.S. v. Philip Morris tobacco litigation.

<sup>8</sup> 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).

<sup>9</sup> Influential Magistrate, Judge John Facciola, mentions concept search in his opinion in *Disability Rights Council of Greater Washington, et. al, v Washington Metropolitan Transit Authority, et. al.* 2007 U.S. Dist. Lexis 39605, citing, George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?* 13 Rich. J.L. & Tech. 10 (2007).



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 or subsequent effort 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.

## **22. How will de-duplication 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 back up. De-duplication is the process by which identical items are reduced to a single instance for purposes of review. De-duplication 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. If production will be made on a custodial basis—and depending upon the review platform employed—it may be desirable to request re-population of content de-duplicated horizontally so each custodian's collection is complete.

## **23. What forms of production are offered or sought?**

Notably, the 2006 Federal Rules amendments gave 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 other forms on the requesting party, but must disclose the other forms before making production so as to afford the requesting party the opportunity to ask the court to compel production in the designated form or forms.<sup>11</sup>

Options for forms of production include native file format, quasi-native forms (e.g., a partial export of data from a database), imaged production (PDF or, more commonly, TIFF images accompanied by load files containing searchable text and metadata), hosted (online) production

---

<sup>10</sup> See, e.g., Paul, George L. and Jason R. Baron, "Information Inflation: Can The Legal System Cope?," 13 Richmond Journal of Law and Technology (2006), <http://law.richmond.edu/jolt/v13i2/article11.pdf>. See also, The Sedona Principles, Second Edition: Best Practices Recommendations & Principles for Addressing Electronic Document Production, Comment 11.a (The Sedona Conference@Working Group Series, 2007), available at [www.thesedonaconference.org](http://www.thesedonaconference.org).

<sup>11</sup> Fed. R. Civ. P. 34(b)

and even paper printouts for 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 or word processed documents, but native forms are best for spreadsheets and essential for audio and video. Quasi-native forms are well-suited to e-mail and databases.

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 trade offs, typically sacrificing utility 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."<sup>12</sup> That is, know the needs and capabilities of the applications or tools you'll employ to index, sort, search and access electronic evidence.

#### **24. How will you handle redaction of privileged, irrelevant or confidential content?**

Defendants often seek to redact ESI in the way they once redacted paper documents: by blacking out text. To make that possible, ESI are 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 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

---

<sup>12</sup> 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. A platform geared to review of ESI in native formats must be able to open the various types of data received without corrupting its content or metadata. 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.

There are many review platforms on the market, including the familiar Concordance and Summation applications, Internet-accessible hosted review environments 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.

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 now 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.

### **25. 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 agreeing to employ a structure geared to review software such as CT Summation or LexisNexis Concordance.

### **26. 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 Bates numbering.

### **27. 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.

Michael Arkfeld warns that, "Some defense attorneys take the position that additional burden or cost associated with any ESI makes it 'not reasonably accessible' and the requesting party must pay for its production." Arkfeld agrees that's a misinterpretation, but one that can prevail when parties or the court don't make the effort to understand the amended rule.

**The meet-and-confer session is an opportune time to resolve inaccessibility claims without court intervention--to secure a commitment that the information at issue will be preserved.**

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.

## **28. 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 Collaboration**

Courts and commentators uniformly cite the necessity for transparency and collaboration 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

**Candor and cooperation in e-discovery isn't a sign of weakness, but a hallmark of professionalism**

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.<sup>13</sup> 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.

---

<sup>13</sup> Courts have sanctioned ESI discovery abuse for actions characterized as intentional deception: *Qualcomm Inc. v. Broadcom Corp.*, 2008 WL 66932 (S.D. Cal. Jan. 7, 2008); gross negligence: *Phoenix Four, Inc. v. Strategic Res. Corp.*, 2006 WL 2135798 (S.D.N.Y. Aug. 1, 2006); "reckless disregard:" *United Med. Supply Co., Inc. v. United States*, 77 Fed. Cl. 257 (2007); "purposeful sluggishness:" *In re Seroquel Prods. Liab. Litig.*, 2007 WL 2412946 (M.D. Fla. Aug. 21, 2007); "foot dragging:" *Toussie v. County of Suffolk*, 2007 WL 4565160 (E.D.N.Y. Dec. 21, 2007) and negligence: *Finley v. Hartford Life and Acc. Ins. Co.*, 2008 WL 509084 (N.D. Cal. Feb. 22, 2008). Increasingly, courts regard the duty to preserve and produce ESI as one mutually shared by client and counsel, and refuse to accept ignorance on either's part as an excuse. See, e.g., *Qualcomm Inc. and Phoenix Four, Inc.*



# Selected Musings on Electronic Discovery for Employment Lawyers

**“Ball in Your Court”  
June 2005 – May 2009**

© Craig Ball

The *Law Technology News* column “Ball in Your Court” is both the 2007 and 2008 Gold Medal honoree as “Best Regular Column” as awarded by Trade Association Business Publications International. It’s also the 2007 and 2009 Silver Medalist honoree of the American Society of Business Publication Editors as “Best Contributed Column” and their 2006 Silver Medalist honoree as “Best Feature Series” and “Best Contributed Column.”

Cowboys and Cannibals .....	104
Locard’s Principle .....	106
A Golden Rule for E-Discovery .....	108
Ten Common E-Discovery Blunders.....	110
Ten Tips to Clip the Cost of E-Discovery .....	112
Copy That? .....	114
In Praise of Hash .....	117
Unlocking Keywords .....	119
Getting to the Drive.....	122
Who Let the Dogs Out? .....	124
Page Equivalency and Other Fables .....	126
Re-Burn of the Native .....	128
The Power of Visuals.....	130
Well Begun is Half Done.....	132
Ask the Right Questions .....	134
Crystal Ball in Your Court .....	136
Redaction Redux .....	139
Trying to Love XML.....	141

The Science of Search .....	143
Dealing with Third-Parties.....	145
Grimm Prognosis.....	147
Brain Drain.....	149
SNAFU .....	151
Problematic Protocols.....	153
What Lies Beneath? .....	156
Don't Touch That! .....	158
Special Masters .....	160
About the Author.....	163

## **Cowboys and Cannibals**

**by Craig Ball**

*[Originally published in Law Technology News, June 2005]*

With its quick-draw replies, flame wars, porn and spam, e-mail is the Wild West boom town on the frontier of electronic discovery--all barroom brawls, shoot-outs, bawdy houses and snake oil salesman. It's a lawless, anyone-can-strike-it-rich sort of place, but it's taking more-and-more digging and panning to get to the gold.

Folks, we need a new sheriff in town.

### **A Modest Proposal**

E-mail distills most of the ills of e-discovery, among them massive unstructured volume, mixing of personal and business usage, wide-ranging attachment formats and commingled privileged and proprietary content. E-mail epitomizes "everywhere" evidence. It's on the desktop hard drive, the server, backup tapes, home computer, laptop on the road, Internet service provider, cell phone and personal digital assistant. Stampede!

There's more to electronic data discovery than e-mail, but were we to figure out how to simply and cost-effectively round up, review and produce all that maverick e-mail, wouldn't we lick EDD's biggest problem?

The e-mail sheriff I envision is a box that pops up when you hit send and requires designation of the e-mail as personal or business-related. If personal, it's sent and a copy is immediately forwarded to your personal e-mail account. The personal message is then purged from the enterprise system. If business related, you must assign the message to its proper place within

the organization's data structure. If you don't put it where it belongs, the system won't send it. Tough love for a wired world. On the receiving end, when you seek to close an e-mail you've read, you're likewise prompted to file it within your organization's data structure, deciding if it's personal or business and where it belongs.

When I first broached this idea to my e-discovery colleagues, the response was uniformly dismissive: "Our people wouldn't do it" being the common reply. Hogwash! They'll do it if they have to do it. They'll do it if there's a carrot and a stick. They'll do it if the management system is designed well and implemented aggressively. I ask them, "Why do you make employees punch in a code to use the photocopier, but require no accountability for e-mail that may sink the company?"

Some claim, "Our people will just call everything personal or file all business correspondence as 'office general.'" Possibly, but that means that business data will be notable by its absence from its proper place. Eventually, the boss will say, "Dammit Dusty, why can't you keep up with your e-filing?" In addition, Dusty won't want the system to report that he characterizes 95% of the at-work electronic communications he handles each day as personal in nature. Certainly, there needs to be audit and oversight, and the harder you make it to for a user to punt or evade the system, the better the outcome. This model worked for paper. It can work for e-mail.

Once, a discovery request sent a file clerk scurrying to a file room set aside for orderly information storage. There, the clerk sought a labeled drawer or box and the labeled folders within. He didn't search every drawer, box or folder, but went only to the places where the company kept items responsive to the request. From cradle to grave, paper had its place, tracked by standardized, compulsory practices. Correspondence was dated and its contents or relevance described just below the date. Files bore labels and were sorted and aggregated within a structure that generally made sense to all who accessed them. These practices enabled a responding party to affirm that discovery was complete on the strength of the fact that they'd looked in all the places where responsive items were kept.

By contrast, the subject lines of e-mails may bear no relation to the contents or be omitted altogether. There is no taxonomy for data. Folder structures are absent, ignored or unique to each user. Most users' e-mail management is tantamount to dumping all their business, personal and junk correspondence into a wagon hoping the Google cavalry will ride to the rescue. The notion "keep everything and technology will help you find it" is as seductive as a dance hall floozy...and just as treacherous.

E-discovery is not more difficult and costly than paper discovery simply because of the sheer volume of data or even the variety of formats and repositories. Those concerns are secondary to the burdens occasioned by the lack of electronic records management. We could cope with the volume if it were structured because we could rely on that structure to limit our examination to manageable chunks. Satirist Jonathan Swift was deadly humorous when, in his 1729 essay,



“A Modest Proposal,” he suggested the Irish eat their children to solve a host of societal ills, but I’m deadly serious when I modestly propose we swallow our reluctance and impose order on enterprise e-mail. The payback is genuine and immediate. Tame the e-mail bronco and the rest of the herd will fall in line.

Does imposing structure on electronic information erase the advantages of information technology? Is it horse-and-buggy thinking in a jet age? No, but it’s has its costs. One is speed. If the sender or recipient of an e-mail is obliged to think about where any communication fits within their information hierarchy and designate a “location,” that means the user has to pause, think and act. They can’t just expectorate a message and hit send. Dare we re-introduce deliberation to communication? The gun-slinging plaintiff’s lawyer in me will miss the unvarnished, *res gestae* character of unstructured e-mail, but in the end, we can do with a little law west of the Pecos.

## **Locard’s Principle** **by Craig Ball**

***[Originally published in Law Technology News, February 2006]***

Devoted viewers of the TV show “CSI” know about Locard's Exchange Principle: the theory that anyone entering a crime scene leaves something behind or takes something away. It’s called cross-transference, and though it brings to mind fingerprints, fibers and DNA, it applies to electronic evidence, too. The personal computer is Grand Central Station for PDAs, thumb drives, MP3 players, CDs, floppies, printers, scanners and a bevy of other gadgets. Few systems exist in isolation from networks and the Internet. When these connections are used for monkey business like stealing proprietary data, the electronic evidence left behind or carried away can tell a compelling story.

Recently, a colleague owning a very successful business called about an employee who’d quit to start a competing firm. My colleague worried that years of collected forms, research and other proprietary data might have gone out the door, too. The departing employee swore he’d taken nothing, but the unconvinced boss needed reassurance that someone he trusted hadn’t betrayed him. He asked me to examine Mr. Not Me’s laptop.

Turning to a forensic specialist was a smart move. Had the boss yielded to temptation and poked around the laptop, Locard’s Principle dictates he would have irretrievably contaminated the digital crime scene. Last access dates would change. Log entries would be overwritten. Some deleted data might disappear forever. More to the point, an unskilled examiner would have overlooked the wealth of cross-transference evidence painting a vivid picture of theft and duplicity.

Stolen data has to be accessed, copied and then find its way out of the machine. Whether it's sent to a printer, e-mailed, burned to optical disk, written to a floppy or spirited away on a thumb drive, each conduit carries data away and leaves data behind as evidence of the transaction.

Forensic analysis of the employee's laptop turned up many examples of Locard's Principle at work. Windows employs a complex database called the Registry to track preferences and activities of the operating system and installed applications. When a USB storage device like a thumb drive connects, however briefly, to a Windows computer, the operating system interrogates the attachment and dutifully records information about the device and the date in the Registry. A moment-by-moment analysis of every file accessed shortly before the employee's departure and of the Registry revealed attachment of a thumb drive—an event reinforced by the system accessing the sound file played when a device attaches to a USB port. “Bonk-bink.” This immediately preceded access to many proprietary files on the network, concluding with the system accessing the sound file signaling removal of the USB device. “Bink-bonk.”

Further examination showed access to other proprietary data in conjunction with use of the system driver that writes data to recordable CDs. This evidence, along with an error log file created by a CD burning application detailing the date and time of difficulty encountered trying to burn particular proprietary files to CD-R, left no doubt as to what had transpired.

The coup de grace demonstrating the premeditated nature of the theft emerged from a review of files used to synchronize the laptop with a “smart phone” PDA. These held records of cell phone text messaging between the employee and a confederate in the firm discussing what files needed to be spirited away. Though the messages weren't created on or sent via the laptop, they transferred to the laptop's hard drive unbeknownst to the employee when he synched his PDA. Armed with this evidence, the boss confronted the still-employed confederate, who tearfully confessed all to the sadder-but-wiser employer. Case closed, but no happy ending.

Computers, like crime scenes, have stories to tell. Data and metadata in their registries, logs, link files and abandoned storage serve as Greek chorus to the tragedy or comedy of the user's electronic life. Most cases don't require the “CSI” treatment, but when the computer takes center stage, don't overlook the potential for computer forensic analysis—and Dr. Locard's Exchange Principle--to wring decisive evidence from the machine.

# A Golden Rule for E-Discovery

by Craig Ball

*[Originally published in Law Technology News, March 2006]*

Albert Einstein said, "In the middle of every difficulty lies opportunity." Electronic data discovery is certainly one of the greatest difficulties facing litigants today. So wouldn't you know some genius would seize upon it as an opportunity for abuse? Perhaps Einstein meant to say, "In the middle of every difficulty is an opportunity for lies."

I'm not talking about the pyrotechnic failures to produce email or account for backup tapes that brought low the mighty in such cases as *Zubulake v. UBS Warburg* and *Coleman (Parent) Holdings v. Morgan Stanley*. Stonewalling in discovery predated electronic discovery and will likely plague our progeny's progeny when they grapple with photonic or neuronal discovery. But while an opponent's "No, we won't give it to you," may be frustrating, it's at least sufficiently straightforward to join the issue and promote resolution. The abuses lately seen make stonewalling seem like fair play.

## Playing the Telephone Game

I'm talking sneaky stuff, like printing electronic information to paper, then scanning and running it through optical character recognition (OCR), or "printing" electronic information to a TIFF image format then OCR'ing the TIFF.

If you've played the parlor game, "Telephone," you've seen how transmitting messages introduces errors. The first listener interprets the message, as does the next listener and the next. Each mangles the message and the errors compound hilariously. "Send reinforcements--we're going to advance" emerges as, "Send three and four pence--we're going to a dance."

When you print electronic evidence, part of the message (e.g., its metadata) is lost in the printing. When you scan the printout, more distortion occurs, and then optical character recognition further corrupts the message, especially if the scanned image was askew, poorly resolved or included odd typefaces. Page layouts and formatting suffer in the translation process, too. If you're lucky, what emerges will bear a resemblance to the original evidence. If not, the output will be as distorted as the Telephone game message, but no laughing matter. Much of its electronic searchability is gone.

Speaking on a panel at New York LegalTech 2006, I groused, "Imaging data to TIFF and then OCR'ing it ought to be a crime in all 50 states." Was I surprised when that drew applause from the EDD-savvy audience! Their enthusiastic response confirmed that others are fighting TIFF/OCR abuse, too.

There's always been gamesmanship in discovery, but it wasn't hard to detect dirty pool with paper. Bad copies *looked* bad. Redaction stood out. Page numbers and dates exposed omission. But e-discovery creates fresh-and-furtive opportunities for shenanigans, and they're harder to detect and prove.

### **Bad OCR**

Take OCR. We tend to think of optical character recognition as a process that magically transforms pictures of words into searchable text. OCR is OCR, right? In fact, error rates for OCR applications vary widely. Some programs are superb, correctly interpreting better than 99% of the words on most pages, even when the page is askew, the fonts obscure and the scan a mess. Other applications are the Mr. Magoo's of the OCR world, misinterpreting so many words that you might as well retype the document. In between are OCR apps that do well with some typefaces and formatting and poorly with others.

The OCR application or service provider that processes electronic evidence has an enormous impact on the usability of the production. Bad OCR insures that text searches will come up short and spreadsheet data will be worthless. But how do you know when a producing party furnishes bad OCR, and how do you know if it's an effort to hamper your investigation? Start by checking whether the other side depends on the same bad data or if they are relying on the pristine originals.

"Even a dog," observed Justice Oliver Wendell Holmes, "knows the difference between being tripped over and being kicked." True, but e-discovery can leave you feeling dumber than a dog when you can't tell if the opposition's messing with you or just plain incompetent. One day, it will be a distinction without a difference for purposes of enforcement--sloppy and slick will both draw sanctions. Until then, courts need to explore whether the data produced is hobbled compared with that used by the producing party and its counsel.

### **Level the Playing Field**

So how do you deal with opponents who convert native data to naked TIF formats and deliver bad OCR? The answer is to insist that the source data stay in its native digital format. That doesn't necessarily mean native file production, but be sure that the text and the relevant metadata are ported directly to the production format *without* intervening OCR. It's cheaper, faster and much more accurate.

A level playing field means that the form in which information's produced to me isn't more cumbersome or obscure than what's available to you. The elements needed to sort, read, classify, search, evaluate and authenticate electronic evidence—elements like accurate text and relevant metadata—should be in my hands, too.

In short, *it shouldn't be much harder to use or understand the information you've produced when it's on my system than when it's on yours.* This digital Golden Rule has yet to find its full

expression in the Sedona Guidelines or the new Federal e-discovery rules, but it's a tenet of fairness that should guide the hand of every Solomon grappling with e-discovery.

## **Ten Common E-Discovery Blunders**

**by Craig Ball**

*[Originally published in Law Technology News, August 2006]*

A colleague recently asked me to list 10 electronic data discovery errors lawyers make with distressing regularity. Here's that list, along with suggestions to avoid making them:

### **1. Committing to EDD efforts without understanding a client's systems or data.**

It's Russian roulette to make EDD promises when you haven't a clue how much data your client has, or what and where it is. Instead, map the systems and run digital "biopsies" on representative samples to generate reliable metrics and gain a feel for how much are documents, e-mail, compressed files, photos, spreadsheets, applications and so on.

It matters. A hundred gigabytes of geophysical data or video may be a handful of files and cost next to nothing to produce. The same 100 gigs of compressed e-mail could comprise tens of millions of pages and cost a fortune.

### **2. Thinking you can just "print it out."**

Even if you've the time and personnel to stick with paper, is it ethical to subject your clients to the huge added costs engendered by your unwillingness to adapt?

### **3. Foolishly believing that enough smart people can take the place of the right technologies or that the right technologies eliminate the need for enough smart people.**

No search tool yet invented finds every responsive or privileged e-document, and no law firm can marshal enough qualified people to manually review 100 million pages. The best outcomes in EDD flow from pairing well-trained people with the right tools.

### **4. Ignoring preservation obligations until the motion to compel.**

The duty to preserve evidence doesn't hinge on a preservation notice or lawsuit. You must advise your client to preserve potentially relevant paper and electronic evidence as soon as they reasonably anticipate a suit or claim. Even if they aren't obliged to produce inaccessible electronic evidence, they're probably obliged to preserve it.

### **5. Thinking that search technology trumps records management.**

Sorry, but Google isn't going to save us. Privileged communications once went straight from the printer into a file labeled "Attorney Correspondence." Now, they're jumbled with Viagra ads and notices about donuts in the coffee room. We need to enforce cradle-to-grave management for electronic records and restore the "power of place" that allows us to once more limit where we

look for responsive data to just those places "where we keep that stuff." Much of the heavy lifting will be over when users must "file" messages in a virtual "file room" when they're sent or received.

#### **6. Hammering out EDD agreements without consulting an expert.**

Just because both sides agree to something doesn't make it feasible, or even a good idea. An agreed order stating that an expert will recover "all deleted files" sounds simple, but it's the sort of muddled directive that needlessly drives up the cost of EDD. The right expert will identify efficiencies, flag pitfalls and suggest sensible, cost-effective search and sampling strategies from the earliest meet-and-confer session.

If your client can't afford an attending expert — though in the end, amateurs costs much more — at least run proposed agreements by someone in the know before they go to the judge.

#### **7. Taking a "peek" at a computer that may contain critical evidence.**

Metadata is the data about data that reveals, inter alia, dates of creation, access and modification. Sometimes it's the "who-knew-what-when" evidence that makes the case. But if you access an electronic document, even for a split second, you irrevocably alter its metadata. So when metadata matters, beware the IT guy who volunteers to "ghost" the drive or run searches. Run—don't walk—to engage a properly trained expert to create a forensically qualified image or clone of the evidence.

#### **8. Failing to share sufficient information or build trust with the other side.**

The judges are serious about this meet-and-confer business. You can't complain about the other side's demand to see everything if you're playing hide the ball. EDD-savvy requesting parties appreciate the futility of "any-and-all" requests, but how can they seek less if you keep them in the dark about the who, what and where of your client's electronically stored information? Surviving the mutually assured destruction scenario for EDD means building trust and opening lines of communication. The EDD meet-and-confer isn't the place for posturing and machismo. Save it for court.

#### **9. Letting fear displace reason.**

Don't let an irrational fear of sanctions rob you of your good judgment. Clients don't have to keep everything. Judges aren't punishing diligent, good-faith efforts gone awry. Your job is to help manage risk, not eliminate it altogether. Do your homework, talk to the right folks, document your efforts and be forthcoming and cooperative. Then, if it then feels right, it probably is.

#### **10. Kidding ourselves that we don't need to learn this stuff.**

O.K., you went to law school because you didn't know enough technology to change the batteries on a remote control. This English major feels your pain. But we can't very well try

lawsuits without discovery, and we can't do discovery today without dealing with electronically stored information.

You don't want to work through an expert forever, do you? So, we have to learn enough about EDD to advise clients about preservation duties, production formats, de-duplication, review tools, search methodologies and the other essential elements of e-discovery. Our clients deserve no less.

## Ten Tips to Clip the Cost of E-Discovery

by Craig Ball

*[Originally published in Law Technology News, September 2006]*

E-discovery costs *less* than paper discovery. Honest. *In comparable volumes*, it's cheaper to collect, index, store, copy, transport, search and share electronically stored information (ESI). But we hoard data with an indiscriminate tenacity we'd label "mental illness" if we were piling up paper. It's not just that we keep so *much*; it's that our collections are so *unstructured*. Squirrel away twenty years of National Geographic with an index and you're a "librarian." Without the index, you're that "crazy cat lady."

So the number one way to hold down the cost of e-discovery is:

### 1. If you don't need to keep it, *get rid of it*

Preservation obligations aside, if you're keeping backup tapes you don't need for disaster recovery or that you can't *read* because you no longer have the hardware or software, *get rid of them*. The same holds for all those old computers, hard drives, floppies, CD-ROMs, Zip disks and former e-mail accounts. Don't stick tapes in a closet intending to someday wipe and sell them on e-Bay. *If they don't hold information you must retain*, wipe them, shred them or pulverize them *now*.

### 2. Get tough on e-mail

E-mail *should* be easy. It's got those handy subject lines. It's electronically searchable. The circulation list's right up front. It's a cinch to file.

In reality, e-mail conversations (*threads*) veer off topic, search is a hit-or-miss proposition (*CUL8R*), addresses are cryptic (*HotBob37@aol.com*) and only the most organized among us (*anal-retentive*) file e-mail with anything like the effort once accorded paper correspondence. Personal messages rub elbows with privileged communications, spam and key business intelligence.

During WWII, everyone knew, "Loose lips sink ships." But does every employee appreciate the risk and cost of slipshod e-mail? Get tough on e-mail through policy, then train, audit and



enforce. Train to manage e-mail, appreciate that *messages never die* and know that hasty words are eaten under oath. Tame the e-mail beast and the rest is easy.

### **3. Have a data taxonomy and standardize storage**

Paper discovery cost less, in part because we generated and retained less paper, but also because we did a better job managing paper. We didn't search everywhere because there was always a file, folder or cabinet where we kept "that stuff." That's the power of place.

Records management isn't a form of personal expression. We must restore the elements of good records management to ESI. Want a desktop background with puppies? *Fine, but you must use the company's folder structure and naming protocols.* Want to send an e-mail? *No problem, but if it's personal, you must designate it as such, and if not, you must assign it a proper place within the company's information management system.*

### **4. Trim ESI requiring attorney review**

The costliest phase of e-discovery is attorney review, so big savings flow from shrinking the volume of ESI reviewed, shifting the review burden to the other side and using cheaper talent.

Pare review volume by filtering and de-duplication to cull non-responsive data *before* attorney review, and work with the other side to identify irrelevant file types and target discovery to specific custodians and date ranges. Discovery rules permit production of ESI as maintained in the usual course of business, so consider leaving review to the opposition, protecting privileged content through claw back agreements. Finally, must a local attorney pore over everything, or can some of the work be done by legal assistants or outsourced to lower-cost lawyers in Indiana or India?

### **5. Keep responsive ESI on the servers**

Between road warriors, at-home workers, local drives and smart phones, ESI has gone off the reservation, straying beyond the confines of the company's servers. Harvesting maverick data is costly, so employ policy and technology to insure that responsive data stays on the servers where it's more efficiently secured, searched and backed up.

### **6. No new gadgets without an e-discovery plan and budget**

Everyone loves new toys, but the price tag on the latest PDA, messaging system or software won't reflect the costs it adds to e-discovery. You don't have to give up gadgets, but factor their impact on e-discovery into the total cost of ownership, and be sure preserving and harvesting their contents is part of your e-discovery plan.

### **7. Build cross-enterprise search and collection capability**

Harvest is e-discovery's second costliest component. Eliminating onsite collection adds up to major savings. Emerging technologies make it possible to remotely search and harvest ESI

from all machines on a network. Though still in its infancy, cross-enterprise search and collection makes sense for serial litigants and large workforces.

#### **8. Develop in-house harvest expertise**

If you want to destroy evidence, ask the IT guy to preserve it. Forensically sound preservation isn't the same as copying, Ghosting or backing up. It demands special tools and techniques. Oil well firefighter Red Adair put it well: "If you think it's expensive to hire a professional, wait until you hire an amateur!"

Learning to be a computer forensic *examiner* is hard, but learning to do forensically sound *acquisitions* isn't. You'll preserve more data than you'll analyze, so having an IT staffer trained in forensically sound preservation saves money on outside experts...and spoliation sanctions

#### **9. Know the component cost of vendor services**

Though e-discovery vendors tout "proprietary technologies," all use pretty much the same prosaic processes. Still, some are especially efficient at particular tasks (like tape restoration or scanning) and price these services competitively. When you understand the pieces of ESI processing and what each adds to the bill, you can match the task to the best-qualified vendor and get the best price.

#### **10. Work cooperatively with the other side**

This tip saves more than the others combined. Being forthright about your ESI and transparent in your e-discovery methodology fosters the trust that enables an opponent to say, "You don't have to produce that." The e-discovery horror stories—the ones that end with sanctions—all start with, "Once upon a time, there was a plaintiff and a defendant who couldn't get along."

### **Copy That?**

**by Craig Ball**

***[Originally published in Law Technology News, October 2006]***

One of the frustrating things about e-discovery is that two lawyers discussing preservation will use the same words but mean entirely different things. Take "copying." When a producing party agrees to copy a paper document, there's rarely a need to ask, "What method will you use," or "Will you copy the entire page?" It's understood they'll capture all data on both sides of the page and produce a duplicate as nearly equivalent as possible to the original.

But when data is stored electronically, "making a copy" is susceptible to meanings ranging from, "We'll create a forensically sound, authenticated image of the evidence media, identical in the smallest detail," to "We'll duplicate some parts of the evidence and change other parts to substitute misleading information while we irreparably alter the original." Of course, nobody defines "making a copy" the latter way, but it's an apt description of most data copying efforts.

Unlike paper, electronically stored information (ESI) always consists of at least two components: a block of data called a file and at least one other block of data containing, inter alia, the file's name, location and its last modified, accessed, and created dates (MAC dates) of the file. This second block, called system metadata, is often the only place from which the file name, location and dates can be gleaned. Anyone working with more than a handful of files appreciates the ability to sort and search by MAC dates. Take away or corrupt system metadata and you've made ESI harder to use.

So, copying a file means more than just duplicating the data in the file. It also means picking up the system metadata for the file stored in the disk's "Master File Table" or "File Allocation Table."

The good news is that Microsoft Windows automatically retrieves both the file and its system metadata when copying a file to another disk. The bad news is that Windows automatically changes the creation date of the duplicate and the last access date of the original to the date of copying. The creation date changes because Microsoft doesn't use it to store the date a user authored the contents of the file. Instead, Creation Date denotes the date on which the file was created on the particular medium or system housing it. Copying a file re-creates it. Spoliation *and* misrepresentation in a click!

### **But wait! It gets worse.**

Floppy disks, thumb drives, CDs, and DVDs don't use the same file systems as hard drives running Windows. They don't record the same system metadata in the same way. If a Windows computer is an old roll-top desk with many small drawers and pigeonholes to hold file metadata, then a thumb drive or recordable CD is a modern desk with just a few. If you try to shift the contents of the roll-top to the modern desk, there aren't as many places to stash stuff. Likewise, file systems for floppy disks, thumb drives, CDs, and DVDs aren't built to store the same or as many metadata values for a file as Windows. So, when a file is copied from a hard drive to a thumb drive, floppy disk or optical media, some of its system metadata gets jettisoned and only the last modified value stays aboard. That's bad.

Now, copy the data from the thumb drive, floppy or optical media back to a Windows machine and the operating system has a bunch of empty metadata slots and pigeonholes to fill. Not receiving a value for the jettisoned system metadata, it simply makes something up! That is, it takes the last modified date and uses it to fill both the slot for last modified date and the slot for last accessed date. That's worse. So, if we can't copy a file by...copying it, what do we do?

The answer is that you have to use tools and techniques designed to preserve system metadata or you must record the metadata values before you alter them by copying. Various tools and techniques exist to duplicate files on Windows systems without corrupting metadata. One that Windows users already own is Microsoft Windows Backup. If you have Windows XP Pro

installed, you'll probably find Windows Backup in Accessories>System Tools. If you use Windows XP Home Edition, Windows Backup wasn't automatically installed, but you can install it from valueadd/MSFT/ntbackup on your system CD.

So far, we've talked only about copying a file and its system metadata. But each file comes from a complex environment containing lots of data illuminating the origins, usage, manipulation and even destruction of files. Some of this information is readily accessible to a user, some is locked by the operating system and much more is inaccessible to the operating system, lurking in obscure areas such as "unallocated clusters" and "slack space." When you copy a file and its metadata, all of this information is left behind. Even if you copy all the active files on the hard drive, you won't preserve the revealing latent data. To do that, you have to go deeper than the operating system and create a forensically sound copy.

The classic definition of a forensically sound copy is that it's an authenticable duplicate of a storage medium by a method that doesn't alter the source and reflects or can reliably reconstruct every readable byte and sector of the source with nothing added, altered or omitted. It's a physical, rather than a logical duplicate of the original.

A forensically sound copy may be termed a clone, drive image, bit stream duplicate, snapshot or mirror. As long as the copy is created in a way that preserves latent information and can be reliably authenticated, the name doesn't matter, though drive image denotes a duplicate where the contents of the drive are stored or compressed in one or more files which can be reconstituted as a forensically sound copy, and some use snapshot to mean a full system backup of a server that doesn't preserve latent data.

Beware the misguided use of the Symantec Corp.'s Ghost or other off-the-shelf duplication programs. Though it's possible to create a forensically sound drive clone with Ghost, I've never seen it done correctly in the wild. Instead, IT personnel invariably use Ghost in ways that don't preserve latent data and alter the original. Usually this flows from ignorance; occasionally, it's an intentional effort to frustrate forensic examination.

There is no single approved way to create a forensically sound copy of a drive. Several hardware and software tools are well suited to the task, each with strengths and weaknesses. Notables include Guidance Software Inc.'s EnCase, the no-cost Linux "dd" (data dump) function, AccessData Corp.'s Forensic Toolkit, X-Ways Software Technology AG's X-Ways Forensics, Paraben Corp.'s Replicator and drive duplication devices from Intelligent Computer Solutions Inc. and Logicube Inc. There are many different types of digital media out there, and a tool appropriate to one may be incapable of duplicating another. You have to know what you're doing and select the correct application for the job.

And there's the takeaway: Not all copies are created equal. Successful preservation of ESI hinges not only on selecting the tools, but also on your planning and process, e.g., defining your

goals, protecting the chain of custody, authenticating the duplicate, documenting the effort and understanding the consequences of your chosen method. Copy that?

## **In Praise of Hash**

**by Craig Ball**

*[Originally published in Law Technology News, November 2006]*

I love a good hash. Not the homey mix of minced meat and potato Mom used to make. I mean *hash values*, the results of mathematical calculations that serve as reliable digital “fingerprints” of electronically stored information. If you haven’t come to love hash values, you will, because they’re making electronic discovery easier and less costly.

Using hash algorithms, any amount of data—from a tiny file to the contents of entire hard drives and beyond—can be uniquely expressed as an alphanumeric sequence of fixed length.

The most common forms of hashing are MD5 and SHA-1. The MD5 hash value of Lincoln’s Gettysburg Address is E7753A4E97B962B36F0B2A7C0D0DB8E8. Anyone, anywhere performing the same calculation on the same data will get the same unique value in a fraction of a second. But change “Four score and seven” 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 “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.

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 and supports processes vitally important to electronic discovery, including identification, filtering, Bates numbering, authentication and de-duplication.

### **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.

Hash algorithms 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. Thus, hashing helps

resolve how to search for stolen data on a competitor's systems without either side revealing trade secrets. It's done by comparing hash values of their files against hash values of your proprietary data. The hash values reveal nothing about the contents of the files except whether they match. It's not a foolproof solution because altered data present different hash values, but it's sometimes a sufficient and minimally intrusive method. A match conclusively establishes that purloined data resides on the competitor's system.

### **Filtering**

Matching to known hash values simplifies e-discovery and holds down costs by quick and reliable exclusion of irrelevant data from processing and search. Matching out-of-the-box values for entire operating systems and common applications like Microsoft Windows or Intuit's Quicken, culls huge chunks of patently irrelevant files from consideration without risk of overlooking relevant information excluded based on location or file extension. Hashing thwarts efforts to hide files by name change or relocation because hash-matching flushes out a file's true nature--so long, that is, as the contents of the file haven't changed.

### **Bates Numbering**

Hashing's ability to uniquely identify e-documents makes it a candidate to replace traditional Bates numbering in electronic production. Though hash values don't fulfill the sequencing function of Bates numbering, they're excellent unique identifiers and enjoy an advantage over Bates numbers because they eliminate the possibility that the same number might attach to different documents. An electronic document's hash value derives from its contents, so will never conflict with that of another document unless the two are identical.

### **Authentication**

I regularly use hashing to establish that a forensically sound duplicate of a hard drive faithfully reflects every byte of the source and to prove that my work hasn't altered the original evidence.

As e-discovery gravitates to native production, concern about intentional or inadvertent alteration requires lawyers to have a fast, reliable method to authenticate electronic documents. Hashing neatly fills this bill. In practice, a producing party simply calculates and records the hash values for the items produced in native format. Once these hash values are established, the slightest alteration of the data would be immediately apparent when hashed.

### **De-duplication**

In e-discovery, vast volumes of identical data are burdensome and pose a significant risk of conflicting relevance and privilege assessments. Hashing flags identical documents, permitting one review of an item that might otherwise have cropped up hundreds of times. This is de-duplication, and it drastically cuts review costs.

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) may frustrate de-

duplication when hashing an entire e-document. An alternative is to hash relevant *segments* of e-documents to assess their relative identicality, a practice called “near de-duplication.”

### **Here’s to You, Math Geeks**

So this Thanksgiving, raise a glass to the brilliant mathematicians who dreamed up hash algorithms. They’re making electronic discovery and computer forensics a whole lot easier and less expensive.

## **Unlocking Keywords**

**by Craig Ball**

*[Originally published in Law Technology News, January 2007]*

The notion that words hold mythic power has been with us as long as language.

We know we don't need to ward off evil spirits, but we still say, "Gesundheit!" when someone sneezes. Can't hurt.

But misplaced confidence in the power of word searches can seriously hamper electronic data discovery. Perhaps because keyword searching works so well in the regimented realm of automated legal research, lawyers and judges embrace it in EDD with little thought given to its effectiveness as a tool for exploring less structured information. Too bad, because the difference between keyword searches that get the goods and those that fail hinges on thoughtful preparation and precaution.

### **Text Translation**

Framing effective searches starts with understanding that most of what we think of as textual information isn't stored as text. Brilliant keywords won't turn up anything if the data searched isn't properly processed.

Take Microsoft Outlook e-mail. The message we see isn't a discrete document so much as a report assembled on-the-fly from a database. As with any database, the way information is stored little resembles the way we see it onscreen after our e-mail program works its magic by decompressing, decoding and decrypting messages.

Lots of evidence we think of as textual isn't stored as text, including fax transmissions, .tiff or PDF documents, PowerPoint word art, CAD/CAM blueprints, and zip archives. For each, the search software must process the data to insure content is accessible as searchable text.

Be certain the search tool you or your vendor employ can access and interpret all of the data that should be seen as text.



## **Recursion**

Reviewing a box of documents that contains envelopes within folders, you'd open everything to ensure you saw everything.

Computers store data within data such that an Outlook file can hold an e-mail transmitting a zip archive containing a PowerPoint with an embedded .tiff image.

It's the electronic equivalent of Russian nesting dolls. If the text you seek is inside that .tiff, the search tool must drill down through each nested item, opening each with appropriate software to ensure all content is searched. This is called recursion, and it's an essential feature of competent search. Be sure your search tool can dig down as deep as the evidence.

## **Exceptions**

Even when search software opens wide and digs deep, it will encounter items it can't read: password protected files, proprietary formats, and poor optical character recognition. When that happens, it's important the search software generates an exceptions log flagging failures for follow up.

Know how the search tool tracks and reports items not searched or incompletely searched.

## **Search Term Tips**

So far, I've talked only about search tools; but search terms matter, too.

You'll get better results when you frame searches to account for computer rigidity and human frailty. Some tips:

**Stemming:** Computers are exasperatingly literal when searching. Though mechanized searches usually overlook differences in capitalization, they're easily confounded by variances in prefixes or suffixes of the sort that human reviewers easily assimilate (e.g., flammable and inflammable or exploded and exploding).

You'll miss fewer variations using stemmed searches targeting common roots of keywords; e.g., using "explod" to catch both exploded and exploding.

But use stemming judiciously as the more inclusive your search, the more challenging and costly the review. Be sure to include the correct stemming operator for the search tool.

**Boolean Search:** Just as with legal research, pinpoint responsive items and prioritize review using Boolean operators to find items containing both of two keywords, or keywords within a specified proximity.

**Misspelling:** It's scary how many people can't spell. Even the rare good speller may hit the wrong key or resort to the peculiar shorthand of instant messaging.

Sometimes you can be confident a particular term appears just one way in the target documents—e-mail addresses are prime examples—but a thorough search factors in common misspellings, acronyms, abbreviations and IM-speak.

**Synonyms:** Your search for "plane" won't get off the ground if you don't also look for "jet," "bird," "aircraft," "airliner" and "crate."

A comprehensive search incorporates synonyms as well as lingo peculiar to those whose data is searched.

**Noise words:** Some words occur with such regularity it's pointless to look for them. They're "noise words," the static on your ESI radio dial.

I recently encountered a situation where counsel chose terms like "law" and "legal" to cull data deemed privileged. Predictably, the results were disastrously overinclusive.

I recommend testing keywords to flush out noise words. There's irrelevant text all over a computer—in spelling dictionaries, web cache, help pages, and user license agreements. Moreover, industries have their own parlance and noise words, so it's important to assess noisiness against a representative sample of the environment you're searching.

Noise words are particularly nettlesome in computer forensic examinations, where searches extend beyond the boundaries of active files to the wilds of deleted and fragmented data. Out there, just about everything has to be treated as a potential hiding place for revealing text.

Because computers use alphabetic characters to store non-textual information, billions or trillions of characters randomly form words in the same way a million typing monkeys will eventually produce a Shakespearean sonnet. The difference is that the monkeys are theoretical while there really are legions of happenstance words on every computer. Consequently, searching three- and four-letter terms in forensic examinations—e.g., "IBM" or "Dell"—can be a fool's errand requiring an examiner to plow through thousands of false hits. If you must use noisy terms, it's best to frame them as discrete occurrences (flanked by spaces) and in a case-specific way (IBM but not iBm).

### **Striking a Balance**

Effective keyword searching demands more than many imagine. You don't have to put every synonym and aberrant spelling on your keyword list, but you need to appreciate the limits of text search and balance the risk of missing the mark against the burden of grabbing everything and

the kitchen sink. The very best results emerge from an iterative process: revisiting potentially responsive data using refined and expanded search terms.

## **Getting to the Drive**

**by Craig Ball**

***[Originally published in Law Technology News, April 2007]***

Traditionally, we've relied on producing parties to, well, *produce*. Requesting parties weren't entitled to rifle file cabinets or search briefcases. When evidence meant paper documents, relying on the other side's diligence and good faith made sense. Anyone could read paper records, and when paper was "deleted," it was gone.

But, as paper's given way to electronically stored information (ESI), producing parties lacking computer expertise must blunder through or depend upon experts to access and interpret the evidence. Lawyers get disconnected from the evidence. When discoverable ESI resides in places the opposition can't or won't look, how can we accept a representation that "discovery responses are complete?" When there's a gaping hole in the evidence, sure, you can do discovery about discovery, but sometimes, you've just got to "get to the drive."

"Getting to the drive" means securing forensically qualified duplicates of relevant computer disk drives used by the other side, and having them examined by a qualified expert. Often lumped together, it's important to consider these tasks independently because each implicates different concerns.

When not writing or teaching, I examine computer hard drives voluntarily surrendered by litigants or pried from their fingers by court order. Serving as neutral or court-appointed special master, my task is to unearth ESI bound up with privileged or confidential content, protecting the competing interests of the parties. The parties can separate wheat from chaff for conventional, accessible data, but when the data's cryptic, deleted or inaccessible, I'm brought in to split the baby.

Increasingly, I see lawyers awakening to the power of computer forensics and wanting access to the other side's drives, but unsure when it's allowed or how to proceed. Some get carried away.

In a recent Federal District Court decision, *Hedenburg v. Aramark American Food Services*, 2007 WL 162716 (W.D. Wash.), the defendant in a discrimination and wrongful termination case suspected the plaintiff's e-mail or internet messaging might be useful for impeachment concerning her mental state. Apparently, Aramark didn't articulate more than a vague hunch, and Hedenburg dubbed it a "fishing expedition."

Judge Ronald Leighton denied access, analogizing that, "If the issue related instead to a lost paper diary, the court would not permit the defendant to search the plaintiff's property to ensure that her search was complete."

True enough, and the right outcome here, but what if a credible witness attested to having seen the diary on the premises, or the plaintiff had a history of disappearing diaries? What if injury or infirmity rendered the plaintiff incapable of searching? On such facts, the court might well order a search.

In weighing requests to access hard drives, judges should distinguish between the broad duty of preservation and the narrower one of production. It's not expensive to preserve the contents of a drive by forensic imaging (comparable in cost to a half-day deposition transcript), and it permits a computer to remain in service absent concerns that data will be lost to ongoing usage.

A drive can be forensically imaged without the necessity of anyone viewing its contents; so, assuming the integrity of the technician, no privacy, confidentiality or privilege issues are at stake. Once a drive image is "fingerprinted" by calculating its hash value (See, LTN Nov. 2005), that value can be furnished to the court and the other side, eliminating potential for undetected alteration.

Considering the volatility of data on hard drives and the fact that imaging isn't particularly burdensome or costly, courts shouldn't hesitate to order forensically-qualified preservation when forensic examination is foreseeable. In contrast, such forensic examination and production is an expensive, intrusive, exceptional situation.

Hard drives are like diaries in how they're laced with intimate and embarrassing content alongside discoverable information. Drives hold privileged spousal, attorney and health care communications, not to mention a mind-boggling incidence of sexually-explicit content (even on "work" computers). Trade secrets, customer data, salary schedules, passwords abound.

So how does a court afford access to the non-privileged evidence without inviting abuse or exploitation of the rest? An in-camera inspection might suffice for a diary, but what judge has the expertise, tools, and time to conduct an in-camera computer forensic examination?

With so much at stake, courts need to approach forensic examination cautiously. Granting access should hinge on demonstrated need and a showing of relevance, balanced against burden, cost or harm. It warrants proof that the opponent is either incapable of, or untrustworthy in, preserving and producing responsive information, or that the party seeking access has some proprietary right with respect to the drive or its contents. Showing that a party lost or destroyed ESI is a common basis for access, as are situations like sexual harassment or data theft where the computer was instrumental to the alleged misconduct.

Of course, parties often consent. Seeking to prove your client has "nothing to hide" by granting the other side unfettered access to computers is playing Russian roulette with a loaded gun. You won't know what's there, and if it's sufficiently embarrassing, your client won't tell you. Instead, the cornered client may wipe information and the case will turn on spoliation and sanctions.

Orders granting examination of an opponent's drive should provide for handling of confidential and privileged data and narrow the scope of examination by targeting specific objectives. The examiner needs clear direction in terms of relevant keywords and documents, as well as pertinent events, topics, persons and time intervals. A common mistake is to agree upon a search protocol or secure an order without consulting an expert to determine feasibility, complexity or cost. The court should encourage the parties to jointly select a qualified neutral examiner as this will not only keep costs down but will also help ensure that the agreed-upon search protocol is respected.

Getting to the drive isn't easy, nor should it be. When forensics may come into play, e.g., cases of data theft, spoliation and computer misuse, demand prompt, forensically-sound preservation. When you want to look, be ready to show good cause and offer appropriate safeguards.

## **Who Let the Dogs Out?**

**by Craig Ball**

***[Originally published in Law Technology News, May 2007]***

What is evidence? I won't quote *Black's Law Dictionary* or *McCormick on Evidence*, partly because I boxed mine when online legal research made my library obsolete, and because my well-thumbed copies inhabited a time when evidence was largely a thing or statement. We examined things. Witnesses made statements.

After law school and apart from the occasional trial, lawyers rarely reflect on the nature of evidence. Like pornography, we know it when we see it. But with electronic evidence, we hardly see it anymore. No longer can we open a file drawer and wade in.

Now, we rely on experts and technicians using searches and filters to troll roiling oceans of data and process the catch of the day. By the time lawyers "see" electronic evidence, it's frozen fish sticks and canned tuna. Sorry, Charlie McCormick, 21st century lawyers don't go near the water.

### **Rethinking Assumptions**

Fundamentals of evidence mastered in law school are still helpful, but some electronically stored evidence is so foreign to traditional assumptions that we need to rethink them. Who is charged

with its content and custody? What's an original? How do we authenticate it? When/how do we allow its use?

We still expect lawyers to know the evidence in their cases and produce it, but electronic evidence forces counsel to rely on crude tools and methodologies and work through technical intermediaries of uneven ability who speak in acronyms and jargon. Lawyers are increasingly so disconnected from the evidence that when we search for evidence, we tend to find only what we seek instead of what's there to be found.

I see this glaringly manifested by colleagues who regard a text search for a handful of keywords as a sufficient effort. Just because Lexis or Westlaw make you feel like the Amazing Kreskin, a seat-of-the-pants keyword search in unstructured data is a whole different kettle of fish.

Ever run a pack of bloodhounds to find a fugitive? Me neither, but we've *seen* it a million times in old movies. Outskirts of city at night. Hardboiled detective hands tattered shirt sleeve to dog wrangler. Ol' Blue sniffs the rag. "Go git 'em, boy." Cut to thick forest. Baleful "roof, roof, a-roof" signals auspicious time to wade down fortuitously encountered stream and throw off scent. Segue to confused hound. Fade to shot of grinning anti-hero sipping Mojitos with Brazilian beauty on Ipanema Beach. Roll credits.

We didn't see Blue bounding by his quarry's e-ticket confirmation to Rio and the thumb drive storing offshore account numbers. It wasn't a bad search, it was just too single-minded.

### **Form Above Substance**

Processing volume in this narrow way without assimilating it is emblematic of the lengths we go to elevate form above substance. Hacking through terabytes of data, we've become the child squinting at the scary parts of the movie through hands over our eyes, looking as narrowly as possible at the content.

Too cavalier about locating responsive evidence, we are disproportionately obsessed with inadvertent production of privileged information—to the point that much of the time and cost of e-discovery is consumed by the effort.

Are confidential attorney-client communications really so much a part of every custodian's data that e-discovery must slow to a costly crawl? If so, we need to encapsulate and tag these privileged items at the time they're created to isolate them from mainstream electronically stored information. Better to treat lawyers like vestal virgins than let the taint of their work bloat the cost and complexity of review.

When will we see that clients self-immolate far more often through incomplete production than inadvertent production?

We need to devote more time to thinking about what the evidence is instead of where it lodges. Too often, we fixate on the containers—the e-mail, spreadsheets and databases—with insufficient regard for the content. This isn't just a rant against producing parties. I see the failure as well in requesting parties determined to get to the other side's tapes and hard drives, but unable to articulate what they're seeking.

Saying, "I want the e-mail" is as meaningless as saying, "I want the paper." E-mail, voicemail, ledgers or lipstick on the mirror are just media used to hold and convey information. It's the transaction and the content that make them evidence.

The form matters, but only for reasons of accessibility (Can I view or hear it?), preservation (How do I protect it?), utility (Can I search and sort it?), completeness (Is something added or absent?) and authentication (Can I rely on it?).

Pondering the essential nature of evidence can't remain the exclusive province of law review commentators and law school professors. As never before, trial lawyers in the trenches must think hard about just what is the evidence? What are we really looking for? What gets us closer to the truth?

## **Page Equivalency and Other Fables**

**by Craig Ball**

*[Originally published in Law Technology News, August 2007]*

When the parties to a big lawsuit couldn't agree on a vendor to host an electronic document repository, the court appointed me to help. Poring over multimillion dollar bids, I saw the vendors were told to assume that a gigabyte of data equals 22,500 pages. If the dozens of entities involved produced their documents in a mix of .tiff images and native formats—spreadsheets, word processed documents, e-mail, compressed archives, maps, photos, engineering drawings and more—how sensible, I wondered, was it to assume 22,500 pages per gig?

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 number of pages unless you know the composition of the data. Even then, it's a leap of faith.

I've been railing against page equivalency claims for years because they're so elusive and often abused to misstate the burden and cost of electronic data discovery.

*"Your Honor, Megacorp's employees each have 80 GB laptops. That means we will have to review 40 million pages per machine. Converting those pages to .tiff images will cost Megacorp*



*4 million dollars per laptop.”*

Nonsense!

If you troll the internet for page equivalency claims, you'll be astounded by how widely they vary, though each is offered with utter certitude. A GB 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, ExxonMobil representatives twice asserted that one GB yields 500,000 typewritten pages. The National Conference of Commissioners on Uniform State Laws proposes to include 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 GB value like a bad cold.

Yet, 500,000 pages per GB isn't right. It's not even particularly close to right.

Several years ago, my friend Kenneth Withers, now with The Sedona Conference and then e-discovery guru for the Federal Judicial Center, 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. Withers, who owned up to the error with his customary grace and candor, has contributed so much wisdom to the bench and bar that he can't be faulted. But the echoes of that innocent thousand-fold miscalculation still reverberate today. 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 GB—we're still talking an equivalency with all the credibility of an Elvis sighting.

Now, with more e-discovery miles in the rear view mirror, it's clear we've got to 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 lifespan of the media and the evolution of particular 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 8½"x11" documents scanned at 300 dpi equals approximately 18,000 pages."

Consider the column you're reading. In plain text, it's a file just 5 kilobytes in size and prints as one to two typewritten pages. As a rich text format document, the file quadruples to 20 KB. The

same text as a Microsoft Word document is 25 KB. Converted to a .tiff image, it's 123 KB without an accompanying load file. Applying a page equivalency of 500,000 pages per GB, a vendor using per page pricing may quote this column as being anything from one page to as many as 61 pages. Billed by the GB, you'll pay almost five times more for the article as two .tiff pages than as a native Word document. A flawed page equivalency hits the bottom line...hard.

So how many pages are in a gigabyte of data? Lawyers know this answer: *it depends*. To know, perform a data biopsy of representative custodians' collections and *gauge*—don't guess—page volume.

## **Re-Burn of the Native**

**by Craig Ball**

***[Originally published in Law Technology News, September 2007]***

I could hear the frustration in her voice. “We keep going back and forth with the plaintiff’s lawyer. I don’t understand what he wants. Can you help us?”

Defense counsel was trying to satisfy an opponent bent on getting e-mail in “native file format.” With each disk produced, the plaintiff’s lawyer demanded, “Where’s the e-mail?” Now he was rattling the sanctions saber. Poring over copies of what she’d produced, defense counsel saw the e-mail. “Why can’t he see it?”

Reviewing the correspondence between counsel, I spotted the problem. The e-mail was there, but in Rich Text Format. Like many lawyers new to e-discovery, defense counsel regarded electronically stored information and native data as one-and-the-same. They’re not.

The IT department had dutifully located responsive e-mail on the mail server and furnished the messages in a generic format called Rich Text Format or “RTF.” It’s a format offering full access to the contents of the messages, and it’s electronically searchable. Any computer can read RTF files. So, it’s a pretty good production format.

But, it’s not the native format.

### **Container Files**

The native format for virtually all enterprise e-mail is a *container file* lumping together relevant, irrelevant, personal and privileged communications, along with calendar data, to-do lists, contact information and more.

The precise native format depends upon the e-mail client and server. The prevailing enterprise e-mail application, Microsoft’s Exchange Server, uses a container file with the file extension

.EDB. Lotus Notes stores its e-mail on a Lotus Domino server in a container file with the extension .NFS. These containers are the “native file format” for server-stored e-mail, but they hold not only all then-existing e-mail for a specific user, but also the e-mail and other data for ALL users. Furnishing these files is tantamount to letting the opposition rifle every employee’s desk.

When enterprise e-mail is stored locally on a desktop or laptop system, it’s almost always in a container file, sometimes called a *compound file*. For users of Microsoft’s Outlook e-mail program (a “client application” in geek speak), the local container file is typically called “Outlook.PST” or “Outlook.OST.” There may also be a file holding older e-mail called “Archive.PST.” Collectively, these data are commonly referred to as a user’s “local PST.”

Like their counterparts on e-mail servers, local container files weave together the user’s responsive and non-responsive items with privileged and personal messages; consequently, they’re more like self-contained communications databases than paper correspondence folders.

### **Conundrum**

Because the native file format for enterprise e-mail is bound up with information beyond the scope of discovery, it’s the rare case where e-mail should be produced in its native format. Litigants must also be wary of producing native e-mail container formats because, until those containers are compacted by the client application, they hold information (like double deleted files) invisible to users but potentially containing privileged and confidential material. It’s possible to “mine” local PSTs for hidden data, and metadata scrubber tools offer no protection.

How, then, do we realize the considerable benefits of native production for e-mail? The answer lies in distinguishing between production of the native container file and production of responsive, non-privileged e-mail in electronically searchable formats that *preserve the essential function of the native source*, sometimes called *quasi-native* formats.

### **Quasi-Native Production**

Chockablock as it is with non-responsive material, there are compelling reasons not to produce “the” source PST. But there’s no reason to refuse to produce responsive e-mails and attachments *in the form* of a PST file, so long as it’s clearly identified as a reconstituted file containing selected messages and the contents fairly reflect the responsive content and relevant metadata of the original. Absent a need for computer forensic analysis or exceptional circumstances, a properly constructed quasi-native production of e-mail is an entirely sufficient substitute for the native container file.

It doesn’t have to be in PST format. There are several generic e-mail formats well suited to quasi-native production (e.g., .MSG and .EML formats). Even RTF-formatted production may suffice when paired with attachments, if the parties don’t need to search by discrete header fields (i.e., to sort by To, From, Subject, Date, etc.).

## Talk to Me

In the case at hand, the problem isn't one of intent or execution. It's miscommunication and misunderstanding. Plaintiff counsel saw only that he hadn't gotten the format he wanted. Defense counsel saw e-mail in an electronic format and assumed that it must be the right stuff. One fixed on form and the other on content. In e-discovery, both matter.

Accordingly, defense counsel will burn new disks containing the responsive e-mail in PST format.

So, talk to each other, and don't rely on buzzwords like "native file format" unless your meaning is clear. You'll be amazed how often the question, "What do you mean by native file format?" will be answered, "I have no idea. I just heard it was something I should ask for."

## The Power of Visuals

by Craig Ball

*[Originally published in Law Technology News, October 2007]*

Are we so up to our necks in electronic alligators that we've forgotten why we're in the swamp?

So it seemed as I spent two days on the stand in a little Texas town. The case concerned the alleged theft of trade secrets by former employees, and though the companies were small, the stakes weren't. It was Dickensian litigation—the sort of bitter, prolonged, expensive showdown where the only surefire winners are the lawyers and experts.

I was the first witness, and my challenge was to distill a wide-ranging computer forensic investigation into a succinct and compelling story—a task complicated by counsel's sketchy description of what he would cover on direct and cavalier approach to evidentiary foundations and the record. Both sides had fine lawyers, but neither appeared to have given much thought to how they would present or attack the electronic evidence.

At one point, the court asked, "Is it always this hard [to present electronic evidence]?"

"Yes," I answered, thinking, "But it doesn't have to be."

Every jury trial is an education. Here are lessons from this one:

**Lesson One:** Plan the direct examination. Tell the expert what you'll cover in time to marshal responsive data. When an expert can take the ball and run with it, just get out of the way and hope opposing counsel doesn't object to the narrative. But if your expert needs direction, or should the court sustain a narrative objection, have questions at hand, and be sure you know the witness' answers. If uncertain about how to elicit a key point, ask the witness to suggest the right question.

**Lesson Two:** Lay the proper foundation for admission. They haven't suspended the rules of evidence for bits and bytes. You still need to follow the MIAO rule (Mark, Identify, Authenticate and Offer) and be ready to meet hearsay objections. U.S. Magistrate Judge Paul Grimm's 101-page opinion in *Lorraine v. Markel American Ins. Co.*, 241 F.R.D. 534 (D. Md. 2007), thoroughly explores common foundations for electronic evidence.

**Lesson Three:** Make it Interesting. Enthusiasm is infectious, so counsel should convey that what the jury will hear and see is exciting, interesting and important. Then, the expert must deliver on counsel's promise, using simple descriptive language to build bridges to complex ideas.

Sure, you want experts credible enough that jurors will take them at their word, but the most effective experts equip the jury to share conclusions, not merely accept them.

Some testimony gets repeated every time an expert takes the stand. For a computer forensics examiner, data carving, hash authentication, metadata, and "why deleted doesn't mean gone" are routine topics. Your expert should be lively, practiced, and polished at explaining such things using incisive analogies and strong visuals.

Nothing hammers home the power of visual evidence like a trial. The most important takeaway:

**Lesson Four:** Engage the jurors visually. Paper records may be tedious, but they're tangible. You can hand them to a witness and wave them around on argument. Electronic evidence is gossamer absent something concrete to convey it.

To anchor electronic evidence, use the visual arsenal: icons, illustrations, time lines, graphs, charts, photos, printouts, animations, and screen shots. The Texas case hinged on the theft of computer-aided drafting and manufacturing data called CAD/CAM files. As a demonstrative, I created visually distinctive two-dimensional illustrations of the contents of key CAD/CAM files. Instead of hearing testimony about a file named abc-123-xyz.dwg jurors "saw" the file onscreen as I testified.

But no good deed goes unpunished. On cross, defense used the 2-D representations to secure my concession that his client "couldn't manufacture the product using just the drawing."

True, you can't build these widgets from the drawings alone, but electronic records go deeper than that. Counsel sought to keep me from adding that CAD/CAM files can contain layers of information detailing, e.g., three-dimensional characteristics, tolerances, and machining instructions—data deeper in the file that may, indeed, be all that's required to fabricate the part. Without re-direct unearthing this buried treasure, the jury may accord little value to the stolen data.

A few compelling visuals are better than a hundred reiterations. I focused on a drawing found on the defendant's computer bearing the plaintiff's logo, then prepared an animated Microsoft PowerPoint slide superimposing defendant's drawing on plaintiff's. The jury could see they were identical.

Be sure experts furnish visuals early enough that they won't unfairly surprise the other side. My night-before-trial exhibits proved invaluable, but they might have been excluded were I not a neutral examiner in the case.

### **Why We're in the Swamp**

Sometimes electronic discovery feels like an end in itself, but remember that it all comes down to trial. As you're identifying, preserving, collecting, searching, and producing electronically stored information, always consider, "*How will I present this in court?*"

## **Well Begun is Half Done**

**by Craig Ball**

*[Originally published in Law Technology News, November 2007]*

It's easy to feel overwhelmed by the daunting complexity of electronic discovery. There's so much to do in an arena where lawyers feel distinctly disadvantaged. We know we've got to hit the ground running, but so often we're paralyzed instead of galvanized. If only lawyers knew what to do first, certain of making the right choice.

Take heart. There is a reliably correct first step, and it's the identification of sources of electronic evidence. Do it well, and much of the fog hiding the hazards of e-discovery lifts. Pitfalls remain, but you're less likely to stumble into them.

Identification of electronically stored information (ESI) involves more than just a head count of machines, backup tapes, custodians, network storage areas, and thumb drives. Certainly, it's important to have a current inventory, but identification of potentially responsive sources of ESI goes deeper. You've got to know what you've got, who's got it, how much they have, where it is, and when it's going away.

Identification anticipates obligations imposed by the Federal Rules of Civil Procedure, such as Rule 26(a)(1)(B)'s requirement that litigants describe and supply the location of ESI going to claims or defenses and Rule 26(f)'s dictate that litigants discuss the forms of ESI. Then there's the duty to identify ESI claimed not reasonably accessible pursuant to Rule 26(b)(2)(B) or as privileged under Rule 26(b)(5)(A). Both must be identified with sufficient particularity to enable your opponent to gauge the merits of the objection.

If you can't properly identify the sources of ESI, you may be compelled to overproduce at enormous cost or run the risk of sanctions for failure to do so. That's not a Catch-22. It's an avoidable consequence of failing to do what the law requires.

Jump start the identification process by obtaining IT asset inventories and system diagrams. Most medium-size to large businesses track the acquisition, deployment, and disposal of computer systems. These assets tend to be depreciated for tax purposes, so the bean counters have to know when they come and go. Follow the money trail.

Similarly, IT departments often track deployment of systems and software for warranty, support and licensure, and they certainly track intranet connections and user privileges, if only to know where the wires from the patch panel lead! Check to see if the IT staff has a network map laying out the relationship between servers, users, business units and backup systems. Even an out-of-date network diagram is a leg up. Now, you're on the hardware and software trail.

Identify potentially responsive ESI along the people trail. Who are the persons most knowledgeable about the matters in contention? Pin down the principal software applications, data storage practices, devices, and media used by these key custodians. A phone call or e-mail may suffice to gather what you need, but better results flow from visits to the custodians' workplace and face-to-face interviews. Using a checklist tailored to the issues and computing environment is desirable, but don't let it get in the way of listening and observing.

It helps to lay eyes on the external hard drive or the mothballed system on the floor beside the desk. Ask about that stack of CDs on the shelf. Probe to find the pack rats. Remember: Even benign ESI hurts if you've sworn it doesn't exist.

Collect machine service tags and serial numbers, e-mail addresses, and user logon IDs. Record the overall capacity of hard drives along with their active data volume. Determine if there are local e-mail stores and archives on the machine, their file types, and sizes. Be sure to inquire about former machines, applications, and e-mail systems and to what extent legacy data migrated to current systems. Meet representations of, "That's gone," with, "How can you be certain?"

While identifying ESI, you're also collecting information about foreseeable threats to its integrity and existence. For backup media, you want to know the rotation cycle and anticipated changes to hardware and software. Explore whether desktop systems, laptops, or portable data storage devices are slated for replacement or modification. For e-mail servers and voicemail systems, pin down purge settings that dictate when and how deleted messages become unrecoverable.

Of course, it's not enough to identify when potentially relevant ESI will disappear. You've got to be poised to preserve it. Ensure that those identifying spoliation hazards are trained to react to them.

The goal of all this is to generate a spreadsheet or database allowing an evolving view of the lay of your client's data landscape by custodian, volume, location, and other criteria. Thus equipped, you can more reliably gauge the cost and complexity of e-discovery and implement right-sized preservation. Plus, you'll be better able to fulfill your "meet and confer" obligations and build trust with the other side.

So have no fear. Identification of ESI is always the right thing to do; and done well, it greases the wheels for the labors to follow.



# Ask the Right Questions

by Craig Ball

*[Originally published in Law Technology News, December 2007]*

Sometimes it's more important to ask the right questions than to know the right answers, especially when it comes to nailing down sources of electronically stored information, preservation efforts and plans for production in the FRCP Rule 26(f) conference, the so-called "meet and confer."

The federal bench is deadly serious about meet and confers, and heavy boots have begun to meet recalcitrant behinds when Rule 26(f) encounters are perfunctory, drive-by events. Enlightened judges see that meet and confers must evolve into candid, constructive mind melds if we are to take some of the sting and "gotcha" out of e-discovery. Meet and confer requires intense preparation built on a broad and deep gathering of detailed information about systems, applications, users, issues and actions. An hour or two of hard work should lay behind every minute of a Rule 26(f) conference. Forget "winging it" on charm or bluster, and forget, "We'll get back to you on that."

Here are 50 questions of the sort I think should be hashed out in a Rule 26(f) conference. If you think asking them is challenging, think about what's required to deliver answers you can certify in court. It's going to take considerable arm-twisting by the courts to get lawyers and clients to do this much homework and master a new vocabulary, but, there is no other way.

These 50 aren't all the right questions for you to pose to your opponent, but there's a good chance many of them are . . . and a likelihood you'll be in the hot seat facing them, too.

1. What are the issues in the case?
2. Who are the key players in the case?
3. Who are the persons most knowledgeable about ESI systems?
4. What events and intervals are relevant?
5. When did preservation duties and privileges attach?
6. What data are at greatest risk of alteration or destruction?
7. Are systems slated for replacement or disposal?
8. What steps have been or will be taken to preserve ESI?
9. What third parties hold information that must be preserved, and who will notify them?
10. What data require forensically sound preservation?
11. Are there unique chain-of-custody needs to be met?
12. What metadata are relevant, and how will it be preserved, extracted and produced?
13. What are the data retention policies and practices?
14. What are the backup practices, and what tape archives exist?

15. Are there legacy systems to be addressed?
16. How will the parties handle voice mail, instant messaging and other challenging ESI?
17. Is there a preservation duty going forward, and how will it be met?
18. Is a preservation or protective order needed?
19. What e-mail applications are used currently and in the relevant past?
20. Are personal e-mail accounts and computer systems involved?
21. What principal applications are used in the business, now and in the past?
22. What electronic formats are common, and in what anticipated volumes?
23. Is there a document or messaging archival system?
24. What relevant databases exist?
25. Will paper documents be scanned, at what resolution and with what OCR and metadata?
26. What search techniques will be used to identify responsive or privileged ESI?
27. If keyword searching is contemplated, can the parties agree on keywords?
28. Can supplementary keyword searches be pursued?
29. How will the contents of databases be discovered? Queries? Export? Copies? Access?
30. How will de-duplication be handled, and will data be re-populated for production?
31. What forms of production are offered or sought?
32. Will single- or multi-page .tiffs, PDFs or other image formats be produced?
33. Will load files accompany document images, and how will they be populated?
34. How will the parties approach file naming, unique identification and Bates numbering?
35. Will there be a need for native file production? Quasi-native production?
36. On what media will ESI be delivered? Optical disks? External drives? FTP?
37. How will we handle inadvertent production of privileged ESI?
38. How will we protect trade secrets and other confidential information in the ESI?
39. Do regulatory prohibitions on disclosure, foreign privacy laws or export restrictions apply?
40. How do we resolve questions about printouts before their use in deposition or at trial?
41. How will we handle authentication of native ESI used in deposition or trial?
42. What ESI will be claimed as not reasonably accessible, and on what bases?
43. Who will serve as liaisons or coordinators for each side on ESI issues?
44. Will technical assistants be permitted to communicate directly?
45. Is there a need for an e-discovery special master?
46. Can any costs be shared or shifted by agreement?

47. Can cost savings be realized using shared vendors, repositories or neutral experts?
48. How much time is required to identify, collect, process, review, redact and produce ESI?
49. How can production be structured to accommodate depositions and deadlines?
50. When is the next Rule 26(f) conference (because we need to do this more than once)?

For alternate views on the EDD topics to be addressed at a Rule 26(f) conference, Magistrate Judge Paul Grimm's committee's "Suggested Protocol for Discovery of ESI," ([www.mdd.uscourts.gov/news/news/ESIProtocol.pdf](http://www.mdd.uscourts.gov/news/news/ESIProtocol.pdf)), and the U.S.D.C. for the District of Kansas'"Guidelines for Discovery of Electronically Stored Information" ([www.ksd.uscourts.gov/guidelines/electronicdiscoveryguidelines.pdf](http://www.ksd.uscourts.gov/guidelines/electronicdiscoveryguidelines.pdf)).

## **Crystal Ball in Your Court**

**by Craig Ball**

*[Originally published in Law Technology News, January 2008]*

I glimpsed the future while mediating database discovery disputes in a recent multidistrict product liability matter. There was shuttle diplomacy over sample sizes and search terms. Collaborative documents memorialized hypertechnical agreements. IT experts darted in and out of arcane discussions about SAP, Oracle, e-rooms, and XML.

Because the parties came together like the happy tangle of cables snaking across the table to routers and outlets, I didn't have to don my Special Master cap and direct the outcome. Yet, I doubt there'd have been the same preparation and cooperation without a neutral presiding. Folks just behave better when company comes.

We will see more expert-mediated conferences as courts grapple with the technical intricacies of EDD and the inflated costs that dog inept efforts. It just makes economic sense. In large cases, EDD expenses alone can dwarf the entire amount in controversy in smaller cases; in any size case, EDD mistakes can determine outcomes. Why wouldn't you resolve foreseeable disputes before you bet the company?

As I gaze into my crystal ball, here are 17 more EDD predictions:

**1. *Virtual machines shine as a form of production for challenging ESI.*** When a level litigation playing field requires one side to see and manipulate ESI just as the other side can, it may seem a virtually impossible undertaking absent identical hardware and software. Now, it's "virtually" possible.

Because software code can emulate hardware, an entire virtual computer can exist within an onscreen window. These virtual machines look and function just like the real thing, at little cost. So tomorrow's e-production challenge—particularly of databases—may be met by delivering a virtual machine file containing relevant, non-privileged content in its native operating environment which the recipient loads and explores like its real-world counterpart.

The hurdles are legal more than technical. Software and operating system licensing must accommodate e-discovery when evidence is bound up with pricy programs, or courts should establish a litigation "fair use" exception.

**2. Fueled by virtualization, thin client computing returns.** Readers over 40 will remember thin client computing 1.0—those "dumb" terminals connected to mainframes. Thin client 2.0 is different because devices will perform some offline tasks; but expect to see local hard drives marginalized by rapid growth of virtualized applications tied to corporate networks and the internet.

**3. Personal data principally resides on portable media and the internet.** Data is the ultimate portable commodity, so it's odd we don't take our computing environments with us. We will. If desktop machines survive, they will be little more than screens with network connectivity temporarily hosting the virtual identities we carry in our pockets or store online. Local hard drives will be an increasingly irrelevant place to search for files as EDD turns to personal storage devices and online storage.

**4. We'll share a common EDD vocabulary.** You say potato and I say quasi-native. With princely sums riding on the outcome, shouldn't we mean exactly the same thing? Thanks to, e.g., The Sedona Conference, EDRM, blogs, and publications, there's progress afoot. Do your part. When someone mistakenly refers to "hash values" as "hash marks," rap them smartly on the snout with a rolled-up newspaper.

**5. Intelligent harvest mechanisms.** Though cross-network search and collection will flourish, expect to see "plug and pry" devices used by support staff (or dispatched to custodians) to suck up potentially responsive information via USB and other connections. Think "Ghostbusters" sans green slime.

**6. It'll cost less to store a terabyte of data than to buy a tank of gasoline.** At the rate these two benchmarks are diverging, expect this prediction to materialize within three years for an online terabyte. For the cost of a local terabyte, you'll fill a Hummer's tank twice.

**7. EDD custodial data volumes swell by three orders of magnitude.** Rocketing data volumes reflect the changing face of messaging, richer content, more complex applications and still-feasible increases in hard drive capacities coupled with still-plummeting cost-per-gigabyte.

**8. Routine production of system metadata.** As if a switch was flipped, we will wake to the realization that system metadata, such as file names, paths, and dates, are essential to managing e-records and wonder why we wasted time fighting about it. We'll bicker about application metadata until the other epiphany kicks in. Then, we'll rue the time and money wasted on .tiff productions when a sensible native or hybrid production would have been better and cheaper.

**9. Generic production containers for native and quasi-native production.** Some argue XML is the answer, and they're partly right; but you also can tuck a native file inside an Adobe .PDF file and enjoy the best of both formats. We need more generic production container options.

**10. Low-cost desktop review tools.** Generic production containers require tools to view, search, annotate, and redact their contents. Today, we buy Concordance and Summation or lease online review tools. Tomorrow, vendors will gravitate to the Adobe Reader model, giving away desktop review tools to profit from collecting and processing the ESI that is filling those containers.

**11. Hosted production takes hold.** We bank and do our taxes online. Soon we'll receive and review ESI the same way. It'll take time to gain lawyers' trust; longer still if a high-profile gaffe makes news.

**12. Widespread use of hashing for authentication and identification.** Better buy the bumper sticker that says, "They'll get my Bates stamp when they pry it from my cold, dead hands," because it's going the way of fax machines. Hashing isn't a complete substitute, but in certain ways, it's superior. Imagine near-instantaneous authentication of e-records of any size or complexity. Native production and hashing go together like cereal and milk.

**13. Data footprints of serial litigants become well-kept and well-known.** Oft-sued companies won't reinvent the wheel discovering their data footprint with each new case. They'll track it on an ongoing basis. Likewise, plaintiffs will share information on corporate ESI much as they share data on product defects and experts.

**14. Key-based encryption demarks and encapsulates privileged communications.** We expend fortunes ginning seeds of privilege from bales of ESI. If securely encrypted when created, privileged communications could be easily quarantined or just left alone. Everyone wants frictionless e-mail, but privileged communications that warrant special status oblige special handling.

**15. Backup tape usage wanes as costs drop and active sources proliferate.** Backup tape has outlived many who predicted its demise; but in five years, it will drop by 30% in favor of network mirroring.

**16. Location data routinely recorded and discovered.** Our cars and phones now track us, and soon GPS will be built into other products. When that data is relevant, we'll need to preserve and produce it.

**17. U.S. data privacy rights move closer to EU model.** In the European Union, where memories of genocide linger, data privacy is a fundamental human right. Stateside, plan on increased privacy push back with respect to harvesting and reviewing employee e-mail and other private ESI.

## **Redaction Redux**

**by Craig Ball**

***[Originally published in Law Technology News, February 2008]***

"The forceps of our minds are clumsy forceps," observed H. G. Wells, "and crush the truth a little in taking hold of it." Clumsier still is a method commonly used to redact information from electronically stored information—one that so crushes truth, it's alarming *anyone* defends it, let alone promotes it as a "standard."

I speak of redacting electronic documents by converting them to .tiff images, blacking out privileged and confidential content, then clumsily attempting to recreate electronic searchability by optical character recognition (OCR). When applied to spreadsheets and databases, it simply doesn't work. Why, then, are we content to spin invisible cloth rather than acknowledge the emperor's privates are on parade?

Good sense and fair play dictate that redaction methods preserve the integrity of unredacted content and the searchability and usability of the document. Instead, expediency and anxiety drive use of .tiff and OCR for redaction, enabling counsel to cling to familiar, if shopworn, "black line" redaction methods out of fear that privileged contents lurk in some dark digital recess.

To appreciate the problem, consider a complex spreadsheet like those routinely encountered in e-discovery. Spreadsheets are data grids made up of "cells" formed at the intersection of rows and columns. Cells contain hidden formulae entered by the user that generate calculated values seen as numbers in the cell. Formulae are what distinguish a spreadsheet from a word-processed table and may be important evidence in that they establish the origins, dependency and sensitivity of the calculated values. Put differently, *formulae make the numbers dance*. Without them, cell values are runes bereft of rhyme or reason.

With its embedded content, page-defying proportions and dynamic functionality, the exemplar spreadsheet fairly cries out for native production. Alas, it also harbors privileged or confidential content that must be excised.

If the requesting party isn't vigilant, here's how redaction goes wrong:

First, the producing party images the spreadsheet in .tiff format. It sprawls beyond the bounds of an 8½ x 11-inch page, so the data spills confusingly across multiple pages of .tiff images, obscuring column and row relationships. It's a mess.

Second, converting the spreadsheet to .tiff strips away all the underlying formulae, destroying spreadsheet function and undermining a key advantage of native production.

Finally, converting to .tiff means the data is no longer intelligible as data—i.e., it's not electronically searchable. A .tiff is just a picture—static ink on a virtual page—and no more electronically searchable than a Gutenberg Bible.

But it gets worse. To this point, the spreadsheet has been folded across unnatural dimensions, stripped of its usability and rendered electronically unsearchable. Now, the producing party redacts objectionable information like it was any 2D paper document—by using a drawing utility to black it out or printing it to paper for obliteration by a trusty felt-tip marker!

The spreadsheet's on life support. Seeking to resuscitate its electronic searchability, the producing party administers OCR.

OCR is inherently error-prone, but when the optically recognized data is text, spell checking corrects egregious recognition errors and restores some of the electronic searchability the federal rules require. When the data is numeric, however, there are no means to spell-check the inevitably myopic OCR. Wrong numbers replace right ones, and the data becomes wholly untrustworthy. By the time the spreadsheet reaches the requesting party, it's a goner:

- Usability: **gone**.
- Searchability: **crippled**.
- Integrity: **destroyed**.
- Content: **affirmatively misrepresented**.

The operation was a success, but the patient died.

If this is an "industry standard" practice, then we must recall that an entire industry can be negligent. As Judge Learned Hand wrote, "Courts must in the end say what is required; there are precautions so imperative that even their universal disregard will not excuse their omission." *The T.J. Hooper*, 60 F.2d 737 (2d Cir. 1932).

Preemptively, requesting parties should hone in on how ESI will be redacted, and if flawed redaction techniques will materially impair usability or searchability, they must act swiftly to combat their use and promote alternatives.

Redaction of ESI should be tailored to the nature of the data, using the right tool for the task. Where once native redaction was daunting, now there are reliable, cost-effective techniques for Adobe Systems Inc. PDF and Microsoft Corp. Office documents, including spreadsheets. For example, Adobe Acrobat 8.0 supports data layer redaction, and the latest release of Microsoft's Office productivity suite stores documents in readily redactable XML formats.

In sum, .tiff-OCR has its place, but when it's the *wrong* approach, don't use it. Opt instead for techniques that preserve the intelligibility and integrity of the unredacted content.

## **Trying to Love XML** by Craig Ball

***[Originally published in Law Technology News, March 2008]***

I *want* to love XML. I want to embrace it with the passion of my wiser colleagues, excited by its schemas, titillated by its well-formed code, flushed from its pull-parsing. I want to love XML as much as the cool kids do. So why does it leave me cold?

I want XML the dragon slayer: all the functionality of native electronic evidence coupled with the ease of identification, reliable redaction and intelligibility of paper documents. The promise is palpable; but for now, XML is just a clever replacement for load files, those clumsy Sancho Panzas that serve as squire to addled .tiff image productions. Maybe that's reason enough to love XML.

XML is eXtensible Markup Language, an unfamiliar name for a familiar technology. Markup languages are coded identifiers paired with text and other information. They can define the appearance of content, like the Reveal Codes screen of Corel Inc.'s WordPerfect documents. They also serve to tag content to distinguish whether 09011957 is a birth date (09/01/1957), a phone number (0-901-1957) or a Bates number. Plus, markup languages allow machines to talk to each other in ways humans understand.

Internet surfers rely on a markup language called HyperText Markup Language or HTML that forms the pages of the World Wide Web. There's a good chance the e-mail you send or receive is HTML, too. If you've tried to move documents between WordPerfect and Microsoft Corp.'s Word, or synchronize information across different programs, you know success hinges on how well one application understands the data of another.



Something as simple as importing day-first European date formats to month-first U.S. systems causes big headaches if the recipient doesn't know what it's getting.

Standardized markup languages alleviate problems by tagging data to describe it (e.g., `<EuropeanDate>`), constraining data by imposing conditions (e.g., restricting dates to U.S. formats: `<xs:pattern value="[0-1][0-9]/[0-3][0-9]/[1-2][0-9]{3}">`) and supporting hierarchic structuring of information (e.g., `<Lawyers><Name="Craig Ball"><EuroBirthDate> 01/09/1957 </EuroBirthDate> </Lawyers>`).

There are so many kinds of data and metadata unique to applications and industries that a universal tagging system would be absurdly complex and couldn't keep pace with technology and business. Accordingly, XML is extensible; that is, anyone can create tags and set their descriptions and parameters. Then, just as persons with different native tongues can agree to converse in a language both speak, different computer systems can communicate using an agreed-upon XML implementation. It's Esperanto for electrons.

In e-discovery, we deal with information piecemeal, such as native documents and system metadata or e-mail messages and headers. We even deconstruct evidence by imaging it and stripping it of searchability, only to have to reconstruct the lost text and produce it with the image. Metadata, header data and searchable text tend to be produced in containers called load files housing 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 agreeing to employ a structure geared to review software such as CT Summation or LexisNexis Concordance. Conventional load files are unforgiving. Deviate from the required sequence, or omit, misplace or include an extra delimiter, and it's a train wreck.

By tagging each value to identify its content and connection to the evidence, XML brings intelligence and resilience to load files. More importantly, XML fosters the ability to move data from one environment to another simply by matching the tags to proper counterparts.

Like our multilingual speakers using a common language, as long as two systems employ the same XML tags and organization (typically shared as an XML Schema Definition or .XSD file), they can quickly and intelligibly share information. Parties and vendors exchanging data can fashion a common schema custom-tailored to their data or employ a published schema suited to the task.

There is no standard e-discovery XML schema in wide use, but consultants George Socha and Tom Gelbmann are promoting one crafted as part of their groundbreaking Electronic Discovery Reference Model project. Socha and Gelbmann have done an impressive job securing commitments from e-discovery service providers to adopt EDRM XML as an industry lingua franca. See <http://edrm.net>.

A mature e-discovery XML schema must incorporate and authenticate native and nontextual data and ensure that the resulting XML stays valid and well formed. It's feasible to encode and incorporate binary formats using MIME (the same way they travel via e-mail), and to authenticate by hashing; but these refinements aren't yet a part of the EDRM schema.

So stay tuned. I don't love XML *yet*, but it promises to be *everyone's* new best friend.

## **The Science of Search**

**by Craig Ball**

***[Originally published in Law Technology News, April 2008]***

Federal Magistrate Judge John Facciola is a remarkable fellow. He hails from Brooklyn, wears bow ties, knows the Bruce Springsteen songbook by heart and doesn't hesitate to bring the White House to heel when the administration gets sloppy in its electronic evidence preservation. But his most heretical act may be his observation in *United States v. O'Keefe*, No. 06-249 (D.D.C. Feb. 18, 2008), that keyword search of electronically stored information is a topic "clearly beyond the ken of a layman." By a layman, he means any lawyer or judge who isn't an expert in computer technology, statistics and linguistics.

Facciola adds that, given the complexity of the science of search, "[F]or lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than [other] terms . . . is truly to go where angels fear to tread."

Heeding the call, the crack team of Forensically-trained Offerers of Legal Services (FOOLS) at Ball Labs have rushed in to formulate 36 search terms guaranteed to grab the smoking gun in any English-language ESI collection. The 36 terms are the letters of the alphabet and the numbers 0-9.

Ridiculous? Sure! But in a case where I serve as special master for ESI, a party proposed that the letter "S" be used as a search term. In another appointment, the plaintiff wanted to search for the number 64.

These earnest requests came from good lawyers offering credible rationales. They saw only that the term would be found within the evidence they sought, not appreciating that it would also appear in just about everything else, too. In the parlance of information retrieval, the terms scored high on recall but failed miserably in precision.

The parties advocating their use failed to appreciate that keyword search in e-discovery is less a means to find information than it is a method to filter it—and a pretty poor one at that. Keyword search of ESI is a sampling strategy—a way to look at less than everything with some assurance that you're examining the parts most likely to hold responsive data.

The notion that lawyers are unqualified per se to concoct keyword searches is likely to shake some sensibilities. Lawyers believe themselves adept at keyword search in e-discovery because they've mastered keyword search in online legal research. The correlation is superficial at best. Unlike the crazy quilt of ESI, the language of reported cases is precise, consistent and structured. Misspellings are rare. Legal research is Disneyland. E-discovery is Baghdad.

Judge Facciola is right to point to lawyers' misplaced reliance on keyword search and lack of expertise. *Search is a science*, yet we approach it on faith, gambling that intuition and luck are enough. Still, noting the profession's lack of expertise doesn't address the knottier problem of *where* to secure the expertise we now must bring to court to establish or challenge the efficacy of search.

The answer isn't to spawn a new breed of self-anointed cyberlinguistics experts for hire. Neither will a wholesale move to concept search tools suffice. Smarter search tools employing algebraic and probabilistic analysis are unquestionably an improvement on the crude tools we employ, but hardly dispense with the need for experts to explain their operation and defend their performance.

The answer is that lawyers need to learn more about the science of search as part of our legal and continuing education. We need to become skilled at tools and methods that help us refine searches and routinely test them against representative data so we can distinguish noisy terms from effective ones and learn to zero in on relevant ESI.

Law schools teach the science and art of legal research when modern methods have all but eliminated the need to navigate the reporter system. Instead, students and lawyers must be afforded the means to master the art and science of digital information. We must dare to tread in these areas, not as fools but as professionals skilled in eliciting, testing and marshaling evidence wherever it may be found.

"The right to practice law is not one of the inherent rights of every citizen . . . [but] is a peculiar privilege granted and continued only to those who demonstrate special fitness in intellectual attainment and in moral character." *Matter of Keenan*, 314 Mass. 544, 546 (1943). So it has been, and so it must remain as evidence takes new forms, if we are to be afforded that peculiar privilege.

## Dealing with Third-Parties

by Craig Ball

*[Originally published in Law Technology News, May 2008]*

Recently, a team of e-discovery consultants called, seeking feedback on a plan to collect responsive data from non-parties. To their credit, they recognized that not all relevant electronically stored information resides on their client's systems. Contractors, agents, vendors, clients, lawyers, accountants, consultants, experts, outside directors and former employees also hold responsive ESI.

Consequently, parties must factor non-parties (over whom they have influence) into litigation hold and production strategies. The consultants had done so, but now wondered how to retrieve relevant data without compromising its integrity and usability.

They planned to send external hard drives loaded with Microsoft Corp.'s Robocopy backup utility to each non-party custodian, asking them to herd responsive ESI into a single folder, then run Robocopy to replicate and return their collection on the external hard drive. They were proud of their plan, noting that use of Robocopy would preserve system metadata values for the files.

*Or would it?* Recall that system metadata is data a computer's operating system compiles about a file's name, size and location, as well as its Modified, Accessed and Created (MAC) dates and timestamps.

Don't confuse hardworking *system* metadata with its troublemaker cousin, *application* metadata. The latter is that occasionally embarrassing marginalia embedded in documents, holding user comments and tracked changes.

By contrast, system metadata values are important, helpful dog tag data. They facilitate searching and sorting data chronologically, and shed light on whether evidence can be trusted. System metadata values present little potential for unwitting disclosure of privileged or confidential information and should be routinely preserved and produced.

But Microsoft makes it tough to preserve system metadata. Open a file to gauge its relevance, and you've changed its access date. Copy a file to an external hard drive, and the creation date of the copy becomes the date copied. *Grrrrr!* Robocopy, a free download from Microsoft's website, does a fine job preserving system metadata, but it can't restore data already corrupted.

When I pointed out that copying the files to assemble them would change their MAC dates before Robocopy could preserve them, one of the consultants countered that he'd thought of that already. Each third-party would be instructed to use the Windows "Move" command to aggregate the data.

They'd thought of everything . . . *or had they?*

An advantage of the Move command is that it preserves a file's MAC dates. But, faithful to its name, Move also relocates the file from the place where the third-party keeps it to a new location. So here, it's like requiring those assembling files for production to dump their carefully ordered records into a sack. Demanding non-parties sabotage their filing systems is a non-starter.

To make matters worse, Robocopy is a *command line* application—more like DOS than Windows—employing six dozen switch options, so it's hardly a tool for the faint of heart. Mistype one of these cryptic command line instructions, and the source data's gone forever. Moreover, Robocopy only runs under Windows. What if the data resides on a Mac or Linux machine?

Finally, the approach wasn't geared to collecting e-mail evidence. Sure, they could copy Outlook .pst files holding complete e-mail collections, but non-parties won't agree to share unrelated personal and confidential data. Instead, they'll need to select responsive messages and save them out to a new container file or as individual messages.

Further, if their Exchange e-mail system doesn't support local .pst container files, or if the system uses a different e-mail application like IBM's Lotus Notes or Novell's GroupWise, an entirely different approach is needed.

The well-intentioned consultants were so enamored of their favored "solution," they lost sight of its utter impracticality. Still, they were on the right track seeking low-cost, out-of-the-box approaches to collection—approaches that preserve metadata and don't require technical expertise.

The consultants went back to the drawing board. Their better mousetrap will incorporate input from the other side, an easier-to-implement collection scheme and the use of experts for the most important data.

Sometimes there's no getting around the need to use properly trained personnel and specialized tools; but, if you decide to go a different way, be sure you:

- 1. Know the systems and applications housing and creating the electronic evidence;**
- 2. Assess the technical capabilities of those tasked to preserve and collect evidence;**
- 3. Understand and thoroughly test collection tools and techniques; and**
- 4. Discuss collection plans with the other side. They may not care about metadata and will accept less exacting approaches.**

## Grimm Prognosis

by Craig Ball

*[Originally published in Law Technology News, July 2008]*

There's a double standard in e-discovery. Keyword search is deemed "good enough" for identifying responsive electronically stored information; yet when privilege is on the line, lawyers insist on page-by-page review. It's a tacit recognition that keyword search is a blunt instrument — a point artfully made twice this year by Magistrate Judge John Facciola in *U.S. v O'Keefe*, 537 F. Supp. 2d 14, 24 (D.D.C. 2008), and *Equity Analytics v Lundin*, 248 F.R.D. 331, 333 (D.D.C. 2008), and emphatically underscored by Magistrate Judge Paul Grimm in *Victor Stanley, Inc. v Creative Pipe, Inc.*, Civil Action No. MJG-06-2662 (D. Md. May 29, 2008).

It's assumed that lawyers are qualified to review documents for relevance, responsiveness and privileged character. But are we qualified to craft *proxies* for our judgment in the form of keyword searches? In *Victor Stanley*, 165 documents slipped by a privilege review employing keyword search and a cursory- sounding "title page" analysis for non-searchable items. Defendants had unwisely abandoned efforts to secure a clawback agreement (a nonwaiver agreement providing that inadvertently produced privileged materials may not be used).

Plaintiff's counsel spotted the documents and dutifully reported their potentially privileged character, but argued defendants waived privilege by using a faulty review process. The court agreed, pointing to defendants' failure to provide information regarding keywords used, how they were selected, steps taken to assess the reliability of the outcome and the qualifications of the attorneys to design an effective and reliable search.

Thus another jurist dismisses the legal profession's ability to search ESI without demonstrated expertise. It's enough to give Perry Mason an inferiority complex!

Do lawyers have so insightful a grasp of the words and semantic relationships behind our relevance and privilege decisions that we can distill the *je ne sais quoi* of our well-honed legal minds into quotidian keyword spotting? We'd like to *think* we do, despite studies showing we possess little ability to frame effective keyword searches. We're shocked when our magic words catch *barely 20%* of responsive documents. We shouldn't be.

Language is deceptively complex, and meaning is an elusive, protean quarry. We depend upon context for meaning, but keyword search ignores context entirely. Boolean search is only marginally better at gleaning context.

That leaves lawyers in a tough spot. Mushrooming volumes of ESI require us to rely more on automated search tools at the same time courts and opposing counsel are less willing to indulge the fiction that these tools perform in unskilled hands. The jig is up, and lawyers are now obliged to *prove* these proxies really work.

How do we meet that burden of proof? Judge Facciola deems both lawyers and judges keyword naïfs, instead summoning a phalanx of linguists, statisticians and computer experts. Though expecting searches to be designed by qualified persons, Judge Grimm leaves the door open to lawyer-initiated keyword search when counsel can demonstrate adequate quality assurance and quality control.

This is a subtle but important distinction. Lawyers can become "qualified persons," though they may never be linguists, statisticians or computer experts. Still, Judge Grimm sets the bar high:

"Use of search and information retrieval methodology, for the purpose of identifying and withholding privileged or work product protected information from production, requires the utmost care in selecting methodology that is appropriate for the task ... [and] careful advance planning by persons qualified to design effective search methodology.

The implementation of the methodology selected should be tested for quality assurance; and the party selecting the methodology must be prepared to explain the rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented."

*Victor Stanley* departs from *O'Keefe* in another subtle way. By emphasizing collaboration, Judge Grimm preserves counsel's ability to negotiate and agree upon search methods. Judge Facciola is no less a proponent of collaboration and transparency in e-discovery, but declaring both counsel and courts unequipped to oversee keyword search without expert assistance imperils the parties' freedom to agree on search methods and the court's authority to ratify such agreements.

What court, admittedly unqualified to weigh such matters, could endorse a search protocol framed by those equally unequal to the task? Thus *Victor Stanley* preserves the litigants' inalienable right to be wrong, so long as everyone agrees that wrong is right. It's a Faustian bargain, but one permitting cases to move forward by simply ignoring pesky questions concerning the integrity and completeness of electronic discovery.

The *Victor Stanley* decision gives teeth to the duty to use better search techniques. Avoiding privilege waiver is a powerful incentive to:

- Get expert help.
- Collaborate on search methods.
- Test your searches.
- Check the discard pile.
- Get that clawback agreement.

## **Brain Drain**

**by Craig Ball**

*[Originally published in Law Technology News, August 2008]*

Want to get a lawyer's attention? Just mention "data wiping" and "litigation" in the same breath. You might need to administer CPR. Yet there are cases where both sides recognize the need to thoroughly eradicate electronic data, such as when an employee has spirited away proprietary information to a new job and the old employer needs assurance it won't be exploited. It's a simple-sounding task that's harder and more expensive than many lawyers and judges appreciate.

Sure, you could wipe every sector on the hard drives or scuttle the machines into the Mariana Trench, but then you'd have no record of what went where or how it was used. Think also of the legitimate business and personal data that would be lost. Shifting non-contraband data to new media might work, but who can be entrusted with that job, and how will they divvy up the contents of e-mail container files and other amalgams tainted by stolen information?

The former employer could supervise the process, but affording a competitor such unfettered access is often out of the question. Even if these issues are resolved, will ordinary deletion be sufficient? What's to prevent the other side from resurrecting the deleted data once the case is dismissed?

Before you include data obliteration as a condition of settlement, be certain you've considered all the steps needed to effectuate reliable eradication, as well as the total cost and potential disruption. Start by determining what's been taken by a focused forensic examination of the ex-employer's machines previously used by the departed employee, a job made harder, but not impossible, if machines have been re-tasked to new users or the employee tried to cover his tracks.

Data enters and leaves computers via a handful of common vectors, such as e-mail, thumb drives, external hard drives, optical media or network transfer. So you'll want to know what files, network areas, internet sites—especially web mail services—and external storage media the employee accessed, especially in the last weeks on the job.

You'll also want to gather the information needed to perform a thorough search of the other side's relevant machines, such as the names, sizes, last modified dates and hash values of stolen files, as well as unique phrases or numerical values within those files. Searching for stolen data by its hash value is useful and cost-effective, but it won't turn up data that's been altered or deleted. For that, forensic examiners must analyze file metadata, carve unallocated clusters, run keyword searches and review content.



Next, you'll want to account for all the media that has housed any of the contraband data. Forensic examination of the former employer's machines can pin down the portable devices employed to transport the data, while analysis of the new employer's systems usually reveals if and when the transport media were connected and whether other portable storage devices helped copies fly the coop.

The trail of stolen data often leads first to home systems, particularly where the errant employee took time off between jobs. It naturally progresses to the new employer's laptop and desktop machines and network storage areas to which the employee had access. These are typically searched for files with matching hash values, similar or identical file names, and files containing distinctive words, phrases or numeric values present in the stolen data.

Machines are analyzed to see if file deletion, data hiding or file wiping were used to conceal the stolen data. Metadata and registry keys are examined to identify notable events (such as the arrival of a large numbers of files, drive swapping or operating system re-imaging). It's a lot of old-fashioned detective work using newfangled technology.

Even when no one has deleted or hidden stolen data, some of it routinely finds its way into the unallocated clusters, a vast digital landfill where operating systems dump transient data like the contents of swap memory or working copies created by word processing applications. Data may also lodge in file slack space, the area between the end of a file and the end of the last cluster in which it's stored. Consequently, a thorough eradication includes identifying any stolen data that's wormed its way into these hard-to-access regions.

It's so important to examine these obvious places where stolen data lodge and determine whether and how the data's been used, abused or disseminated because that knowledge guides resolution of a costly, contentious issue: Where do you stop?

Victims of data theft understandably fret about the potential for missed or hidden copies of contraband data and demand the broadest and most exacting search, especially when they bear none of the cost and regard the new employer as complicit in the theft. However, extending search beyond machines with a clear connection to the former employee should be based on evidence signaling their involvement or a sensible sampling protocol.

Courts and counsel should be wary of imposing or agreeing to a search and eradication method that's so wide-ranging, costly and disruptive as to be unintentionally punitive.

Once found, it's fairly easy to delete and overwrite contraband active data files and the entirety of the unallocated clusters and slack space (the contents of which have no value to the user). However, separating contraband transmittals and attachments from e-mail containers is a laborious process necessitating selective deletion, compaction and/or re-creation of the container files on local hard drives, as well as agreement concerning the handling of server mail stores and back up media. These enterprise storage areas don't lend themselves to piece-meal deletion, necessitating considerable effort, ingenuity and expense to purge contraband data.

Employee data theft is a common, costly and growing problem, so lawyers handling these cases must understand the expense and complexity attendant to expunging purloined data and recognize that an agreement to "delete it" sounds straightforward but may be biting off more than the client intends to chew.

## **SNAFU**

**by Craig Ball**

***[Originally published in Law Technology News, September 2008]***

On September 2, 1945, my father was ordered to fashion nine impregnable containers to carry the just signed Japanese surrender documents to the President of the United States, the King of England and other heads of state. Dad earned his law degree from Harvard in 1932; so naturally, the Navy made him a gunnery officer.

Good thing, because I can't imagine there's much Lt. Commander Herbert Ball took from Langdell Hall that equipped him to convert five-inch powder charge casings into watertight containers. His ingenuity helped the important V-mail (Victory mail) make it to Mr. Truman, safe and sound.

I proudly share this family lore because a very different war requires me to deconstruct electronic containers carrying missives from the front. Safe in my lab, thousands of miles from IEDs and insurrection, I'm grappling with wacky date values on thousands of e-mail messages from Iraq. It brings to mind that wonderful WWII acronym: SNAFU, for "Situation Normal: All Fouled Up," though no sailor ever said "fouled."

When e-mails originate around the globe on servers from Basra to the Bronx, they seem to travel back in time. Replies precede by hours the messages they answer. Such is the discontinuity between the languorous rotation of the earth and the near light speed of e-mail transmission. A message sent from Baghdad at dinner arrives in Austin before lunch. E-mail client applications dutifully—some might say stupidly—report the time of origin. The confusion grows when receiving machines apply different time zone and daylight savings time biases. It gets even more fouled up when a user in Iraq sends mail via a stateside server. In the end, it's tough to figure out who said what when.

What's needed is time and date normalization; that is, all dates and times expressed in a single consistent way called UTC for Temps Universel Coordonné or 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. Peace in our time.

My mission was to convert all messages to UTC, changing Situation Normal: All Fouled Up into Situation Normalized: All Fixed Up.

This requires going deeper than the date and time information displayed by Microsoft Corp. Outlook, down to the header data in the message source. There you find a time-stamped listing of servers that handed off the message and the message's time of receipt, expressed in hours plus or minus UTC.

Of course, you've got to have header data to use header data. But when e-mail is produced as .tiff or PDF images, header data is stripped away. The time seen could indicate the time at the place of origin or at the place of receipt. It could reflect daylight savings time ... or not.

Absent header data or the configuration of the receiving machine, you just don't know. So reasonably usable production necessitates a supplemental source for the UTC values and offsets (such as a spreadsheet, slip sheet or load file); otherwise, messages should be reproduced in a native or quasi-native format (e.g., .pst, .msg or .eml).

If you're the party gathering and producing e-mail from different time zones, make it a standard part of your electronically stored information collection protocol to establish and preserve the relevant UTC and daylight savings time offsets for the custodial machines. On Microsoft Windows devices, this data can be readily ascertained by clicking on the clock in the System Tray. It can also be gleaned by examination of the System Registry hives if the boot drive was preserved in a forensically sound fashion.

E-mail threads pose additional challenges because erroneous time values may be embedded in the thread. It's important that production include not only the threaded messages, but also each of the constituent messages in the thread.

Don't underestimate the importance of date and time normalization when the timing of events and notices may prove key issues. In a flat world, or one at war, keeping communications on a common clock is a necessity.

## Problematic Protocols

by Craig Ball

*[Originally published in Law Technology News, November 2008]*

Forensic examination lays computer usage bare. Personal, confidential and privileged communications, sexual misadventure, financial and medical recordkeeping, proprietary business data and other sensitive information are all exposed. In the white picket fenced places where active data lives, you can tiptoe around privileged and private information, but deleted data hails from the wrong side of the digital tracks, where there are no names, no addresses and no rules.

Forensic examiners see it all, including confidential materials that can't be shared. Courts impose examination protocols to limit the intrusiveness, scope and conduct of the work and establish who can see the outcome. It takes technical expertise to design a good protocol. Without it, you get protocols that are forensic examinations in name only, impose needless costs and cumbersome obligations or simply elide over what the examiner is expected to do.

The perils of protocols are seen in *Ferron v. Search Cactus, LLC, et al.* 2008 WL 1902499 (S.D. Ohio Apr. 28, 2008), a *pro se* action by an attorney seeking damages for unsolicited spam e-mail. The defendants claimed Ferron solicited the spam via his web surfing and wanted their computer forensic examiner to inspect Ferron's home and office computers. As these systems held privileged and irrelevant confidential material, the court needed to restrict the scope of the examination and sharing of recovered information.

The defendants wanted to know whether plaintiff visited to particular websites or deleted information. Tracking Internet usage sounds simple because tech-savvy user can access certain information about Internet usage history. Internet History files, Temporary Internet File cache and a user's cookie directory are reasonably accessible without specialized tools and not difficult to understand.

But in a Windows/Internet Explorer environment, the most revealing and complete Internet activity data isn't accessible without the tools and training to locate and interpret it. Both IE and the Windows System Registry maintain detailed, time-stamped records of Internet surfing, IE in files named Index.dat and the registry as weakly encrypted data within an obscure area storing "User Assist" keys. If you've never heard of these forensically-significant records, you're not alone. No user could reasonably be expected to access and preserve these files--let alone know they exist--without technical expertise or assistance.

Thus, it's breathtaking that the *Ferron* court found the plaintiff breached a duty to preserve this hard-to-handle information in anticipation of litigation, *i.e.*, with no express request to preserve or produce vestiges of Internet activity. Is this now the standard in every case involving e-mail or

just cases where other forms of Internet activity may be at issue? Either way, Judge Frost expands the role of forensically sound drive imaging for preservation far beyond custom and practice and positions forensic examination of drives holding confidential and privileged information as a suitable response to failure to meet the preservation duty.

Judge Frost wisely recognized that courts permitting forensic examination must balance the need for access against privacy rights to insure appropriate safeguards in the examination protocol, *e.g.*,

- Use of a neutral examiner,
- Restricting an examiner's disclosure of protected material,
- Initial review of examiner work product by producing counsel,
- Collaboration between opposing examiners,
- "Attorneys' eyes only" inspection, and
- *In camera* review of findings.

The optimum approach depends upon the sensitivity of the information, risk of waiver attendant to third-party exposure, level of trust between counsel, examiners' expertise and the budget. The court's protocol in *Ferron* was an eight-step amalgam of safeguards based on protocols from *Playboy Ent., Inc. v. Welles*, 60 F.Supp.2d 1050 (S.D. Cal. 1999), but adding an unprecedented "data removal" step that, in my view, shouldn't serve as precedent.

The first three steps of the protocol can be summarized as:

1. Plaintiff's expert images Plaintiff's hard drives and preserves the images.
2. Plaintiff's expert removes Plaintiff's confidential personal information from the images, detailing the removal protocol.
3. Plaintiff then affords Defendants' expert access to his hard drives.

The court surely intended that the *hard drives*, not the *images*, be purged of confidential information since it's silly to purge copies while affording Defendants access to unexpurgated originals. Yet even harmonizing the protocol this way, selectively purging the source hard drives is an awful idea.

Anyone who's pored over the endless expanse of unallocated clusters and file slack space of a well-used hard drive knows the convolution and chaos there. Thorough, selective deletion of anything but the narrowest and most clearly defined items is impractical and prohibitively expensive. So, it's doubtful that Plaintiff's expert can remove *all* instances of personal information from the drives without also destroying discoverable data. Restrict how *results* are used, *but don't mess with the digital evidence*.

The remaining steps call for defendants' expert to image and analyze the sterilized hard drives then review his findings in confidence with plaintiff. Before sharing information with the

defendants, the defendants' expert must remove from his work product any information plaintiff asserts is privileged. Both sides' experts are designated as officers of the court.

Having opposing experts work cooperatively as officers of the Court helps assure that information won't be improperly concealed or revealed. However, it's more costly than employing a single neutral examiner and puts the requesting party's expert in the untenable position of being privy to confidential and privileged information yet forbidden to share that knowledge or allow it to influence further work, advice or testimony. It effectively disqualifies the expert going forward.

Another case, *Coburn v. PN II, Inc.*, 2008 WL 879746 (D. Nev. Mar. 28, 2008), exemplifies the "empty" forensic protocol. While there's much to commend the court's detailed five-step protocol in terms of addressing the choice of expert, privilege concerns, confidentiality and convenience, the order is silent as to whether or how the forensic expert will recover information or analyze the data.

In the end, all the independent expert does is image Coburn's hard drive and hand the image back to Coburn's counsel for printing and review of any "recovered" documents. Trouble is, the expert isn't permitted to *recover* any documents, and presumably counsel is ill-equipped to do so without expert assistance. The "forensic" nature of the process is illusory. The party seeking examination is in no wise aided because the process isn't calculated to expose new evidence. Coburn already had access to the contents of her own drive, and Coburn's counsel was already under an obligation to search same for responsive ESI. It's an empty, expensive exercise.

I have a love-hate relationship with examination protocols. The lawyer in me knows there must be constraints, but wearing my forensic examiner's hat, the ideal protocol is, "Here's what we want to know. Go where the evidence leads."

Lifting a protocol from a reported decision is no assurance of success. A capable expert—better yet, collaborating experts on both sides--can draft a protocol that's calculated to get to the information sought without revealing undiscoverable material.

## What Lies Beneath?

by Craig Ball

*[Originally published in Law Technology News, February 2009]*

In something of an impromptu Philip K. Dick film festival, I had a chance to revisit the still-inspired 1982 *Blade Runner* and the still-disappointing 1990 *Total Recall*.

The first showed a billboard for Pan American Airlines, circa 2019; the other for a Sharper Image store, circa 2084. Of course, Pan Am (its lunar shuttle also figured prominently in *2001: A Space Odyssey*) collapsed just 10 years after *Blade Runner*'s release, and Sharper Image closed all stores last year.

Perhaps there is a lesson here: These cinachronisms bear out the folly of assuming too much about the future. Wall Street's crystal ball is no better than Hollywood's.

Going back to 1896 and the 12 companies on the original Dow Jones Industrial Average, nearly all were broken up or absorbed generations ago. Only General Electric remains a part of the DJIA. The pundits of the Roaring '20s no doubt took it for granted that U.S. Leather would stay on top — surely industrious America would always need miles and miles of leather belts to transfer power to machinery!

Packard Motor Car Co., F.W. Woolworth Co., Trans World Airlines, Arthur Andersen, Enron Corp., Lehman Brothers, Washington Mutual Inc., Heller Ehrman ... all gone. To quote Donald Rumsfeld, an expert on unforeseen calamities, "Stuff happens"—and there's more to come.

If you think the markets and indicators have bottomed out, think again. There's more smoke to clear, more mirrors to break. The electronic data discovery industry and, to a lesser extent, the legal services industry are microcosms of the broader wounded economy.

Marc Dreier, the nabob New York lawyer accused of peddling hundreds of millions of dollars of bogus commercial paper, is our industry's Mini-Me to Bernard Madoff's Dr. Evil. Tinfoil titans laid low overnight. So, write this on your hand and don't wash it off: Nothing is sacred. No one is safe. Anyone can disappear ... fast.

No matter who you are using for EDD services, now is the time to assess your exposure, mobility and disaster recovery strategy.

Ponder these questions about your EDD vendors:

- How long do we anticipate the necessity of vendor involvement?
- Do they have the only accessible copy of any evidence?

- Do we have a complete copy of our vendor's work product in a format we can use?
- How will we handle the inevitable delay occasioned by the flight of key personnel or outright failure?

What becomes of our data in their hands in the event of bankruptcy or failure?

You'd be smart to plan for these eventualities and careless not to.

Remember when law students would attend their first class, to hear: "Look to your left and right — only one of you will graduate"?

I suspect vendors entering the "EDD Class of 2011" are in the same boat. If you're one of those providers, guard against being squeezed by slow-to-pay customers. Don't let outstanding accounts get too far ahead of services rendered, and dare to demand pre-payment from customers with dodgy payment histories.

Sure, you want to compete, but bad business is worse than no business. Bad business costs you money.

In times of tight credit, customers will impose on your goodwill to finance litigation. If you want to be their bank, be sure you're adequately compensated and acting in compliance with credit regulations, then be prepared for default. A mechanic has a lien on the car being repaired, but you can't sell client data to defray unpaid bills.

Because any customer can disappear, vendors can find themselves an unsecured claimant in bankruptcy. Have an action plan, and understand the difference between a clawback in EDD and a clawback in bankruptcy— in the latter, you have to refund monies previously collected. How's that for adding insult to injury? Finally, though financial-ratings companies have lately cast themselves as fools on the world's stage, periodically pulling a report on key customers may help you dodge a bullet.

Mired in credit crunches and client collapses are the lawyers. Like the feckless brokers who steered billions to Bernard Madoff, we lawyers have due diligence obligations. We must do better than those dozing brokers — not by eliminating risk, but by managing it: doing our homework, asking hard questions, educating our clients and planning for the foreseeable and the formidable.

The potential for titanic failure must be on our radar — and stay there for quite a while.

It may sound like an ad campaign for Murder Inc., but make it your mantra for 2009: **Anyone can disappear ... fast.**



## Don't Touch That!

by Craig Ball

*[Originally published in Law Technology News, April 2009]*

Why do people who know better than to traipse through crime scenes blithely muck about with digital smoking guns? With computers, it seems we must trip over the *corpus delicti* and grab the knife before we realize we're standing in a pool of blood!

Sometimes a computer *holds* evidence, and sometimes a computer *is* evidence. It's a distinction with a difference when deciding whether to act in ways that will stomp on data essential to computer forensic examination.

In most e-discovery efforts, computers are just digital file cabinets, and the evidence is the e-mail and files stored within. Just as paper records require a modicum of care to avoid ripping and staining, digital documents require preservation of basic metadata akin to date stamps and margin notes on paper documents. But, we needn't go to extraordinary lengths to protect this information. It's either embedded in the files and e-mail messages as application metadata, or stored by the operating system as accessible system metadata—such as file names, folder locations, and the dates files were created, modified, and accessed. We use such stuff every day, so preserving it isn't rocket science and needn't be expensive or cumbersome.

But computers aren't always simply repositories of evidence. They may be the instrumentalities of a crime, tort, or conduct under investigation, or carry clues to the origins and integrity of suspect electronic evidence. In these instances, the computers, too, are evidence—virtual crime scenes where careless conduct compromises outcomes, and diligence demands scrupulous protection and analysis of the revealing, complex and obscure data about data they hold. Now, we *do* have to go to extraordinary lengths to protect the information.

In civil litigation, computer forensic examiners often see the evidence only after some well-meaning soul has poked around and unwittingly changed last access dates and registry values. That's the trade-off: Without that first look, the misconduct might have been discovered too late or overlooked altogether.

There's precedent for this in other forensics work. If a victim might still have a pulse, good Samaritans and EMTs are coming through, fingerprints, fibers, and DNA be damned!

Crime scene investigation offers another parallel, this one worth emulating for digital evidence. Some crimes—e.g., murder, sexual assault, kidnapping—are so heinous that bringing in the CSI is standard practice, and first responders know they must secure these scenes.

Likewise, some situations in civil practice are so likely to be bound up with electronic evidence requiring computer forensics that improvident metadata mauling could easily be avoided by applying the following rule of thumb:

Before allowing anyone untrained in digital forensics to access a computer that may be evidence, consider:

1. Does the computer's user occupy so crucial a position that an accusation of data tampering or destruction could hurt the company?
2. Is the user suspected of stealing trade secrets, or poaching customers or employees?
3. Is a suspected forged computer-generated document or communication involved?
4. Does inappropriate e-mail or internet use figure into the suspected misconduct?
5. Did a departing employee bring a personal laptop, external hard drive, or thumb drive to work?
6. Did the size of the user's server e-mail stores suddenly and significantly diminish, or are messages believed to be missing from the user's server stores?
7. Do server logs or indicators reflect atypical access to data areas?
8. Has the user been notably secretive using company computers or been observed using other users' machines without permission?
9. Has the user recently requested that IT reinstall the operating system on his or her machine?
10. Has the user asked about data destruction techniques, or been observed with wiping software?

I've heard lawyers claim, "Metadata doesn't matter." Their myopic view stops at application metadata; that is, tracked changes, embedded commentary, and other potentially privileged or prejudicial information they fear opponents will dredge up. But in many cases — especially those involving allegations of data theft — it's the system metadata, particularly the file dates and paths, that matter most. And it's the system metadata that eager explorers fail to protect.

When you open or even preview a file, you alter its last access date and make it harder for forensic examiners to assess what previous users have done, and when. When you copy a file, it typically changes the creation date on the copy. When you save a file—even without making apparent changes—you alter its last modified date. Because it's easy to copy the contents of

huge folders or trigger antivirus applications that "touch" every file, even brief, well-intentioned peeks wreck havoc with thousands of files.

Messing with system metadata isn't just a concern for computer forensics. We also depend on file names, dates, and folder structures to search, sort, and make sense of electronically stored information in e-discovery.

Making it harder to use electronic evidence is less egregious than destroying the evidence, but both bad outcomes can be avoided by resisting the impulse to poke around.

"Write protecting" a drive to safeguard metadata isn't difficult, and tools run from free to a couple of hundred dollars.

If the IT person or your EDD service provider need to look at electronic evidence, be sure they have the tools and know-how to protect it; and where computer forensic examination is foreseeable, treat the computer like evidence at a crime scene and call in the pros.

## **Special Masters**

**by Craig Ball**

*[Originally published in Law Technology News, May 2009]*

I frequently field this question: "Hi Craig. I'm a tech-savvy lawyer and want to serve as an e-discovery special master. What advice can you offer me? And what does a special master do exactly?"

A special master for electronically stored information is a technical expert—ideally a lawyer—appointed by the court to manage and resolve discovery disputes involving electronic evidence.

Governed by FRCP Rule 53 in the federal courts, an SM-ESI enjoys such powers as the court delegates, subject to de novo review by the judge. Courts may turn to special masters when the judge lacks the technical expertise or time to address complex or contentious e-discovery disputes.

An SM-ESI may sort out search terms, fashion collection protocols, investigate spoliation, resolve privilege concerns, arbitrate forms of production, suggest sampling scenarios, apportion costs, and make sanctions recommendations. It's fascinating, challenging, creative work.

But there's more to being an effective SM-ESI than legal and technical know-how. Special masters don't have skills training courses such as those available to lawyers, judges and mediators. We learn by doing and from our mistakes.

## TIPS & TECHNIQUES

Here are some lessons I've learned in the trenches.

Special masters are often appointed because the parties won't cooperate. Discussions are ugly, angry, and petty. Demand that backbiting and snide comments cease. When recriminations fly, give them no quarter. Professionals should act professionally, and compulsory courtesy fosters the real thing.

The special master, too, needs to be courteous and patient. The times I rue most are those where I lost my cool. Once, when their sniping and pettiness wouldn't stop, I lost my temper and called the lawyers "nattering nabobs." I regret my incivility almost as much as I regret quoting Spiro Agnew!

Litigators love to talk. They need to be heard, again and again... and again. A lawyer is more likely to believe in the tooth fairy than accept that you understood something the first time.

Be patient. Force yourself to listen. Then, when enough is enough, recap the point quickly, have the lawyer confirm you got it, and move on. Meters are running. Someone's paying for all that jawing. As SM-ESI, if you're not fostering efficiency, you're not doing your job.

Being neutral isn't the same thing as being evenhanded. I've been appointed to serve as special master in cases where one side proved incapable of meeting its discovery obligations. When your mandate is to fix a problem and one side's at fault, the parties don't start even. The errant party has to get back to good stead, and it's the special master's job to help them find the way. Nothing's as corrosive to cooperation as the charge that counsel reneged on a commitment. So, claims of that nature should be met with a request that the side making the claim produce a written record of agreement. I routinely remind counsel that the rules of procedure dictate how lawyers must memorialize their agreements, and I require the parties to abide by the rules.

It's unwieldy to put every representation and agreement in writing, especially on prolonged conference calls. I found having one side act as recording secretary led to more squabbles, but I didn't want the cost and formality of a court reporter on every call. The resolution proved amazingly simple.

My conference service supports call recording at no cost, so I now record each conference call and make the recordings available to those who participated on the call. No one is permitted to share the recording with persons who weren't actually on the call or offer any part of it into

evidence; however, a participant can testify about the proceedings after refreshing his or her memory from the recordings.

At first, the parties groused, but the results were splendid. Disputes about what was said or promised ceased. The ability for each side to hear their own words left no room for doubt.

In an ESI meet-and-confer conference, the technical personnel are at the top of my pecking order, so I turn the caste system upside down. I treat technical personnel with utmost respect and deference. It encourages them to help me find the right results, and it sets the right example for the lawyers.

One of the smartest things an SM-ESI can do is get the geeks together. IT specialists are natural problem solvers who speak a language all their own. Lawyer intermediaries can just add friction, and when they do, I convene conferences of just the IT folks from each side and me. No lawyers allowed. So far, no one's objected, and it works. (Of course, the lawyers are never shut out of substantive legal discussions.)

An SM-ESI stands in the shoes of the court, so be vigilant about ex parte contact with counsel. A special master's integrity and credibility matter more than technical expertise or legal prowess. But special master work parallels mediation in certain ways and, like mediation, *ex parte* communications can be conducive to forging a compromise.

I secure the court's authorization of such contact in my appointment order or seek the parties' agreement. Either way, my rule is that the fact and timing of all ex parte contacts must be promptly disclosed in writing to the other side.

Be considerate. I recall several three and four-hour conference calls where I neglected to call a recess. Don't make the lawyers and support personnel have to ask for a break. Invite them at the start to "do you a favor" and remind you to call a recess after 90 minutes or so. Likewise, have food available at face-to-face meetings.

Clearly, the qualities and practices of an effective SM-ESI have much in common with those of a good judge or mediator. Being tech-savvy is important. Being people-savvy, cost-conscious and keeping your ego in check, matter more.

For further discussion of the role of ESI special masters, I recommend Shira Scheindlin & Jonathan Redgrave, "*Special Masters and E-Discovery: The Intersection of Two Recent Revisions to the Federal Rules of Civil Procedure*," *Cardozo Law Review*, Volume 30, Issue 2 (Nov. 2008).

## About the Author



Craig Ball, of Austin is a Board Certified Texas trial lawyer and an accredited computer forensics expert, who's dedicated his career to teaching the bench and bar about forensic technology and trial tactics. Craig hung up his trial lawyer spurs to till the soils of justice as a court-appointed special master and consultant in electronic evidence, as well as publishing and lecturing on computer forensics, emerging technologies, digital persuasion and electronic discovery. Fortunate to

supervise, consult on or serve as Special Master in connection with some of the world's largest electronic discovery projects and most prominent cases, Craig also greatly values his role as an instructor in computer forensics and electronic evidence to the Department of Justice and other law enforcement and security agencies.

Craig Ball is a prolific contributor to continuing legal and professional education programs throughout the United States, having delivered over 500 presentations and papers. Craig's articles on forensic technology and electronic discovery frequently appear in the national media, including in American Bar Association, ATLA and American Lawyer Media print and online publications. He also writes a multi-award winning monthly column on computer forensics and e-discovery for Law Technology News and Law.com called "Ball in your Court." Rated AV by Martindale Hubbell and named as one of the Best Lawyers in America and a Texas Superlawyer, Craig is a recipient of the Presidents' Award, the State Bar of Texas' most esteemed recognition of service to the profession.

Craig's been married to a trial lawyer for 22 years. He and Diana have two delightful teenagers and share a passion for world travel, cruising and computing.

Undergraduate Education: Rice University, triple major, 1979  
Law School: University of Texas, 1982 with honors

 <p><b>Craig D. Ball, P.C.</b></p>	<p><b>Craig Ball</b> Trial Lawyer Technologist</p> <p>E-Mail: <a href="mailto:craig@ball.net">craig@ball.net</a></p> <p>Tel: 512/ 514.0182 Mobile: 713/ 320.6066</p>
<p><i>Helping Lawyers Master Technology</i></p> <p>Computer Forensic Examiner E-Discovery Special Master</p>	<p>3723 Lost Creek Blvd. Austin, Texas 78735</p> <p><a href="http://www.craigball.com">www.craigball.com</a></p>