Musings on Electronic Discovery

“Ball in Your Court”
April 2005 – July 2013

The American Society of Business Publication Editors and Trade Association Business Publications International have chosen The *Law Technology News* column “Ball in Your Court” as their Gold or Silver Medal honoree for “Best Regular Column,” “Best Contributed Column” or “Best Feature Series” every year since 2006.

The blog, Ball in Your Court, is published at www.ballinyourcourt.com.

The author gratefully acknowledges the skilled editorial contributions of Diana Ball and Monica Bay to the print works, and accepts sole responsibility for the blog posts.

2005

2006

2007

2008

2009

2010

2011

2012

2013

About the Author

Craig Ball, of Austin is a Board Certified Texas trial lawyer, law professor 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 Adjunct Professor teaching Electronic Discovery and Digital Evidence at the University of Texas School of Law and as a lecturer in computer forensics and electronic evidence for the Department of Justice, National Judicial Center, Texas Office of the Attorney General 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 1,000 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 26 years. He and Diana have two delightful collegians 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

The DNA of Data
by Craig Ball

[Originally published in Law Technology News, April 2005]

Discovery of electronic data compilations has been part of American litigation for two generations, during which time we've seen nearly all forms of information migrate to the digital realm. Statisticians posit that only five to seven percent of all information is "born" outside of a computer, and very little of the digitized information ever finds its way to paper. Yet, despite the central role of electronic information in our lives, electronic data discovery (EDD) efforts are either overlooked altogether or pursued in such epic proportions that discovery dethrones the merits as the focal point of the case. At each extreme, lawyers must bear some responsibility for the failure. Few of us have devoted sufficient effort to learning the technology, instead deluding ourselves that we can serve our clients by continuing to focus on the smallest, stalest fraction of the evidence: paper documents. When we do garner a little knowledge, we abuse it like the Sorcerer's Apprentice, by demanding production of "any and all" electronic data and insisting on preservation efforts sustainable only through operational paralysis. We didn't know how good we had it when discovery meant only paper.

However, electronic evidence isn't going away. It's growing…exponentially, and some electronic evidence items, like databases, spreadsheets, voice mail and video, bear increasingly less resemblance to paper documents. Proposed changes in the rules of procedure wending their way through the system require lawyers to discuss ways to preserve electronic evidence, select formats in which to produce it and manage volumes of information dwarfing the Library of Congress. Litigators must learn it or find a new line of work.

My goal for this column is to help make electronic discovery and computer forensics a little easier to understand, never forgetting that this is exciting, challenging—and very cool—stuff.

Accessible versus Inaccessible

You can't talk about EDD today without using the "Z" word: Zubulake (pronounced "zoo-boo-lake"). Judge Shira Scheindlin's opinions in *Zubulake v. UBS Warburg, L.L.C.,* 217 F.R.D. 309 (S.D.N.Y. 2003) triggered a whirlwind of discussion about EDD. Judge Scheindlin cited the "accessibility" of data as the threshold for determining issues of what must be produced and who must bear the cost of production. Accessible data must be preserved, processed and produced at the producing party's cost, while inaccessible data is available for good cause and may trigger cost shifting.

But what makes data "inaccessible?" Is it a function of the effort and cost required to make sense of the data? If so, do the boundaries shift with the skill and resources of the producing party such that ignorance is rewarded and knowledge penalized? To understand when data is truly inaccessible requires a brief look at the DNA of data.

Everything's Accessible

Computer data is simply a sequence of ones and zeroes. Data is only truly inaccessible when you can't read the ones and zeroes or figure out where the sequence starts. To better grasp this, imagine you had the unenviable responsibility of typing the complete works of Shakespeare on a machine with only two keys, "A" and "B," and if you fail, all the great works of the Bard would be lost forever. As you ponder this seemingly impossible task, you'd figure out that you could encode the alphabet using sequences of As and Bs to represent each of the twenty-six capital letters, their lower case counterparts, punctuation and spaces. The uppercase "W" might be "ABABABBB" and the uppercase "S," "ABABAABB." Cumbersome, but feasible. Armed with the code and knowing where the sequence begins, a reader can painstakingly reconstruct every lovely foot of iambic pentameter.

This is just what a computer does when it stores data in ones and zeroes, except computers encode many "alphabets" and work with sequences billions of characters long. Computer data is only "gone" when the media that stores it is obliterated, overwritten or strongly encrypted without a key. This is true for all digital media, including backup tapes and hard drives. But, inaccessibility due to damage, overwriting or encryption is rarely raised as grounds for limiting e-discovery or shifting costs.

Just Another Word for Burdensome?

Frequently, lawyers will couch a claim of undue burden in terms of inaccessibility, arguing that it's too time-consuming or costly to restore the data. But, burden and inaccessibility are opposite sides of the same coin, and "inaccessibility" adds nothing to the mix but confusion. Arguing *both* burden and inaccessibility is two bites at the apple.

Worse, there is a risk in branding particular media as "inaccessible." Parties resisting discovery shouldn't be relieved of the obligation to demonstrate undue burden simply because evidence resides on a backup tape. We must be vigilant to avoid a reflexive calculus like:

All backup tapes are inaccessible
▼
Inaccessible means undue burden presumed
▼
Good cause showing required for production
▼
Requesting party pays cost of conversion to "accessible" form.

Zubulake put EDD on every litigator's and corporate counsel's radar screen and proved invaluable as a provocateur of long-overdue debate about electronic discovery. Still, its accessibility analysis is not a helpful touchstone, especially in a fast-moving field like computing. Codifying it in proposed amendments to F.R.C.P. Rule 26(b)(2) would perpetuate a flawed standard. Even if that occurs, don't be cowed by the label, "inaccessible," and don't shy away from seeking discovery of relevant media just because it's cited as an example of something inaccessible. Instead, require the producing party to either show that the ones and zeroes can’t be accessed or demonstrate that production entails an undue burden.

Unclear on the Concept
by Craig Ball

[Originally published in Law Technology News, May 2005]

A colleague buttonholed me at the American Bar Association's recent TechShow and asked if I'd visit with a company selling concept search software to electronic discovery vendors. Concept searching allows electronic documents to be found based on the ideas they contain instead of particular words. A concept search for "exploding gas tank" should also flag documents that address fuel-fed fires, defective filler tubes and the Ford Pinto. An effective concept search engine "learns" from the data it analyzes and applies its own language intelligence, allowing it to, e.g., recognize misspelled words and explore synonymous keywords.

I said, "Sure," and was delivered into the hands of an earnest salesperson who explained that she was having trouble persuading courts and litigators that the company's concept search engine worked. How could they reach them and establish credibility? She extolled the virtues of their better mousetrap, including its ability to catch common errors, like typing "manger" when you mean "manager."

But when we tested the product against its own 100,000 document demo dataset, it didn't catch misspelled terms or search for synonyms. It couldn't tell "manger" from "manager." Phrases were hopeless. Worse, it didn't reveal its befuddlement. The program neither solicited clarification of the query nor offered any feedback revealing that it was clueless on the concept.

The chagrined company rep turned to her boss, who offered, "100,000 documents are not enough for it to really learn. The program only knows a word is misspelled when it sees it spelled both ways in the data it's examining and makes the connection."

The power of knowledge lies in using what's known to make sense of the unknown. If the software only learns what each dataset teaches it, it brings nothing to the party. Absent from the application was a basic lexicon of English usage, nothing as fundamental as Webster's Dictionary or Roget's Thesaurus. There was no vetting for common errors, no "fuzzy" searching or any reference foundation. The application was the digital equivalent of an idiot savant (and I'm taking the savant on faith because this application is the plumbing behind some major vendors' products).

Taking the Fifth?

In the Enron/Andersen litigation, I was fortunate to play a minor role for lead plaintiff's counsel as an expert monitoring the defendant's harvesting and preservation of electronic evidence. The digital evidence alone quickly topped 200 terabytes, far more information than if you digitized all the books in the Library of Congress. Printed out, the paper would reach from sea-to-shining sea several times. These gargantuan volumes — and increasingly those seen in routine matters — can't be examined without automated tools. There just aren't enough associates, contract lawyers and paralegals in the world to mount a manual review, nor the money to pay for it. Of necessity, lawyers are turning to software to divine relevancy and privilege.

But as the need for automated e-discovery tools grows, the risks in using them mount. It's been 20 years since the only study I've seen pitting human reviewers against search tools. Looking at a (paltry by current standards) 350,000 page litigation database, the computerized searches turned up just 20 percent of the relevant documents found by the flesh-and-bone reviewers.

The needle-finding tools have improved, but the haystacks are much, much larger now. Are automated search tools performing well enough for us to use them as primary evidence harvesting tools?

Metrics for a Daubert World

Ask an e-discovery vendor about performance metrics and you're likely to draw either a blank look or trigger a tap dance that would make the late Ann Miller proud. How many e-discovery products have come to market without any objective testing demonstrating their efficacy? Where is the empirical data about how concept searching stacks up against human reviewers? How has each retrieval system performed against the National Institute of Standards and Technology text retrieval test collections?

If the vendor response is, "We've never tested our products against real people or government benchmarks," how are users going to persuade a judge it was a sound approach come the sanctions hearing?

We need to apply the same Daubert-style standards [*Daubert v. Merrell Dow Pharmaceuticals* (92-102) 509 U.S. 579 (1993)] to these systems that we would bring to bear against any other vector for junk science: Has it been rigorously tested? Peer-reviewed? What are the established error rates?

Calibration and Feedback

Like the airport security staff periodically passing contraband through the x-ray machines and metal detectors to check the personnel and equipment, automated search systems must be periodically tested against an evolving sample of evidence scrutinized by human intelligence. Without this ongoing calibration, the requesting party may persuade the court that your net's so full of holes, only a manual search will suffice. If that happens, what can you do but settle?

Thanks to two excellent teachers, I read Solzhenitsyn in seventh grade and Joyce Carol Oates in the ninth. I imagine that if I re-read those authors today, I'd get more from them than my adolescent sensibilities allowed. Likewise, if software gets smarter as it looks at greater and greater volumes of information, is there a mechanism to revisit data processed *before* the software acquired its "wisdom" lest it derive no more than my 11-year-old brain gleaned from One Day in the Life of Ivan Denisovitch? What is the feedback loop that ensures the connections forged by progress through the dataset apply to the entire dataset?

For example, in litigation about a failed software development project, the project team got into the habit of referring to the project amongst themselves as the "abyss" and the "tar baby." Searches for the insider lingo, as concepts or keywords, are likely to turn up e-mails confirming that the project team knowingly poured client monies into a dead end.

If the software doesn't make this connection until it processes the third wave of data, what about what it missed in waves one and two? Clearly, the way the data is harvested and staged impacts what is located and produced. Of course, this epiphany risk—not realizing what you saw until after you've reviewed a lot of stuff—afflicts human examiners too, along with fatigue, inattentiveness and sloth to which machines are immune.

But, we trust that a diligent human examiner will sense when a newly forged connection should prompt re-examination of material previously reviewed.

Will the software know to ask, "Hey, will you re-attach those hard drives you showed me yesterday? I've figured something out."

Concept Search Tools

Though judges and requesting parties must be wary of concept search tools absent proof of their reliability, even flawed search tools have their place in the trial lawyer's toolbox.

Concept searching helps overcome limitations of optical character recognition, where seeking a match to particular text may be frustrated by OCR's inability to read some fonts and formats. It also works as a lens through which to view the evidence in unfamiliar ways, see relationships that escaped notice and better understand your client's data universe while framing filtering strategies.

I admire the way EDD-savvy Laura Kibbe, in-house counsel for pharmaceutical giant Pfizer, Inc., uses concept searching. She understands the peril of using it to filter data and won't risk having to explain to the court how concept searching works and why it might overlook discoverable documents. Instead, Laura uses concept searching to brainstorm keywords for traditional word searches and then uses it again as a way to prioritize her review of harvested information.

For producing parties inclined to risk use of concept searching as a filtering tool, inviting the requesting party to contribute keywords and concepts for searching is an effective strategy to forestall finger pointing about non-production. The overwhelming volume and the limitations of the tools compel transformation of electronic discovery to a collaborative process. Working together, both sides can move the spotlight away from the process and back onto the merits of the case.

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.

Give Away Your Computer

by Craig Ball

[Originally published in Law Technology News, July 2005]

With the price of powerful computer systems at historic lows, who isn't tempted to upgrade? But, what do you do with a system you've been using if it's less than four or five-years old and still has some life left in it? Pass it on to a friend or family member or donate it to a school or civic organization and you're ethically obliged to safeguard client data on the hard drive. Plus, you'll want to protect your personal data from identity thieves and snoopers. Hopefully you already know that deleting confidential files and even formatting the drive does little to erase your private information—it's like tearing out the table of contents but leaving the rest of the book. How do you be a Good Samaritan without jeopardizing client confidences and personal privacy?

Options

One answer: replace the hard drive with a new one before you donate the old machine. Hard drives have never been cheaper, and adding the old hard drive as extra storage in your new machine ensures easy access to your legacy data. But, it also means going out-of-pocket and some surgery inside both machines—not everyone's cup of tea.

Alternatively, you could remove or destroy the old hard drive, but those accepting older computers rarely have the budget to buy hard drives, let alone the technician time to get donated machines running. Donated systems need to be largely complete and ready to roll.

Probably the best compromise is to wipe the hard drive completely and donate the system recovery disk along with the system. Notwithstanding some largely theoretical notions, once you overwrite every sector of your hard drive with zeros or random characters, your data is gone forever. The Department of Defense recommends several passes of different characters, but just a single pass of zeros is enough to frustrate all computer forensic data recovery techniques in common use.

Free is Good

You can *buy* programs to overwrite your hard drive, but why do so? Effective erasure tools are available as free downloads from the major hard drive manufacturers, and most work on other manufacturers' drives. Western Digital offers its Data Lifeguard Diagnostic Tool at http://support.wdc.com/download. Seagate's DiscWizard Starter Edition is found at www.seagate.com/support/disc/drivers/discwiz.html and Maxtor's PowerMax utilities is found by drilling down from www.maxtor.com/support. DBAN (for Darik's Boot and Nuke), a free Linux program, will also obliterate all data on a Windows system and is available at http://dban.sourceforge.net/. Each application offers bells-and-whistles, but all you're seeking is the ability to create a boot floppy that can write zeroes to the hard drive. If your system has no floppy drive, each site also offers a boot CD image download.

Why a boot floppy or CD? Because no wiping program running under Windows can erase all of the data on a Windows drive. Running under DOS (or, in the case of DBAN, Linux) insures that no file is locked out to the wiping utility while it does its job. To this end, check to be sure that whatever wiping application you select "sees" the entire hard drive. If it only recognizes, say, the first 32 GB of a 40 GB drive, check your settings or use a different utility. Fortunately, these utilities are user-friendly and report what they see and do.

Careful!

Wiping every sector on a hard drive is a time consuming process. Allow hours of (largely) unattended operation to get the job done, and if it's an option, be sure to select a full overwrite (or "low level format") and not a quick version. There are no shortcuts to overwriting every sector to sterilize a drive. Check to be sure there is only one hard drive in the system. If multiple drives are present, wipe each of them. Above all, understand that *there is no turning back from this kind of data erasure*. No Recycle Bins. No Undo command. No clean room magic. Be absolutely certain you have another working copy of anything you mean to keep.

An Important Courtesy

When you sterilize a drive, your privileged data obliterated along with the operating system and all applications. A wiped drive can’t boot a computer, but can return to service if you remember to donate the system restore disk with the hardware. For computers lacking restore disks, supply the operating system installation disk and any application disks you wish to donate. As long as you're not continuing to use the same applications loaded from the same disks (or copies) on your new machine, your end user license is likely to be freely transferable. If the donated system came without disks, you or your recipient will need to contact the manufacturer and request a restore disk. If, as is often the case in larger firms, the operating systems are site licensed, it may be a violation of that license to share them. Your recipient will then need to purchase their own license or seek out someone who'll donate an operating system. School districts typically have their own site licenses.

Dodging Blasts from The Past

Be sure to caution your recipient that it's very important to promptly download critical security patches and service packs for the restored operating system and applications. A restored machine is like a step back in time to when many now-closed security holes were wide open, so the recipient needs to slam these vulnerabilities shut at the very first connection to the Internet.

Help for the Helper

Worries about data security needn't keep you from helping others by donating your used computer. For additional guidance, contact TechSoup (www.techsoup.org) or the National Cristina Foundation (www.cristina.org), and seek out — or organize — the computer donation program in your community.

Breaking News:
Clearing your donated, sold or discarded hard drives of sensitive information isn't just good practice, it's now also required by law. Effective June 1, 2005, the Federal Trade Commission's Disposal Rule 16 CFR Part 682, requires businesses—including lawyers and law firms—to take reasonable measures to dispose of sensitive information derived from credit reports and background checks so that the information cannot practicably be read or reconstructed. The Rule, which applies to both paper and digital media, requires implementing and monitoring compliance with disposal policies and procedures for this information. Comments to the rule suggest using disc wiping utilities, but also offer that electronic media may be economically disposed of by "simply smashing the material with a hammer." Sounds like a great stress reliever, but don't forget your safety goggles!

Don't Try This at Home
by Craig Ball

[Originally published in Law Technology News, August 2005]

The legal assistant on the phone asked, "Can you send us copies of their hard drives?"

As court-appointed Special Master, I'd imaged the contents of the defendant's computers and served as custodian of the data for several months. The plaintiff's lawyer had been wise to lock down the data before it disappeared, but like the dog that caught the car, he didn't know what to do next. Now, with trial a month away, it was time to start looking at the evidence.

"Not unless the judge orders me to give them to you," I replied.

The court had me act as custodian because the discoverable evidence on a hard drive lives cheek by jowl with all manner of sensitive stuff, such as attorney-client communications, financial records and pictures of naked folks engaged in recreational activity. In suits between competitors, intellectual property and trade secrets such as pricing and customer contact lists need protection from disclosure when not evidence. As does all that full-of-surprises deleted data accessible by forensic examination.

"Even if the court directs me to turn over the drive images, you probably won't be able to access the data without expert assistance."

I explained that, like most computer forensic specialists, I store the contents of hard drives as a series of compressed image files, not as bootable hardware that can be attached to a computer and examined. Doing so is advantageous because the data is easier to access, store and authenticate, as well as far less prone to corruption by the operating system or through examination. Specialized software enables me to assemble the image files as a single virtual hard drive, identical in every way to the original. On those rare occasions when a physical duplicate is needed, I reconstitute those image files to a forensically sterile hard drive and use cryptographic algorithms to demonstrate that the restored drive is a faithful counterpart of the original. Of course, putting the digital toothpaste back in the tube that way takes time and costs money.

"Do we ask the court for a restored drive?"

"You could," I said, "and you might get it if the other side doesn't object."

Incredibly, lawyers who'd never permit the opposition to fish about in their client's home or office blithely give the green light when it comes to trolling client hard drives. No matter how much you want to demonstrate good faith or that your client has "nothing to hide," be wary of allowing the other side to look at the drives.

Even when you've checked the contents, you can't see all that a forensic exam can turn up, and your client may not tell you about all those files she deleted last night.

"But," I warned, "as soon as you attach the drive to your computer and start poking around, you'll alter the evidence."

Microsoft Windows acts like a dog marking territory. As soon as you connect a hard drive to Windows, the operating system writes changes to the drive. Forensic examiners either employ devices called "write blockers" to intercept these alterations or perform their examination using operating systems less inclined to leave their mark all over the evidence. Without similar precautions, opening files, reading e-mail or copying data irretrievably alters file metadata, the data-about-data revealing, inter alia, when a file was last modified, accessed or created. You may find the smoking gun, but good luck getting it into evidence when it emerges you unwittingly altered the data! This is why smart lawyers never "sneak a peek" at digital evidence.

"It'd be a violation of the software licensing to use the programs on the duplicate so you'll need to have the right software to read the e-mail and other documents and to crack any passwords you run into. However, you can't load your software on the duplicate drive because that will overwrite recoverable deleted files. Don't forget to take steps to isolate the system you'll use for examination from your office network and the internet as well as to…."

She stopped me. "We shouldn't be doing this ourselves, should we?"

"Not unless you know what you're doing. Anyway, I doubt the court will allow it without a showing of good cause and some provision to protect privileged and non-discoverable confidential data."

Now I got the question I was waiting for: "What should we do?"

"As the court's neutral," I answered, "I'm not in a position to answer that question, but before I'd burn a lot of time and money pursuing electronic discovery of particular media, I'd work out the answers to, 'What's this case about, and what am I really looking for?'"

What I wanted to add is that electronic discovery is no more about hard drives than traditional discovery was "about" paper. The hard drive is just a gigantic file cabinet, locked up like some Houdini vanishing act and packed with contents penned in Sanskrit. We don't gear discovery to metal boxes, big or small.

Sure, it's smart to focus on specific media and systems when you seek preservation, but when your goal is discovery, media ceases to be an end in itself. Then, the objectives are the e-mail, documents and other digital evidence relating to the issues in the case, narrowly targeted by time, topic, and custodian. Sorry Marshall McLuhan, it's not the medium. It's the message.

Yours, Mine and Ouch!

by Craig Ball

[Originally published in Law Technology News, September 2005]

When star performer Sarit Shmueli was fired from her real estate agent job with The Corcoran Group, she returned to her desk to find that she'd been locked out of the company computer system. Shmueli was barred from retrieving virtual property, in particular, her client list. When demands for her data were unavailing, Shmueli sued Corcoran for three million dollars.

Though opinions vary on whether to drop the axe on Friday or Monday, human resource experts agree that immediate steps must be taken to block the fired individual's access to company computers. That's wise, considering more than a few heading to the door have trashed important files or attempted to sabotage entire networks. But, what about personal computer data? Fired workers are routinely furnished a cardboard box and the supervised opportunity to collect personal belongings before being escorted to the door. Is denying access to personal data stored on a company computer a violation of the discharged party's property rights?

Overruling a motion for summary judgment, a recent decision in Sarit Shmueli's case holds that when The Corcoran Group blocked Shmueli from accessing her records on the company computer system, Corcoran may have been guilty of conversion. "There should be no reason why [a] practical view should not apply equally to the present generation of documents—electronic documents—which are just as vulnerable to theft and wrongful transfer as paper documents, if not more so," reasoned New York Supreme Court Justice Herman Cahn. *Shmueli v. The Corcoran Group*, 104824/03 (N.Y.S.Ct. July 25, 2005) online at http://decisions.courts.state.ny.us/fcas/FCAS_docs/2005JUL/30010482420033SCIV.PDF.

Of course, Justice Cahn is right to ascribe value to electronic documents, but what are the ramifications of affording a conversion cause of action to employees and contractors against companies that refuse to recognize personal property interests in data stored on their systems?

The bulk of my work as a special master in computer forensics revolves around employees who've allegedly purloined company data to build rival businesses. If the employees haven't already jumped ship, they're canned as soon as the boss realizes they're playing for the other team. Then, much time and money goes into assessing what information was taken and what tracks were covered. Initial reports of the Shmueli decision made me wonder if workers being shown the door might now be entitled to access the company network for the purpose of copying information they deem to be their own property. Are they perhaps at liberty to delete such "personal" items, too? What of the longstanding notion that anything stored on the company computer belongs to the company?

The Order makes clear that Shmueli's unique employment status as independent contractor and not an employee played a decisive role in the court's view that Corcoran may be liable for converting its former agent's digital property just as if the data had been printed to paper. Though Corcoran asserted its ownership of the computer trumped the plaintiff's rights, the court countered that because Shmueli worked with Corcoran and not as an employee of Corcoran, the computer was "licensed" for plaintiff's use to facilitate the independent contract.

The problem with the court's distinction is that employers enjoy no privilege to confiscate and convert personal property. Just because you have your spouse's photo on your desk at work doesn't mean the boss can keep it when you're fired. Instead, the employer's right to retain the information on the computer must be grounded on a presumption that information stored on an employer's computer is either company property in the first instance or acquires that character by virtue of being the fruit of an employee's labors.

The commingling of personal and business property on company computers is a growing concern. From personal e-mail to screenplays written on the lunch hour, employers should anticipate the obligation to identify, segregate and return personal data "belonging" to fired employees. Likewise, employees need to tailor their personal use of company systems to the possibility of lock out.

This is going to be harder than it sounds because employers aren't disposed to grant "computer visitation rights" after a firing to let former employees—particularly ones now working for the competition—clean out their digital lockers. But without such access, how well could any of us identify from memory all personal items on our office computers, and what fired employee wants a former boss or co-worker sifting through their personal information?

A well-drafted acceptable use policy signed by the employee helps by defining rights and responsibilities in the use of business computer systems during employment. However, if employers are lax in their enforcement of the policy such that employees harbor a reasonable expectation of privacy in their use of company systems, terminations may entail the added ugliness of a custody battle over data and potential liability for conversion. Even where a violation of an AUP is clear, will courts decide that personal data escheats to a former employer simply because it's found on their virtual premises? That is, which schoolyard canon will prevail: the altruistic "If it's not yours, give it back," or the Draconian "You shouldn't have brought it here, so now it's mine?"

The Path to E-Mail Production

(Part I of IV)

by Craig Ball

[Originally published in Law Technology News, October 2005]

Asked, "Is sex dirty," Woody Allen quipped, "Only if it's done right." That's electronic discovery: if it's ridiculously expensive, enormously complicated and everyone's lost sight of the merits of the case, you're probably doing it right.

But it doesn't have to be that way. Over the next few issues, we'll walk a path to production of e-mail — perhaps the trickiest undertaking in EDD. The course we take may not be the shortest or easiest, but that's not the point. We're trying to avoid stepping off a cliff. Not every point is suited to every production effort, but all deserve consideration.

Think Ahead

EDD missteps are painfully expensive, or even unredeemable, if data is lost. Establish expectations at the outset.

Will the data produced:

- Integrate paper and electronic evidence?
- Be electronically searchable?
- Preserve all relevant metadata from the host environment?
- Be viewable and searchable using a single application, such as a web browser?
- Lend itself to Bates numbering?
- Be easily authenticable for admission into evidence?

Meeting these expectations hinges on what you collect along the way through identification, preservation, harvest and population.

Identification

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

Consider a recent case where I was asked to assess whether a departing associate stole files and diverted cases. The firm used a Microsoft Exchange e-mail server, so I could have collected or searched the associate's e-mail there. Had I looked only at the server, 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. Or the

Microsoft Outlook archive file (.pst) and offline synchronization file (.ost), both stored on a laptop hard drive, and holding thousands more e-mails.

Just looking at the server wouldn't have revealed the stolen data or the diverted business—searching elsewhere uncovered a treasure trove of damning evidence.

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

Easily Accessible:

- 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)
- E-mail stored in active files on local or external hard drives and network shares: User workstation hard drives (e.g., .pst, .ost files for Outlook and .nsf for Lotus Notes), laptops, "local" e-mail data files stored on networked file servers, mobile devices, and home systems, particularly those with remote access to networks.
- Nearline e-mail: Optical "juke box" devices, backups of user e-mail folders.
- Offline e-mail stored in networked repositories: e.g., Zantaz EAS, EMC EmailXtender, Waterford MailMeter Forensic, etc.

Accessible, but Often Overlooked:

- E-mail residing on remote servers: ISPs (IMAP, POP, HTTP servers), Gmail, Yahoo Mail, Hotmail, etc.
- E-mail forwarded and cc'd to third-party systems: Employee forwards e-mail to self at personal e-mail account.
- E-mail threaded behind subsequent exchanges: Contents diverge from earlier exchanges lodged in body of e-mail.
- 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.
- E-mail saved to other formats: PDF, .tiff, .txt, .eml, 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.)
- Print outs to paper.

More Difficult to Access:

- Offline e-mail on server back up media: Back up tapes (e.g., DLT, AIT)
- E-mail in forensically accessible areas of local hard drives: Deleted e-mail, internet cache, unallocated clusters.

The issues in the case, key players, relevant times, agreements between the parties and orders of the court determine the extent to which locations must be examined; however, the failure to identify all relevant e-mail carries such peril that caution should be the watchword. Isn't it wiser to invest more to know exactly what the client has than concede at the sanctions hearing the client failed to preserve and produce evidence it didn't know it had because no one bothered to look for it?

Electronic evidence is fragile and ever changing, so once you've found the e-mail evidence, you must guard against its loss or corruption.

Next month, we'll walk through preservation thicket.

The Path to Production: Retention Policies That Work

(Part II of IV)
by Craig Ball

[Originally published in Law Technology News, November 2005]

In this second in a series, we continue down the path to production of electronic mail. Last month, I reminded you to look beyond the e-mail server to the many other places e-mail hides. Now, having identified the evidence, we're obliged to protect it from deletion, alteration and corruption.
Preservation

Anticipation of a claim is all that's required to trigger a duty to preserve potentially relevant evidence, including fragile, ever-changing electronic data. Preservation allows backtracking on the path to production, but fail to preserve evidence and you've burned your bridges.

Complicating our preservation effort is the autonomy afforded e-mail users. They create quirky folder structures, commingle personal and business communications and — most dangerous of all — control deletion and retention of messages.

Best practices dictate that we instruct e-mail custodians to retain potentially relevant messages and that we regularly convey to them sufficient information to assess relevance in a consistent manner. In real life, hold directives alone are insufficient. Users find it irresistibly easy to delete data, so anticipate human frailty and act to protect evidence from spoliation at the hands of those inclined to destroy it. Don't leave the fox guarding the henhouse.

Consider the following as parts of an effective e-mail preservation effort:

- Litigation hold notices to custodians, including clear, practical and specific retention directives. Notices should remind custodians of relevant places where e-mail resides, but not serve as a blueprint for destruction. Be sure to provide for notification to new hires and collection from departing employees.
- Suspension of retention policies that call for purging e-mail.
- Suspension of re-use (rotation) of back up media containing e-mail.
- Suspension of hardware and software changes which make e-mail inaccessible.
- Replacing backup systems without retaining the means to read older media.
- Re-tasking or re-imaging systems for new users.
- Selling, giving away or otherwise disposing of systems and media.
- Preventing custodians from deleting/ altering/corrupting e-mail.
- Immediate and periodic "snapshots" of relevant e-mail accounts.
- Modifying user privileges settings on local systems and networks.
- Archival by auto-forwarding selected e-mail traffic to protected storage.

- Restricting activity like moving or copying files tending to irreparably alter file metadata.
- Packet capture of Instant Messaging (traffic or effective enforcement of IM prohibition.
- Preserve potential for forensic recovery.
- Imaging of key hard drives or sequestering systems.
- Suspension of defragmentation.
- Barring wiping software and encryption, with audit and enforcement.

Threshold issue

A threshold preservation issue is whether there is a duty of preservation going forward, e.g., with respect to information created during the pendency of the action. If not, timely harvest of data, imaging of drives and culling of relevant backups from rotation may sufficiently meet the preservation duty so as to allow machines to be re-tasked, systems upgraded and back up tape rotation re-initiated. Securing guidance from the court and cooperating with opposing counsel to fashion practical preservation orders help insulate a producing party from subsequent claims of spoliation.

The Knowledge Hurdle

Thanks to a string of recent, high profile decisions, litigants are gradually awakening to their obligation to preserve electronic evidence. Still, attitudes often range from insufficient ("We'll just stop rotating backup tapes") to incredulous ("Why would we need to preserve voice mail?").

One hurdle is the lack of knowledge on the part of those charged with the responsibility to design and direct preservation efforts: too many don't understand what and how data change or what triggers those changes. They fail to appreciate how the pieces fit together.

For example, in a lawsuit concerning a plant explosion, the defendant, a major oil company, preserved monthly "full" backups of its e-mail server but failed to hang on to four weeks of incremental backups immediately preceding the blast.

A full back up is a snapshot of the e-mail system at a single point in time. An incremental back up records changes to the e-mail system between snapshots. Did someone think that full back ups were cumulative of the incremental sessions? If so, they missed the fact that any e-mail received and deleted between snapshots might exist on the incremental backups but be absent from the monthly tapes. They didn't consider how the pieces fit together.

If you've done a good job identifying where e-mail lives, preservation is largely a matter of duplicating the e-mail without metadata corruption or shielding it from subsequent loss or alteration. Both demand technical competence, so you'll need expert help the first time or two. If you ask questions and seek out reasons behind actions, knowledge gained from one effort will guide you through the next.

Minimize Burden and Cost

With digital storage costs at all time lows, it's tempting to minimize spoliation risks by simply keeping everything. ***Don't***. Keeping everything merely postpones and magnifies the cost and complexity of production. Yet, you can suspend document retention and tape rotation without triggering a costly data logjam, if you adapt your preservation from reflexive to responsive.

Reflexive preservation describes steps you take while figuring out what's relevant and what's not. It's immediate and encompassing action to preserve the status quo while you sift the facts, forge agreements with opponents or seek guidance from the court. Calling a halt to back up tape rotation or suspending retention policies is reflexive preservation.

Reflexive preservation is a triage mechanism and a proper first response; but it's too expensive and disruptive for the long haul. Instead, convert reflexive preservation to responsive preservation by continually tweaking your preservation effort to retain only what's relevant to claims or necessary to meet business and regulatory obligations. Narrow the scope of preservation by agreement, motion practice and sound, defensible judgment.

Having identified the e-mail evidence and preserved it, we need to collect it and make it accessible for review and searching. Next month, we hike up harvest hill and perambulate population pass. Wear sensible shoes!

The Path to Production: Harvest and Population

(Part III of IV)
by Craig Ball

[Originally published in Law Technology News, December 2005]

On the path to production, we've explored e-mail's back alleys and trod the mean streets of the data preservation warehouse district. Now, let's head to the heartland for harvest time. It's data harvest time.
After attorney review, data harvest is byte-for-byte the costliest phase of electronic data discovery. Scouring servers, local hard drives and portable media to gather files and metadata is an undertaking no company wants to repeat because of poor planning.

The Harvest
Harvesting data demands a threshold decision: Do you collect all potentially relevant files, then sift for responsive material, or do you separate the wheat from the chaff in the field, collecting only what reviewers deem responsive? When a corporate defendant asks employees to segregate responsive e-mail, (or a paralegal goes from machine-to-machine or account-to-account selecting messages), the results are "field filtered."

Field filtering holds down cost by reducing the volume for attorney review, but it increases the risk of repeating the collection effort, loss or corruption of evidence and inconsistent selections. If keyword or concept searches alone are used to field filter data, the risk of under-inclusive production skyrockets.

Initially more expensive, comprehensive harvesting (unfiltered but defined by business unit, locale, custodian, system or medium), saves money when new requests and issues arise. A comprehensive collection can be searched repeatedly at little incremental expense, and broad preservation serves as a hedge against spoliation sanctions. Companies embroiled in serial litigation or compliance production benefit most from comprehensive collection strategies.

A trained reviewer "picks up the lingo" as review proceeds, but a requesting party can't frame effective keyword searches without knowing the argot of the opposition. Strategically, a producing party requires an opponent to furnish a list of search terms for field filtering and seeks to impose a "one list, one search" restriction. The party seeking discovery must either accept inadequate production or force the producing party back to the well, possibly at the requesting party's cost.

Chain of Custody
Any harvest method must protect evidentiary integrity. A competent chain of custody tracks the origins of e-evidence by, e.g., system, custodian, folder, file and dates. There's more to e-mail than what you see on screen, so it's wise to preempt attacks on authenticity by preserving complete headers and encoded attachments.

Be prepared to demonstrate that no one tampered with the data between the time of harvest and its use in court. Custodial testimony concerning handling and storage may suffice, but better approaches employ cryptographic hashing of data — "digital fingerprinting" — to prove nothing has changed.

Metadata

There's more to an e-mail than its contents: there's metadata, too. Each e-mail is tracked and indexed by the e-mail client ("application metadata") and every file holding e-mail is tracked and indexed by the computer's file system ("system metadata"). E-mail metadata is important evidence in its own right, helping to establish whether and when a message was received, read, forwarded, changed or deleted. Metadata's evidentiary significance garnered scant attention until Williams v. Sprint, 2005 W.L. 2401626 (D. Kan. Sept. 29, 2005), where in a dispute over production of spreadsheets, the court held that a party required to produce electronic documents as kept in the ordinary course of business must produce metadata absent objection, agreement or protective order.

System metadata is particularly fragile. Just copying a file from one location to another alters the file's metadata, potentially destroying critical evidence. Ideally, your data harvest shouldn't corrupt metadata, but if it may, archive the metadata beforehand. Though unwieldy, a spreadsheet reflecting original metadata is preferable to spoliation. EDD and computer forensics experts can recommend approaches to resolve these and other data harvest issues.

Processing and Population

However scrupulous your e-mail harvest, what you've reaped isn't ready to be text searched. It's a mish-mash of incompatible formats on different media: database files from Microsoft Exchange or Lotus Domino Servers, .PST and .NSF files copied from local hard drives, HTML fragments of browser-based e-mail and .PDF or .tiff images. Locked, encrypted and compressed, it's not text, so keyword searches fail.

Before search tools or reviewers can do their jobs, harvested data must be processed to populate the review set, i.e., deciphered and reconstituted as words by opening password-protected items, decrypting and decompressing container files and running optical character recognition on image files. Searching now will work, but it'll be slow going thanks to the large volume of duplicate items. Fortunately, there's a fix for that, too.

Next month: de-duplication, deliverables, documentation and the destination on the path to production.

The Path to Production: Are We There Yet?

(Part IV of IV)

by Craig Ball

[Originally published in Law Technology News, January 2006]

The e-mail's assembled and accessible. You could begin review immediately, but unless your client has money to burn, there's more to do before diving in: de-duplication. When Marge e-mails Homer, Bart and Lisa, Homer's "Reply to All" goes in both Homer's Sent Items and Inbox folders, and in Marge's, Bart's and Lisa's Inboxes. Reviewing Homer's response five times is wasteful and sets the stage for conflicting relevance and privilege decisions.

Duplication problems compound when e-mail is restored from backup tape. Each tape is a snapshot of e-mail at a moment in time. Because few users purge mailboxes month-to-month, one month's snapshot holds nearly the same e-mail as the next. Restore a year of e-mail from monthly backups, and identical messages multiply like rabbits.

De-Duplication

De-duplication uses metadata, cryptographic hashing or both to exclude identical messages. De-duplication may be implemented vertically, within a single mailbox or custodian, and horizontally, across multiple mailboxes and custodians. When questioning or prepping a witness, you'll want to see all relevant messages in the witness' mailbox, not just unique messages; so track and log de-duplication to facilitate re-population of duplicated items. De-duplication works best when unique messages and de-duplication logs merge in a database, allowing a reviewer to reconstruct mailboxes.

Be wary of "horizontal" de-duplication when discovery strategies change. An e-mail sent to dozens of recipients de-duplicated from all but one custodian's mailbox may be lost forever if the one custodian's e-mail ends up not being produced.

Review Tools

Rather than plow through zillions of e-mails for responsive and privileged items, reviewers often turn to keyword or concept search tools. Automated search tools make short work of objective requests for "all e-mail between Simpson and Burns," but may choke on "all e-mail concerning plant safety." To frame effective keyword searches, you have to know the lingo describing events and objects central to the case. Even then, crucial communiqués like, "My lips are sealed" or "Excellent" may be missed.

Are tireless black box tools an adequate substitute for human review? The jury's still out. In a seminal study, keyword searching fared poorly, finding only about one-fifth of relevant items identified by human reviewers. However, litigation management consultant Anne Kershaw looked at an advanced search tool and found machines performed almost twice as well as humans. The safest course is to arm conscientious, well-trained reviewers with state-of-the-art search tools and work cooperatively with opposing counsel to frame searches. Even then, examine the mailboxes of key witnesses, message-by-message.

Redaction
Paper redaction was easy: We concealed privileged text using a black marker and photocopied. It's trickier to eradicate privileged and confidential information at the data layer of document image files and within encoded attachments and metadata. Run your approach by an expert.

Re-population
For production, should you re-populate to restore relevant, non-privileged items previously de-duplicated, or will the other side accept a de-duplication log? Never produce de-duplicated e-mail without memorializing that opposing counsel knows of the de-duplication and waives re-population.

Deliverables
There isn't just one "right" media or format for deliverables. Options for production media include network transmittal, external hard drives, optical disks, tape, online repositories and hard copies. Formats range from native (.pst), exported (.eml), text (.txt), load files (Concordance, Summation), image files with or without data layers (.pdf, .tiff) and delimited files. Evidence ill suited to .tiff production (databases, some spreadsheets, etc.), compels native production.

Documentation
Inevitably, something will be overlooked or lost, but sanctions need not follow every failure. Document diligence throughout the discovery effort and be prepared to demonstrate why bad decisions were sound at the time and under the circumstances. Note where the client looked for responsive information, what was found, how much time and money was expended, what was sidelined and why. Avoid sanctions by proving good faith.

Are We There Yet?
The path to production is a long and winding road, but it's heading in the right direction. Knowing how to manage electronic evidence is as vital to trial practice as the ability to draft pleadings or question witnesses. Don't forget what happened on Main Street when they built the Interstate. Paper discovery's the old road. E-discovery's the Interstate.

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 back up 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. Magoos 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.

Data Recovery: Lessons from Katrina

by Craig Ball

[Originally published in Law Technology News, April 2006]

When the sea reclaimed New Orleans and much of the Gulf Coast, hundreds of lawyers saw their computers and networks submerged. Rebuilding law practices entailed Herculean efforts to resurrect critical data stored on the hard drives in sodden machines.

Hard drives operate within such close tolerances that a drop of water or particle of silt that works its way inside can cripple them; yet, drives aren't sealed mechanisms. Because we use them from the beach to the mountains, drives must equalize air pressure through filtered vents called "breather holes." Under water, these breather holes are like screen doors on a submarine. When Hurricane Katrina savaged thousand of systems, those with the means and motivation turned to data recovery services for a second chance.

Data recovery, in the words of John Christopher, a veteran data recovery engineer at DriveSavers Inc., (www.drivesavers.com) is "open heart surgery" for hard drives. Companies such as Novato, Calif.-based DriveSavers and Ontrack Data Recovery (a division of Kroll Ontrack Inc., www.ontrack.com) are the courts of last resort for damaged drives. DriveSavers worked on dozens of Katrina-damaged drives, some submerged for weeks. Drive housings were full of crud, and recovery required finding identical drives and sacrificing them for compatible parts. DriveSavers reported that it was able to resurrect data from about two-thirds of the Katrina drives sent in.

Keep Them Wet

Ontrack's vice president of operations Todd Johnson reports that his company recovered useable data from about 70 percent of the 425 Katrina-damaged drives they received. All the drives required clean room treatment, with the best outcomes seen in those kept immersed in water or sealed in airtight plastic bags until delivery.

"Don't dry them out," Johnson warned, because that causes the heads that read data to become affixed to the platters.

Another factor favoring recovery was quick action. Whether you proceed with full-scale data recovery or not, promptly getting a drive cleaned and processed by a professional keeps your options open.

DriveSavers' Christopher echoed the need to move quickly and resist turning on the power to "see what works." He lamented that too many dim their prospects for recovery by letting a tech-savvy relative or electronics superstore take a stab at it.

Back It Up and Lock It Down

Despite the miracles performed by professional disk doctors, data recovery is unpredictable and very expensive. Add the cost of business interruption and frustrated clients, and the IT lesson from Katrina is **back it up and lock it down**. Even when systems survive, they may be inaccessible for prolonged periods due to closed or clogged roadways, hazardous conditions, areas cordoned off to prevent looting or loss of basic services, like electricity and telecommunications. You've got to have an accessible back up.

Katrina forced firms across the Gulf Coast to come to grips with flawed backup practices. Many had no backup system at all. Others were horrified to discover that never-tested backup tapes were useless. The proliferation of data on desktop drives and laptops off the backup grid meant that even those diligent about backup suffered data loss. Still others found to their dismay that backups stored in the same city were kept too close.

Lessons Learned

Whether the risk is hurricane, earthquake, fire, flood, terrorism, theft or disgruntled IT person, no firm is beyond disaster's reach. Here are steps to help weather the storm:

1. Back up critical data…regularly, thoroughly, obsessively.

2. Do periodic test restores of backed up data.

3. Ensure that key data on laptops and desktops is captured.

4. Mass disasters claim entire regions, so store backed up data out of harm's way. Consider online back- up, which safely ensconces data in distant servers, accessible via high-speed net connection from anywhere.

5. Know the answer to, "What would I grab if I had to leave right now?" Prepare for "grab and go" emergencies by using removable disk drive drawers or external hard drives.

Keep anti-static drive packaging and watertight containers on hand. For desktops, consider simple and inexpensive RAID configurations to make grab and go practical (see "Peace of Mind for a Pittance," in the March issue of Law Technology News, March 2005).

6. Encrypt the back up. Recent high-profile breaches of data security stemmed from poor management of backup media. Be sure your data back- ups are safe from prying eyes and that several in the firm know the encryption key. A backup you can't decrypt might as well have washed away.

Do-It-Yourself Digital Discovery

by Craig Ball

[Originally published in Law Technology News, May 2006]

Recently, a West Texas firm received a dozen Microsoft Outlook PST files from a client. Like the dog that caught the car, they weren't sure what to do next. Even out on the prairie, they'd heard of online hosting and e-mail analytics, but worried about the cost. They wondered: Did they really *need* an e-discovery vendor? Couldn't they just do it themselves?

As a computer forensic examiner, I blanch at the thought of lawyers harvesting data and processing e-mail in native formats. "Guard the chain of custody," I want to warn. "Don't mess up the metadata! Leave this stuff to the experts!" But the trial lawyer in me wonders how a solo/small firm practitioner in a run-of-the-mill case is supposed to tell a client, "Sorry, the courts are closed to you because you can't afford e-discovery experts."

Most evidence today is electronic, so curtailing discovery of electronic evidence isn't an option, and trying to stick with paper is a dead end. We've got to deal with electronic evidence in small cases, too. Sometimes, that means doing it yourself.

The West Texas lawyers sought a way to access and search the Outlook e-mail and attachments in the PSTs. It had to be quick and easy. It had to protect the integrity of the evidence. And it had to be cheap. They wanted what many lawyers will come to see they need: the tools and techniques to stay in touch with the evidence in smaller cases without working through vendors and experts.

What's a PST?

Microsoft Outlook is the most popular business e-mail and calendaring client, but don't confuse Outlook with Outlook Express, a simpler application bundled with Windows. Outlook Express stores messages in plain text, by folder name, in files with the extension .DBX. Outlook stores local message data, attachments, folder structure and other information in an encrypted, often-massive database file with the extension .PST. Because the PST file structure is complex, proprietary and poorly documented, some programs have trouble interpreting PSTs.

What about Outlook?

Couldn't they just load the files in Outlook and search? Many do just that, but there are compelling reasons why Outlook is the wrong choice for an electronic discovery search and review tool, foremost among them being that it doesn't protect the integrity of the evidence. Outlook changes PST files. Further, Outlook searches are slow, don't include attachments and can't be run across multiple mail accounts. I considered Google Desktop--the free, fast and powerful keyword search tool that makes short work of searching files, e-mail and attachments--but it has limited Boolean search capabilities and doesn't limit searches to specific PSTs.

Non-Starters

I also considered several extraction and search tools, trying to keep the cost under $200.00. One, a gem called Paraben E-Mail Examiner ($199.00), sometimes gets indigestion from PST files and won't search attachments. Another favorite, Aid4Mail Professional from Fookes Software ($49.95), quickly extracts e-mail and attachments and outputs them to several production formats, but Aid4Mail has no search capability. I looked at askSam software ($149.95), but after studying its FAQ and noodling with a demo, askSam proved unable to access any PST except the default profile on the machine—potentially commingling evidence e-mail and the lawyer's own e-mail.

dtSearch

The answer lay with dtSearch Desktop, a $199.00 indexed search application offering a command line tool that extracts the contents of PST files as generic message files (.MSG) indexed by dtSearch. In testing, once I got past the clunky command line syntax, I saved each custodian's mail to separate folders and then had dtSearch index the folders. The interface was wonderfully simple and powerful. Once you select the indices, you can use nearly any combination of Boolean, proximity, fuzzy or synonym searches. Search results are instantaneous and essential metadata for messages and attachments are preserved and presented. It even lets you preview attachments.

dtSearch lacks key features seen in products designed as e-discovery review tools, like the ability to tag hot documents, de-duplicate and redact privileged content. But you can copy selected messages and attachments to folders for production or redaction, preserving folder structures as desired. You can also generate printable search reports showing search results in context. In short, dtSearch works, but as a do-it-yourself e-mail tool, it's best suited to low volume/low budget review efforts.

Wave of the Future?

Any firm handles a fifty-page photocopy job in-house, but a fifty *thousand*-page job is going out to a copy shop. Likewise, e-discovery service providers are essential in bigger cases, but in matters with tight budgets or where the evidence is just e-mail from a handful of custodians, lawyers may need to roll up their sleeves and do it themselves.

Tips for Doing It Yourself

If you'd like to try your hand, dtSearch offers a free 30-day demonstration copy at www.dtsearch.com. Practice on your own e-mail or an old machine before tackling real evidence, and if you anticipate the need for computer forensics, leave the evidence machines alone and bring in an expert.

Whether e-mail is stored locally as a PST, in a similar format called an OST or remotely on an Exchange server depends on the sophistication and configuration of the e-mail system. To find a local PST file on a machine running Windows XP, NT or 2000, look for C:\Documents and Settings*Windows user name*\Local Settings\Application Data\Microsoft\Outlook\Outlook.pst. Archived e-mail resides in another file typically found in the same directory, called Archive.pst. Occasionally, users change default filenames or locations, so you may want to use Windows Search to find all files with a PST extension.

When you locate the PST files, record their metadata; that is, write down the filenames, where you found them, file sizes, and dates they were created, modified and last accessed (right click on the file and select Properties if you don't see this information in the directory). Be sure Outlook's not running and copy the PST files to read-only media like CD-R or DVD-R. Remember that PSTs for *different* custodians tend to have the *same* names (i.e., Outlook.pst and Archive.pst), so use a naming protocol or folder structure to keep track of who's who. When dealing with Outlook Express, search for messages stored in archives with a DBX extension.

Though dtSearch will index DBX files, PSTs must first be converted to individual messages using the included command line tool, mapitool.exe. For DOS veterans, it's old hat, but those new to command line syntax may find it confusing. To use mapitool, you'll need to know the paths to mapitool.exe and to the PSTs you're converting. Then, open a command line window (Start>Run>Command), and follow the instructions included with mapitool.

When mapitool completes the conversion, point the dtSearch Index Manager to the folder holding the extracted messages and index its contents. Name the index to correspond with the custodian and repeat the process for each custodian's PST files.

.

Function Follows Form

by Craig Ball

[Originally published in Law Technology News, June 2006]

The federal rules amendments governing discovery of electronically stored information have sailed through the U.S. Supreme Court and are now before Congress. Assuming passage, they'll be effective this December.

Though all bolts aren't tight and a few sections of track are missing, we're lining up to board the e-discovery roller coaster. It's going to be a wild ride.

As we countdown to the new rules, we should use the time to explore what powerful tools they'll be and acquire the skills to use them artfully while avoiding the sharp edges.

Rule 34(b): Have It Your Way

My favorite amendment—and let's not tarry over what sort of loon has a "favorite"—is FRCP 34(b), which empowers a requesting party to specify the form or forms in which electronically stored information (ESI) is to be produced.

Form didn't matter for paper production, but it makes all the difference in the ability to manage and search ESI.

The producing party must deliver ESI in the specified form or make an objection stating the reasons it won't and the form or forms it intends to provide. Alternate forms must be either those in which the ESI is ordinarily maintained or that are "reasonably usable." This is a giant leap forward for requesting parties, who get ESI their way or at least in a way that's electronically searchable.

The Committee Notes bear this out. "If the responding party ordinarily maintains the information it is producing in a way that makes it searchable by electronic means, the information should not be produced in a form that removes or significantly degrades this feature."

That means no more "naked" .tif or PDF files stripped of searchable data layers. No more blowbacks to paper. Even printouts of e-mail won't cut it… unless that's what the requesting party wants.

Having the power to specify the forms of production presupposes the ability to make an informed choice; plus changes to Rule 26 (f) (3) require parties to discuss forms of production in the pre-discovery meet-and-confer.

We're all going to have to know this stuff.

Five Forms

One of the biggest mistakes a requesting party makes is requesting or accepting production of electronic evidence in a format ill-suited to their needs. ESI production takes five principal forms:

1. Hard copies;
2. Paper-like images of data in, e.g., Adobe's Portable Document Format (PDF) or in one of the Tagged Image File Formats (.tif);
3. Quasi-native data exported to "reasonably usable" electronic formats like Access databases or load files;
4. Native data; and
5. Hosted data.

Your format specification hinges on both the nature of the data and your in-house capabilities for dealing with it.

In a perfect world, you'd want everything in native electronic format. But in the real world, you may lack the systems, software or expertise to access native data and preserve its evidentiary integrity. Plus, concerns about redaction, alteration and Bates numbering mean your opponents may be unwilling to produce native data.

Hard Copies

Converting searchable electronic data to costly and cumbersome paper is usually a step backwards, but paper still has its place.

In a case where the entire production consists of a few hundred e-mails and several thousand e-documents, searching and volume aren't a problem and paper remains as good a medium as any. But once the volume or complexity increases beyond that which you can easily manage by memory, you're better off insisting on production in electronically searchable forms.

Image Production

Here, production consists of files that are digital "pictures" of the documents, e-mails and other electronic records, typically in accessible file formats (PDF or .tif). As long as the information lends itself to a printed format and is electronically searchable, image formats work reasonably well; but for embedded information (such as the formulae in spreadsheets) or when the evidence moves beyond the confines of printable information (e.g., voicemail, databases or video), image production breaks down.

Requesting parties must ensure that electronically searchable data layers and relevant metadata accompany the images. Beware those who try to pawn off "naked" .tif images (devoid of searchable information and metadata) as responsive.

Exported Formats

Some electronic evidence adapts to multiple production formats, so sometimes you'll want exported, delimited data in order to work with it in the compatible application of your choice.

For example, e-mail may be readable in any of several programs or in generic e-mail formats (e.g., .eml, .msg). The contents of simple databases like contact lists can be exported to generic formats (e.g., comma or tab-delimited output) and imported into compatible applications, such as Microsoft Corp.'s Excel spreadsheets or Access databases.

The key is to be sure that important data or the ability to manipulate it isn't lost in the export/ import process.

Native Production

As data structures grow more complex, it's much harder to present exported data in an accurate or complete way.

In native production, the producing party furnishes duplicates of the actual data files containing responsive information and a requesting party with copies of the software programs used to create and manipulate the data (or compatible viewers) has the ability to see the evidence more-or-less exactly as it appears to the other side.

Sounds great, but native production is not without its problems. The native applications required to view the data in its native format may be prohibitively expensive or difficult to operate without extensive training (e.g., Oracle Corp. or SAP America Inc. databases).

Additionally, care must be taken not to change the native data while viewing it. Native production is best, but only when you have the experience, expertise and resources to manage native data.

Producing parties often fight native production because of difficulty in redacting privileged information. An Outlook post office (.pst) file can hold both discoverable e-mail and privileged attorney-client communications, but as it's a unified and complex database file, it's challenging to separate the two.

Another (largely overblown) risk to defendants is that native data (like Microsoft Office files) can contain embedded, revealing metadata. Where native files are concerned, metadata is evidence, too. See, e.g., *Williams v. Sprint/United Management Co*., 230 F.R.D. 640 (D. Kan. 2005).

Hosted Data

This is production without production in that the information produced resides on a controlled-access website. The requesting party reviews the data through an online application (similar to a web browser) capable of displaying information from a variety of electronic formats. More commonly, hosted data and online review tools are used by counsel for the producing party to search the production set for privileged and responsive items rather than as a means to afford access to the requesting party. The items identified are then burned to CD or DVD and produced, usually in image formats as discussed above.

Next Month: Specifying the right form of production for the most common ESI.

Rules of Thumb for Forms of ESI Production

by Craig Ball

[Originally published in Law Technology News, July 2006]

Come December 2006, amended Rule 34(b) of the Federal Rules of Civil Procedure has a gift for requesting parties both naughty and nice. It accords them the right to specify the form or forms of production for electronically stored information (ESI) sought in discovery. Though December may seem remote in these dog days of July, litigators better start making their lists and checking them twice to insure that, come December, they'll know what forms are best suited to the most common types of ESI.

Last month, I covered the five principal forms ESI can take:

1. Hard copies;
2. Paper-like images of data in, e.g., TIFF or PDF;
3. Data exported to "reasonably usable" electronic formats like Access databases or load files;
4. Native data; and
5. Hosted data.

This month, we'll look at considerations in selecting a form of production for the kinds of data most often seen in e-discovery.

Word Processed Documents

In small productions (e.g., less than 5,000 pages), paper and paper-like forms (.PDF and .TIFF) remain viable. However, because amended Rule 34(b) contemplates that producing parties not remove or significantly degrade the searchability of ESI, both parties must agree to use printouts and "naked" image files in lieu of electronically searchable forms. When the volume dictates the need for electronic searchability, image formats are inadequate unless they include a searchable data layer or load file; otherwise, hosted or native production (e.g., .DOC, .WPD, .RTF) are the best approaches. Pitfalls in native production include embedded macros and auto date features that alter the document when opened in its native application. Moreover, word processor files can change their appearance and pagination depending upon the fonts installed on, or the printer attached to, the computer used to view the file. Be careful referring to particular pages or paragraphs because the version you see may format differently from the original.

Consider whether system and file metadata are important to the issues in your case. If so, require that original metadata be preserved and a spreadsheet or other log of the original system metadata be produced along with the files.

E-Mail

Again, very small productions may be managed using paper or images if the parties agree on those forms, but as volume grows, only electronically searchable formats suffice. These can take the form of individual e-mails exported to a generic e-mail format (.EML or .MSG files), image files (i.e., .PDF or TIFF) coupled with a data layer or load file, hosted production or native production in one of the major e-mail storage formats (.PST for Outlook, .NSF for Lotus Notes, .DBX for Outlook Express). While native formats provide greatest flexibility and the potential to see far more information than hard copies or images, don't seek native production if you lack the tools and skill to access the native format without corrupting its contents or commingling evidence with other files.

All e-mail includes extensive metadata rarely seen by sender or recipient. This header data contains information about the routing and timing of the e-mail's transmission. Require preservation and production of e-mail metadata when it may impact issues in the case, particularly where there are questions concerning origin, fabrication or alteration of e-mail.

Spreadsheets

Even when spreadsheets fit on standard paper, printed spreadsheets aren't electronically searchable and lack the very thing that separates a spreadsheet from a table: the formulae beneath the cells. If the spreadsheet is just a convenient way to present tabular data, a print out or image may suffice, but if you need to examine the methodology behind calculations or test different theories by changing variables and assumptions, you'll need native file production. Hosted production that allows virtual operation may also suffice. When working with native spreadsheets, be mindful that embedded variables, such as the current date, may update automatically upon opening the file, changing the data you see from that previously seen by others. Also, metadata about use of the spreadsheet may change each time it is loaded into its native application. Once again, decide if metadata is important and require its preservation when appropriate.

PowerPoint Presentations:

You can produce a simple PowerPoint presentation as an electronically searchable image file in PDF or TIFF, but if the presentation is animated, it's a poor candidate for production as an image because animated objects may be invisible or displayed as incomprehensible layers. Instead, native or hosted production is appropriate. Like spreadsheets, native production necessitates preservation of original metadata, which may change by viewing the presentation.

Voice Mail

Often overlooked in e-discovery, voice mail messages and taped conversations (such as recorded broker-client transactions) may be vitally important evidence. As voice mail converges with e-mail in so-called integrated messaging systems, it's increasingly common to see voice mail messages in e-mail boxes. Seek production of voice mail in common sound formats such as .WAV or .MP3, and be certain to obtain voice mail metadata correlated with the audio because information about, e.g., the intended recipient of the voice message or time of its receipt, is typically not a part of the voice message.

Instant Messaging

Instant messaging or IM is similar to e-mail except that exchanges are in real-time and messages generally aren't stored unless the user activates logging or the network captures traffic. IM use in business is growing explosively despite corporate policies discouraging it. In certain regulated environments, notably securities brokerage, the law requires preservation of IM traffic. Still, requests for discovery of IM exchanges are commonly met with the response, "We don't have any;" but because individual users control whether or not to log IM exchanges, a responding party can make no global assertions about the existence of IM threads without examining each user's local machine. Although IM applications use proprietary formats and protocols, most IM traffic easily converts to plain text and can be produced as an ASCII- or word processor-compatible files.

Databases

Enterprises increasingly rely on databases to manage business processes. Responsive evidence may exist only as answers obtained by querying a database. Databases present enormous e-discovery challenges. Specify production of the underlying dataset and application and you'll likely face objections that the request for production is overbroad or intrudes into trade secrets or the privacy rights of third parties. Producing parties may refuse to furnish copies of database applications arguing that doing so violates user licenses. But getting your own license for applications like Oracle or SAP and assembling the hardware needed to run them can be prohibitive.

If you seek the dataset, specify in your request for production the appropriate back up procedure for the database application geared to capture all of the data libraries, templates and configuration files required to load and run the database. If you simply request the data without securing a back up of the entire database environment, you may find yourself missing an essential component. By demanding that data be backed up according to the publisher's recommended methodology, you'll have an easier time restoring that data, but be sure the backup medium you specify is available to the producing party (i.e., don't ask for back up to tape if they don't maintain a tape backup system).

An approach that sometimes works for simpler databases is to request export of records and fields for import to off-the-shelf applications like Microsoft Access or Excel. One common export format is the Comma Separated Variable or CSV file, also called a Comma Delimited File. In a CSV file, each record is a single line and a comma separates each field. Not all databases lend themselves to the use of exported records for analysis, and even those that do may oblige you to jump through hoops or engage an expert.

If you aren't confident the producing party's interrogation of the database, will disgorge responsive data, consider formulating your own queries using the application's query language and structure. For that, you'll need to understand the application or get expert help, e.g., from a former employee of the responding party or by deposing a knowledgeable employee of your opponent to learn the ins-and-outs of structuring a query.

Summer Reading

ESI. CSV. WAV. It's a new language for lawyers, but one in which we must be fluent if we're to comply with amended Rule 26(f)(3) and its requirement that parties discuss forms of production in the pre-discovery meet-and-confer. So, this summer, lay down that Grisham novel in favor of a work that has us all in suspense: *The Rules.*

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.

Santa@NorthPole.com

by Craig Ball

[Originally published in Law Technology News, December 2006]

Dear Santa,

I've been a good boy this year. I spent all my time helping lawyers and judges with electronic discovery and studying really, really hard about ESI, data harvest, spoliation, de-duplication, meet-and-confer, search tools, forms of production and computer forensics. I didn't use the word "solution" in a single column.

Please leave these presents under my tree:

1. I want a container file format for electronically stored information (ESI). We are gathering all this discoverable data but corrupting its metadata in the process. Plus, it's so hard to authenticate and track ESI. The container would safely hold the evidence as we harvest, search and produce it. It would include hash verification of all its parts, a place to store both an image of the document and its native content and even a special pocket to hold an overlay of all that helpful stuff we used to stamp onto paper documents, like Bates numbers and confidentiality warnings. And Santa--this is really important--it needs to be open sourced so no one has to pay to use it and extensible so we can keep using it for a very long time.

2. I want integrally write-protected external hard drives with removable electronic keys. Producing ESI on optical disks is nice because they're read-only media and you can't intentionally or inadvertently corrupt their contents. But nowadays, there's just too much ESI to hand over on optical disks. I want external hard drives designed for e-discovery such that a producing party can fill them with information then remove a USB key or snap off a tab to insure that nothing else can be written to or changed on the drive. If it hashed its contents and burned that hash value to an onboard write-once chip, that would be pretty cool, too.

3. May I have information technology training courses designed expressly for lawyers and litigation support, offering real depth and serious accountability for mastering the subject matter? Lawyers and their staff are waking up to the need to learn this stuff, but the traditional CLE and CPE paths don't offer or demand enough. We don't need another 10,000-foot "certification" course. We need Parris Island.

4. While we're at it big guy, how about making electronic discovery and digital evidence a discrete part of law school curriculum? I understand that teaching the *practice* of law is looked down upon at the best schools, but the assumption that young lawyers who grew up with computers automatically "get it" is misguided.

5. Could there also be licensure for computer forensic examiners geared to insuring genuine expertise and experience? Putting computer forensic examiners under the jurisdiction of the state boards that regulate private investigators and security guards

is like putting the football coach in charge of the Physics Department. Weeding out unqualified computer forensic examiners is a worthwhile goal, but can't legislatures put the task in the hands of those best qualified to judge?

6. Since we're regulating the forensics side of e-discovery, how about a code of ethics for electronic discovery vendors and experts, too? One that's not just lipstick on a pig! All parties need to be confident that information in a vendor's custody, including how that material is reviewed, is secure and that vendors are keeping their software current and adhering to other sound practices.

7. Santa, I'm still hoping to get what I asked for last year, like e-mail clients that compel immediate filing of messages within an information taxonomy and published standards and definitions for metadata fields of common file types. I do hope the elves are working on those, too.

Thanks. Fly carefully. Love, Craig

Thanksgiving

As this goes to press, the e-discovery amendments to the Federal Rules of Civil Procedure finally take effect--a milestone culminating six years of hard work by the Rules committee. We owe them a huge debt of gratitude even as we greet with trepidation the consequences of what they've wrought

Like Y2K, there will be no falling of the sky or trembling of the ground. But like Y2K, the post-FRCP amendments world will never be quite the same. The Y2K apocalypse never materialized, but the world quietly changed as dramatically as if it had. Enormous sums were plowed into computing infrastructure, and the clamor for programming talent threw wide the doors to India and the world, transforming the global economy in the many ways New York Times columnist Thomas Friedman insightfully describes in his bestseller, "The World is Flat."

The changes to the Federal rules are modest. The added language probably amounts to fewer words than this column. But those Amendments are already driving massive investment in infrastructure, training, services and personnel. It's just the tip of the iceberg. Litigants and their lawyers won't feel anything on December 1. Most will escape the maelstrom for another week, month, even a year or two. But it's coming, as inevitably as death and taxes, ready…or not.

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 over inclusive.

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.

Climb the Ladder

by Craig Ball

[Originally published in Law Technology News, February 2007]

Though computer forensics is a young discipline, it's not the exclusive province of new graduates of computer forensics degree programs. It's a natural career extension for IT and law enforcement professionals and peripatetic lawyers with a dominant geek gene. Expertise in litigation and computer forensics also opens the door to lucrative opportunities in electronic data discovery consulting. Here are "The Eight Es" to becoming a skilled CF expert:

1. Exploration...The lion's share of CF knowledge is self-taught. The best examiners are insatiably curious and voraciously read about software, hardware, registry keys, root kits, etc. They live for figuring out how it all fits together. Fortunately, there's a wealth of information: in books (search Amazon.com for "computer forensics") and online (www.e-evidence.info) in discussion forums, product FAQs, user groups and confabs.

2. Education...A computer science or law degree is nice, but you can study animal husbandry so long as you go on to study CF in a comprehensive way. Professional certifications that legitimately demonstrate training, testing and practical experience have value in helping courts, clients, and potential employers assess your qualifications. Supplement your college degree with as many courses and certifications as your time and budget allow.

Excellent programs are offered by universities, vendors, professional associations, and the government, such as ***New Technologies Inc.,*** (www.forensics-intl.com), ***Guidance Software*** (www.guidancesoftware.com), ***Access Data*** (www.accessdata.com), the ***International High Technology Crime Investigation Association*** (www.htcia.org), the ***International Association for Computer Information Systems*** (www.iacis.org), and the ***Federal Law Enforcement Training Center*** (www.fletc.gov). But don't fool yourself into thinking that a weeklong boot camp will qualify you as a CF expert. In a battle between an experienced examiner and one with an advanced degree, juries may defer to the latter. Some jurisdictions require licensure to perform forensic investigations.

3. Experimentation...The ability to construct illuminating experiments and the patience to elicit data are hallmarks of a skilled examiner. If you need to know how metadata changes when a user touches a file, you'll be prepared to testify if you've proven your theory by competent experimentation. Experiment with systems, applications and operating systems to understand how they work.

4. Experience...There's no substitute for applying your skills and testifying in real cases. How can you get that experience? Apprentice to a veteran examiner or offer to perform a "shadow exam," to see if you find something he or she missed. Assist attorneys or local law enforcement at little or no cost.

5. Exchange...Every examiner benefits from the exchange of ideas with colleagues. Join industry associations, go to meetings, subscribe to online discussion groups and unselfishly share what you learn. Caveat: the CF community is very supportive, but other examiners may justifiably regard you as a competitor, so don't expect them to reveal all. Show respect by doing your homework. Be a learner, not a leech.

6. Equipment...Learn the tools and techniques suited to the task, and invest in them. Use quality hardware and properly license software. Keep applications up-to-date, test tools to insure they're reliable. Cross-validate results. Too many people confuse buying tools with acquiring skills. A well-trained examiner can do the job with a hex editor and a viewer. We use forensic suites, such as Guidance Software's EnCase or Access Data's FTK, to automate routine tasks, improve efficiency, and lower costs—but buying a program doesn't make you a ready expert.

7. Earning...The demand for examiners is growing, but it takes marketing skill and financial acumen to create a thriving business. You must attract and serve quality clients, and make ends meet, to transform opportunity into achievement. Consider a first job with established CF companies or law enforcement, not only for a steady income, but also for the training. Starting salaries average $50,000 to $75,000, but in the private sector, quickly rise to six figures as you gain experience and responsibility. (Examiners with J.D.s or network security skills command higher salaries.)

Many CF firms charge clients $250 to $600 per hour, so it's not unrealistic for entrepreneurial examiners to hang out their shingles after learning the ropes. Expect $25,000 in minimum startup costs for hardware, software and training. Overhead will vary on whether you operate from your home or offsite.

8. Essential Element—Character...The final "E" is the "essential element"—*character*. A successful examiner is at once, teacher and student, experimenter, skeptic, confidante, translator, analogist, and raconteur. He or she unearths the human drama hidden in the machine. So many qualities distinguish the best examiners—integrity, tenacity, technical skill, imagination, insatiable curiosity, patience, discretion, attention to detail and the ability to see both the forest and the trees. Ultimately, it's your character that will determine if you'll be a top computer forensics expert.

Vista Changes the View

by Craig Ball

[Originally published in Law Technology News, March 2007]

Vista, Microsoft Corp.'s long awaited re-invention of its Windows operating system, finally premiered with little fanfare. No Rolling Stones theme music this time, though users frustrated with ineradicable security holes in Windows XP could have made the case for "19th Nervous Breakdown" or "(I Can't Get No) Satisfaction."

While most businesses are taking a wait-and-see attitude about migrating, sooner or later, they'll make the move. Within two years, Vista will have made significant inroads against XP on business desktops and laptops, and in the home, Vista will dominate. Many Windows users will also upgrade to Office 2007, the latest release of Microsoft's four horsemen, Word, Outlook, Excel and PowerPoint. What does this inexorable Vista and Office creep mean for electronic data discovery and the nerdy little corner of EDD called computer forensics? Only time will tell, but dramatic changes are in store.

Versions

Remember that deluxe Crayola box you longed for as a kid—the one with the sharpener and colors like "flesh" and "periwinkle?" Well, Vista has nearly as many versions as that box had crayons. There's Vista Home Basic, Home Premium, Business, Enterprise, Ultimate and Vista with Retsyn (okay, I made that last one up). Though all affect EDD to some extent, as you move higher up the evolutionary ladder of Vista versions, you'll bump into features like BitLocker volume encryption that really complicate EDD and forensics.

The big news in Vista is security, especially against prying eyes and careless keystrokes. Business, Enterprise and Ultimate editions include an automatic backup feature called Shadow Copy that invisibly saves your work to unused disk space to protect you from "Oh, No!" moments—like saving over an important file. Sounds great, except it salts away all prior versions, including those you *don't* intend to keep.

Unlike the hidden, fragmented forensic data the federal rules call examples of inaccessible ESI, the Vista shadow copy is a hardy survivor: complete, coherent and readily accessible. Vista grows the volume of discoverable ESI, perhaps significantly.

Little Brother

Not only does Vista do a better job hanging on to your work, it also keeps tabs on users as they work, through a feature called Transactional NTFS, or TxF. Bid goodbye to Last Access Times corrupted by peeking at the evidence or antivirus scans.

By default, Vista quits tracking access times as a file property. Instead, TxF logs file system activities, and a counterpart called TxR logs Registry activity. The bottom line is that a user's activity will be closely and constantly tracked, step-by-fateful-step. From the standpoint of investigating claims of evidence destruction, it's less a piecing together of fragments and more a "Let's look at that again in instant replay" situation.

This means a heck of a lot of new digital evidence out there to preserve and discover.

BitLocker

At the top of the Vista food chain, the Enterprise and Ultimate editions include a drive volume encryption feature called BitLocker—giving greater protection against data breaches from lost and stolen laptops.

BitLocker makes users access their data like $20 bills at the ATM. But to work, the protected machine must be equipped with a microchip called a Trusted Platform Module, or unlocked by a USB flash drive that serves as a key. Maybe Microsoft will hook up with the Stones again to market Jumpin' Jack Flash and Under My Thumb Drives.

The same encryption hardening a machine against identity theft and competitive intelligence can hopelessly frustrate forensic examination and emerging remote search and collection tools for EDD.

Absent a robust key escrow program, companies will have a harder time enforcing acceptable use policies or stealthily inspecting machines to identify candidates for litigation hold.

Rumors of a back door for law enforcement abound, but so far, Microsoft vehemently denies the existence of a way around BitLocker. Instead, the feature is reserved to only the most costly versions. Budget-minded criminals beware.

New Folders & Formats

Updates large and small will impact e-discovery. For example, Vista banishes spaces from standard folder names, so the Windows "My Documents" folder is now "Documents." Even so trivial a change can wreck havoc with automated or scripted collection protocols and trigger expensive do-overs if EDD systems and personnel aren't adaptable and vigilant.

There's some encouraging news, too. Vista and Office introduce fundamental changes to file formats that will, in time, dramatically impact EDD by lowering cost and complexity of both review and production. One of the persistent objections to native file production of ESI is the difficulty of redacting privileged content from native formats. Consequently, Microsoft's decision to store Office 2007 files in XML bodes a sea change for EDD, not only because it facilitates review but particularly because of the ease with which privileged content can be identified and redacted in XML. It opens the door to widespread use of native file formats in production and consigns the claim "native files can't be redacted" to "urban legend."

Search That Works?

A big EDD question mark hangs over Vista's enhanced search capabilities. Search in previous Windows versions was literally and figuratively a dog. But Vista's search tool rivals Google Desktop in speed, text search and metadata filtering.

It may yet prove an ally to litigation hold efforts if counsel circulates not only retention instructions describing targeted information, but also specific queries to be undertaken in Vista Search—a task potentially facilitated by Vista's ability to store programmatic searches in so-called "Search Folders."

The Big Picture

I've focused on a few of the many features of Windows Vista and Office 2007 likely to impact e-discovery, but these products are just the most visible components in a roll out of more than 30 Microsoft products that collectively promise to transform e-discovery and digital evidence in ways lawyers must anticipate and address. More wrenching changes will flow from Microsoft's embrace of collaboration and integrated messaging. The static, single-author printable document is evolving into something entirely new, organic and multiplayer. Collaborative construction means much more metadata playing a much more crucial role as the glue that holds "documents" together.

Integrated messaging shifts IM and voice to center stage and further undercuts paper and .tiff as viable forms of review and production. Both developments signify fresh challenges in every discovery phase.

For lawyers and litigants who feel like EDD has crept up and kicked them, you ain't seen nothing yet. Vista et al. paint a broad new horizon over rough seas.

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?

Do-It-Yourself Forensics

by Craig Ball

[Originally published in Law Technology News, June 2007]

All over America, vendors stand ready to solve the e-discovery problems of big, rich companies. But here's the rub: Most American businesses are small companies that use computers—and along with individual litigants, they're bound by the same preservation obligations as the Fortune 500, including occasionally needing to preserve forensically significant information on computer hard drives. But what if there's simply no money to hire an expert, or your client insists that its own IT people must do the job?

THE D-I-Y CHALLENGE

I challenged myself to come up with forensically sound imaging methods for conventional IDE and SATA hard drives—methods that would be inexpensive, use off-the-shelf and over-the-net tools, yet simple enough for nearly anyone who can safely open the case and remove the drive. In that vein, the safest way to forensically preserve evidence is to employ a qualified computer forensics expert to professionally "image" the drive and authenticate the duplicate. No one is better equipped to prevent problems or resolve them should they arise.

Further, when you open up a computer and start mucking about, plenty can go awry, so practice on a machine that isn't evidence until you feel comfortable with the process.

FORENSICALLY SOUND

When you empty deleted files from your computer's recycle bin, they aren't gone. The operating system simply ceases to track them, freeing the clusters the deleted data occupies for reallocation to new files. Eventually, these unallocated clusters may be reused and their contents overwritten, but until that happens, Microsoft Corp.'s Windows turns a blind eye to them and only recognizes active data. Because Windows only sees active data, it only copies active data. Forensically sound preservation safeguards the entire drive, including the unallocated clusters and the deleted data they hold.

Even lawyers steeped in electronic data discovery confuse active file imaging and forensically sound imaging. You shouldn't. If someone suggests an active data duplicate is forensically sound, set them straight and reserve "forensically sound" to describe only processes preserving all the information on the media.

PRIMUM NON NOCERE

Like medicine, forensic preservation is governed by the credo: "First, do no harm." Methods employed shouldn't alter the evidence by, e.g., changing the contents of files or metadata. But that's not always feasible, and the first method described departs from the forensic ideal.

METHOD 1: THE DRIVE SWAP COMPROMISE

Pulling the plug and locking a computer away is a forensically sound preservation method, but rarely practical. By the same token, imaging programs such as Symantec

Corp.'s Ghost (www.ghost.com) or Acronis Inc.'s True Image (www.acronis.com) leave unallocated clusters behind and may alter the source. Our first do-it-yourself approach strikes a balance between practical and perfect by recognizing that users obliged to preserve the contents of unallocated clusters have no use for those contents. They use only active data. So, the first method employs off-the-shelf cloning software to copy just active files from the original evidence drive to a duplicate of equal or greater capacity. The forensic twist is that you preserve the original drive and put the duplicate back into service.

Be sure that the drive you swap has the same size enclosure as the original (typically 2.5 inches for laptops and 3.5 inches for desktops) and that it connects to the computer in the same way, e.g., parallel ATA (a.k.a. "IDE") or Serial ATA. Pull the plug (for laptops, remove the battery too), then open the case to determine the type of drive interface before heading to the store. Buy the proper replacement internal drive in a gigabyte capacity at least as large as the original. Greater capacity is fine.

Accessing a laptop drive can be tricky, so check the manufacturer's website if you're uncertain how to remove and safely handle the drive. Another hurdle: laptops lack cabling to add a second internal drive, so you'll need an adapter to connect the target drive via USB port. A Vantec Thermal Technologies' (www.vantecusa.com) CB-ISATAU2 adapter cable runs about $25 at www.newegg.com, or find other adapters and suppliers by web searching "sata/ide usb adapter."

Follow the software's instructions, but never install the duplication software to the drive you're preserving because that overwrites unallocated clusters. Instead, run the application from a CD, floppy or thumb drive. It's critically important that you don't inadvertently copy the contents of the blank drive onto the original, so check settings, and then check them again before proceeding.

When the imaging completes, label the original drive with the date imaged, name of the user, machine make, model and serial number, and note any inaccuracy in the BIOS clock or calendar. Secure the original drive in an anti-static bag and install the duplicate drive in the machine. Confirm that it boots. The user should see no difference except that the drive offers more storage capacity.

Done right, this method hews close to a forensically sound image, the qualifier being that the cloning software and the operating system may make some (typically inconsequential) alterations to the source drive. The method combines the advantages of Ghosting (speed and ease-of-use) with the desirable end of preserving the original digital evidence with [most] metadata and unallocated clusters intact. Best of all, it employs tools and procedures likely to be familiar to the service techs at your local electronics superstore. Be sure they adhere to the cautions above.

Next month, I'll describe a do-it-yourself approach to *true* forensically sound imaging.

Do-It-Yourself Forensic Preservation (Part II)

by Craig Ball

[Originally published in Law Technology News, July 2007]

How does a non-expert make a forensically sound copy of a hard drive using inexpensive, readily available tools? That's the D-I-Y challenge. Last month, we discussed a nearly perfect way to forensically preserve hard drives that entails swapping the original drive for a Ghosted copy containing just active files.

But when it comes to crucial evidence, nearly perfect doesn't cut it. Last month's method made minor changes to the source evidence, didn't grab unallocated clusters (necessitating we sequester the original drive) and offered no means to validate the outcome.

Because a forensically sound preservation protects all data and metadata along with deleted information in unallocated clusters, think of the Three Commandments of forensically sound preservation as:

1. Don't alter the evidence;
2. Accurately and thoroughly replicate the contents; and
3. Prove the preceding objectives were met.

This month's method employs write blocking to intercept changes, software that preserves every byte and cryptographic hash authentication to validate accuracy.

Write Blocking

Computer forensics experts use devices called "write blockers" to thwart inadvertent alteration of digital evidence, but write blockers aren't sold in stores (only online) and cost from $150-$1,300. Hardware write blocking is best if timetable and budget allow. Manufacturers include Tableau, LLC (www.tableau.com), WiebeTech, LLC (www.wiebetech.com), Intelligent Computer Solutions, Inc. (www.ics-iq.com) and MyKey Technology, Inc. (www.mykeytech.com).

If you're running Windows XP or Vista, you may not need a device to write protect a drive. To hinder data theft, Windows XP Service Pack 2 added support for software write blocking of USB storage devices. A minor tweak to the system registry disables the computer's ability to write to certain devices via USB ports. To make (and reverse) the registry entry, you can download switch files and view instructions explaining how to manually edit the registry at http://www.lawtechnews.com/r5/showkiosk.asp?listing_id=1560974 (the contents of this web link follow on page 69).

You'll also need:

- **Imaging Machine**--a computer running Windows XP with Service Pack 2 and equipped with both USB 2.0 and IEEE 1394 (aka Firewire or i.Link) ports.

• **Forensic Imaging Application**--though forensic software companies charge a pretty penny for their analysis tools, several make full-featured imaging tools freely available. Two fine Windows-compatible tools are Technology Pathway's Pro-Discover Basic Edition (in the Resource Center at http://www.techpathways.com) and AccessData's FTK Imager (http://www.accessdata.com/support/downloads/). I prefer FTK Imager for its simplicity and ability to create images in multiple formats, including the standard Encase E01 format.

• **Target Drive**--a new, shrink-wrapped external hard drive to hold the image. It should be larger in capacity than the drive being imaged and, if using software write blocking, choose a drive that connects by IEEE 1394 Firewire(as USB ports will be write blocked).

• [Software write blocking only] A **USB bridge adapter cable or external USB 2.0 drive enclosure** matching the evidence drive's interface (i.e., Serial ATA or Parallel ATA). Though you'll find drive enclosures at your local computer store, I favor cabling like the Vantec Thermal Technologies' (www.vantecusa.com) CB-ISATAU2 adapter cable because they connect to 2.5", 3.5" and 5.25" IDE and SATA drives and facilitate imaging without removing the drive.

Imaging the Drive

Here is a step-by-step guide:

1. It's important to carefully document the acquisition process. Inspect the evidence machine and note its location, user(s), condition, manufacturer, model and serial number or service tag. Photograph the chassis, ports and peripherals.

2. Disconnect all power to the evidence machine, open its case and locate the hard drive(s). If more than one drive is present, you'll need to image them all. Accessing a laptop drive can be tricky, so check the manufacturer's website if you're uncertain how to safely remove and handle the drive. Take a picture of the drive(s) and cabling. If you can't read the labeling on the face of the drive or comfortably access its cabling, uninstall the drive by disconnecting its data and power cables and removing mounting screws on both sides of the drive or (particularly in Dell machines) by depressing a lever to release the drive carriage.

Handle the drive carefully. Don't squeeze or drop it, and avoid touching the circuit board or connector pins. If using a hardware write blocker, connect it to the evidence drive immediately and leave it in place until imaging is complete and authenticated.

3. Download and install FTK Imager on the imaging machine. If using software write blocking, initiate the registry tweak, reboot and, using a thumb drive or other USB storage device, test to be sure it's working properly.

4. Connect the evidence drive to the imaging machine through the hardware write block device or, if using software write protection, through either the USB drive enclosure or

via bridge cable connected to a software write blocked USB port. ***Above all, be sure the evidence drive connects only through a write blocked device or port***.

5. If USB ports are software write blocked, connect the target drive via the IEEE 1394 port. Optionally, connect via USB port if using hardware write blocking.

6. Run FTK Imager, and in accordance with the instructions in the program's help file for creating forensic images, select the write protected evidence drive as the source physical drive, then specify the destination (target) drive, folder and filename for the image. I suggest incorporating the machine identifier or drive serial number in the filename, choosing "E01" as the image type, accepting the default 650MB image fragment size and opting to compress the image and verify results.

Hash Authentication

Creating a forensically sound compressed image of a sizable hard drive can take hours. FTK Imager will display its progress and estimate time to completion. When complete, the program will display and store a report including two calculated "digital fingerprints" (called MD5 and SHA1 hash values) which uniquely identify the acquired data. These hash values enable you to prove that the evidence and duplicate data are identical. Hash values also establish whether the data was altered after acquisition.

7. When the imaging process is done, label the target drive with the date, the names of the system user(s) and machine identifier. Include the model and serial number of the imaged drive.

8. With the evidence drive disconnected, reconnect power to the evidence machine and boot into the machine's setup screen to note any discrepancy in the BIOS clock or calendar settings. Disconnect power again and re-install the evidence drive, being careful to properly reconnect the drive's power and data cables.

Whether you return the evidence machine to service or lock it up depends on the facts of the case and duties under the law. But once you've secured a forensically sound, authenticated image (along with your notes and photos), you've got a "perfect" duplicate of everything that existed on the machine at the time it was imaged and, going forward, the means to prove that the data preserved is complete and unaltered.

The safest way to forensically preserve digital evidence is to engage a qualified computer forensics expert because no one is better equipped to prevent problems or resolve them should they arise. But when there's no budget for an expert, there's still an affordable way to meet a duty to forensically preserve electronic evidence: ***do-it-yourself***.

Enabling and Disabling USB Write Protection in Microsoft Windows XP P2 and Vista

(This is the target page for the link in the preceding BIYC July 2007 column)

Windows XP machines updated with Service Pack 2 (SP2) acquired the option to enable write protection for removable storage devices connected to the machine via USB. You can still read from the devices, but you can't write to them. In my testing, it works as promised, preventing changes to the data and metadata of external USB hard drives and thumb drives. Though the Windows cache may make it seem that data has been written to the protected device, subsequent examination demonstrated that no changes were actually made. And you can't beat the price: it's free.

Still, software write protection has its ardent detractors (See, e.g., *The Fallacy of Software Write Protection in Computer Forensics*, Menz & Bress 2004), and because there's no outward manifestation that software write blocking is turned on and working, there's none of the reassurance derived from seeing a hardware write blocker play burly bodyguard to an evidence drive. Other downsides are that software write protection requires a geeky registry hack and lacks the selectivity of hardware write blocking. That is, when you implement software write blocking, it locks down all USB ports, including the one you'd hoped to use to connect an external USB hard target drive. Write blocked for one is write blocked for all.

Caveat: Software write protection of the USB ports only works in Windows XP with Service Pack 2 and Windows Vista. It can be implemented only by users with Administrator level privileges on the machine. Failing to disable write blocking may cause the loss of data you seek to store on external USB storage devices.

The Easy Way

To simplify software write protection, you can download a file from http://www.craigball.com/USB-WProtect.zip containing two .REG files that, when run (i.e., double clicked), serve as switches to enable and disable software write protection of the USB ports.

The Geeky Way

If you'd rather make the registry changes manually, here's how:

Caveat: It's prudent to create a system restore point before editing the registry. To do so, click Start > All Programs > Accessories > System Tools > System Restore. Select "Create a restore point," then click "Next." Type a brief description for your restore point (e.g., "Before adding write protection"), then click "Create."

Enabling Write Protection

To block the computer’s ability to write to a removable storage device connected to a USB port, begin by calling up a Windows command dialogue box:

Press the Windows key + R to bring up the Run dialogue box (or click Start > Run).

Type regedit and click “OK” to activate the Windows Registry Editor.

Click the plus sign alongside HKEY_LOCAL_MACHINE, then drill down to SYSTEM\CurrentControlSet\Control. [Fig 1.]

Examine the tree under Control to determine if there is a folder called “StorageDevicePolicies.” If not, you need to create it by right clicking on Control and selecting New > Key. [Fig. 2]

Name the key “StorageDevicePolicies,” (All one word. Match capitalization. Omit quotation marks) then right click on the key you’ve just created and select New > DWORD value [Fig. 3]

Name the new DWORD “WriteProtect” and hit Enter.

Right click on the new DWORD value and select “Modify.” Set the WriteProtect DWORD value to 1. [Fig. 4]

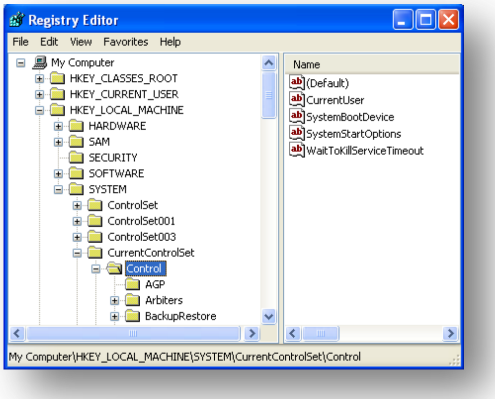
Exit the Registry Editor and reboot the machine. The USB ports should now be write protected.

Disabling Write Protection

To restore the system’s ability to write to USB media, navigate to the WriteProtect key as above and either delete it or change its value to 0.

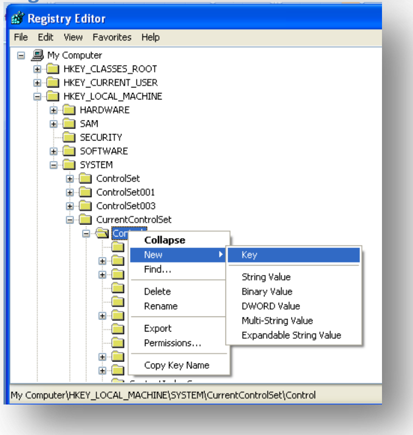
Reminder: **WriteProtect = 1 [ON]**
WriteProtect = 0 [OFF]

Figure 2



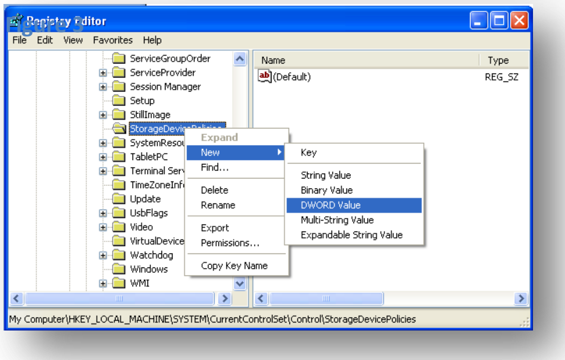
My Computer\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control

Figure 1



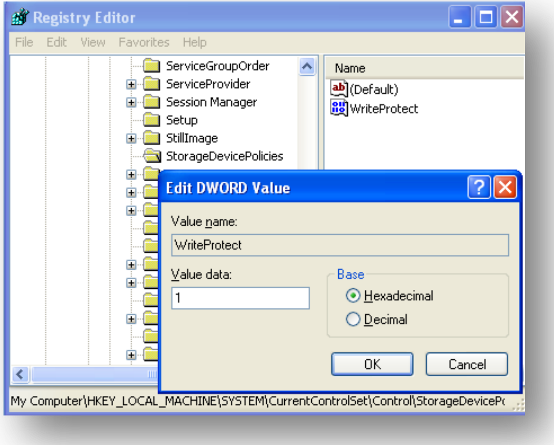
My Computer\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control

Figure 3



My Computer\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\StorageDevicePolicies

Figure 4



Value name: WriteProtect

Value data: 1

Base: ● Hexadecimal ○ Decimal

My Computer\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\StorageDevicePo

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 <u>not</u> 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), 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).

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.

Tumble to Acrobat as an E-Discovery Tool

by Craig Ball

[Originally published in Law Technology News, June 2008]

When the time comes to turn over e-data uncovered by forensic examination, it's hardly surprising that e-mail makes up a big chunk of the evidence. Notwithstanding its prevalence, e-mail is among the more challenging evidence types to share with clients in ways they can readily review messages and attachments without corrupting the metadata.

I've tried nearly everything, including converting messages to web formats and furnishing a browser-based viewer. That proved easy to run and navigate, but offered no search tools. Imaged formats (e.g., .tiff and .jpg files) also weren't searchable without load files and demanded that my clients have an EDD review platform on hand.

Some lawyers don't have the budget for .tiff conversion and load file generation, let alone a recent copy of Concordance or CT Summation. I've furnished native formats (e.g., .pst or .nsf), quasi-native formats (.eml, .msg) and even Access or NTSearch databases, but there are many pitfalls to trying to review e-mail using desktop applications. And if you need to engage in even the tiniest bit of techno-tinkering it turns lions of the courtroom to jelly. Nothing was quite easy enough.

So, the challenge was to convert e-mail into something I could give to a client with confidence that they could:

1. *Easily open the e-mail evidence on any machine without buying software.*
2. *Search messages quickly and powerfully, with full-text indexing and Boolean support.*
3. *View the messages in a way that faithfully preserves their appearance.*
4. *Print e-mail in a consistent way no matter what printer they used.*
5. *Enjoy document security, authentication and reliable redaction, too.*

While I'm at the wishing well, it would be nice if I could accomplish all this with software I already owned and something that could effortlessly handle the volume of e-mail I come across in computer forensic examinations.

Wouldn't you like to know what wondrous tool fills the bill? *So would I*, because I've yet to find it!

But, the happy news is I got *darn close* to the ideal using the latest version of Adobe System, Inc.'s Acrobat.

Yes, Adobe Acrobat 8.0, that utilitarian tool used to prepare documents for e-filing and keep secrets from sneaking off as Word metadata. Who knew that when this dowdy librarian of a program lets her hair down the results are easy, agile and gorgeous?

Despite a few drawbacks, Acrobat 8.0 turned out to be a nifty way to deliver moderate volumes of e-mail to technophobes and provide a way to search message text with instantaneous results.

Cons:

1. It's slow, taking hours to convert and index about 15,000 messages, even on a fast machine.
2. It refuses to even attempt conversion if you point it at more than 10,000 e-mails. So, for big collections, you must convert the data in chunks and stitch up the results as best you can.
3. Though it retains attachments in native formats with transport messages, the sole attachment type it can search is PDF. Microsoft Corp.'s Outlook, too, has long suffered from an inability to search within attachments. That's a serious shortcoming in both applications, but it's a shortcoming slated to improve in Acrobat's next release.
4. You can redact PDF documents beautifully within Acrobat 8, but not other formats; so, be wary of the potential for privileged data slipping out via an attachment.

Pros:

1. Anyone can review the resulting collection or "PDF Package" on any operating system using the ubiquitous, free Adobe Reader.
2. The search is fast and allows for fine tuning by, inter alia, Boolean operators, stemming, whole word search and case sensitivity.
3. Browsing messages is speedy, and image quality is excellent (screen or printed).
4. It supports annotation and book marking, so it's not a bad review platform for the price.
5. The Acrobat interface is instantly familiar and unintimidating.

To give credit where it's due, I was pointed in the right direction by Rick Borstein, an Adobe business development manager. He's the perfect public face for Adobe because he loves the product and enjoys teasing out its hidden joys without overselling its virtues. He has a fine blog called Acrobat for Legal Professionals (http://blogs.adobe.com/acrolaw/).

Unfortunately, you can't simply point Acrobat to an e-mail container and convert it. Acrobat must run as a PDF Creator toolbar within Outlook or Lotus Notes. The e-mail container must be in either .pst or .nsf format and must be accessible via Outlook or Notes. You can set up a dummy user account for conversion to prevent mixing your mail with evidence mail—a big no-no. The hurdle the first time I used Acrobat for e-mail production was that the evidence e-mail was in Eudora, so I had to apply another tool, **Aid4Mail** from Fookes Software, to convert the Eudora mail to an Outlook-compatible .pst format. This was easy, and the nifty Aid4Mail program costs less than $25, so it paid for itself on first usage.

Adobe Acrobat holds enormous promise as an EDD processing and review platform in smaller cases. It’s not all it can or will be, but each new version brings us closer to the goal of effective, affordable electronic discovery for everyone.

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.

Car 54, Where Are You?

by Craig Ball

[Originally published in Law Technology News, October 2008]

A reporter recently interviewed me about in-car GPS navigation systems as evidence. Aside from vehicle tracking devices planted on suspect vehicles, neither of us could point to more than a few matters where GPS evidence played a role in court; yet, its untapped value to criminal and civil cases is enormous. Think how many murders, rapes, burglaries, robberies, thefts, kidnappings and drug deals could be solved—and innocent persons exonerated--by reliably placing suspects in space and time. DNA just puts the accused at the scene. Reliable GPS data puts the suspect there between 9:42 and 10:17 p.m. and reveals where he came from and where he went next.

GPS-enabled “personal travel assistants” store both waypoints and typed destinations, distinguishing a suspect who claims coincidental presence from one who entered the address of the crime scene as his objective. Some units offer hands-free phone interfaces, recording frequently called numbers and holding unique identifiers for each linked telephone, enabling prosecutors to more persuasively tie the navigation system to a particular user.

With many navigation units costing under $150.00, these marvelous devices are appearing on more dashboards and will be appearing in more courtrooms; but the reporter was thinking too small. There will soon be a location-enabled device in near-constant use by almost every man, woman, tween and teen in the U.S. And it won’t be on dashboards. It’s as close as your cell phone.

Late in 2005, the FCC mandated that cell phones must be capable of geolocation to support 911 emergency calls as part of the first phase of its Enhanced 911 (E911) initiative. By September 11, 2012, phase II of the FCC's E911 rules kick in, requiring cell service providers to deliver extremely precise location-enabled data. By law, your phone is going to know exactly where it is and will be transmitting that information to the network. Most already do.

For example, the latest Apple iPhone joins a long list of phones supporting advanced GPS location technology that delivers amazingly precise geolocation without line-of-sight connection to satellites. Push a button, and your phone locates you on the map, even indoors. Add some software, and the phone shows you the locations of your spouse, kids or employees, within a few meters and updating in real time.

It’s wonderful…and scary…and evidence.

Expectations of personal privacy will likely yield to consumer demand for location-enabled services and devices. Marketers want to put themselves on the map--literally--and will pay handsomely for the privilege. Users will contentedly sacrifice geographic anonymity for the location of the nearest Starbucks. It's a profit center for cell carriers, so resistance is futile.

As for those who think that criminals will respond by simply turning off their phones or leaving them behind, don’t kid yourselves. The criminal genius is a creature of comic books and movies. There are thankfully few Lex Luthors out there, and even those who disable location awareness leave a trail marked by the absence of their customary data stream.

Geopositioning will aid civil cases, too. The day is not long off when we will be able to place price fixers or cheating spouses in the same room, or calculate the speed, path and braking action of colliding drivers. We’ll gauge exposure to environmental toxins, challenge a witness’ ability to observe, calculate wage and hour abuses, prove an employee was asleep at the switch and precisely determine a claimant’s proximity to an explosion.

The practical challenge posed by dashboard navigators, cell phones and the host of coming location-aware devices is that the systems are proprietary and fast-changing. Even devices that use identical chipsets to collect geospatial data often process and store that data differently. Forensic examiners must grapple with obscure interfaces, oddball cabling and few proven software tools. In sum, the data is there, and it promises to be compelling evidence; but, getting to it in time and with the right tools and expertise militates against our seeing it in court on a routine basis for some time. It may prove to be like the black box data in airbag deployment modules--sometimes the most revealing and reliable data is right under our noses but we just fail to grab it.

From the standpoint of e-discovery, whether geopositioning data must be preserved and produced depends upon the issues in the case and the accessibility of the data, considering the burden and cost of its acquisition. Today, it’s exceptional and ignored, but so was e-mail a few years ago. As you run through your list of potential sources of ESI, keep geopositioning devices in mind, and ask, “How might I use locator data to make my case or clear my client?”

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.

Right from the Start

by Craig Ball

[Originally published in Law Technology News, December 2008]

Certainly it's smart to prepare for electronic data discovery—to be "proactive" about electronically stored information, 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 EDD efforts. Before the ESI experts arrive, there are things you can and must do. Here's a quick compendium:

1. *Apply the five Ws of journalism (who, what, when, where, and why) to get a handle on your core preservation duties*. Immediately list the people, events, time intervals, business units, records, and communications central to the case.

a. List apparent key players (don't forget assistants who, for example, handle the boss' e-mail 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 that is diligently implemented will be 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? Voicemail, 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

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 the memo," they say, "but I didn't know where to start."

Tell custodians 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 helpful—it's what courts expect.

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 servers or on databases pose special challenges that are made more difficult 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*. Making sense of ESI, 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.

Crystal Ball

by Craig Ball

[Originally published in Law Technology News, January 2009]

It's January, and the pundits are atwitter with predictions. Every prognostication I've seen is colored by the nation's recent financial reverses. Certainly in recessionary times, businesses and firms contract, and air will escape from the electronic data discovery bubble, too. Painful as that is, I predict it's for the best.

Prosperity begets bad habits. Cheap capital promotes throwing more money at problems than sense, and that's nowhere more evident than in EDD. Austere budgets and skeletal workforces are no fun, but they make efficiency as important as oxygen.

Lacking budgets for experts, lawyers will be forced to grapple with nuts-and-bolts issues and learn about the technology. Clients unable to shell out fortunes to pay brigades of reviewers will warm to clawbacks, quick peeks for data unlikely to hold privileged material, off-shoring, in-housing, cooperation and whatever shortcuts they decide carry more benefits than risks.

Some shortcuts will be train wrecks (case in point: the company that tried to use Google Desktop as its review platform for upwards of 50 machines), but creativity spurred by tight money will breed simpler, lower-cost methods.

Have no fear: the *Fortune 500* will come to grips with EDD. They have the impetus, the means and as lately noted, they're "too big to fail." Instead, I predict the focus will shift to EDD for small and medium-sized business. In 2009, we probably won't see the emergence of the QuickBooks-like desktop tool needed to serve smaller clients, but the impetus to build it will grow.

The federal EDD rules are two years old. As every parent knows, the "terrible twos" are when toddlers, frustrated by their lack of language skills, turn to hitting, biting and tantrums.

Pediatrician Vincent Iannelli, writing for About.com, advises parents coping with the terrible twos to:

- Establish and stick to regular routines;
- Limit choices to exclude unacceptable alternatives, but share the decision-making process;
- Make the environment safer to explore by removing what's not needed;
- Discipline by taking away privileges; and
- Expect to see the limits tested by those exploring what they can get away with.

Perhaps I see e-discovery in everything, but isn't that a list of EDD best practices? If the analogy holds, 2009 will be the year we see improved language skills. Lawyers acclimating to EDD are starting to frame better questions and employ more precise terminology.

But 2009 will bring more tantrums, like the embarrassing poll where members of the American College of Trial Lawyers raged against EDD and the courts. Only about half of the 1,382 respondents had received or served a request for ESI under the rules, and some 40% had no EDD experience whatsoever. Tellingly, the respondents had practiced law 38 years on average, and because the survey was conducted via e-mail, the responses skewed toward the youngest members.

Ken Withers, the sage of Sedona, captured the essence of the survey in a recent speech. He conjured up old men shaking their fists, shouting, "Hey you kids — get off of my litigation*. And take your Googles with you.*" Attitudes matter, but it's a shame to clothe ignorance with authority.

In that vein, I predict 2009 will see the fitful demise of the Bobby Ewing EDD movement. Anyone old enough to remember the TV show *Dallas*—certainly that includes the ACTL—will recall how the show made Bobby Ewing's death just Pam's "bad dream," and the events of the prior season went down the shower drain.

No amount of wailing and gnashing of teeth will make EDD just a bad dream. In fact, 2009 will be the year that EDD really starts to work its nose under the tent of state practice. More states are enacting versions of EDD rules, and state judges are giving greater deference to the federal model.

In sum, 2009 will be more a time of assimilation than innovation. Look for further consolidation among vendors and firms as those crippled by cash flow problems combine or are gobbled up by competitors. The projected spike in litigation won't materialize because the rush to sue for financial malfeasance will be tempered by financial realities. But for many directors, corporate officers, banks and insurers, 2009 will be the year of the ugly e-mail and the occasional perp walk.

One encouraging sign for 2009 is Georgetown University Law Center's five-day "E-Discovery Training Academy" that debuts in February. I'm proud to serve on its volunteer faculty. It remains to be seen whether we teach the students what they need or simply teach them what we know, but it's an ambitious leap in the right direction.

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.**

The Multipass Erasure Myth

by Craig Ball

[Originally published in Law Technology News, March 2009]

Ambling along the back roads of listservs and blogs, I often come upon a flea-bitten claim that, "Top notch computer forensic examiners have special tools and techniques enabling them to recover overwritten data from a wiped hard drive so long as the drive was wiped less than 3 or 7 or 35 times."

Nonsense!

I think I know where this persistent fairy tale started. In 1996, a smart New Zealander, Peter Gutmann, published a paper, "Secure Deletion of Data from Magnetic and Solid-State Memory" [http://tinyurl.com/Gutmann].

Gutmann explained how cool toys such as magnetic force scanning tunneling microscopes and ferromagnetic fluids could serve as a Ouija to dear departed data. "Even for a relatively inexperienced user, the time to start getting images of the data on a drive platter is about five minutes."

The good doctor went on to prescribe a regimen of 35 varied overwriting passes to thoroughly erase data — a so-called Gutmann Method erasure.

It's all a lot of hogwash, at least with respect to any drive made this century.

To his credit, Gutmann awoke to his folly of '96. In an epilogue added years later, he marveled that so many came to regard his erasure scheme as "a kind of voodoo incantation to banish evil spirits" from hard drives, conceding that "performing the full 35-pass overwrite is pointless for any drive."

Yet, like a horror film zombie, the Gutmann Wipe lives on as a feature of modern drive erasure tools. Because it takes days to Gutmann erase them, big drives that should be wiped aren't and find their way onto eBay.

So what's the truth about multipass erasure?

In the years since Gutmann's article, the amount of data that can be packed onto a hard drive (its "areal density") has increased 10,000 fold.

So, hoary notions of data remanence like "offtrack persistence" and "additive and subtractive voltage thresholds" hold no hope of resurrecting overwritten data.

One point that hits home when we make the leap from the simplistic way we imagine hard drives work to the way they really track and encode information is this: All the anecdotal wiped data recovery stuff we've heard about is completely bogus. So stop folks when they say, "I know a guy who has a cousin who recovered overwritten data using EnCase by tweaking the frazzle setting and putting the drive in the freezer." It just ain't so.

You only need one complete pass to eviscerate the data (unless your work requires slavish compliance with obsolete parts of Department of Defense Directive 5220.22-M and you make two more passes for good measure).

No tool and no technique extant today can recover overwritten data on 21st century hard drives. Nada. Zip. Zilch.

My friend and fellow computer forensic examiner Dave Kleiman and his colleagues Craig Wright and Shyaam Sundhar actually used a magnetic force electron microscope to painstakingly analyze one-pass overwritten data.

"In many instances, using a MFM to determine the prior value written to the hard drive was less successful than a simple coin toss," they concluded. "The fallacy that data can be forensically recovered using an electron microscope or related means needs to be put to rest."

But even as the Jason Voorhees of multipass erasure settles to the bottom of Crystal Lake, two data security risks still bear mention.

The most egregious is the assumption that formatting a hard drive is the same as wiping its contents. In fact, formatting obliterates almost none of a drive's contents. Any eBay purchaser of a formatted drive can easily restore its contents.

Second, and principally of interest to three-letter agency types and paranoiacs, user data resides in areas of a hard drive that no wiping tool can reach: the so-called G List sectors.

Over time, one or many 512 byte sectors of a hard drive lose their ability to hold information reliably.

When detected, the drive copies the contents of the ailing sectors to spare sectors each drive manufacturer provides. This remapping of contents occurs automatically and without notice to the user. The drive inventories remapped sectors in the G (for Growth) List. The G List is stored in the drive's system area, stashed in the negative tracks no user can touch.

When you wipe a drive that contains remapped sectors, only one of the two data iterations will be overwritten. The other, camouflaged by G List redirection, remains untouched. But lose no sleep. The chances that remapped data will be important and intelligible, or that anyone will ever see them, are exceedingly small.

Remarkably, nearly all hard drives manufactured after 2001 incorporate the ability to rapidly and securely self-erase everything, including the G List; but drive and computer manufacturers are so petrified you'll mess that up, they don't offer an easy way to initiate a self-destruct sequence.

For those at ease with command line interfaces, the Secure Erase commands can be run using free tools developed for the NSA and available at http://tinyurl.com/serase. But be careful with these as there's no road back.

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 delecti* 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?

5. Does inappropriate e-mail or internet use figure into the suspected misconduct?

6. Did a departing employee bring a personal laptop, external hard drive, or thumb drive to work?

7. 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?

8. Do server logs or indicators reflect atypical access to data areas?

9. Has the user been notably secretive using company computers or been observed using other users' machines without permission?

10. Has the user recently requested that IT reinstall the operating system on his or her machine?

11. 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).

Surefire Steps to Splendid Search-Part 1

by Craig Ball

[Originally published in Law Technology News, June 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 insight 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 carefully craft appropriate keywords or implement quality control testing for searching ESI. 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. This month and next, I'll lay out ten steps I promise 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 RFP

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 background 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 provider analyzes the complement of file types and flags any that can't be searched. Unless you make it clear that certain files 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.

Next month, I'll share the remaining steps to splendid search.

Surefire Steps to Splendid Search-Part 2

by Craig Ball

[Originally published in Law Technology News, July 2009]

Last month, I showed how judges are waking up to the senseless, slipshod way lawyers approach keyword search and setting new standards for quality. I also laid out the first four of ten steps guaranteed to 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**

This month, we cover the next six steps.

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, ESI fairly crawls with misspellings that complicate keyword search. If you don't search for common spelling variants and errors, you'll overlook responsive items. "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 by substitution, insertion, deletion and transposition, often producing an unacceptably high level of false hits.

While every word processor application flags misspelled terms, how do you convert keywords to their most common misspellings and variants? A linguist could help, or you might probe for propensities by having key custodians type keywords as they are read aloud. More likely, you'll turn to the web. Until an online tool emerges that lists common variants and predicts the likelihood of false hits, you might visit a site like www.dumbtionary.com that checks a keyword against more than 10,000 common misspellings or consult the Wikipedia list of over 4,000 common misspellings.

If you've ever played the board game Taboo, you know there are many ways to communicate an idea without using obvious word choices. Searches for "car" or" automobile" will miss documents about someone's "wheels" or "ride." You've got to consult the thesaurus, but don't go hog wild with Dr. Roget's list. Identify and include likely alternatives for critical keywords. Also, question key players about alternate terms, abbreviations or slang used internally to reference the same topics as the search terms.

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 by de-NISTing for known hash values and culling by binary signatures, file extensions, paths, dates or sizes.

The exponential growth in the volume of information seen in e-discovery 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. Be sure any method of deduplication employed tracks the origins of suppressed iterations so that repopulation can be accomplished on a per custodian basis, if needed.

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 something 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 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 values.

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, a company may want to test search terms against data samples collected from employees or business units having nothing to do with the events of concern to determine whether search terms are too generic to be of value.

8. Review the Hits

My practice when testing keywords is to generate spreadsheets letting me preview search hits in context; that is, flanked by perhaps 20-30 words on each side of the hit. It's very efficient and illuminating to scan a column of hits for searches gone awry while selecting 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 unable to be searched 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, you're looking 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 by spaces).

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. This 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, and 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. Be sure to 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 tweak them.

Effective searches are tailored to the data under scrutiny. For example, it's silly to run a custodian's name or address against their 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, 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-mail address than by name. In a company with dozens of Raj Patel's, 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.

I guarantee these steps will improve your keyword searches but…

If you tell a court, "Craig Ball said to do it this way," you might hear, "Who?" or "Who cares?" Yet, until we know exactly what courts regard as sufficient and whose imprimatur matters, these techniques will help wring more quality and trim some of the fat from text retrieval. Don't forget, the least costly approaches to e-discovery are those done right from the start.

All Wet

by Craig Ball

[Originally published in Law Technology News, August 2009]

I was once trial counsel for the water authority of a Mexican city seeking damages for delay in the mapping of a water system serving three million customers. I learned that most water entering the pipes never reached consumers because the patchwork system was riddled with leaks. The leaks were difficult to repair because the water authority didn't know where its pipes were buried! Repair crews made Swiss cheese of streets, but the massive leakage limited water service to just a few hours a day. Those who could afford it erected tanks to hoard water. The rest suffered.

Until *Servicios de Agua y Drenaje* learned where its pipes lay, staunched the leaks and addressed local hoarding, the system stayed broken. *¡Ay, caramba!*

At the faucet, the thirsty señora didn't care how hard or costly it was to collect, filter and deliver the water. She couldn't tell the water company what reservoirs and wells to tap, purification techniques to employ or pipes to use to route the water. She certainly didn't want to hear that she didn't *need* the water or hadn't used the spigot correctly. *She wanted a drink*, and felt it should flow to her in a timely and adequate way.

A judge could have ordered the water company to pump, but the cost in terms of wasted agua would have been astronomical and unsustainable. Telling the consumer to, "Find your own water or do without," was likewise untenable.

An apt metaphor for e-discovery, don't you think?

Litigants harbor immense reservoirs of ESI. Servers, like lakes and rivers, are evident and expansive. Databases and archives are vast subterranean aquifers. Information puddles in desktops, portable devices and online storage. It's costly to preserve, tap and process, and after all that effort, much is lost to leaky mains:

- We don't know where our pipes are buried (lax records management);
- We let sources evaporate and sour (poor preservation);
- We poison the well (spoliation);
- We use sieves to dip and dowsing rods to explore (careless collection and search);
- We fill the tub when a tumbler would do (overbroad requests for production); and
- We bathe in Perrier (conversion of ESI to image formats for manual review).

Through education, cooperation and improved tools and techniques, these holes are slowly getting plugged. Good thing, too, because our thirst for electronic evidence is growing fast.

Still, there's a leak in the pipes that draws no attention. Sometimes it yields just a trickle, other times it's a gusher; but if we don't find and gauge the loss, how will it ever get fixed?

This leak is blind reliance on text extraction and indexing engines as principal tools of ESI search.

Many think of electronic search in linear terms--as something that surfs across the connected and collected sources of ESI comparing words and phrases to queries. Indeed, that's the way we search files on our computers and how computer forensic tools typically operate.

But most e-discovery search efforts aren't linear explorations. Instead, they run against an index of words extracted from the source data.

So, is that really different? Quite.

It may take hours or days to extract text and create the index, but once complete, searches run against indices are lightning fast compared to plodding linear search. That's the upside. But there's a noteworthy trade off to using indices: you may not find what you seek even though it's in the collection and you've chosen the right keyword.

Why? There are several reasons text extraction and indexing let data evaporate. To start, text extraction tools parse data for sequences meeting the rules by which they define words. Is L33T a word? Is .DOC a word? How about 3.14159?

A simple parser might define a word as, "more than 4 but less than 14 contiguous alphabetic characters flanked by a space or punctuation." Parsers also employ rules barring certain combinations. Numbers, most punctuation and symbols are typically ignored, and common terms called "stop words" are sidelined, too. The very popular MySQL database excludes over 500 common English words, and DTSearch excludes more than 120; so, Shakespeare buffs can forget about finding "to be or not to be."

A more insidious shortcoming flows from failure to include encoded text in the index.

ESI is encoded in many different ways, and encoded objects are often nested like Russian matryoshka dolls. Consider this frequent scenario: a Word document and a PowerPoint inside a Zip archive attached to an e-mail message within a compressed Outlook PST container file. Each nested object is encoded differently from its parent and child objects, and encoding may vary within the body of an object. Encoding is critical. In fact, next to metadata, encoding may be the most important thing many people don't understand about e-discovery.

When a parser processes encoded ESI, it must apply the appropriate filter to the data to convert it to plain text so it that can be indexed. If the data is encoded in multiple ways, multiple filters must be applied in the correct sequence to cycle through all different forms of encoding to reach textual content. If no filter or the wrong filter is applied, the text isn't indexed. This occurs various ways, e.g., the encoding isn't recognized, the tool doesn't support the encoding, the content isn't text or the file is corrupted, encrypted or password protected.

If a parser doesn't recognize the encoding, it may default to applying the most common textual encoding schemes to the unrecognized content in a last-ditch effort to find intelligible text. But that doesn't always work. Foreign alphabets employ many more than our paltry 26 letters. Ideographic languages like Chinese and Japanese don't separate words with spaces. Even in English, you don't want to miss finding "résumé" when you search for "resume," so success hinges on whether the index is accent-sensitive or insensitive. Text parsers work around these challenges in various ways, but not all perform in the same way. There's many a slip 'twixt cup and lip.

Another reason data goes down the drain in text extraction and indexing is that failure is hard-coded into indexing applications. They're programmed to ignore file types deemed unlikely to hold text or apply only rudimentary text extraction methods. For example, Microsoft's Windows Search and Index Server have a limited capacity to index the contents of Access databases.

Finally, text extraction tools can't capture what they don't see as text. Facsimile or TIFF images are classic examples of text-laden documents not captured. These ESI sources, as well as documents storing text as vector graphics, must undergo optical character recognition to expose text. Although the same issue surfaces in linear search, you can subsequently run OCR against source data. You can't go back to the well with an index.

Maybe that's the ultimate failing of indices: *They're just a shadow of the evidence*. Because an index isn't the data, you can't apply new and better ways to wring out the truth.

Is it wrong to employ indexed searches in e-discovery? Certainly not, but it's all wet to select a tool to perform a task it can't accomplish. So, plumb the depths of your parser and indexer, then test them against representative samples of the data in the case and evaluate the search results for both data recovery and leaks. Be prepared to identify which encoded formats, file types and stop words are absent from the index.

In short, you need to know the capabilities and limits of the text extraction and indexing engines you deploy. Because if the index won't hold water, you're up a creek.

Tell Ol' Yahoo, Let my e-Mail Go

by Craig Ball

[Originally published in Law Technology News, September 2009]

A voice came from on high and said unto me, "Go forth and harvest the clouds." Well, not a *voce in excelsis* exactly, but a court order directing I gather up parties' webmail. The task seemed simple enough: The litigants would surrender their login credentials, and I'd collect and process their messages for relevance while segregating for privilege review.

Their data lived "in the cloud," and considering its celestial situation, I might have taken a cue from Ecclesiastes 11:4: "Whoever looks at the clouds shall not reap." So it was, I nearly got smote--not by Yahweh but by Yahoo!

Cloud computing refers to web-based tools and resources that supplant local applications and storage. It's called "the cloud" because of the cloud-shaped icon used to signify the Internet in network schematics.

Cloud computing lets companies avoid capital expenditure for hardware and software. Instead, they scale up or down by renting "virtual machines" as needed, connecting to them via the Internet. Cloud computing also encompasses Software as a Service (SaaS), where users “lease” programs via the Internet--think Google Apps or SalesForce.com--along with the much-touted Web 2.0--a catchall for Internet-enabled phenomena like social networking, blogs, wikis, Twitter, YouTube and arguably any web-centric venture that survived the dot-com apocalypse.

Such cloud-based services aren't new--my e-mail's been in the cloud for five years and twice that for my calendar. But cloud computing is big news in today's economy as companies great and small seek savings by migrating data services to the ether. For the rest of us, accessing and searching our e-mail from anywhere, coupled with near-limitless free storage, makes webmail irresistible. No surprise, then, that Yahoo! Mail's estimated 260 million users make it the largest e-mail service in the world. Add Hotmail and Gmail, and we're talking half a billion webmail users!

The silver lining for e-discovery is that all those candid, probative revelations once the exclusive province of e-mail now flood social media like FaceBook and Twitter. But cloud computing poses e-discovery challenges of near-Biblical proportions because it’s harder to access, isolate and search ESI without physical dominion over the data. Moreover, repatriation of cloud content depends on the compatibility of cloud formats with local storage formats and tools, including the ability to preserve and produce relevant metadata.

Consider the unique way Gmail threads messages into conversations. How do you replicate that structure in the processing and presentation of ESI? You can say, "We don't care about structure;" but increasingly, the *arrangement* of information is vital to full comprehension of the information. Such meta-information is key to a witness' ability to identify and authenticate evidence, especially when it's culled from collaborative environments like virtual deal rooms and Microsoft Corporation's popular SharePoint products.

Crafting protocols to reliably collect ESI from the cloud isn't tomorrow's problem. Today, it’s the rare e-discovery scenario that doesn’t involve webmail, and the court appointing me demanded action now.

I wasn't about to employ Yahoo! Mail's rudimentary search tools to tackle tens of thousands of messages and attachments. I needed a local collection amenable to indexing, search and de-duplication.

Yahoo! Mail lets users download messages and attachments using the common Post Office Protocol (POP), but only from the Inbox folder! *Thou shalt not download from Sent items, custom folders or Drafts.*

I'd either have to forgo multitudes of messages or find a workaround that would make Yahoo! let my e-mail go. I investigated third-party applications like Zimbra and YPOPS that claim to download from beyond the Inbox and tried them without success.

The workaround I devised required multiple steps and careful accounting. The initial set-up involved three steps:

1. I created a pristine user account in a local e-mail client to receive the messages. This can be done using Microsoft Outlook, but I turned to something every Windows user already owns: "Windows Live Mail."
2. I next downloaded the entire contents of the user's Yahoo! Mail Inbox to the Windows Live Mail Inbox, checking to be certain that message counts matched.
3. Then, I created a Live Mail folder called "Hold Inbox" and moved the downloaded messages to it. I did the same thing on the Yahoo! Mail side; that is, created a folder to temporarily hold the contents of the Inbox, then relocated those contents.

Now, the Inboxes were empty and available to serve as conduits to transfer the contents of other folders. In turn, I moved each folder's contents to the empty Yahoo! Mail Inbox, downloaded those items to the local Live Mail Inbox and shifted them to a like-named counterpart folder. After I'd captured all the folders of interest, I replaced the temporarily relocated Inbox contents on both sides and deleted the "Hold Inbox" folders.

Finally, I had a local counterpart of the Yahoo! Mail collection complete with matching folder structure. Using Live Mail, I could even export it as an Outlook PST for processing. Handled with care, the user should see no change to their Yahoo! Mail. But if you try this, be sure that the collecting POP client is set to leave messages on the server and that any Yahoo! Mail that arrives during the collection process makes its way to the local and Yahoo! Mail Inboxes.

This process worked, but it felt like that riddle where the man with the rowboat has to get a duck, a fox and a bag of corn across a river, transporting only one at a time. It's a reminder to consider more than cost savings alone when making the jump to cloud computing. It pays to know how much control you're ceding and how quickly and easily you can harvest your data, for "He that reapeth receiveth wages." [John 4:36]. Amen to that!

Jolly Roger Justice

by Craig Ball

[Originally published in Law Technology News, October 2009]

It's fitting that my friend (and author/blogger) Ralph Losey, hails from Orlando--the House of the Mouse--because reading his posts on *EDD Update* (www.eddupdate.com) is like a ride on one of the really good Disney attractions once called "E-ticket" rides. Losey's animated prose takes wonderful twists and turns, punctuated by delightfully silly visuals--and all steeped in solid American values. I always glean something good from Ralph's scholarship, even if only a different, well-argued point of view.

Losey and I have a playful wager respecting the viability of Judge Nuffer's opinion in *Phillip M. Adams & Associates, L.L.C., v. Dell, Inc., et al.,* 2009 WL 910801 (D. Utah March 30, 2009). I think the judge's opinion will stand (though pushing the outer bounds of preservation), but Ralph anticipates an appellate slap down.

Losey recently posted about *KCH Services, Inc. v. Vanaire, Inc.,* 2009 WL 2216601 (W.D.Ky. July 22, 2009), and kindly noted that where he disagreed with me on similar issues in Adams, we were of one mind on Vanaire. Hearing that I'd stumbled onto an acorn of rectitude moved me to actually read the opinion. And, indeed, Losey is right to side with the judge. (In fact, one can make a pretty good living siding with judges.)

But for one fateful misstep by the defendant, *KCH v. Vanaire* might have been the rare case where what seemed like spoliation was actually the decent thing to do. It wasn't, as it turns out, and Vanaire deserved the upbraiding it got. But let me explain why doing wrong might have been the right thing. Since it's a morality play, I'll tell it as a fairy tale.

Once upon a time a pirate named Scott the Freeman sailed from the shores of KCH and dropped anchor at arch rival, Vanaire. Legend has it that some proprietary software stowed away in Freeman's buckler, and made a home for itself on Vanaire's computers where, like all those birds and mice that help Cinderella dress for the ball, it spent happy years toiling through the night to spin magically efficient pollution control equipment layouts.

One crisp October in 2005, Kenny the Vengeful, King of KCH, called Guillermo the Elder, King of Vanaire, and said, "Fe fi fo fum, I smell the blood of a thieving scum" (or words to that effect).

Learning that his computers were enchanted with stolen software, Guillermo called for his pipe, and called for his bowl, and called for his fiddlers three, and then issued a proclamation:

"Henceforth, let it be known throughout the land that I banish from Vanaire any software that we did not purchase or do not own" (or words to that effect).

Scott the Freeman reported to the King and court that he was working to "insure there is nothing left on the computers."

And so it was, that instead of locking the stolen software in the dungeon--*poof!*--it vanished into thin air. (Hmmm…Vanaire, should have seen that one coming.)

RIGHT & WRONG

This is where our story comes to those fateful crossroads, Right and Wrong.

Would Guillermo the Elder advise Kenny the Vengeful that he'd found and banished the stolen software from his realm and humbly beg King Kenny's pardon? Or would King Guillermo take the dark and scary road through the fire swamp and claim that Vanaire didn't use or possess any pirate booty?

Vanaire got Scott Freeman, but it didn't get off scot-free. Based on the court's sanctions, Vanaire opted to go with a fairy tale instead of the truth.

And this, boys and girls, is where wrong could have been right. I can envision an old-fashioned, honest man saying, "I don't want any stolen software around here. Get rid of it, *now*!"

I can feel his anger and shame at being chastised by a rival and his zeal to purge the cause. Such intense emotions might, in a person of pride and honor, trump the deliberation and judgment needed to preserve the evidence.

Eradicating the software might have been the just thing to do, if not the wisest. Alas, truth did not out, and now we travelling minstrels strum our lutes and sing of Guillermo the Vain and Gassy, instead of Guillermo the Just.

Once, I knew an able lawyer who advised a client to delete stolen data anywhere it resided on the company's network after first copying it to a thumb drive. Sound advice? Afraid not. But decent and well-intentioned advice? Yes; especially when the other side demands an immediate cease and desist and you seek to mitigate damages. Wrong can feel right.

Judges are loathe to impose sanctions. They bend over backwards not to do it. So, if the facts bear out that Vanaire deleted the evidence just to foster a deception—to claim there was no stolen software because they'd made it disappear—then we should rejoice that another court has had the courage to lay on the lash for e-discovery abuse. That helps us all live a bit more happily ever after.

The ESIs of Texas

by Craig Ball

[Originally published in Law Technology News, November 2009]

My home state of Texas was the first to enact a discovery rule dealing with electronically stored information. Years before the federal rules amendments, and in four simple sentences, Rule 196.4 addressed a litigant's right to discover ESI, the scope of e-discovery, forms of production and cost shifting. The rule was either so completely successful or so utterly ignored that it wasn't cited in a published decision for nearly a decade.

So, when the Texas Supreme Court--the state's highest tribunal--issued its first e-discovery opinion, I listened to oral arguments. *In re: Weekley Homes*, 52 Tex. Sup. Ct. J. 1231 (2009), concerned a litigant's right to directly access an opponent's storage media. The plaintiff wanted to run 21 search terms against the hard drives of four of defendant's employees in an effort to find deleted e-mails from 2004. I eagerly anticipated insightful arguments by advocates who grasped the important technical and legal issues afoot, but what I heard would make a hearse horse snicker. Judge for yourself by listening to the arguments at http://tinyurl.com/weekleyhomes.

Fortunately for Texans and all e-discovery practitioners inspired by well-reasoned opinions, the lawyers' confusion didn't infect the Court's decision. The *Weekley Homes* standards that emerged from the Court's remand serve as a sensible guide to those seeking to compel an opponent to recover and produce deleted email, to wit:

1. Parties seeking production of deleted emails should specifically request them and specify a form of production;

2.Responding parties must produce reasonably available information in the format sought. They must object if the information is not reasonably available or if they oppose the requested format.

3. Parties should try to resolve disputes without court intervention; but if they can't work it out, either side may seek a hearing at which the responding party bears the burden to prove that the information sought is not reasonably available because of undue burden or cost;

4. If the trial court determines the requested information is not reasonably available, the court may still order production if the requesting party demonstrates that it's feasible to recover deleted, relevant materials and the benefits of production outweigh the burden, i.e., the responding party's production is inadequate absent recovery;

5. Direct access to another party's storage devices is discouraged; but if ordered, only a qualified expert should be afforded such access, subject to a reasonable search and production protocol protecting sensitive information and minimizing undue intrusion; and

6. The requesting party pays the reasonable expenses of any extraordinary steps required to retrieve and produce the information.

The Texas Supreme Court further articulated a new duty: Early in the litigation, parties must share relevant information concerning electronic systems and storage methodologies to foster agreements regarding protocols and equip courts with the information needed to craft suitable discovery orders. That's a familiar--though poorly realized--obligation in federal practice, but one largely absent from state court practice nationwide.

Weekley Homes brings much-needed discipline to the process of getting to the other side's drives, but scant guidance about what's required to demonstrate feasible recovery of deleted e-mail or what constitutes a proper protocol to protect privilege and privacy. Something that sounds simple to counsel can enormously complicate forensic examination and recovery, at great cost. A sound protocol balances what lawyers want against what forensic experts can deliver.

Because everyone uses e-mail, everyone has a little knowledge *about* e-mail. A little knowledge is a dangerous thing. Most assume their e-mail experience is universal, transferable and relevant. "When I delete a message," an opponent may say, "it goes into that trash bin, and I just look in there to find it." Much of what even tech-savvy lawyers and judges understand about deletion of data doesn't apply to e-mail.

For example, one of the lawyers arguing the *In re Weekley Homes* case claimed his client wasn't seeking electronic data like databases and spreadsheets. They were just seeking documents, i.e., deleted e-mail.

The problem with that distinction is that most email systems <u>are</u> databases. And not simple sorts either, but poorly-documented, proprietary, compressed, encrypted, pull-your-hair-out-complicated databases.

To illustrate, when you delete messages from the Deleted Items folder in a Microsoft Exchange mail server or Microsoft Outlook mail client ("double deletion"), they don't go to the Recycle Bin. They don't even go to the same place documents go when you empty the Recycle Bin. They don’t just slink off to the unallocated clusters, and they don't simply "*lose* their address in the file directory" as counsel claimed in the *Weekley Homes* arguments. Because individual messages aren't tracked by a computer's file system before deletion, they *never had* an address in the file directory!

Double deleted Outlook messages lurk locally inside the mail container file, invisible to the user, until the container file is compacted. Maybe in a day. Maybe two weeks. Maybe never.

If deleted messages were stored on an Exchange server or back up media, everything changes--the potential for recovery, the places an examiner looks, the encoding of the messages and even the tools and techniques employed are different. For example, Microsoft's Exchange Server includes a deleted item recovery feature inelegantly named "the dumpster." Messages purged from a user's Deleted Items folder are gone insofar as the user is concerned; however, those double deleted messages remain in the Exchange dumpster for a period of (typically) 14 to 30 days after deletion or for any interval set by the server administrator.

If the deleted messages were webmail, the leftovers lodge in entirely different forms and venues!

Again, e-mail are entries in a database, and thus, they reside within a world all their own. They are like the bottled city of Kandor in the Superman comics or domed Springfield in The Simpsons Movie--encapsulated and isolated from the outside world of the computer's operating system.

Deleted messages may serve as linchpins of liability and be well worth the cost and effort of recovery; but recovery methods and expectations must be calibrated to particular systems and applications, as well as to the needs and the budget of the case.

Recovering deleted e-mail is one of the most challenging tasks in computer forensics. If someone assures you it's easy or cheap, they've either never done it, or they're not doing it very well.

Geek's Gift Guide

by Craig Ball

[Originally published in Law Technology News, December 2009]

I love tools. My most fondly cherished Christmas presents as a boy were the Heathkits I used to build my first oscilloscope and frequency counter. Forty years on, tools are still the most tantalizing treasures under the tree, whether it's a digital level for the shop or a well-crafted whisk for the kitchen. Tools are empowering, enabling us to take things apart, divine their secrets and put them back together again--not unlike electronically stored information in discovery.

So, with the holidays at hand, I offer some favorites from my e-discovery toolbox. A few are pricey, but there are stocking stuffers, too- maybe something you'll want to whisper in Santa's ear.

Two admirable tools for comprehensive desktop e-discovery are from Australia. The first and most polished software package is called Nuix (www.nuix.com, $15,000 per annum). It's an inspired approach to indexing and searching the most-commonly encountered ESI, especially Microsoft Outlook and Lotus Notes e-mail.

Sic Nuix on a motley collection of e-mail containers and documents, and in no time, it opens, analyzes and indexes contents, dutifully flagging files that are unrecognized, encrypted or not amenable to text search. Nuix careens through big e-mail collections, and because the items are fully indexed, search results are instantaneous.

Nuix supports complex Boolean queries, as well as fuzzy and proximity searching, along with easy filtering by, e.g., evidence item, file type, classification and hash value. It capably de-duplicates and generates detailed reporting, but really shines when it comes to generating production sets. Need to build a PST from Lotus Notes messages? No problem! Want to produce Concordance, Ipro, Ringtail, or Summation load files? It's nearly as easy as point and click. Nuix isn't perfect, but it's as close to perfect as anything geared to mass market electronic data discovery I've seen.

The second wonder from down under is a newcomer called Intella (Vound Software, $2,695 with one year support). Despite a rough-around-the-edges interface, Intella is remarkably nimble at making common ESI formats amenable to search, and it delivers most of the same "must have" e-discovery features seen in the more expensive Nuix.

Intella offers arresting visual analytic capabilities not found in comparably priced offerings. Some will find visual analytics invaluable for teasing out relationships between issues or custodians. I found them handy for exposing connections based on as many as four or five search parameters, but they quickly grew too hard to follow using more.

It remains to be seen whether Intella will catch or surpass Nuix, but the promise is there, and price may prove the critical differentiator.

MUST HAVE TOOLS

Topping the category of "Tools I Couldn't Do Without" is a brilliant example of German engineering called X-Ways Forensics (X-Ways Software Technology, $1,169).

The simplest way I can describe XWF is to say that it does pretty much everything its better-known competitors--Guidance Software Corp.'s EnCase and Access Data Corp.'s Forensic Tool Kit--can do, but at far lower cost. Plus, it's capable of digital derring-do the other forensic suites can't match.

The caveat is, XWF is so powerful and incredibly feature-rich, it's not a tool that belongs in untrained hands. Training adds another $2,000 or more to your investment, but the return is formidable.

Unlike other tools here, XWF is not for the litigator's desktop. It's an expert's tool, and for the cost-conscious power user and frugal aficionado of binary bliss, it's X-Ways über alles!

Compared to the power tools just mentioned, Acrobat 9 Pro Extended (Adobe Systems, $635) seems as exciting as a sweater from Aunt Edna, but did you know Acrobat can convert up to 10,000 Microsoft Outlook and Lotus Notes e-mail messages to nicely-indexed, easy-to-use PDF Portfolios? Acrobat even embeds message attachments in the PDFs, so everything you need for a modest EDD production is at hand.

If e-mail conversion doesn't spike your eggnog, then consider that Acrobat 9 is also a capable and secure e-document redaction tool. It will custom Bates number electronic documents and even functions as a rudimentary tiff-ing engine for imaged productions.

STOCKING STUFFERS

In the stocking stuffer category, some of my most indispensible e-discovery tools can be had for $100 or less. One tool I wouldn't be without is a Swiss import, Aid4Mail Forensic (Fookes Software, $99), a stunningly simple and effective e-mail extraction and conversion tool that can parse an Outlook PST container file into its constituent messages, filtering by date or text; extract e-mail addresses or attachments; and reconstitute common message formats as a PST.

On corporate laptops, Outlook e-mail is often stored as an OST, or offline synchronization file, instead of as the more-familiar PST. Not every e-discovery tool can handle OST files, so when I need to convert an OST to a PST, I turn to a simple application called Recovery Toolbox for Outlook (Recovery Toolbox, $74) that works every time.

If I need to examine a machine remotely, a remarkable tool called F-Response (F-Response; Trial Edition, $100 for 15-day license) makes the distant machine available over a network or the internet in as complete a way as if it were right in the lab with me, including areas accessible only with specialized forensic tools. It's like astral projection for e-discovery.

Then again, it's not all high-tech. Sometimes, you just want to slap on quick, legible adhesive label. I use my P-Touch QL-500 PC label printer (Brother International, $59) to identify evidence, flag machines for preservation and print shipping labels. Labels are cheap; chain-of-custody errors costly.

They say the best things in life are free, and that's certainly true with respect to FTK Imager (AccessData, free). This easy-to-use program creates and accesses forensic images, hashes files, exports detailed media contents listings and even fashions forensically-sound containers for data collection without mangling metadata.

Happy holidays!

E-Discovery Bill of Rights

by Craig Ball

[Originally published in Law Technology News, January 2010]

There's a move afoot to revamp the e-discovery rules. When it comes to electronic evidence, some want to strip the comma from the mandate that litigation be "just, speedy and inexpensive."

Dig beneath the efforts to "reform" e-discovery, and you'll find familiar corporate interests dedicated to closing the courthouse doors. Their rallying cry: "Let's do things as we've always done them." Even trial lawyers, erstwhile champions of discovery rights, are so cowed and confused by e-discovery, they're ready to trade the cow for magic beans enabling them to dodge the hard and humbling task of acquiring new skills.

True, there's waste and inefficiency in e-discovery, largely driven by fear and ignorance. Requesting parties are struggling to adapt, and their demands for the moon and stars would be silly if they weren't so serious.

But requesting parties have rights. If there were a Bill of Rights protecting parties seeking electronic discovery, it might read like this:

I am a requesting party in discovery. I have rights. I am entitled to:

1. Production of responsive ESI in the format in which it's kept in the usual course of business. A producing party's fear of alteration, desire to affix Bates numbers or preference for TIFF images doesn't trump my right to receive the evidence in its native or near-native form.
2. Clear and specific identification of any intentional alteration of ESI made in the discovery process. If, e.g., a producing party omits attachments or redacts content or metadata, the producing party must promptly disclose the alteration with sufficient detail to permit me to assess whether such action was warranted.
3. Production of relevant metadata when I can promptly and specifically identify the metadata fields sought and articulate a reasonable basis for the production.
4. Discover the methodology employed to either select ESI for production or cull ESI from production whenever the method employed was automated, i.e., something other than manual review for responsiveness. This includes disclosure of the relevant capabilities and limitations of electronic search and indexing tools employed to produce or exclude ESI.
5. A detailed explanation of costs when a producing party asserts cost as a basis to resist e-discovery.
6. Put my technical advisor in direct communication with a knowledgeable counterpart for the producing party when technical issues arise, with reasonable and appropriate limits to protect legitimate privilege or confidentiality concerns.

7. Assume a producing party is preserving ESI that I specifically requested be preserved absent timely notice to the contrary.
8. Rely on the use of an iterative approach to electronic search, whereby the production of ESI from an initial search and review informs at least one further electronic search effort.
9. Adequate preservation and complete production, both in proportion to the amount in controversy and importance of the matters at issue.
10. Competence, candor and cooperation from producing party's counsel and support personnel commensurate with the competence, candor and cooperation extended by my counsel and support personnel.

11. These rights come coupled with duties. Requesting parties have a parity obligation to learn this new craft, work cooperatively and let relevance and reasonableness bound their actions.

I am a requesting party in discovery. I have duties. I am obliged to:

1. Anticipate the nature, form and volume of the ESI under scrutiny and tailor my requests to minimize the burden and cost of securing the information I seek.
2. Clearly and promptly communicate my expectations as to the forms of ESI and fields of metadata sought and be prepared to articulate why I need a specified form of production or field of metadata.
3. Work cooperatively with the producing party to identify reasonable and effective means to reduce the cost and burden of discovery, including, as appropriate, the use of tiering, sampling, testing and iterative techniques, along with alternatives to manual review and keyword search.
4. Know the tools I expect to use for review and processing of ESI produced to me and whether those tools are suited to the forms of ESI sought.
5. Work cooperatively with the producing party to minimize the burden of preservation and to agree promptly to release from a preservation obligation any sources that do not appear likely to hold responsive ESI.
6. Accommodate requests to produce ESI in alternative forms when such requests won't materially impair my ability to access relevant information or use the material produced.
7. Accede to reasonable requests for clawback and confidentiality agreements or orders when to do so won't materially impair my rights or those of others similarly situated.
8. Direct requests for production first to the most accessible sources, and to consider responsive information produced and available to me in framing subsequent requests for production.
9. Make available a competent technical advisor to communicate directly with a knowledgeable counterpart for the producing party concerning technical issues and accommodate reasonable and appropriate limits to protect legitimate privilege or confidentiality concerns.
10. Employ counsel and support personnel who possess a level of e-discovery competence, candor and cooperation commensurate with the competence, candor and cooperation I expect from producing party's counsel and support personnel.

James Madison, author of the U.S. Bill of Rights, wrote, “Knowledge will forever govern ignorance; and a people who mean to be their own governors must arm themselves with the power which knowledge gives.” It takes years to learn the law and become an able litigator. It will take time for lawyers to arm themselves with the novel skills e-discovery requires. There are no shortcuts, and none to be found by “reforming” that which is not yet fully formed in support of ignorance.

Are We Just “Makin' Copies?”

by Craig Ball

[Originally published in Law Technology News, February 2010]

Whether as a punitive measure or to achieve an equitable result, courts have the power to shift the costs of discovery from responding to requesting parties. For lawyers obliged to advance litigation expenses, the prospect of handing a blank check to the other side is so chilling, it frequently forces them to quit the case. That’s a powerful strategic weapon, and one that might have gutted historic cases had they been brought in today's wired world. *"If you want the Board of Education's e-mail, Mr. Brown, you'll have to bear the cost of restoration."*

Cost shifting is never more fraught with peril than when costs flow from electronic discovery because lawyers have proved themselves spectacularly profligate in their approach to digital evidence. A predominant e-discovery strategy entails throwing massive amounts of money at any task in lieu of thought or skill, with much of that money gravitating to attorney’s fees for review of items not requiring lawyers’ eyes. Where else in law can ineptitude *both* stoke revenues *and* deliver a strategic advantage?

Of course, litigants harmed by poor stewardship of shifted costs can mount a challenge, but it’s hard for a judge to distinguish necessary from needless when faced with invoices couched in dollars-per-gigabyte and technobabble. “Your Honor,” the other side counters, “they asked us for *any* and *all* responsive ESI, so we processed anything and everything.”

Too, how can we expect judges to know what e-discovery services cost when they’re regularly confronted with improbably varying assessments? More than one EDD vendor has confided they offer *two* estimates of their cost of services: the big one they use to help a client prove the request is too expensive and the little one they use to get the work.

Though cost-shifting during discovery is problematic, some e-discovery disbursements should be routinely shifted when a case concludes, not to punish, but as a sensible evolution of the procedural rules—one that properly acknowledges ESI’s ascendency over paper.

I’m referring to taxing certain e-discovery expenses as the “court costs” a prevailing party is entitled to recover. Court costs aren't the same as costs of litigation and don't include attorney's or expert's fees. "Court costs" are narrowly defined by each jurisdiction's rules of procedure, but generally include filing, service and witness fees and costs of preparing transcripts. They may also include required printing and photocopying costs and fees of court-appointed officers, like special masters and interpreters. Because they’re “taxed” against the losing party, they’re usually called “taxable costs.”

In federal practice, taxable costs are governed by Rule 54(d) of the Rules of Civil Procedure and 28 U.S.C. § 1920, which dictate that the prevailing party shall recover, among other charges, disbursements for printing, for "exemplification and copies of papers necessarily obtained for use in the case” and for interpreters and court-appointed experts. Unchanged since 1978, the rules governing taxable court costs hearken back to a time when most information was on paper and have fallen out of step with modern discovery practice. For example, though requesting parties may now specify the forms in which they receive ESI, producing parties are not uniformly permitted to treat the expense of producing in the specified forms as taxable costs. Such a right would be a powerful incentive to producing parties to respect the requestor's choice, reducing discovery disputes.

Several recent cases address efforts to tax e-discovery expenses as costs, with mixed results. Some courts, as in *Klayman v. Freedom's Watch, Inc.,* 2008 WL 5111293 (S.D. Fla. Dec. 4, 2008), treat e-discovery as synonymous with the work of lawyers and legal assistants and refuse to tax e-discovery expense as costs because of prohibitions against the taxing of attorney fees. Other courts analyze specific processes with an eye to whether they are more like photocopying than legal services—even to the point of equating optical scanning of paper records to copying (taxable) but optical character recognition or electronic Bates labeling to lawyer work (not taxable). *See, e.g., Fells v. Virginia Dept. of Transp.*, 2009 WL 866178 (E.D. Va. Mar. 25, 2009) and *Rundus v. City of Dallas*, No. 3-06-CV-1823-BD (N.D. Tex. Nov. 2, 2009). The case law splits without clear guidelines, seeming to turn variously on whether the party seeking discovery was the principal beneficiary of the expense or upon the judge's attitudes about e-discovery and the merits of the failed claim.

One exasperated judge recently dismissed arguments that the work of a consultant to collect, search, identify and help produce electronic documents is akin to legal work, noting that EDD services are "highly technical" and "not the type of services that attorneys or paralegals are trained for or are capable of providing." *CBT Flint Partners, LLC v. Return Path, Inc*., *et al.,* No. 1:07-CV-1822-TWT, 2009 WL 5159761 (N.D. Ga. Dec. 30, 2009). Taxing $243,453.02 incurred for EDD services as costs, the Court concluded that electronic discovery is “the 21st Century equivalent of making copies.”

Lawyers seeking to tax e-discovery expenses as court costs should segregate and document disbursements that most closely correspond to § 1920 categories of taxable costs. It’s useful to establish that the producing party incurred the expenses in response to a request by or agreement with an opponent or an instruction from the court (*e.g.,* “produce as TIFF images with load files”) and that the cost was necessarily incurred to fulfill that request. It's key to show that the benefits reached the requesting party or the court. Be specific, and closely track the statutory language, likening to printing, copying, exemplifying or even translating as appropriate. Studiously exclude costs principally benefitting the producing party or merely for counsel’s convenience, and consider whether using a court-appointed neutral for e-discovery enhances your ability to tax such expenses as costs.

In the final analysis, Congress must amend 28 U.S.C. § 1920 to mesh with 21st century practice, ending tortured efforts to define EDD as photocopying, printing or exemplification. Till then, the courts must delineate the electronic counterparts for authorized § 1920 taxable costs, and the e-discovery bar and industry should work to develop guidelines on what components of EDD are properly taxable as costs and how these should be tracked and authenticated.

The Lowdown on Backups

by Craig Ball

[Originally published in Law Technology News, March 2010]

Backup is the Rodney Dangerfield of information technology. It don't get no respect. Or maybe it's Milton, the *Office Space* sad sack with the Coke-bottle glasses and the red stapler. Backup is pretty much ignored...until headquarters burns to the ground, and it turns out the tapes hold the last copies of the TPS reports.

But to lawyers, backup is the fearsome Lord Voldemort of ESI; so much so that the two imperatives every lawyer accepts about e-discovery are: (1) Instruct clients to preserve backup tapes, and (2) Resist all efforts to obtain information from those tapes.

Backup tape has long been the poster child for ESI deemed “not reasonably accessible.” After all, you can't search backup tapes unless you restore them, and *everyone knows* it's a slow, laborious and expensive task. Doesn't restoration require companies to re-create entire server environments just to have a place to return restored data? And what about redundancy? Because each backup set is a snapshot of data on a particular day, the information captured from one backup to the next is 90% the same. Reviewing such massively duplicative data is a costly, risky proposition. That's all well settled, right?

Except it's not.

In recent years, while the legal profession resigned itself to backup tapes being out-of-bounds in discovery and access was generally secured only via sanctions, technology pressed forward and turned much of the bench and bar's well-settled assumptions about backup tapes on their ear.

It's heresy, but sometimes, backup tapes will be the easiest, most cost-effective source of ESI.

To start, the role of tape in backup is changing. It's not the disaster recovery (DR) staple it once was. Drive-based backup once thought too costly or disruptive is now feasible and commonplace because great strides in data deduplication and compression, along with faster networks, lower hard drive costs and leaps in capacities, eroded the advantages of tape-based storage. Hard disk arrays now cost-effectively hold months of DR data, and for DR, that's long enough.

D2D (for Disk-to-Disk) backup made its appearance wearing the sheep's clothing of tape. The first disk arrays, called Virtual Tape Libraries or VTLs, were designed to emulate tape drives so that existing software and programmed backup routines needn't change. The move to VTLs undercuts the rationale for making DR data off-limits in discovery. VTLs are as accessible as computer hard drives because they *are* computer hard drives.

Even as D2D supplants tape for backup, there's still a need for a stable, low-cost, portable medium for long-term retention of data too old to be of value for DR but comprising the digital annals of the enterprise. Some of this is still met by tape, giving rise to a new acronym: D2D2T, for Disk-to-Disk-to-Tape. Tape holds the company's archives by design, prompting courts to revisit the equities of denying access based on burden and companies to maintain detailed, readily-accessible information about the contents of archival tapes.

Too, companies will increasingly outsource long-term retention of archival data to low-cost, secure cloud-based storage providers in the same way they now outsource physical storage to Iron Mountain, Inc. More data online equals more accessible sources of ESI.

The biggest game changer for backup tape is "non-native" or "virtual" restoration, which permits data on tape to be searched and specific files extracted without "landing" the contents of the tape. Virtual restoration dispenses with the need to obtain copies of obsolete backup software and the time, cost and aggravation of recreating a sometimes decades-old computing environment. All major vendors of tape restoration services offer non-native restoration services, and software is available for those seeking in-house virtual restoration capabilities.

The most striking progress in working with data on tape is seen in tools like those from Index Engines, Inc. (www.indexengines.com), which index and deduplicate tape on-the-fly. Now, it’s possible to search and reconstruct the content of documents and messages on tape from a database.

Yes, we may have reached the point where backups are not that much harder or costlier to deal with than dispersed active data, and they're occasionally the smarter *first* resort in e-discovery.

For example, if the issue in the case turns on e-mail communications between Don and Elizabeth during the last week of June of 2007, but Don's no longer employed and Elizabeth doesn't keep all her messages, what do we do? We could pursue a forensic examination of Elizabeth's computer (cost: $3,000-$10,000) and collect and search the server accounts and local mail stores of ten other employees who might have been copied on the missing messages (cost: $5,000-$10,000).

Or, we could go to the July 1 backup set for the company's e-mail server and recover just Don's and Elizabeth's mail stores (cost: $1,000-$2,500). By zeroing in on the right source for the right time, we get exactly what we need.

So you see, the conventional wisdom to fight any effort to go to the tapes as being too burdensome and too costly is outmoded thinking. On the right facts and with the right tools, tape can be the fastest, cheapest, most reliable source of ESI.

Sidebar to 3/2010 Ball in Your Court Column
Tape: It's About Time

Though backup tape seems antiquated, tape technology has adapted well to modern computing environments. Consider that those reel-to-reel tapes in 1980's computer rooms held 240 feet of ½-inch tape on 10.5-inch reels. Their 9 tracks of data stored a then-impressive 100 megabytes of information traveling at 1.2 megabytes per second. Compare them to the LTO-4 tapes in common use today. Within a 4-inch square cartridge, 2600 ft of ½-inch tape is divided into 896 tracks and holds 800 gigabytes of information traveling at 120 megabytes per second.

That’s 100 times as many tracks, 100 times faster data transfer and *8,000 times greater* data storage capacity. Clearly, tape is a remarkable technology that’s seen great leaps in speed and capacity. The LTO-5 format arriving any day will natively hold 1.6 terabytes of data at a transfer rate of 180 megabytes per second.

Still, there are those pesky laws of physics. Time is our principal adversary when dealing with tape.

Remember auto-reverse tape decks that eliminated the need to turn over an audiocassette? Many backup tapes use a scaled-up version of that back-and-forth or “linear serpentine” recording scheme. “Linear” because it stores data in parallel tracks running the length of the tape, and “serpentine,” because its path snakes back-and-forth. Sixteen of an LTO-4 cartridge’s 896 tracks are read or written as the tape moves past the heads, so it takes *56 back-and-forth passes* or “wraps” to read or write one tape.

That’s about *28 miles* of tape passing the heads!

All that shuttling back and forth through the tape is a mechanical process, occurring at a glacial pace relative to the speed with which processors or hard drives move data. It takes time to traverse 28 miles of tape, sometimes stopping along the way to re-read flaky parts.

The big Atlanta-based tape house, eMag Solutions, LLC, recently examined the difference between the time it *should* take to restore a backup tape based on its stated capacity and data transfer rate versus the time it *really* takes considering factors that impact restoration like tape format, device interface, compression, tape condition, data block size and file size.

They found that it takes *about twice as long* to restore a tape under real world conditions than the media's stated capacity and transfer rate alone would suggest. Even to generate a catalog or use non-native restoration tools, a tape must be read in its entirety at least once. Each read takes hours, and even the largest tape houses can only go so wide in terms of simultaneous reads using multiple readers. Consequently, it's not feasible to deliver large numbers of tapes to a vendor on Friday and expect a catalog or searchable database to be generated by Monday. The *price* to do the work has dropped dramatically, but the *time* has not. When tape is part of the e-discovery plan, get on it early, and use tools that won't require tape to be re-processed for subsequent searches.

“Quickbooks” for E-Discovery

by Craig Ball

[Originally published in Law Technology News, April 2010]

As a special master investigating spoliation and abuse issues, I directed a big firm partner to turn over three years of e-mail. Of course, I wanted the mail in a searchable electronic format. Counsel estimated it would take three days to locate and segregate the messages. That didn't include time for IT to work its magic before the lawyer put eyes on the mail and, afterward, to put the mail in a container file for me.

"C'mon," I thought, "Twenty-four gilt-edged lawyer hours just to separate one case's e-mail from another's? And why all the IT time?" Shouldn't any lawyer be able to search, select and produce his or her own e-mail?

Certainly. But for most of us, despite our years of e-mail experience, gathering all the messages related to a particular matter is still a slog. If we didn't folder a message when it came in or went out, we'll be whacking our foreheads for not adding some unique matter identifier to every message. Because we didn't invest a second or two then, we're wasting hours or days now.

The big firm's first mistake was to leave the subject and content of case-related e-mail entirely to the discretion of the correspondent — so rapid, reliable electronic division was impossible. The "file" was no help, because to get a message in the file, it had to be printed, the case name added by hand, and then filed.

At best, this was a hit-or-miss operation. Over time, few lawyers bothered to print messages because few used the paper files. The paper file was still the official historical record of the matter, but as more of its contents migrated to electronically stored information, the file lacked the "who, what, when, where and why" of the case.

To meet my request, the big firm partner had to work with the messages, not the file, and that's where things went off the rails. Apart from Microsoft Outlook, the firm had no desktop tools that allowed lawyers to work with ESI — nothing to sort, search, and segregate with the ease of paper.

My request would be addressed in the same way the firm dealt with e-discovery: by hiring a vendor, processing a whopping slug of data into images and load files, then loading it all into a big ticket review tool at a per gigabyte price. It was using elephant guns to kill a fly.

By the time the dust settled, a request that should have cost next-to-nothing — Let me see your e-mail on this case — would cost tens of thousands of dollars. That's not an e-discovery problem. It's a tools and training problem. Lawyers lack desktop tools and skills to handle even their own ESI.

The reason I thought it would be simple is because I have all sorts of programs capable of slurping up an e-mail container, quickly indexing and deduplicating the contents and letting me gather, parse, and tag messages by date, keyword, sender, recipient, subject, attachment, whatever. I can view and browse messages and attachments as fast or faster than paper, and when I've made my selections, I can uniquely identify each item and export it to a variety of common formats.

Another reason I thought the task no big deal is because it's something I do all the time. The tools I use to manage everyday volumes of ESI are as familiar to me as Word, PowerPoint, and Excel are to other lawyers. They're not intimidating anymore, so it's economical and reliable for me to jump in and do it, even at lawyer billing rates.

I've lived through a lot of law office technology: from carbon paper and dictation belts, to word processing and online legal research, to e-discovery and telepresence. At each hill, there were those who insisted lawyers weren't going over. "Lawyers don't type." "Online legal research is for librarians." "My assistant handles e-mail."

It seems like yesterday I was going coast-to-coast telling lawyers they'd someday use the internet from their PCs and mobile phones. "Crazy talk," some said. And they were right, until the tools became easy and cheap, and until we stopped saying, "We won't," and started thinking, "We'd better."

Cost is the first hurdle. My desktop tools cost thousands, yet most lawyers don't own software that costs more than $500. What's the priciest program on your machine — Microsoft Office? Adobe Acrobat? Would lawyers have laid hands on keyboards if word processing programs still cost $3,000 per user? Would e-mail have taken off at $3 a message?

We need a "QuickBooks" for e-discovery: a tool that runs on an ordinary PC or Mac and slices and dices e-mail and common file formats like a champ. It'll be inexpensive enough to be ubiquitous, like Office or Acrobat, so we can teach it at CLE and know the other side has it, too. Then, like me, you'll have it. You'll use it. You'll wonder how you ever lived without it.

I thought this was just a solo and small firm need, but now I see that big firm lawyers need desktop ESI tools as well.

To Have and To Hold

by Craig Ball

[Originally published in Law Technology News, May 2010]

Several years ago, DLA Piper counsel and e-discovery luminary Browning Marean observed that, "Knowing how to draft a proper litigation hold might be a litigator's most important skill." Browning had a clear-eyed vision of the future. Few cases ever make it to trial, so jury skills are needed less often than the ability to fashion a successful legal hold, something a lawyer must accomplish in every case. Plus, in U.S. District Judge Shira Scheindlin's court in lower Manhattan, *not* knowing how to implement a proper litigation hold invites strict liability for gross negligence and severe sanctions.

Judge Scheindlin's terrifying 88-page opinion in *Pension Comm. of the Univ. of Montreal Pension Plan v. Banc of America Sec. LLC*, No. 05 Civ. 9016 at 10, 2010 WL 184312 (S.D.N.Y. Jan. 15, 2010), gives no quarter: "[T]he failure to issue a *written* litigation hold constitutes gross negligence because that failure is likely to result in the destruction of relevant information." As a standard of care denoting strict liability, that pronouncement goes too far. Certainly, there can be diligent, responsible preservation efforts absent written hold notices. But the opinion is well-reasoned overall, and the first 41 pages are compulsory reading.

It's a lawyers inclination to distill a decision like *Pension Committee* down to a black letter proposition and *do* something: develop a checklist, draft a form or tweak their discovery boilerplate. What they *don't* want to do is parse the holding for every new case. Modern lawyering is programmatic; necessarily so when non-hourly billing arrangements or insurance companies are involved. Thinking is a liability when carriers cap billable hours. Thus, the matter-specific instructions essential to an effective, efficient litigation hold quickly devolve into boilerplate so broad and meaningless as to serve no purpose but to enable the lawyer to say, "I told you so," if anything falls through the cracks.

How can we insure that the legal hold doesn't become just another formulaic, omnibus notice--so general as to confuse and so broad as to paralyze?
Realistically, we can't. The use of forms is too ingrained. But we can tweak our reliance on forms to avoid the worst abuses and produce something that better serves *both* lawyer and client. Accordingly, this column is not about "best practices." More like, "*not awful* practices." If you must use forms, here are some bespoke touches to consider:

Ask Why, Not Why Not: Lawyers don't eliminate risk, they manage it. Over-preservation saddles your client with a real and immediate cost that must be weighed against the potential for responsive information being lost. Your hold notice goes too far when it compels a client to "preserve everything." That's gross negligence, too--except the "sanction" is immediate and self-inflicted.

Get Real: It's easy to *direct* clients to segregate responsive matter, but the work could take them hours or days--*boring* days--even assuming they have adequate search tools and know how to use them. Some clients won't be diligent. Some will be tempted to euthanize compromising material. Naturally, you'll tell them not to deep-six evidence; but, anticipate *real* human behavior. Might it be safer and cheaper to shelve a complete set of their messages and lock down a copy of the user's network share?

Focus on the fragile first: Despite Judge Scheindlin's tough talk, you can't get in trouble for a botched legal hold if the information doesn't disappear. Fortunately, electronically stored information is tenacious, thanks to cheap, roomy hard drives and routine backup. There's little chance the company's payables or receivables will go to digital heaven. The headaches seem wedded to a handful of dumb mistakes involving e-mail and re-tasked or discarded machines. Manage these risks first.

Key custodians must get e-mail and messaging hold notices, and IT and HR must get machine hold notices. Is it really so hard to put stickers on implicated devices saying, "SUBJECT TO LITIGATION HOLD: DO NOT REIMAGE OR DISCARD?" It's low tech, low cost and fairly idiot proof. Deciding whether to pull backup tapes from rotation entails a unique risk-reward assessment in every case, as does deciding whether it's safe to rely on custodians to segregate and preserve ESI. Live by these words: "Trust everybody, but *cut the cards*." If there's a technology in place like journaling that serves as a backstop against sloth, sloppiness and spoliation, a supervised custodial preservation may be fine.

Forms Follow Function: Consider the IT and business units, then tailor your forms to their functions. What's the point directing a salesperson to preserve backup tapes? That's an IT function. Why ask IT to preserve material about a certain subject or deal? IT doesn't deal with content. Couch preservation directives in the terms and roles each recipient understands. Design a form for each constituency instead of trying to cram it all into one monstrous directive every recipient ignores as meant for someone else.

Get Personal: Add a specific, personal instruction to each form notice--something that demonstrates you've thought about each custodian's unique role, i.e., "Jane, you were the comptroller when these deals went through, so I trust you have electronic spreadsheets and accounting data pertaining to them, as well as checks and statements." Personalization forces you to think about the witnesses and evidence, and personalized requests prompt diligent responses.

Don't Sound Like a Lawyer: An effective legal hold prompts action. It tells people what they must do, how to get it done and sets a deadline. If it's a continuing hold duty, make sure everyone understands that. Get to the point in the first paragraph. Gear your detail and language to a bright 12-year-old. Give relevant examples of sources to be explored and material to be preserved.

Sidebar:
I've been putting out this column for five years. Thank you, Dear Reader, for the hours,

interest and feedback you've generously shared. When, at a trade show or conference, you see my name tag and ask if I'm "Craig Ball the writer," I'm so flattered. You inspire me to earn that status. Writing for you is a privilege I don't take for granted. We're old friends now. If I get something wrong, tell me. If you liked a piece, or if it helped you, I'd love to hear that, too. Thanks!

Show No Fear

by Craig Ball

[Originally published in Law Technology News, June 2010]

How many times have you heard a lawyer tell a court that he or she doesn't "understand computer stuff?" Can you imagine a lawyer confiding that he or she doesn't "understand document stuff?" The single greatest problem posed by ESI isn't its volume or complexity. It's the reluctance of lawyers to exert the time and effort required to understand it.

Knowing how to identify, preserve, collect, search, review, interpret and present the predominant form of evidence in lawsuits is as much a measure of lawyer competency in this century as the ability to search the reporter system was in the last. Clients must trust that their litigation counsel possesses the same fluency and competency with electronic evidence as with paper evidence. So why is it so daunting for us?

A lawyer spends the better part of a quarter- century acquiring the skills needed to deal with paper records. Think about it: we have to master the alphabet, language, reading, writing, general and specialized vocabularies and "thinking like a lawyer." Because all but the last are routine parts of formal education, we don't credit the investment required to competently demand, find, interpret and produce paper evidence.

When we search paper records, we apply our innate understanding of how to open a file drawer, extract a folder and turn a page. We grasp what dates, page numbers and file names signify using just our knowledge of the calendar, number system and alphabetical order. We understand that the requesting party doesn't care to know the weight, color, composition or watermark of the paper or whose latent fingerprints might be lifted from the surface. We don't report the font used, the page size or what the ink smells or tastes like. A lifetime of training and experience equip us to effortlessly, unconsciously separate informational wheat from chaff.

By contrast, when information is stored electronically, we face a host of unfamiliar cues to meaning, order and relevance. Information is encoded in dozens of different ways. We're unsure what metadata matters. Date values have different meanings than we accord them in common usage. Everyone uses their own idiosyncratic approach to e-mail, and personal messages are mixed up with business correspondence. Because we leapt to using information technology without learning its ABCs, we lack the training, experience and tools to readily distinguish ESI wheat from chaff.

Perhaps the profession mistakenly assumed that ESI was just another flavor of document, and that its contents could be shifted to paper or electronic page images to continue using the old, familiar ways.

Only lately has it dawned on the bar how different digital information is in volume and complexity. It's not only grossly inefficient and obscenely expensive to deal with ESI as paper, but content is missed, too. In the end, mass conversion is simply not feasible for terabytes of information; moreover, it's not ethical to visit the immense cost on our clients.

An argument advanced against lawyers learning e- discovery is that it's not prudent for attorneys to grapple with e- discovery tasks at their high billing rates. "E-discovery," they say, "is work best suited to lower paid employees or contractors."

The appeal of this argument is that it makes shirking sound like altruism. Imagine the contention, "Talking is something anyone can do, so jury arguments should be handled by the lowest-paid talkers." Nonsense. If you don't appreciate the perception, judgment, preparation and skill that goes into jury argument, it's just talking. Likewise, if one doesn't appreciate that competent litigators must know where a client's ESI resides, the forms it takes and information it conveys, the cost and means to preserve and collect it, the risks of spoliation and the capabilities and limitations of automated search, then it's deceptively easy to dismiss e- discovery as a delegable responsibility. It's not.

A second, petulant argument for delegation much in vogue is that lawyers "shouldn't have to do it." They've studied hard, passed the bar and become adept at marshaling paper discovery. With all the demands lawyers face, how can they be expected to master information technology? So ESI becomes someone else's problem: a delegated task, like typing, filing or copying. But e- discovery isn't that sort of task. It's more like reading Chinese.

Imagine you don't read or speak Mandarin, yet your firm transfers you to its Beijing office to try cases before Chinese courts. Aided by translators and interpreters, you might get by. But, consider the cost, confusion, frustration and delay!

Clearly, you'd need to learn to speak and read the language. Chinese first graders do it, so how hard can it be for a talented Juris Doctor to master Mandarin?

Okay, *hard*; but unlike Mandarin, electronic discovery is not difficult to learn at any age.

The Good News

Again, e-discovery is not hard to learn. We're all citizens of Dataland now, and we need to learn its language. The first hurdle is accepting that there is no alternative but to learn it, then believing that you can. Other hurdles are identifying what you need to know and the right sources to study.

You could master the essential case law of e-discovery in a day, and you'd pick up a lot of useful information about the ways e-discovery efforts fail. But you'd still be ill-equipped to lead an effective, efficient e-discovery effort.

Most CLE about e-discovery is unhelpful because it's taught by lawyers teaching law. Instead, we need to learn the technology side of e-discovery.

An online computer forensics course or a community college class on networks or e-mail systems are great ways to get started. You can pick up a lot from self-study, but don't begin with books on e-discovery--most of them are written by lawyers. Instead, *focus on the technology first*, learning how computers work, how data "lives" and "dies," forms of ESI and roles of servers, networks and databases.

A simple, deftly-illustrated volume like *How Computers Work* (White, Ron, 9th Ed. 2007) is time well spent, and though it hasn't been updated in years, the clearest, most comprehensive free guide to everything related to personal computing remains *The PC Guide* web site (http://www.pcguide.com). Read Law Technology News (it's free).

Pick out one thing you want to understand each month, e.g., forms of production, application metadata or indexed search, and then hit Wikipedia and Google. Take one of your client's IT folks to lunch and ask how things work. Don't be afraid to ask the dumbest questions that come to mind. Listen, question, follow up and experiment.

You learned the Rule Against Perpetuities. You can learn this stuff.

Traffic Jam

by Craig Ball

[Originally published in Law Technology News, July 2010]

E-mail is such seductive, powerful evidence. It's personal, plentiful and candid. For most adults, e-mail is their primary means of written communication. When lawyers think "e-discovery," it's the e-mail they crave. No surprise, then, that e-mail traffic is the most sought-after and fought-over ESI.

So why are litigants and lawyers still so flummoxed by the preservation, collection and production of e-mail?

Part of the problem stems from the easy familiarity we feel towards e-mail. Using it every day, we assume we understand it. E-mail's avatar has been an envelope for so long, we fool ourselves into thinking it's like snail mail. That is, we imagine that inside that Outlook PST container file on our machine lies a stack of letters like those the postal office brings, and that when we read e-mail, we're just sifting through that stack.
But that's not how it works. And because understanding how it works is crucial to avoiding discovery mistakes, we need to recalibrate our thinking to see e-mail for what it really is: *a custom report queried from a complex database*.

Sorry. I used the dirty D word. *Database*. And you want to flee. No wonder, when you consider how Microsoft defines its personal e-mail container file:

"The Outlook Personal Folders File Format (PST file format) is a stand-alone, self-contained, structured binary file format that does not require any external dependencies. Each PST file represents a Message store that contains an arbitrary hierarchy of Folder objects, which contains Message objects, which can contain Attachment objects. Information about Folder objects, Message objects, and Attachment objects are stored in properties, which collectively contain all of the information about the particular item."

That's clear, right? Bill Gates is not a lawyer's kid for nothing.

This month's column begins a series on understanding e-mail that's geared to the not-too-technical reader. My goal is to instill the "e-mail is a database" mindset that will help you meet the challenges of collecting, searching, reviewing and producing e-mail in electronic discovery. Together, we'll peer into the gooey guts of messages and poke around the confounding containers they call home. We'll talk about time stamps, headers, message identifiers, threading, Base64, MAPI, RFCs and how to specify the right form or forms for production. Along the way, I'll share tips and tricks from the trenches--my own and those of electronic evidence gurus around the world.

We'll focus on "Exchange environments," so-called because they employ Microsoft's Exchange Server software to manage users' mail, calendars, contacts and more. When a business speaks of its "mail servers," most often, they mean Exchange servers.

Other systems, notably IBM's Lotus Notes/Domino and Novell's GroupWise, are out there; but, because Microsoft Exchange accounts for a projected 300 million mailboxes, compared to roughly half that number for other on-premises systems, understanding Exchange is crucial. Fortunately, the similarities between these enterprise, on-premises e-mail systems outstrip their differences.

We'll also talk about web-based mail providers, like Google, Yahoo and emerging cloud service providers that dwarf Microsoft Exchange's user base and handle most personal e-mail. Hosted e-mail services are rapidly chipping away at the on-premises approach to corporate e-mail.

So how much do you think you know about e-mail? Let's start with a six-question pop quiz:

1. You demanded "all metadata" for each Outlook e-mail message. How many fields of metadata is that?
2. Where do you look in your copy of a received message to determine if it was also blind carbon copied to someone else?
3. Your Outlook screen indicates that a crucial message sent to you from the same time zone was "Received" ten minutes *before* it was "Created." How is that possible?
4. You can read your company e-mail using Outlook when not connected to a network, but there's no PST file anywhere on your machine. How is that possible?
5. You collected e-mail as individual .EML messages from all the key custodians and calculated a hash value--a digital fingerprint--for each message. You want to use these to deduplicate, that is, to identify messages that were sent to more than one custodian. Can you do that?
6. Yesterday, you collected a group of Outlook messages, once as .EML files and again as .MSG files. Today, you collected the *same* messages in *exactly* the *same* way. All of the .EML messages from yesterday have the same hash values today, but *none* of the .MSG files do. The two sets look exactly the same. What's going on?

ANSWERS:

1. Metadata means "data about data." Outlook 2007 tracks 54 fields of information about each message. These include not only the familiar To, From, Size, Date, Subject, Cc and Bcc fields, but also a host of other descriptors, such as whether the message has been read, its importance, various date values, foldering information and flag status. Some of these fields reflect information arriving with the message and others describe actions by the recipient. Still others reflect the

version of Outlook in use or designate whether the message is slated for archival.
But that's just a fraction of message metadata. There's metadata from the Mail Application Protocol Interface (MAPI) and message routing, timing, formatting and attachment information. Attachments themselves hold a complement of metadata. Messages forward other messages with--you guessed it--still more metadata. All these messages, attachments and metadata nest within a PST container file, which resides within a folder within a volume within a file system, like Russian matryoshka dolls. There's metadata every step of the way.

So, demanding "all metadata" is as absurd as demanding "all information." Instead, consider what you're trying to establish, and seek just probative fields of metadata.

2. Don't bother looking. Only the sender's copy of an e-mail carries information about blind carbon copies.

3. It's possible because your e-mail arrives at your mail server before you retrieve it with Outlook. The Received date and time reflect when your server got the message, but the Created date and signify the time the message was stored on your local machine; that is, when Outlook retrieved it from your server.

4. If can read your e-mail in Outlook without a network connection or PST file, your machine is probably configured to store messages in an OST file. Such offline storage files are common in Exchange environments, especially on laptops. You need to consider them as part of a thorough collection of e-mail for e-discovery.

5. It won't work. Every e-mail message traverses a unique network path at a unique time to a unique e-mail address. These differences, along with (generally) unique message identifiers and segment separators embedded in each message, mean that every e-mail has a different hash value when the hash value is calculated for the entire message. Instead, hash values must be calculated for various sections of the messages, and these are compared to establish that the messages are the "same" in the ways that matter most for deduplication.

6. The .MSG and .EML formats for individual messages aren't just different ways to store the same information; each carries a different assemblage of information. When a .MSG message is exported, it contains an encoded date and time value indicating when it was exported, but the .EML file for the same message holds no export date and time. Consequently, the hash value of a message stored in the .MSG format will be different moment-to-moment, even if the contents of message exported hasn't changed a bit.

The Miracle of E-Mail

by Craig Ball

[Originally published in Law Technology News, August 2010]

This is Part Two in a series on e-mail in e-discovery

What you see when you open a message in Outlook or Gmail isn't just a snapshot of what someone sent to you. *It's a report*. It's generated by an invisible query and built of select fields of information culled from a complex dataset, then presented to you in an arrangement determined by your e-mail client's capabilities and user settings.

Dude, your e-mails are a *database*, and so are mine…and his…and hers. *Epic.*

And for most corporate e-mail users, their messages and attachments implicate at least *two* databases: the big one housed on a server and storing e-mail records for many users, and smaller, local counterparts residing on employees' desktop computers, laptops, cell phones, iPads and other email client devices.

E-mail Data and Metadata

E-mail databases do more than simply store and transmit messages and attachments; they add information, too.

When a user opens a message, their email client changes the message's appearance to indicate it's been read. When the user flags a message for follow up, moves messages to folders or deletes certain items, the e-mail client records these changes as data *about* data, i.e., *metadata.* This metadata, and pieces of information transmitted within the messages, are *fields* in a database that collectively comprise *records* users query to display what they see onscreen as e-mail messages.

Users rarely see all of the metadata that an e-mail server or local client stores about messages. Instead, they're given a nicely formatted presentation of just the data and metadata their e-mail client software is configured to display. That is, they see the fields in the default "report" that the message database writes to the screen. But, it's easy to see more--*much* more.

If you're an Outlook user, find the pane that lists your e-mail, and note the columns in your current view. You're certain to have From, Subject and Received among them. You may also spot columns for flagging messages or displaying their importance or size. Now, right click on the column title bar and select "Fields." In the "Show Fields" menu that appears, choose "All Mail Fields" from the Available Fields submenu, and you'll have dozens of additional fields from which to choose. Want to display whether a message was read or carbon-copied? Add those columns. Want to report if a message has been opened or flagged? Add those columns. With each column you add, you're revising the 'report' that Outlook displays about your e-mail.
Some of the data you're seeing was delivered in the e-mail, but the rest of it got tacked on by your server, your mail client or you. Actually, the data's not really "tacked on" to

the messages received, because the e-mails received are gone. Instead, all the data about the e-mail is stored in a database and retrieved and displayed when all the various parts of the message and its metadata are assembled onscreen for your viewing pleasure.

Wait a second! The e-mails I received are *gone?!?*

Chop Shop

It's true. In Outlook, when e-mail arrives via the Internet, the messages are ripped apart like hot cars at a chop shop. Headers come apart. Attachments are transformed. Dates and times get scraped off and grafted on. Hardly anything remains intact from the original message. It's all just parts in bins. But, like a chop shop, if you collect the right parts and reassemble them properly, it's a car again.

Whoa. If the messages we see onscreen aren't what traversed the web, just what *do* the original e-mails look like?

For that, we have to look somewhere other than Outlook because, chop shop that it is, the message source gets stripped as quickly as it arrives. Let's turn to an e-mail client that retains the source: Gmail. If you don't have a Gmail account, you can get one free at http://mail.google.com.

Open any message in Gmail, and look for the Reply button. See the arrow alongside that opens a pull down menu? Select the option, "Show original" from this menu.

The box that appears contains the message source. This is the *only* information that came across the Internet. It's the pure, unadulterated essence of e-mail. Take in its beauty, and marvel at the fact that ***every message and attachment that traverses the global Internet must conform to the brilliantly simple layout before you.***

The Miracle of Email

It may not look simple or elegant at first blush, but consider the countless languages it accommodates, the varied attachments it carries and the range of computing equipment it supports —from that old AT clone in the garage to the iPhone on your hip.

The most astonishing thing about the source e-mail is that you can read it. It's not ones and zeroes or hieroglyphs. It's just plain text. Even if the sender attached a picture of their kids or a recording of that drunken karaoke night you'll never live down, the attachments, too, are just plain text.

This is the miracle of e-mail. Because it's just text, it's compatible with any e-mail system invented in the last 40 years. Think about that the next time you come across a floppy disk and wonder how you're going to read it.

Internet e-mail as we know it today was born in 1971, when a researcher named Ray Tomlinson sent a message to himself, first using the "@" sign to distinguish the addressee from the machine. That first message was something like "qwertyuiop," not quite, “Mr. Watson, come here. I need you," but Tomlinson didn’t know he was changing the world. He was just killing time. Also back when the nascent Internet consisted of four university research computers, UCLA student Stephen Crocker originated the practice of circulating proposed technical standards (or “protocols” in geek speak) as publications called “Requests for Comments” or RFCs. They went via snail mail because there was no such thing as e-mail. Ever after, proposed standards establishing the format of e-mail were promulgated as numbered RFCs. So, when you hear an e-discovery vendor mention “RFC2822 content,” fear not, it just means plain vanilla e-mail.

Next month, we’ll deconstruct a message and quietly discuss MIME to explore why something as simple, durable and elegant as e-mail is so challenging in electronic discovery.

Magical, Marvelous E-mail

by Craig Ball

[Originally published in Law Technology News, September 2010]

This is Part Three in a series on e-mail in e-discovery

E-mail is simple. But because there's so much of it in so many different locations, and because enterprise e-mail resides in complex database environments integrating layer-on-layer of useful metadata, it's easy to lose sight of e-mail's inherent simplicity.

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

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

The essential point is that everything in any e-mail is plain text *no matter what was transmitted* .

And by plain text, I mean the *plainest* English text, called 7-bit ASCII in geek speak, lacking even the diacritical characters required for accented words in French or Spanish. It is text so simple any letter can be stored in a single byte of data.
The dogged adherence to plain English text stems in part from the universal use of the *Simple Mail Transfer Protocol* or SMTP to transmit e-mail.

SMTP only supports 7-bit ASCII characters, so sticking with SMTP preserved compatibility with older, simpler systems, but didn't accommodate multimedia attachments or the billions who don't communicate in English.

Somehow, the standards governing the structure of an e-mail needed to be extended to adapt to new content and new users without excluding existing users and legacy systems. Enter MIME.

MIME, which stands for *Multipurpose Internet Mail Extensions*, is a seminal internet standard supporting non-English character sets, multimedia attachments and message bodies with multiple parts (e.g., text and HTML). Virtually all e-mail today is transmitted in MIME format.

In simplest terms, MIME defines a set of message headers that divide and describe the content of enhanced messages without confusing older systems. It starts by announcing its presence with a line stating "MIME-Version: 1.0." It then denotes the content type and subtype, e.g., "Content-Type: multipart/mixed," signaling that the message will have multiple parts including text plus an attachment.

Each part of the message will have its own Content-Type header. Nested parts in plain text will indicate "text/plain," and parts structured like web pages will indicate "text/HTML."

When the message contains the same message in *both* plain text and HTML, the Content-Type header preceding the alternate formats will state "Content-Type: multipart/alternative" and the simplest versions (i.e., plain text) will precede richer, more-complex versions. Content-Type headers for attachments will typically reflect the type of content and the specific file type, e.g., "image/jpeg" or "application/msword."

To separate all these parts, e-mail clients set a boundary value, typically a long string of characters that won't inadvertently occur within the message body, like "boundary="Part_09011957_zarf." Despite their apparent complexity, boundaries are simply separators and no different than drawing a line between the sections. The final boundary ends with two hyphens.

Take a look at the source of an e-mail. Find the boundaries and Content-Type headers. See? It's not really so complicated after all.

The longest and least intelligible parts of any e-mail tend to be encoded attachments. Here, you'll encounter two more MIME headers: Content-Disposition and Content-Transfer-Encoding. If you've ever wondered why corporate logos appear in the body of a message and others pictures only show up as attachments, the answer lies in the Content-Disposition header. Content may be designated "inline" and show up in the body or tagged as an "attachment" and require action by the recipient to display. The Content-Disposition header is also where the name of, and (occasionally) date information about, the attachment is transmitted.

Because any file attached to an e-mail must first be converted to letters, numbers and a couple of punctuation marks for transmission, the Content-Transfer-Encoding header indicates what form of encoding was used in the conversion. The most common forms of attachment encoding are *7-bit, quoted-printable* and *base64*.

The first is plain text, the second is an enhanced form of text that supports a broader alphanumeric character set and the last is capable of transmitting any binary content stored on a computer.

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

But turning all that binary data into letters has a dark side in e-discovery reminiscent of those apocryphal monkeys at typewriters who will, in the fullness of infinite time, type Shakespeare's sonnets. Billions of seemingly random alphabetic characters necessarily form words, including keywords sought in discovery. Very short keywords occur with alarming frequency. If the indexing tool employed in e-discovery treats encoded base64 attachments as text or if your search tool doesn't decode base64 content before searching it, noise hits may be a significant problem.

At first blush, message headers seem esoteric. Why should a lawyer care about the alphanumeric ID assigned to a message, the message's offset from Greenwich Mean Time or the names of the servers it traversed before delivery?

The answer lies in how you plan to use the information in discovery. The Message ID enables threading messages and replies into coherent conversations. The offset from GMT allows messages to be organized chronologically, a task otherwise enormously complicated by the tendency of e-mail clients to apply local time zone and daylight savings time settings. Learning the names of e-mail servers assists in efficiently selecting and sampling backup tapes in discovery. This isn't just geek stuff. It's lawyer stuff, too.

Next month, we will look at these issues, along with the challenge of e-mail de-duplication and various systems used to receive e-mail.

Executing E-Mail

by Craig Ball

[Originally published in Law Technology News, October 2010]

This is Part Four in a series on e-mail in e-discovery

In previous installments of this four-part series, we explored the simple, elegant structure of messages and why e-mail systems are actually complex databases. We wrap up with a discussion of e-mail collection, message IDs, threading, deduplication, and forms of production.

Austin attorney Tom Watkins calls e-mail messages "the cockroaches of litigation. You can't get rid of them, and they always manage to turn up when company comes."

E-mail holds the revealing *res gestae* statements of the wired world. It's the evidence litigants crave and fear. It's been the lifeblood of white collar business days for years. Yet this friendly form of information still confounds us in electronic data discovery.

KNOWING WHERE TO LOOK

We can't manage what we can't find, so it behooves lawyers to know the customary places where e-mail resides. In a **Microsoft Exchange**/**Outlook** environment, the starting point for collection is usually the Exchange Server, a name applied both to the messaging system hardware (the "box" in a client's offices, or virtualized in the cloud), and to the software that runs on that hardware.

An Exchange Server aggregates mail, calendars, contacts, and other data from multiple users (a department or an entire company) into a single massive database. A hold may be applied to all contents of the Exchange Server or just selected user accounts. It's been common to extract the contents of key custodians' mailboxes for search and processing; but Exchange Server 2010 natively supports search of and collection within Exchange.

"ExMerge" is what IT calls the **Exchange Server Mailbox Merge Wizard**. It's a simple, free utility for exporting server-stored e-mail of individual custodians to separate PST container files. ExMerge supports rudimentary filtering, allowing IT staff to cull by message dates, folders, attachments, and subject line content at the time of export. Using ExMerge intelligently is an effective, *cost-free* way to trim electronically stored information volume early, before processing and review.

A little known feature of Exchange Server is the aptly named "Dumpster" that retains double-deleted messages. A message is double-deleted when it's deleted and then purged from the Deleted Items folder. By default, Exchange 2007 keeps Dumpster items for 14 days, but it can be configured from zero days to indefinitely. Prior to 2010, the Dumpster could be circumvented by users intent on deleting their files, and Dumpster contents couldn't be indexed for search. The Dumpster was redesigned in

Exchange Server 2010 to reliably thwart user deletion and support the search of Dumpster contents.

Though we think of the server as the best source of e-mail in EDD, Microsoft's marketing suggests that as much as 90% of enterprise e-mail resides *outside* the Exchange Server, making it just the first stop on an e-mail scavenger hunt.

When responsive e-mail can't be found on Exchange and journaling servers, local machines must be scoured for Outlook containers and message files with PST, OST, or MSG extensions. Outlook container files typically reside six folders deep, in a location unique to the machine's operating system. On a **Windows XP** system, each user's PST or OST container files will likely be found in C:\Documents and Settings\ *User Name* \LocalSettings\Application Data\Microsoft\Outlook. On a Vista or Windows 7 machine, the default path is C:\Users\ *User* Name\AppData\Local\Microsoft\Outlook.

But container file paths, and even the names of the container files themselves, can vary user-to-user. Users may create *ad hoc* message collections anywhere on their drives, and it's not uncommon for users to store messages or Outlook container files in a *file share* , the area allocated for their use on the file server.

WEB MAIL & THE CLOUD

Question: How do you pull backup tape, image drives or conduct forensic examinations in the cloud? *Answer*. You don't.

Web mail poses unique EDD problems stemming from the lack of physical dominion over systems and storage devices, narrowing the techniques and tools that can be used to preserve, collect, and search web mail. It's not just a home user and small business issue. Moved to trim IT costs, even large companies are looking to host Microsoft Exchange Servers in the cloud — a migration that fortunately coincides with the addition of litigation hold and EDD search capabilities to the latest release of Exchange Server. *Hint*: Look for the OST container file to eclipse the PST in importance, as synchronization with the cloud takes the place of local storage.

THREADING

The message threading enjoyed by **Gmail** users and others has long distinguished web mail from enterprise e-mail. I asked Peter Mercer of **Vound Software** to dissect message threading. He explains that three values in a message's header contribute to building a thread. The first is the Message-ID value added when a user hits Send. "When the recipient replies, the recipient's e-mail client inserts the sender's Message-ID into the `In-Reply-To' and `References' fields of the reply. The reply is then assigned its own unique Message-ID," he says.

"The In-Reply-To field is well named. It means, 'This replies to a message with the Message-ID of unique@domain.' The References field tracks the lineage of the messages in the thread by appending the Message-ID values from all messages in the thread. As people reply, the References field grows to reflect the Message-ID of each message in historical order," he continues.

Mercer counsels, "It's important that lawyers become comfortable with how Message-ID values contribute to e-mail threads. After all, Message-ID values were custom designed for the job." So the next time you need to identify an e-mail or thread, use Message-ID values as keyword searches. "If you're given an e-mail header with no Message-ID, there's a good chance it was never sent," he explains.

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 "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 allows hashing to be used to deduplicate 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 technology, it has a few shortcomings. Because the tiniest change in a file alters that file's hash value, hashing is of little value in comparing files that have any differences, even if they have no bearing on the substance of the file. The unique identifiers, time stamps, and routing data frustrate efforts to compare one message to another using their hash values, even for "identical" messages dispatched to different people.

Instead, deduplication of e-mail messages is accomplished by calculating hash values for selected segments of the messages and comparing those segment values. Message identifiers and transit data aren't hashed because, if legitimate, they will always be unique. Instead, the To, From, Subject, and Date lines, message body and encoded attachments are hashed. If these match, the message can be said to be sufficiently identical.

Hash deduplication is tricky. Time values and addressee aliases may vary. These variations, along with formatting discrepancies, may serve to prevent the exclusion of items that should be defined as duplicates. When this occurs, delve into the reasons why apparent duplicates aren't deduplicating, such errors may be harbingers of processing problems.

FORMS FOR REVIEW AND PRODUCTION

Imagine the e-discovery process as a strainer and a funnel. We use a host of objective characteristics (e.g., file type, date interval and path location) and subjective assessments (key word search and lawyer review) to cull and distill vast volumes of ESI to the smallest possible concentration of responsive material.

At two or more points in this process — when we marshal it for review and when we package it for production — we face a decision whether to convert information from its ordinary and customary forms. It's a critical decision in terms of cost, risk and utility; yet lawyers rarely weigh the alternatives carefully or competently. How many simply do as they've always done?

Conversion is sometimes unavoidable. E-mail often requires a modicum of conversion for its use in e-discovery. For example, an Exchange Server may hold the e-mail of entire companies or departments. It makes sense to segregate mailboxes of key custodians by converting Exchange server content to more compact, user-centric formats like PST or to discrete message formats such as EML or MSG. Rarely encountered message formats (e.g., **Novell GroupWise**) may need to be converted to more common messaging formats for processing.

Converting data is costly both in the information stripped away in the process and in the time and out-of-pocket expense required to accomplish the conversion. As ESI moves further from its native forms, it loses content, typically from a failure to convert the metadata undergirding the searchability, usability, structure, and evidentiary integrity of the source evidence.

On the cost front, returning to our funnel analogy, any process priced on a per-item or per-gigabyte basis should occur as close to the narrow end of the funnel as possible, making it prudent to push conversion as late in the process as possible or eliminate it altogether.

There are various forms in which to review and produce e-mail in discovery. Printing individual messages, either onto paper or by conversion into an image file format such as .tiff or PDF, is eminently reasonable when applied to very small and thus readily searchable e-mail collections of dozens or hundreds of messages.

However, as a collection grows to thousands or millions of messages, printing to paper or conversion to image formats is wasteful and unwieldy. It's especially extravagant when conversion occurs before review, i.e., at the wide end of the funnel.

Further, a .tiff image file is not electronically searchable; so, to replicate the native searchability of the source ESI, parties must fashion "load files" carrying the metadata and full text of the imaged message.

As applied to e-mail, "native processing, review and production" signifies working with ESI in the form or forms most closely approximating the contents and usability of the source. Often, this will be a form of production identical to the original (e.g., PST or NSF for Lotus Notes) or a single message format (like MSG or EML) that shares many of the characteristics of the source and can deliver comparable usability. In most cases, native processing, review and production is by far the better way.

Native production has its challenges, too. For example, when e-mail is produced in single message formats, the folder structures may be lost, and with it, important context. Different container formats support different complements of metadata applicable to the message, e.g., a PST container may carry information about whether a message was opened, flagged or linked to a calendar entry.

Though there's no perfect means of production, here's a rule of thumb that comes close:

Absent agreement of the parties or court order specifying the form of production, the form of production for e-mail should be either the mail's native (or near native) form or a form or forms that will:

1. Enable the complete and faithful reproduction of all information available to the sender and recipients of the message, including layout, bulleting, 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.

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.

8. Enable reliable date and time normalization across the messages produced.

Remember, you needn't settle on a single form of production when a mix of forms is better. For example, employ native forms for production of messages and attachments *not* needing redaction and limit the use of .tiff images solely to redacted production.

Is Producing ESI as Images Malpractice?

by Craig Ball

[Originally published in Law Technology News, November 2010]

In e-discovery, it's astonishing how often requesting parties fail to designate the forms in which they want ESI produced. Equally galling is how often producing parties ignore such designations and convert ESI to costly, cumbersome TIFF image formats just to accommodate creaky review tools and antiquated workflows or to indulge an extravagant affection for Bates numbers.

Maybe it's time we ask: Is producing ESI as images malpractice? At what point is wasting a client's money actionable or unethical?

Many approach e-discovery by the conversion of native electronic information to TIFF images. Then, because TIFF images don't include the searchable text of electronic originals, text must be extracted and stored in files accompanying the TIFF images. When loaded into review tools like Summation or Concordance, these text files are indexed or searched to enable counsel to select document images to read. Typically, a third file serves as a means to correlate each image to its counterpart text file when loading the data; hence, these ancillary files are called "load files."

Sound cumbersome? It is. Expensive? Oh yeah.

The burden and cost might be justified if turning to TIFF brought efficiencies or improved performance; but, TIFFs are a mediocre, error-prone medium for e-discovery. Pages go missing. Attachments stray from transmittals. Spreadsheets lose formulae and explode messily across TIFFs like staked vampires. Sound files and video are relegated to slip sheet "tombstones" marking their demise. Metadata often evaporates. In the end, requesting parties may reject unwanted TIFF productions and force expensive do-overs.

Employing TIFF for review ramps up expenditures in several ways. First, it costs money and time to convert ESI to TIFF. Pennies a page sounds trivial, but millions of pennies are tens of thousands of dollars. Further, TIFF imaging is confounded by data much beyond the confines of a letter-size page, resulting in imaging charges for pages without purpose. If you've ever mistakenly printed a big spreadsheet, you've felt the frustration of rows and columns splayed uselessly across multiple pages.

But hefty conversion costs pale in comparison to the princely price of loading TIFF images into a review tool on a pay-per-gigabyte basis. Compared to native forms, production of images costs at least twice as much. TIFFs are simply "fatter" files than their native counterparts; ironically so, because TIFF conversion strips away useful data and metadata, obliging litigants to pay to restore utility via clumsy workarounds like extracted text and load files.

The supersized pricing attendant to TIFF conversion doesn't stop at the cost to create and load chubby files. Bigger files cost more to host, transmit and deduplicate. Then, bigger files slow page displays, running up review costs.

How bloated are TIFF productions?

As an experiment, I grabbed two weeks of e-mail from Microsoft Outlook and processed messages and attachments into a variety of common forms for review, including PST, MSG, EML, MBOX, MHTML and TIFF. I employed both simple and sophisticated e-discovery processing software for the conversions to minimize tool anomalies and then compared the output. Whether stored as a single container file, like PST and MBOX, or as individual messages with embedded attachments, like MSG, EML and MHTML, the 830 messages (about 2,000 printed pages) consumed between 137MB and 148MB of disk space. PST was the leanest and MBOX the fattest, with the discrete message formats in a dead heat around 138MB.

But not even Jenny Craig could keep those portly TIFFs from tipping the scales at a corpulent 262MB. What's more, in order for the TIFFs to be usable in a review tool, they'd need added text and load files, packing on another four to six megabytes.

Best case scenario: TIFF roughly doubles the gigabyte size. In past tests with different ESI, TIFF productions were as much as *five times fatter*. Talk about empty calories! Bottom line, with TIFF production, you spend more and get less.

Proponents of TIFF productions claim three advantages. First, because TIFF is electronic paper, it forces ESI into paged formats, sometimes with ridiculous results. TIFF conversion thus supports embossing Bates numbers onto every page rather than simply assigning a Bates-style identifier to each electronic file and message. Lawyers love Bates numbers, but at what crazy cost? If we must paginate particular items for use at deposition or in Court, why not image and number *only those items* and supply marked sets for ready reference? To minimize cost, keep native data in native formats as long as possible.

Second, TIFF permits continued adherence to tools and techniques geared to brute force review and spawns higher attorney fees and vendor charges versus native review. Yet, attorneys and vendors often guide the choice of format. If not a frank conflict of interest, it's a circumstance that hardly inspires lawyers and vendors to be agents for change.

A third advantage voiced in support of TIFF image conversion is that it prevents alteration of evidence. Yet, TIFF images are easily altered. In fact, their easy alterability is the principal reason why TIFF images are the preferred medium of choice for redaction. Moreover, it's harder to detect fraudulent alteration of TIFF images than of native electronic files because the latter can be easily and quickly hashed to discern the slightest change.

The lunacy of image-and-load file productions lately hit home when I was asked to reproduce the native electronic file of a plat because it was claimed to be illegible. Though I'd furnished a high resolution, native PDF file with color-coded content, the recipients imaged the production to a lower resolution, black-and-white format for review, rendering the contents a grainy, gray mess. Useless, but Bates numbered.

Is it malpractice for lawyers to convert ESI to images? Perhaps the answer is, *not so as long as lawyers disclose that doing so more than doubles certain costs without an offsetting gain in performance or productivity*. Then, if clients consent to indulge their lawyers' penchant for inefficiency and waste, they do so with open eyes.

Ubiquitous Databases

by Craig Ball

[Originally published in Law Technology News, December 2010]

Databases touch our lives every day. Our computers, phones and e-mail are databases. Google, Westlaw, Craigslist, Amazon, E-Bay, Facebook: all big databases.

We can't web surf, make a phone call, use an ATM, charge a meal, buy groceries, get a driver's license or book or board a flight unless a database makes it happen.

Databases run the world. Yet, when it comes to e-discovery, we tend to fix our attention on documents, without appreciating that what "feel" like documents exist only as a flash mob of information assembled and organized on the fly from dozens, thousands, or millions of sources drawn from multiple fast-changing systems, locations, and formats. In our zeal for documents over data, we make discovery harder, slower, and costlier. Understanding databases and acquiring the skill to peruse and use their contents gets us to the evidence better, faster, and cheaper.

Databases demand that we re-examine our thinking about discovery. Historically, parties weren't obliged to create documents for production. They produced what they had. Today, documents don't exist until generated.

Tickets, bank statements, websites, price lists, phone records, and register receipts are all just ad hoc database reports. Only one-tenth of one percent of documents exist as ink on paper, obliging litigants to become adept at crafting queries to elicit responsive data and master ways to interpret and use that data.

THINKING INSIDE THE BOX

Databases employ database management software (DBMS) to enter, update and retrieve data. Though DBMS serves many purposes geared to indexing, optimizing and protecting data, the most familiar role of DBMS software is as a user interface for forms and reports.

SQL

There's little difference between forms and reports. We tend to call the interfaces used to input and modify data, "forms" and those that extract data, "reports." Both are merely user-friendly ways to execute "query language."

Query language is the set of commands used to communicate with databases. The best known query language is Structured Query Language or SQL, whose sole purpose is the creation, management, and interrogation of databases. Officially, it's pronounced "ess-cue-ell," but everyone says "sequel."

SQL is simple and powerful; but, even the simplest query language is daunting for most users. So, databases employ graphical user interfaces to put a friendly face on SQL. When you enter data into a form or run a search, you're triggering a series of preprogrammed SQL commands.

REACHING BEYOND REPORTS

When standard database reports are complete and precise enough to retrieve the information needed in discovery, parties need only agree upon reporting criteria and a suitable form of production. But, companies design databases for operations not litigation; so standard reports may not suffice. Then, you need to plumb deeper into the database.

GETTING THE MAP

Had Tolstoy been a database administrator, he might have said, "Great databases are all alike, every ordinary database is ordinary in its own way." You can only assume so much about the structure of an unfamiliar database; after that, you need the manual and a map.

The "map" is the database's schema, reflected in its data dictionary. A logical schema describes how the database is designed: tables, attributes, fields, relationships, joins and views. A physical schema describes the hardware and software implementation of the database: machines, storage devices and networks. As Tolstoy might have remarked, "A logical schema explains death; but, it won't tell you where the bodies are buried."

Entity Relationship Modeling (ERM) is a system and notation used to graphically lay out database schemas. The resulting flow chart-like diagrams are Entity Relationship Diagrams or ERDs.

Information in a database is mostly gibberish without the metadata that gives it form and function. In an SQL database, that metadata lives in the system catalog. The terms system catalog, schema and data dictionary are often used indiscriminately. They are databases about databases: metadatabases. They're the maps, and database jockeys need them.

LESSONS FROM THE TRENCHES

The value of schema, manuals, data dictionary and ERDs was borne out by my experience as special master for electronically stored information in a drug product liability action.

I was tasked to expedite discovery from as many as 60 different databases, each more complex than the next. Both sides had top-flight counsel and technically astute support teams, but the parties were at loggerheads, and serious sanctions were in the offing.

The plaintiffs insisted the databases would yield important evidence and were willing to narrow the scope of their database discovery; but first, they needed to know the systems.

For each system, we faced the same questions:

- What does the database do?
- What is it built on?
- What information does it hold?
- What forms does it take?
- What are its reporting capabilities?
- How can it be searched effectively, using what query language?
- What content is relevant, responsive, privileged?

•What form of production will be functional, searchable and cost-effective?

It took a three-step process to turn things around. The plaintiffs were required to do their homework, and the defense supplied the curriculum.

First, each system had to be identified. The defense furnished spreadsheets detailing, *inter alia* :

- Names of systems.
- Applications.
- Date range of data.
- Size of database.
- User groups.
- Available system documentation (including ERDs and data dictionaries).

This enabled plaintiffs to target requests to relevant systems, and I directed defendants to furnish operator's manuals, schemas and data dictionaries.

Next came narrowly-focused telephone meet-and-confer sessions between technical personnel.

The defense was required to put knowledgeable personnel on the calls, and plaintiffs were required to confine their questions to the nuts-and-bolts of particular databases. Afterwards, plaintiffs were required to revise their requests for production. Sometimes, the plaintiffs learned enough about the databases that they could proffer SQL queries.

This should have been sufficient, but the matter was especially contentious. The final step needed to break the database discovery logjam was mediation. Counsel, database administrators, and I met face-to-face for two days. We worked through each database and arrived at specific agreements concerning the scope of discovery, including searches to be run, sample sizes employed, and the timing and form of production. The devil is in the details, and the goal was to nail down every detail.

It took two sessions; but in the end, disputes over databases ceased, the production changed hands smoothly, and the parties refocused on the merits.

The heroes were the technical personnel who collaborated to share information and find answers when the lawyers could only quarrel.

The lesson: *Get the geeks together, and get out of their way.*

In another special master appointment, the court questioned the adequacy of defendants' search of the many databases used to run their far-flung operations, ranging from legacy mainframe systems housed in national data centers to homebrew applications cobbled together using Access or Excel.

With disturbing regularity, the persons tasked to query the systems didn't know how to search or lacked rights to access the data they were obliged to search.

Databases employ techniques to optimize performance and protect confidentiality that can result in responsive data being missed, even by an apparently competent operator:

- Older records are purged from indices.

- Users need system privileges to access all potentially responsive records.
- Queries are systemically restricted to regions, business units or intervals.
- Tables aren't joined in the ways required to include the data sought.

The lesson: *Never assume that a query searches all of the potentially responsive records, and never assume that the operator knows what they are doing.*

Establishing operator competence is tough. If you ask those tasked with running queries if they have the proper privileges, you'll draw some dirty looks. They have the privileges they need to do their jobs; but those may be insufficient to elicit all the system can yield. Chances are, operators don't know their limitations.

Database discovery is growing in importance, and will soon be as central to e-discovery as e-mail is today. It's not too soon to learn the language of database discovery or consider how you will pursue or resist it.

Get Back in the Race

by Craig Ball

[Originally published in Law Technology News, February 2011]

This year marks the thirtieth birthday of the first IBM PC. Yet, like the hare in that fabled race, litigators napped through much of the information revolution. The tortoise is so far down the road that, instead of sprinting to catch up, the trial bar nervously reassures itself that yesterday's skills will surely be good enough for tomorrow's challenges. Too many grouse, "I went to law school so I wouldn't have to deal with computers. Just give me a damn form and checklist."

We've got to stop kidding ourselves. It's too late for shortcuts and half measures.

There are no forms or checklists that can take the place of understanding electronic evidence any more than a Polish phrasebook will equip you to try a case in Gdańsk. But there are a few rules of thumb that, *applied thoughtfully*, will get you back in the race. Let's start with the Big Four and work through some geek speak as we go.

The Big Four
Without knowing anything about corporate IT systems, you can safely assume there are four principal sources of digital evidence that may yield responsive ESI:

1. **Key Custodians' E-Mail (server, local, archived and cloud):** Corporate computer users will have a complement of e-mail under one or more **e-mail aliases** (i.e., addresses) stored on one or more **e-mail servers**. These servers may be physical hardware managed by IT staff or **virtual machines** leased from a **cloud provider, e**ither likely running mail server software called **Microsoft Exchange** or **Lotus Domino.** A third potential source is a **Software as a Service (SaaS)** offering from a cloud provider (webmail)**.**

   Users also tend to have a different, but overlapping complement of e-mail stored on desktops, laptops and handheld devices they've regularly used. On desktops and laptops, e-mail is found **locally** (on the user's hard drive) in **container files** with the file extensions **.pst** and **.ost** for Microsoft Outlook users or **.nsf** for Lotus Notes users. Finally, each user may be expected to have a substantial volume of **archived e-mail** spread across several on- and offline sources, including backup tapes, **journaling servers** and local archives on workstations and in network storage areas called **shares** (discussed below).

These locations are the "*where*" of e-mail, and it's crucial to promptly pin down "where" to ensure that your clients (or your opponents) don't overlook sources, especially any that may spontaneously disappear over time through **purges** (automatic deletion) or backup media **rotation** (reuse by overwriting).

Your goal here is to determine for each key custodian what they have in terms of:

- *Types of messages* (did they retain both Sent Items and Inbox contents? Have they retained messages as foldered by users?);
- *Temporal range of messages* (what are the earliest dates of e-mail messages, and are there significant gaps?); and
- *Volume* (numbers of messages and attachments versus total gigabyte volume—not the same thing).

Now, you're fleshing out the essential *"who, what, when, where and how"* of ESI.

2. Key Custodians' Document Storage Areas (local folders and network shares): Apart from e-mail, custodians generate most work product in the form of **productivity documents** like Microsoft Word documents, Excel spreadsheets, PowerPoint presentations and the like. These may be stored locally, *i.e.*, in a folder on the C: or D: drive of the user's computer. More often, corporate custodians store work product in an area reserved to them on a network **file server** and **mapped** to a drive letter on the user's local machine. The user sees a lettered drive indistinguishable from a local drive, except that all data resides on the server, where it can be regularly backed up. This is called the user's **share** or **file share**.

3. Multi-user Document Storage Areas and Workrooms (server and cloud): Just as users have file shares, work groups and departments often have network storage areas that are literally "shared" among multiple users depending upon the access privileges granted to them by the network administrator. These shared areas are, at once, everyone's data and no one's data because it's common for custodians to overlook **group shares** when asked to identify their data repositories. Still, these areas must be assessed and, as potentially relevant, preserved, searched and produced. Group shares may be **hosted** on company servers or "in the cloud," which is to say, in storage space of uncertain geographic location, leased from a service provider and accessed via the Internet. Enterprises employ virtual workspaces called **deal rooms** or **work rooms** where users "meet" and collaborate in cyberspace. Deal rooms have their own storage areas and other features, including message boards and communications tools--they're like Facebook for business.

4. Databases (server, local and cloud): From Access databases on desktop machines to enterprise databases running multinational operations (think UPS or Amazon.com), databases of every stripe are embedded throughout every company. Other databases are leased or subscribed to from third-parties via the cloud (think Salesforce.com or Westlaw). Databases hold so-called **structured data**, a largely meaningless distinction when one considers that the majority of data stored within databases is unstructured, and much of what we deem unstructured data, like e-mail, is housed in databases. The key is recognizing that databases exist and must be interrogated to obtain the responsive information they hold.

The initial goal for e-discovery is to identify the databases and learn what they do, who uses them and what types and ranges of data they hold. Then, determine what standard reports they can generate in what formats. If standard reports aren't sufficient to meet the needs in discovery, determine what **query language** the database supports, and explore how data can be extracted. Databases tend to be updated and purged frequently, so getting to relevant data from the past may entail a costly and difficult restoration of the database from backups.

The Big Four don't cover the full range of ESI, but they encompass *most* potentially responsive data in *most* cases. A few more thoughts worth nailing to your forehead in 2011:

Pitfalls and Sinkholes

Few organizations preserve all **legacy data** (information no longer needed in day-to-day operations); however, most retain large swaths of legacy data in backups, archives and mothballed systems. Though a party isn't obliged to electronically search or produce all of its potentially responsive legacy data when to do so would entail undue burden or cost, courts nonetheless tend to require parties resisting discovery to ascertain what they have and quantify and prove the burden and cost to search and produce it. This is an area where litigants often fail.

A second pitfall is that lawyers too willingly accept "it's gone" when a little wheedling and tenacity would reveal that the information exists and is not even particularly hard to access. It's an area where lawyers must be vigilant because litigation is regarded as a sinkhole by most everyone except the lawyers. Where ESI is concerned, custodians and system administrators assume too much, do too little or simply say whatever will make the lawyers go away.

Lather, Rinse and Repeat

So long as potentially responsive data is properly preserved, it's not necessary or desirable in a high-volume ESI case to seek to secure all potentially relevant data in a single e-discovery foray. It's more effective to divide and conquer. First, collect, examine and produce the most relevant and accessible ESI from what I like to call the über-key custodians; then, use that information to guide subsequent discovery requests. Research from the NIST TREC Legal Track proves that a two-tiered e-discovery effort produces markedly better results when the parties use the information gleaned from the first tier to inform their efforts through the second.

In a bygone era of e-discovery, Thomas Edison warned, "We've stumbled along for a while, trying to run a new civilization in old ways, but we've got to start to make this world over."

A century later, lawyers stumble along, trying to deal with new evidence in old ways. We've got to start to make ourselves over.

Viva Las Metadata

by Craig Ball

[Originally published in Law Technology News Online, March 2011]

It was a Vegas-sized lawsuit on a fast track, so the court appointed a discovery committee of smart, young lawyers. The committee sought a "simple as paper" means to process, review and produce everyone's electronically stored information. Poised to pull the trigger on a multi-million dollar electronic repository, I was appointed to serve as ESI Special Master and fashion an electronic data discovery protocol.

Studying the proposed repository, I found that assumptions about data volumes bore little relation to reality. Also, much structured data would shift from multiple databases where it was usable into one where it would be all-but-useless. Settlements decimated the pool of potential repository users. The clincher: the lead parties wouldn't use it — preferring to stick with their own systems.

We needed a fresh approach. With a rocket docket scheduling order, we had to move fast, and faced several hurdles:

- Parties with smaller claims needed easy ways to meet e-discovery obligations and access data produced to them without the cost and burden of EDD eating up the value of their claims. In other words, the protocol had to be **proportional**.
- The protocol had to work for parties with little electronic evidence as well as those with vast volumes spanning hundreds of systems and servers. The protocol had to be **scalable**.
- Parties shouldn't incur hefty processing charges for changing data from one form to another. The protocol had to be **lightweight**.
- Finally, the protocol had to be **simple** enough to be implemented without armies of vendors and experts, i.e., an IT person could understand what to do.

All these factors favored production of ESI in largely native forms. So we began:
"To reduce cost and preserve maximum utility of electronically stored information, production in native electronic formats shall be the required form of ESI production in response to discovery in this cause." Adding: "Other than paper originals and images of paper documents (including redacted ESI), no ESI produced in discovery need be converted to a paginated format nor embossed with a Bates number."

It's one thing to say "produce in native and forget Bates numbers." It's another to make it work. Parties still need a way to identify production items, ideally one compatible with the inevitable use of printouts.

Consequently, we required that each native file, scanned document or e-mail message be identified by a Unique Production Identifier consisting of a four letter party designation and nine digit numeric sequence for the item. Pretty standard. The next five characters were reserved for the pagination of the item when printed to paper or converted to an image format. Only one printed version was likely to be used in any proceeding, so everyone would (literally) be on the same page.

We had our cake and could eat it, too. We could uniquely identify native ESI, yet had a way to correlate the native file with pages when printed.

For e-mail production, we settled on the single message MAPI-compliant MSG format, with the odd message from webmail, Lotus Notes or GroupWise converted to MSG. We resolved the loss of mail foldering by including mail foldering data as a field in an accompanying load file.

The MSG format natively embeds attachments as compressed, MIME-compliant, Base-64 content; accordingly, the protocol doesn't require parties to produce attachments separately from transmitting messages. "Parent-child relationships" are preserved because child attachments stay with the parent message. That's radical considering the "process everything" approach much in vogue; but it's an elegant way to trim gigabyte volume, lower cost and support both old and new approaches to de-duplication and review.

Finally, we tackled "the metadata." Demanding "the metadata" in discovery is like demanding "the information about information." Only the requesting parties know what they mean (and sometimes, they only know they're supposed to get something called "the metadata").

Every electronically stored file has some metadata, and some have lots of metadata. Application metadata (such as tracked changes and collaborative commentary) is part of the file and moves with it. System metadata (such as file name, dates, location and custodian of the file) is stored apart from the file and must be collected. Different ESI has different metadata, e.g., e-mail messages have "to" and "from" metadata fields that aren't present in text documents, spreadsheets or images.

A key advantage of native production is that it spares parties the burden of selecting and collecting from among hundreds of application metadata fields. Each file produced natively carries its peculiar complement of application metadata.

System metadata, essential for classifying and sorting large volumes of ESI, would be produced in load files. We settled on the following fields parties were required to furnish for all ESI:

- Identifier: The unique production identifier (UPI) of the item.
- Source Name: The original name of the item or file when collected from the source custodian or system.
- MD5 Hash: The MD5 hash value of the item as produced.
- Custodian: The name of the custodian or source system from which the item was collected.
- Source Path: The fully qualified file path from the root of the location from which

the item was collected.

- Production Path: The file path to the item from the root of the production media.
- Modified Date: The last modified date of the item when collected from the source custodian or system.
- Modified Time: The last modified time of the item when collected from the source custodian or system.
- UTC Offset: The coordinated universal time/Greenwich Mean Time offset of the item’s modified date and time.

Additional fields were required to accompany production of e-mail messages, including “to,” “from,” “cc,” “bcc,” “date sent,” “time sent,” “subject,” “date received,” and “time received.”

Also required for images of paper documents:

- Beginning Identifier: The beginning unique production identifier for the first page of the document.
- Ending Identifier: The ending unique production identifier for the first page of the document.
- Page Count: The total number of pages in the document.
- Location: The source box or other location identifier needed to trace the document to its source.

Some might feel we settled on too few or too many fields, but virtually identical fields have lately been characterized by Judge Shira Scheindlin as “the minimum fields of metadata that should accompany any production of a significant collection of ESI” in *National Day Laborer Organizing Network v. United States Immigration and Customs Enforcement Agency*, 2011 WL 381625 (S.D.N.Y. Feb. 7, 2011).

Antiforensics

by Craig Ball

[Originally published in Law Technology News, April 2011]

Antiforensics isn't a word in most vocabularies; yet, for something so unfamiliar, it's distressing how many become antiforensic aficionados when they know their data will be scrutinized. Antiforensics describes efforts to frustrate the tools and methods of computer forensics, encompassing deliberate efforts to hide data, destroy or alter artifacts and cast doubt on forensic examination.

While the delete key and "double deletion" (emptying the Deleted Items folder or Recycle Bin) is the humblest form of antiforensics, data custodians seduced by promises of privacy protection may turn to free and low-cost "cleaning" programs.

Nothing serves to deflect a case from its merits faster than proof that a party has intentionally destroyed or altered evidence. It's alleged and proven with disturbing frequency. Persons who would never shred paper evidence have no qualms about running an evidence elimination tool before the data collectors arrive.

Some target relevant evidence for destruction; but, digital pornography, electronic mail, chat, Internet gambling and even online shopping have made it common for employees to compromise themselves in ways marked by a myriad of electronic footprints. Covering those tracks can seem like the only way to avoid humiliation or termination.

Employers face huge risks when rogue employees destroy data. If an employee unwittingly destroys discoverable data while intentionally destroying irrelevant data, courts are unlikely to afford the bad actor (or his employer) the benefit of the doubt. Antiforensics tools are indiscriminate in their swath of destruction, making it difficult to distinguish antiforensics deployed to protect personal privacy from that used to destroy relevant evidence

Accordingly, counsel must be vigilant to guard against, and prepared to respond to, instances of antiforensics. Never be so naïve as to think your clients wouldn't do "that sort of thing." Almost everyone has something to hide.

Be clear and strong about the need to refrain from the use of such tools. Assure clients that usage of cleaning and wiping tools will be found out, and help them appreciate that nothing they erase could be as bad as what the judge and jury will imagine to fill the void. The cover up is always worse than the crime.

In a recent case, one side learned the other was trying to persuade the IT guy to "fix" the server so as to prevent the lawyers from collecting information for discovery. After a trip to court for a restraining order, the warring parties brought me in as an independent examiner to peruse terabytes of data and determine whether there had been an intentional effort to destroy evidence.

Fortunately for computer forensic examiners, people who resort to spoliation to get a leg up aren't the sharpest knives in the drawer. I figured something was up when the newest icon on the office manager's system desktop was for a program called Crap Cleaner (henceforth "CCleaner," to preserve delicate sensibilities). The investigation that followed illustrates some ways examiners make the leap from a suspicion of antiforensic behavior to establishing intentional spoliation.

CCleaner is a tool designed to obliterate much of the data relied upon by forensic examiners, including artifacts like prefetch data, link files and user assist histories unknown to most IT specialists. CCleaner's presence signaled that someone had obtained and installed a tool built to make data go away. But, so what? There's nothing evil about a computer user seeking to protect privacy. That is, nothing evil absent awareness of a duty to preserve ESI.

Timing is everything, and I needed to reliably determine when the user learned that his machine would be examined, when he installed CCleaner, what it was configured to do and when it was run.

But for CCleaner's scrubbing, I could have traced the source of the download and tracked the user's browsing behavior click-by-click; but, CCleaner is pretty effective at wiping Internet tracks. This data was gone along with a host of other useful information. I considered the artifacts and metadata still available to me. The desktop shortcut held information about the executable file to which it points. The computer's file system stores metadata about both. The version of CCleaner I saw is installed by a program called ccsetup236.exe, which I found in the user's My Downloads folder. Like most such applications, CCleaner is better at eradicating data than at covering its tracks. The computer still held metadata about the installer file. In turn, the installer created a program folder with its own complement of metadata. With all this temporal information, I could precisely establish when the program was obtained and installed.

Armed with e-mail recovered from the user's machine and the Exchange server, I determined that the user's acquisition and installation of CCleaner followed notice of the inspection, including a written reply promising to not to delete any data.

By default, CCleaner is configured to do a basic-but-thorough data cleaning, purging Internet history, cache, cookies and other records of online activity. But, users have the option of initiating a broader data demolition by manually tweaking the program's settings. Should a user claim not to understand the destructive purpose of the tool, it's useful to show he manually changed the settings to maximize the spoliation.

Windows stores configuration data for systems, users and software in files called hives that comprise a database called the Registry. I located CCleaner's configuration data in the NTUSER.dat Registry hive for the user. Obligingly, the data was in plain English, with active features labeled as "True" and those enabled and subsequently disabled as "False." This copy of CCleaner had been manually configured to undertake an especially broad range of antiforensic tasks.

I wondered why one of the most aggressive antiforensic features—wiping free space—had been first enabled and then disabled by the user. A sudden attack of conscience, perhaps?

I used the installer file recovered from the user's system to create an exemplar installation for testing. I could now document the default configuration and each change in the Registry as I configured the program, as well as capture the screens and warnings the user would have seen. When I enabled the option to wipe free space, a warning popped up stating that doing so significantly increases the amount of time for cleaning, recommending users leave the feature disabled. Perhaps it was conscience; but as likely, impatience. The user had a plane to catch.

I'd established the user downloaded, installed and configured the tool. Now, I had to prove the user ran the program. That data customarily seen was absent certainly pointed to use of CCleaner; but, I wanted something tending to show whether the user clicked the "Run Cleaner" button.

For that, I turned to prefetch data that Microsoft Windows uses to optimize the performance of frequently used programs. Prefetch chronicles the date and time of a program's first and last execution. Fortunately, CCleaner didn't clean out the recent prefetch information, so I was able to see when CCleaner was run—twice, in fact--and reconfirm date and time of installation from the installer prefetch.

This investigation entailed a lot more work and evidence than this glimpse into the world of computer forensics allows. The moral of the story is that antiforensics is counterproductive and commonplace. In e-discovery, anticipate human frailty, and head it off at the pass.

Double Exposure

by Craig Ball

[Originally published in Law Technology News, June 2011]

A forward-thinking plaintiffs' lawyer posed this courageous question: "Yesterday, a new client in a sexual harassment case brought in seven different electronic platforms. We routinely image our clients' hard drives for preservation, and it can be a substantial expense to our clients to send this work out. I realize that it is technically quite different than simply making a photocopy, but, at the end of the day, it is copying. My question is: Why do I have to outsource imaging to a vendor?"

I replied: "If the person doing the work does it capably, documents a reasonable chain of custody, and verifies the image by hashing, I see no reason why you would need to outsource forensic imaging for preservation. When all goes well, it's a simple task. In those rare instances where it doesn't, you bring in an expert. Forensic analysis is a wholly different situation; but single drive imaging is (and should be) a ministerial task when performed by a reasonably competent person in a sensible way."

The chief objection to drive-imaging by law firms is the specter that firms risk conflict and disqualification by becoming witnesses in their cases. For the most part, that trope exemplifies the cocktail of fear and ignorance regularly served up to protect vendors' turf and sustain the high price of electronic data discovery (EDD).

Forensically imaging a drive isn't comparable to a lawyer's role in drafting instruments or advising clients. There is *less* discretion or judgment exercised in the routine imaging of electronic media for preservation than in photocopying paper documents (where you might have to decide whether to copy both sides of a page or how to account for foldering or unitization).

How many times have you run across absent or mangled pages in a photocopy job? It happens all the time, yet we don't outsource all photocopying because we might have to testify about it. Lawyers make photocopies without fear because everyone understands the process. With modern drive-imaging tools, you'll push fewer buttons forensically imaging a drive than setting up a photocopy job. Plus, electronic authentication by hashing assures you've copied everything faithfully.

The genesis of the advocate-as-witness bogeyman is Rule 3.7 of the American Bar Association's Model Rules of Professional Conduct barring lawyers from acting as advocates in matters in which they are likely to be witnesses, "except where the testimony relates to an uncontested issue." The comments to the rule make clear that the term "uncontested issue" is merely redundant of the predecessor language of Disciplinary Rule DR 5-101(B)(2), which allowed an advocate to testify, "if the testimony will relate solely to a matter of formality and there is no reason to believe that substantial evidence will be offered in opposition to the testimony."

Drive imaging is a matter of formality. Imaging for preservation in e-discovery is principally prophylactic, and rarely prompts a challenge to chain of custody or data integrity. Most drives imaged for preservation aren't processed for discovery.

"But imaging drives is *technical*," reluctant lawyers reply. "It requires *expertise*."

The same could be said of photocopying in the '60s, faxing in the '70s, word processing in the '80s and internet use in the '90s. Anytime a lawyer plays a role in collecting, preserving, or producing evidence, someone may claim the lawyer's conduct is more than mere formality and seek to engineer disqualification. But is it *likely* to occur? Is there reason to believe that *substantial* evidence will be offered in opposition?

Drive imaging is a mechanical, ministerial task that affords no opportunity to alter the evidence when basic protocols are observed. If a law firm doesn't risk disqualification using its photocopier, then it has little to fear from routine drive imaging by competent personnel.

Because the goal of imaging in-house is to keep costs down, it makes sense only if in-house imaging truly costs less than outsourcing. The cost to outsource entails more than just the contractor's bill. Locating, vetting, and engaging a service provider is time-consuming. Delaying preservation or inconveniencing a client can be costly, too. Going in-house affords greater control over workflow, quality assurance, and confidentiality.

For firms that image 10 or more drives a year, chances are it will prove less costly to use in-house IT or litigation support for routine drive imaging than sending out the work. This projection factors in the cost of tools and training but necessarily assumes that in-house personnel cost less than contractors and have the time to do the work. It further assumes that the cost of in-house imaging can be passed on to clients in more or less the same manner as billings of a service provider. That's an issue best addressed in the firm's retention agreement.

As I described in "Do-It-Yourself Forensics" (*Law Technology News*, July '07), you can use a PC and free software to forensically image drives for the cost of the target drive and your time, but with greater complexity and a slower pace.

For ease of use and speed, I recommend investing in a dedicated forensic duplication device that *images* drives rather than *clones* them. Imaging stores the contents of the evidence drive in one or more files, while cloning creates an operational duplicate of the source drive. Imaging is preferable to cloning because attaching a target drive holding images to a Windows machine won't change the evidence, where connecting a clone will. A downside to imaging is that, if the tool employed doesn't support data compression, you must image to a drive larger than the source (i.e., one terabyte drives must be imaged to 1.5 terabyte or larger targets). The additional space is needed to store data documenting the acquisition. It's a tiny file, but it's just big enough to prevent the source data from being stored on a like-sized drive.

A crucial constraint to forensically sound drive imaging is that the evidence drive be protected from alteration by write-blocking techniques. Devices designed for forensic drive imaging should integrate write protection of the source and incorporate fail-safe features to guard against inadvertent alteration or destruction of the source evidence.

The persons tasked with imaging should be schooled in evidence handling, storage, observation, and documentation. They'll need a reasonably secure workspace to protect the evidence from curious or light-fingered passersby, a digital camera to document the evidence, and an ample supply of hard drives.

A label printer and sets of jeweler's (Stanley 66-039) and tamper-proof screwdrivers (Boxer TP-30) come in handy. A locking evidence cabinet, closet, or safe is a must.

I see a market for a one- or two-day course in electronically stored information preservation where IT and litigation support personnel learn to image, authenticate, and document a forensically sound acquisition of commonly encountered storage media. It's an entrepreneurial opportunity for tool vendors and forensics service providers and a path to new job skills and security for attendees.

For in-house imaging to be cost-effective, you've got to keep it simple. Those tasked with imaging drives must recognize when they're out of their depth and call in a pro. For malfunctioning drives, unfamiliar interfaces (SCSI, SAS, ZIF), server RAID arrays, SAN and NAS devices, handhelds and smart phones, encrypted media, and jobs demanding delicate disassembly, the risks rise and the cost savings evaporate.

Having more tools, a professional can speed acquisition through parallel processing; that is, imaging multiple drives at the same time. Parallel processing significantly lowers the cost of on-site acquisition (where the technician charges for every minute), but has only a modest impact on cost for lab acquisitions (where hourly charges are only assessed for time expended setting up, monitoring, and documenting the acquisition, not for the idle time of data transfer).

For firms preferring to stick with service providers because they feel they realize greater value for the higher cost or because they don't expect to save enough to warrant the in-house effort, forensic examiners are happy to get your call. But for firms that want to bring forensic imaging capabilities in-house, you're not risking conflict and disqualification. Go for it!

HARDWARE OPTIONS

Intelligent Computer Solutions: Image MASSter Solo-4 ($2,900)
Voom Technologies: HardCopy 3P ($1,410)
Guidance Software: Tableau TD1 ($1,249)
Logicube: Forensic Quest-2 ($999)

Hey! You! Get Onto my Cloud

by Craig Ball

[Originally published in Law Technology News, August 2011]

“This cloud thing is really just the internet, right?”

Lawyers want reassurance the “cloud” is something they can ignore. No such luck. In fact, the cloud is re-inventing electronic data discovery in marvelous new ways while most still grapple with the old.

Hey! You! Get onto my cloud.

The cloud will make e-discovery easier and cheaper while improving the quality and efficiency of preservation, search, review and production. So what, exactly, is this amazing cloud?

It’s a buzzword for three on-demand service models delivered via a network (i.e., the internet). The name comes from depicting networks as a cloud on system schematics. There’s so much cloud hype in the market that the three service models warrant brief explanation.

The first, most familiar cloud model is called **Software as a Service** (SaaS). It’s doing things you once did with installed applications using your web browser. When you search Lexis, Tweet, post on Facebook, check webmail or keep a Google calendar, you’re using someone else’s software on someone else’s machine. It feels like desktop computing, but software stays up-to-date without installing updates, and data is backed up more regularly and reliably than if you did it yourself. As for price, you can’t match the service and features for less. SaaS answers the question, “Can I use your program?”

The second model is **Platform as a Service** (PaaS). Here, you rent virtual machines as needed to run your own software. Deploy as many VMs as you need and pay as you go. Such rapid, on-demand elasticity positions PaaS to revolutionize e-discovery. PaaS answers the question, “Can I run my program on your machine?”

The third model is **Infrastructure as a Service** (IaaS). Here, you get the ability to provision and configure remote machines, including choice of operating systems. IaaS answers the question, “Can I run your machine?”

PaaS and IaaS shift the cost of computer rooms (servers, air conditioning, power, space) to trusted names like Amazon.com (AWS/EC2), Microsoft (Windows Azure) and Google (App Engine). Backup becomes their problem.

The cloud changes the forms of electronically stored information, the places we store it and the tools and methods used to preserve, search, process and produce it.

As data moves to the cloud from the existing patchwork of local servers, devices and media now serving as primary storage, there are fewer sources to search and preserve. Bringing ESI beneath one big tent narrows the gap between retention policy and practice and fosters compatible forms of ESI across web-enabled applications. Moving ESI to the cloud also spells an end to computer forensics. Multi-tenant architecture means there are no hard drives to image or artifacts to examine. “Deleted” finally means “gone.” The most exciting changes heralded by the cloud are the elimination of collection for preservation and the availability of sophisticated, scalable search and review tools.

Today, we collect ESI. But, if the evil twins of e-discovery are daunting volume and runaway replication, why are we creating *still more* copies through collection?

We do it to minimize the spoliation risk of leaving ESI in users’ custody, aggregate data from multiple sources and transfer ESI to the systems and service providers that process it.

Cloud computing makes collection unnecessary. Because users have no physical dominion over storage devices and only such control of data as their cloud access privileges allow, users can’t destroy or alter cloud data put on hold. Suddenly, data can be preserved *in situ* with little risk. Why consolidate data from multiple devices and sources? All that’s needed are pointers to the data; often just to a single instance for identical items.

We won’t bring data to tools when it’s faster and cheaper to bring tools to data. Instead of repatriating ESI by sipping it through an internet straw, data and processing tools will sit on the same fat pipe. Processing will be faster and cheaper because you’ll deploy all the computing power required at the business end of the pipe, paying only for what’s required for as long as needed.

Here’s how it works: Imagine the cloud as a water tower. Over time, it fills to a great volume, and the size of the pipe to our homes limits what we can draw. The flow’s fine for a shower, but not enough to fight a fire.

To beat back a blaze, you need big pipes. In the cloud, the “big pipes” are the fiber optic backbones linking servers and data centers. To tap these mains, you bring tools to data, not data to tools. That means deploying processing and search applications alongside data in the cloud. Now, the tools you’ll use won’t run on your machines or be hosted on machines owned by EDD service providers. You won’t know where the machines are, and, for the most part, you won’t care.

E-discovery processing entails a lot of number crunching. Whether extracting text, running searches, parsing messages, cracking passwords or generating tiffs, machines handle the computational heavy lifting. Speed is a function of bandwidth and computing horsepower. That is, fat pipes and fast processors.

One way to speed processing is to break big tasks into smaller tasks parceled out to many machines. The technical term for this is “distributed computing.” I call it, “going wide.”

Remember that scene in *The Matrix* where Neo and Trinity arm themselves from gun racks that appear out of nowhere? That's what it's like to go wide in the cloud. Cloud computing makes it possible to conjure up hundreds of virtual machines and make short work of complex computing tasks. Need a supercomputer-like array of VMs for a day? *No problem*. When the grunt work's done, those VMs pop like soap bubbles, and usage fees cease. There's no capital expenditure, no amortization, no idle capacity. Want to try the latest concept search tool? There's nothing to buy! Just throw the tool up on a VM and point it at the data.

Realizing the promise of the cloud in e-discovery awaits tool sellers and service providers re-engineering their wares for quick and economical deployment in the cloud. Applications must support distributed computing, and licensing models must adapt to short-term deployments and going wide. Customers won't buy a thousand annual licenses when all they need are a thousand virtual iterations for an hour. Only nimble vendors who adapt to the demands of EDD in the cloud will be left standing.

Despite clients mothballing servers and migrating systems into the cloud at breakneck speed, some lawyers dismiss the cloud on security and data privacy grounds. "All it will take is one high-profile breach, and companies will run back to local storage and on-premises data centers." Don't bet on it. Data availability (uptime) and information security are vital; but neither local storage nor on-premises data centers have proved immune to failure and breach. When risks and capabilities are comparable, cost is king. The cloud wins on cost, hands down.

That e-discovery will live primarily in the cloud isn't a question of *whether* but *when*. Considering how much discoverable data already resides in the cloud as webmail, social networking, voicemail, banking and brokerage data, *when* is starting to look a whole lot like *now*.

Hey! You! Get onto my cloud.

Data Mapping
by Craig Ball

[Originally published on the Ball in Your Court blog, September 23, 2011]

Data mapping is one of those nimble e-discovery buzz words–like ECA and Predictive Coding–that take on any meaning the fertile minds in the Marketing Department care to ascribe.

I use "data mapping" to encompass methods used to memorialize the **identification** of ESI–an essential prerequisite to everything in the EDRM east of Information Management. Of course, like Nessie and Bigfoot, Information Management is something many believe exists but no one has ever shown to be anything but a myth. Consequently, identification of ESI, *viz*. data mapping, is the *de facto* entry point for all things e-discovery.

Data mapping is an unfortunate moniker because it suggests the need to generate a graphical representation of ESI sources, leading many to assume a data map is synonymous with those Visio-style network diagrams IT departments use to depict, *inter alia*, hardware deployments and IP addresses.

Unless created expressly for e-discovery, few companies have any diagram approaching what's required to serve as an EDD data map. Neither network diagrams from IT nor retention schedules from Records and Information Management are alone sufficient to serve as an EDD data map, but they contribute valuable information; clues, if you will, to where the ESI resides.

Thus, a data "map" isn't often a map or diagram, though both are useful ways to organize the information. A data map is likely a list, table, spreadsheet or database. I tend to use Excel spreadsheets because it's easier to run totals. A data map can also be a narrative. John Collins, a J.D. with The Ingersoll Firm in Illinois recently shared a data map in the form of a 64-page narrative report describing an enterprise e-mail environment in exacting detail. Whatever the form employed, your client doesn't have a data map lying around somewhere. It's got to be built, usually from scratch.

What your data map looks like matters less than the information it contains. Again, don't let the notion of a "map" mislead. The data map is as much about *what* as *where*. If the form chosen enables you to quickly and clearly access the information needed to implement

Identification
Preservation
Review
Production

defensible preservation, reliably project burden and accurately answer questions at meet-and-confer and in court, then it's the right form even if it isn't a pretty picture.

Scope

The duty to identify ESI is the most encompassing obligation in e-discovery. Think about it: You can't act to *preserve* sources you *haven't found.* You certainly can't collect, review or produce them. The Federal Rules of Civil Procedure expressly impose a duty to identify all potentially responsive sources of information deemed "not reasonably accessible." So even if you won't search potentially responsive ESI, you're bound to identify it.

A "data map" might be better termed an "Information Inventory." It's very much like the inventories that retail merchants undertake to know what's on their shelves by description, quantity, location and value.

Creating a competent data map is also akin to compiling a history of:

- Human resources and careers (after all, cases are still mostly about people);
- Information systems and their evolution; and
- Projects, facilities and tools.

A data map spans both logical and physical sources of information. Bob's e-mail is a logical collection that may span multiple physical media. Bob's hard drive is a physical collection that may hold multiple logical sources. Logical and physical sources may overlap, but they are rarely exactly the same thing.

As needed, a data map might encompass:

1. **Custodian and/or source of information;**
2. **Location;**
3. **Physical device or medium;**
4. **Currency of contents**;
5. **Volume** (e.g., in bytes);
6. **Numerosity** (e.g., how many messages and attachments?)
7. **Time span** (including intervals and significant gaps)
8. **Purpose** (How is the ESI resource tasked?);
9. **Usage** (Who uses the resource and when?);
10. **Form;** and
11. **Fragility** (What are the risks it may go away?).

This isn’t an exhaustive list because the information implicated changes with the nature of the sources being inventoried. That is, you map different data for e-mail than for databases.

A data map isn’t a mindless exercise in minutiae. The level of detail is tailored to the likely relevance and materiality of the information.

Tips for Better Data Mapping

- Custodial interviews are an essential component of a sound data map methodology; but, custodial interviews are an unreliable (and occasionally even counterproductive) facet of data mapping. Custodians will know a lot about their data that will be hard to ferret out except by questioning them. Custodians will not know (or will misstate) a lot about their data that must be supplemented (or corrected) objectively, though, e.g., search or sampling.
- Do not become so wedded to a checklist when conducting custodial interviews that you fail to listen to the subject or use common sense. When a custodian claims they have no thumb drives or web mail accounts, don’t just move on. *It’s just not so*. When a custodian claims they’ve never used a home computer for work, don’t believe it without eliciting a reason to trust their statement. Remember: custodians want you *out of their stuff and out of their hair*. Even those acting in complete good faith will say what promotes that end. Trust, but verify.
- Don’t be so intent on minimizing sources that you foster reticence. If you really want to find ESI, use open-ended language that elicits candor. ” Avoid leading questions. *You didn’t take any confidential company data home, did you*?” isn’t likely to stir a reply of “*Sure, I did!*” Offer an incentive to disclose (“*It would really help us if you had your e-mail from 2009*“).
- Legacy hardware grows invisible, even when it’s right in front of you. A custodian can’t see the old CPU in the corner. The IT guy can’t see the box under his desk filled with backup tapes. You must bring a fresh set of eyes to the effort, and can’t be reluctant to say, “What’s in there?” or “Let me see please.” Don’t be blind leading the blind.
- Companies don’t just buy costly systems and software and expense it. They have to amortize the cost over time and maintain amortization and depreciation schedules. Accordingly, the accounting department’s records can be a ready means to identify systems, mobile devices and even pricey software applications that are all paths to ESI sources.

Three Pressing Points to Ponder

If you take nothing else away from this, please consider these three closing comments:

1. **Accountability is key every step of the way**. If someone says, “that’s gone, ” be sure to note who made the representation and test its accuracy. Get their skin in the game. Ultimately, building the data map needs to be one person’s hands-on, buck-stops-here responsibility, and that person needs to give a hot damn about the quality of their work. Make it a boots-on-the-ground duty devolving on someone with the ***ability, curiosity, diligence and access*** to get the job done.
2. **Where you start matters less than when and with whom**. Don’t dither! Dive in the deep end! Go right to the über key custodians and start digging. Get eyes on offices, storerooms, closets, servers and C: drives, and go where the evidence leads.
3. **Just because your data map can’t be perfect doesn’t mean it can’t be great.** Don’t fall into the trap of thinking that, because no data mapping effort can be truly complete and current, the quality of the data map doesn’t matter. Effective data mapping is the bedrock on which any sound e-discovery effort is built.

The Shadow Knows

by Craig Ball

[Originally published on the Ball in Your Court blog, September 24, 2011]

“You can get anything back from a computer, can’t you? Even the deleted stuff!”

I get that that a lot, and tend to respond, “Pretty much.” My lawyer side wants to add, “but it depends.” Like most in computer forensics, I tend to downplay the challenges and uncertainties of data recovery, not so much to promote forensic examination as to discourage data destruction. Until a forensic examiner processes the evidence, it’s hard to say whether we can recover particular deleted data; but dollars-to-diamonds, a forensic exam will shed light on the parties and issues.

Lately, the likelihood of recovering deleted files on late-model Windows systems has gone way, way up, even if the data’s been thoroughly flushed from the Recycle Bin. Microsoft has been gradually integrating a feature called Volume Snapshot Service (a/k/a Volume Shadow Copy Service) into Windows since version XP; but until the advent of Windows 7, you couldn’t truly say the implementation was so refined and entrenched as to permit the recovery of anything a user deletes from a remarkable cache of data called **Volume Shadow Copies**.

Volume shadow copies are old news to my digital forensics colleagues, but I suspect they are largely unknown to the e-discovery community. Though a boon to forensics, volume shadow copies may prove a headache in e-discovery because their contents represent reasonably accessible ESI; that is, much more potentially probative evidence that you can’t simply ignore. So, for heaven’s sake, *don’t tell anybody*. 😉

I recently returned from presenting my Computer Forensics Jeopardy program at the HTCIA’s annual meeting and training conference in Palm Springs, CA. The HTCIA is the High Technology Crime Investigation Association, boasting the largest membership of computer forensic examiners. The training options offered are always pretty good, but the best reason to attend is the exchange of information that occurs in the bars, lounges and pool areas. The chance to let your hair down with colleagues and share new ways to get to the digital evidence is invaluable. If I come away with a nugget or two, it’s worth the time and travel.

This year some of my best nuggets came from Brandon Fannon and Scott Moulton. Scott was my instructor when I traveled to Atlanta several years ago to study extreme data recovery techniques like hard drive head replacements and platter swaps. Scott is a pioneer in making mere mortals privy to the deepest, darkest secrets of data resurrection.

Sitting around a fire pit on a beautiful desert night afforded me the chance to pick their brains on preferred approaches to processing volume shadow copies in forensic exams. Accessing shadow copy data is easy on a live machine–the user can roll back in a few clicks using the Previous Versions feature–but it's harder for an examiner working from a static image of the drive. This is an introductory treatment of the topic, so I'll leave discussion of those emerging techniques to a later post.

What you need to know now is that much of what you might believe about file deletion, wiping and even encryption goes out the window when a system runs any version of Windows 7 or Vista Business, Enterprise or Ultimate editions. Volume Shadow Copies keep *everything,* and Windows keeps up to 64 volume shadow copies, each made at (roughly) one week intervals for Windows 7 or daily for Windows Vista. ***These aren't just system restore points: volume shadow copies hold user work product, too.*** The frequency of shadow copy creation varies based upon multiple factors, including whether the machine is running on A/C power, CPU demand, user activity, volume of data needing to be replicated and changes to system files. So, the 64 "weekly" shadow volumes could represent anywhere from two weeks to two years of indelible data.

How indelible? Consider this: most applications that seek to permanently delete data at the file level do it by deleting the file then overwriting its storage clusters, which still hold the file but which have been tagged as unallocated clusters as a consequence of the deletion. These are called "unallocated clusters," because they are no longer allocated to storage of a file within the Windows file system and are available for reuse. But, the Volume Shadow Copy Service (VSS) monitors *both* the contents of unallocated clusters and any subsequent efforts to overwrite them. Before unallocated clusters are overwritten, VSS swoops in and rescues the contents of those clusters like Spiderman saving Mary Jane.

These rescued clusters (a/k/a "blocks") are stored in the next created volume shadow copy on a space available basis. Thus, each volume shadow copy holds only the *changes* made between shadow volume creation; that is, it records only *differences* in the volumes on a block basis in much the same way that incremental backup tapes record only changes between backups, not entire volumes. When a user accesses a previous version of a deleted or altered file, the operating systems instantly assembles all the differential blocks needed to turn back the clock. It's all just three clicks away:

1. Right click on file or folder for context menu;
2. Left click to choose "Restore Previous Versions;"
3. Left click to choose the date of the volume.

It's an amazing performance…and a daunting one for those seeking to make data disappear.

From the standpoint of e-discovery, responsive data that's just three mouse clicks away is likely to be deemed fair game for identification, preservation and production. Previous versions of files in shadow volumes are as easy to access as any other file. There's no substantial burden or collection cost for the user to access such data. But, as easy as it is, I expect few (if any) of the EDD collection tools or protocols have been configured to identify, grab or search the previous versions in volume shadow copies. It's just not a part of vendor work flows yet.

Ask your e-discovery service provider about it, and pray they reply, "*Don't fret about that. We haven't run into any Vista or Win 7 machines in your cases, but we'll come up with something before we do*." What's the hurry, after all? Win7 and Vista only run nearly half of all computers in the world!

If your vendor or expert says, "volume shady whatzit?" avoid eye contact, back away slowly, then run like hell!

Doubtlessly, there will be situations where identification, collection and search of previous versions of responsive documents in VSS is excessive in scope and disproportionate to the case; but, there are plenty of instances where deleted documents or prior versions are relevant and material–where these volume shadow copy versions are the smoking guns. Right now, only forensic examiners acknowledge them, and not universally. The question we in e-discovery face as we increasingly process machines running operating systems with the volume shadow snapshot service is this: ***Can we risk pretending that evidence that's instantly available on three mouse clicks is not reasonably accessible?***

If I know about VSS, and now you know about VSS, how long until the other side gets wise?

A Changing Definition of Deletion

by Craig Ball

[Originally published on the Ball in Your Court blog, September 30, 2011]

They're talking about changing the federal e-discovery rules to lessen the fear and loathing attendant to preservation of ESI.

The unstated impetus is that federal judges can't be trusted to weigh preservation and mete out sanctions in ways fairly attuned to facts and culpability. The proposed amendments seek to wrest the gavels from cranky judges whose 20/20 hindsight and outsize expectations operate to impose an impossible, perilous standard nationwide. Or so goes the rhetoric.

It's a crock. We give federal judges a job for *life,* but can't trust them to do that job wisely and well?!? Did we not learn *anything* from the debacle of mandatory sentencing guidelines?

The proposed changes are driven by the second silent goal of sparing litigants (really their technologically challenged counsel) the chore of knowing enough about electronic evidence and information technology to make defensible decisions about preservation. "Don't make us learn anything," they plead, "just make rules specific enough to protect us from not knowing." The rub with grafting such specificity onto e-discovery is that information technology moves far more swiftly than rule making, such that amendments like those proposed principally benefit those who can't or won't keep up.

Case in point: One proposed amendment would create a presumption that certain data is excluded from the preservation duty, to wit:

(A) Deleted, slack, fragmented or unallocated data on hard drives;
(B) Random access memory (RAM) or other ephemeral data;
(C) On-line access data such as temporary internet files;
(D) Data in metadata fields that are frequently updated, such as last opened dates;
(E) Information whose retrieval cannot be accomplished without substantial additional programming, or without transferring it into another form before search and retrieval can be achieved;
(F) Backup data that substantially duplicate *[sic]* more accessible data available elsewhere;
(G) Physically damaged media;
(H) Legacy data remaining from obsolete systems that is unintelligible on successor systems [and otherwise inaccessible to the person]; or
(I) Other forms of electronically stored information that require extraordinary affirmative measures not utilized in the ordinary course of business.

Starting with the word, “deleted,” it’s clear that this list is driven by an outdated understanding of information systems. ”Deleted” in 2011 bears only a passing resemblance to “deleted” circa 2001. A decade ago, recovering deleted data entailed expense, expertise and effort. Back then, you needed someone like me–a forensic examiner–to resurrect deleted data with specialized tools and techniques, or an IT specialist to laboriously restore backup media.

Today, operating systems, like Vista and Windows 7, preserve data even after it’s been deleted, stomped on and ignominiously ejected from the Recycle Bin. Its restoration is now an instantaneous, no cost, three-mouse-click task. [For details, see the immediately preceding article "The Shadow Knows"]. Deleted never meant gone; but now, it doesn’t even mean difficult to get back.

A revised definition of “deleted” applies to e-mail, too. The last release of the ubiquitous enterprise e-mail application, Microsoft Exchange Server, builds the ability to recover double deleted data (i.e., messages purged from the user’s Deleted Items folder) right into the application, as the Recoverable Items feature.

More-and-more, information is only ‘deleted’ in the same sense we might say the world ‘disappears’ when you close your eyes. It’s all still right there–all you have to do is look.

There are other problems with the proposed amendment, born of loose language and an outdated perspective on ESI:

1. The proposed language gives a “get out of jail free” card to “fragmented” data on hard drives, apparently unaware that “fragmented data” is a term of art for storage of data in non-contiguous clusters. Much active, accessible,*responsive* data on hard drives is fragmented. It’s no more difficult to access fragmented data than de-fragmented data; users do it every day without knowing whether the file they’ve opened was fragmented or not. No doubt the Committee was trying to describe digital forensic *artifacts*accessible only through data recovery techniques. That’s one of the problems with getting very specific: You’ve got to get your terms right, and insure they mean what you intend.

2. Another outdated notion is the special dispensation made for loss of random access memory (RAM). There was a time when it was generally understood that RAM meant volatile, ephemeral storage. But today, thumb drives are RAM; and soon, all laptops will use RAM (SSDs or solid state drives) in place of mechanical hard drives, much as your iPhone and iPad use RAM for non-volatile storage. Ephemeral? Hardly!

3. Temporary internet files? Okay, I grant they are evidence in only a narrow range of cases; but such cache content isn’t all that “temporary,” nor is it less accessible than other files. But because it’s ESI we prefer didn’t exist, we pretend it’s different so we’re not obliged to fashion a more rigorous rationale for

why we ignore it. As more and more evidence resides in the Cloud and on social networking sites, is it sensible to summarily dismiss records of online activity as spurious and inconsequential?

4. When you exclude "metadata in fields that are frequently updated," how does that foster predictability and defensibility? Referencing "last opened dates" is little help; not only because *no such metadata field by that name* exists (they probably meant the last *accessed* date), but also because the currency and function of the last accessed date changed significantly with Windows Vista. The last accessed date doesn't signify the same events it tracked ten years ago. Here again, ***precision matters***: *if you're going to make rules for information technology, then use the language information technologists use*.

5. How frequently must a metadata value be updated in order for it to be okay to discard? Does the rule apply with equal force to application metadata (which is embedded in a file and may record communications between collaborators) as it does to system metadata (which is all stored outside the file)? Should the rapidity of change trump the relevancy and reliability of the information? Shouldn't the feasibility and ease of preservation matter most?

6. The disconnect between the intent and the expression of the amendment comes into sharp relief in the exception carved out for, "Information whose retrieval cannot be accomplished without…transferring it into another form before search and retrieval can be achieved." That's <u>all</u> ESI. How do we access ESI without converting physical to logical and encoded to decoded content? That's what computers *do*. How do we text search efficiently without the creation of indices? How do we retrieve without transfer between media? How do we review without converting bytes into pixels or droplets? Yes, I'm almost getting metaphysical here; but if the goal is clarity and predictability, why employ language that is so susceptible to a multitude of interpretations…unless ambiguity is intentional?

7. Finally, we come to the general exclusion for legacy data and backup data. Once these sources are gone, it's easy (leastwise *irrefutable*) to claim they were merely cumulative. Too, it's easy to migrate between systems during the preservation interval, rendering data inaccessible because the proposed rule change doesn't impose a concomitant duty to maintain accessibility.

This would all be so disheartening were it not for the glimmer of hope inspired by a footnote to the list of ESI excepted from preservation. It reads, **"This specific listing is taken from submissions to the Advisory Committee. Besides asking whether it is sensible and complete, one might also ask whether a list this specific is likely to remain current for years."**

Years?!? Heck, it's not current *now*. As to sensible, it's barely comprehensible. But, the footnote gives me hope that the Committee is asking the right questions and this misbegotten mess won’t make it into the FRCP. I have good friends on the Committee working on these proposed amendments; whip smart, fair minded folks, to boot. So, I have reason to trust they’ll ultimately get it right by bringing as much caution to this round of ESI rulemaking as they brought to the last.

My take: We don’t need more specific ESI rules. We need to become competent implementing the good ones we’ve got.

Dancing in the Dark

by Craig Ball

[Originally published in Law Technology News, October 2011]

At a recent discovery conference, opposing counsel asked why I wanted ESI produced in native and near-native formats. Her question struck me as curious. She intended to work with native formats to perform her own review. But when it came time to produce the gigabytes of e-mail, productivity documents, spreadsheets, and database exports, she meant to convert everything to .tiff images. "Where electronically stored information is concerned," I thought, "lawyers make the Amish look wired."

Still, "why native?" was a fair question. Assuming she'd commit to capture certain metadata fields, expose tracked changes before imaging, and produce load files with extracted text, we could make .tiff images work for everything but spreadsheets and databases.

But it's just ... so ... ungainly. It seems pointless to strip out utility and then try to graft it back like a mangled limb on a ballerina.

I pondered what it is about native and near-native forms that makes them prime for production. It's not just that native production forestalls the need for conversion and lowers hosting and processing expense. Or that native files hold embedded data and metadata rich with meaning and serving as aids to authentication.

No, the principal attraction for me is that native production works — and, being at the top of the informational hierarchy, it's the most flexible format, i.e., you can convert native files to other forms, but usually not the reverse.

Thus counsel and I were locked in a *pas de deux*. I wanted functional evidence. She wanted Bates numbers on every page. I wanted the embedded communications. She wanted to be sure I couldn't see anything she'd overlooked. We danced around for a while, but I knew when the music stopped, I might be stuck holding a disk full of .tiffs, partly due to a crucial flaw in the federal rules governing forms of ESI production.

Back when most information changed hands on paper, forms of production hardly mattered. Single- or double-sided, canary or buff. Who really cared? The printed page carried the complete informational payload, so opponents with identical boxes of paper (or sets of images) were fairly matched for usability and content.

But that was long ago. Today, paper accounts for only a fraction of discoverable data. Instead, most information is recorded and communicated electronically, and discoverable data has outgrown the confines of the printed page. It's tweets, threads, searchable text beneath page images, formulae in spreadsheets, animations in digital presentations and collaborative environments. So, the form of production governs whether evidence is functional, searchable, and complete.

The drafters of the 2006 federal rules amendments recognized that there would be fewer battles if lawyers understood the forms of ESI in their cases and agreed upon the forms in which ESI changes hands. Their goal was to insure litigants resolved forms disputes well before production, obviating costly do-overs. Opponents who couldn't work out forms disputes were expected to get them in front of the court quickly.

The Federal Rules of Civil Procedure lay out five steps in the forms of production cha-cha-cha:

Step One: Before the first pretrial conference, parties are obliged to hash out issues related to "the form or forms in which [ESI] should be produced." Fed. R. Civ. P. 26(f)(3)(C). Some lawyers still leave meet and confer sessions content to have selected CDs as the form of production. That should be funny because optical disks are production media, not production forms. If you're not laughing, you may be one of those lawyers.

Step Two: A requesting party specifies the form or forms of production for each type of ESI sought. Paper aside, these break down to native, near-native, imaged formats or a mix of same. Fed. R. Civ. P. 34(b)(1)(C).

Step Three: If the responding party is content to deliver what the requesting party seeks, the forms dance is done. But if the specified forms aren't acceptable, the responding party must object and designate the forms in which it intends to make production.

Even if the requesting party fails to specify the form or forms sought, the responding party must state the form or forms it intends to use. Fed. R. Civ. P. 34(b)(2)(D). Might this be the most overlooked rule in the book?

Step Four: If the requesting party won't accept the forms the producing party's designates, the requesting party must confer with the producing party in an effort to resolve the dispute. Fed. R. Civ. P. 37(a)(1).

Step Five: If the parties can't work it out, the requesting party files a motion to compel, and the court selects the forms to be produced, unconstrained by the choice specified by either side.

But you can't dance if you don't feel the beat, and the ESI rules forgot the beat. That is, they set no discrete deadline for the producing party to object to the requested forms or identify the forms that will be produced.

It's an omission responsible for costly false starts and frustration, and one that increases the risk I'll be stuck with .tiffs even though native forms are the better, cheaper choice.

The rules simply provide that the producing party answer the request in writing within 30 days. Consequently, the sensible choreography hoped for devolves into foot dragging. Objections get filed with responses — usually on the last day — and ESI is produced in forms the requesting party didn't seek and doesn't want.

By the time the dispute gets in front of the judge, the producing party howls that it already made production in one form and shouldn't have to produce another, pointing to the single form of production provision of Fed. R. Civ. P. 34(b)(2)(E)(3). They also argue it's unduly expensive and burdensome to start over in order to produce the same ESI in a different form.

The requesting party counters that the forms produced weren't the forms sought. The court demands to know why the forms produced aren't reasonably usable.

In the end, the requesting party's right to seek preferred forms of production gets short shrift — largely because no deadline requires the responding party to make objection and designate intended forms before proceeding with processing and production. The requesting party loses the ability to act before the die is cast.

The Notes to Rule 34(b) of the 2006 Rules amendments make clear that the Advisory Committee appreciated the risk: "A party that responds to a discovery request by simply producing electronically stored information in a form of its choice, without identifying that form in advance of the production…runs a risk that the requesting party can show that the produced form is not reasonably usable and that it is entitled to production of some or all of the information in an additional form."

The risk of additional production has proven insufficient to promote good practice. Five years on, disputes about forms of production are commonplace.

In retrospect, the wiser approach would have been to impose an early deadline to object and specify intended forms of production — perhaps within seven days of service. That is, join the issue before the lion's share of the cost is incurred.

Hopefully, a change like this finds its way into future rules amendments, but judges shouldn't wait. Through standing orders and local rules, courts can implement a deadline requiring responding parties to object to forms sought and designate the form or forms they intend to produce within seven days after service of the request. Fights over forms of production waste too much time and money. A sensible deadline will help.

Requesting parties, too, shouldn't wait until the response date to know if an opponent refuses to furnish the forms sought. Press for a commitment; and, if not forthcoming, move to compel ahead of the response date. Don't wait to hear, "Why didn't you raise this before they spent all that money?"

The Tyranny of the Outlier

by Craig Ball

[Originally published in Law Technology News, December 2011]

Have you noticed how corporate counsel are flexing their collective muscle in an effort to rein in e-discovery? Their rallying cry is that plaintiffs have begun to "weaponize" preservation. That is, plaintiffs are demanding preservation of ESI with such breadth that corporations are settling just to avoid the cost of finding and protecting their own discoverable data.

The weapon of mass (self) destruction is a preservation letter reminding defendants of the common law duty to preserve relevant evidence. It typically mixes sweeping "any and all" generalities with litanies of specifics that "include, but are not limited to" every form of newfangled storage since Edison recorded "Mary had a little lamb" on tinfoil. The similarity of these preservation demands--down to the typos--is no surprise. Like the larger bar, plaintiffs' lawyers dream of reducing e-discovery to a few forms and checklists. They are no more willing than their opponents to dig into and dirty their hands with data, so they trot out the same filched forms with little thought as to their suitability or scope. To my chagrin, some of the language is lifted straight from papers I wrote years ago that caution against such boilerplate preservation demands.

The fearsome power of the preservation letter hinges on the absurd notion, "If you remind me what the law requires, I might have to comply." In fact, demand letters forge no new duties. The preservation obligations the law imposes can't be broadened by an opponent's demand. At best, a preservation demand fixes the time at which the common law duty attaches and undercuts claims of innocent oblivion to sources of relevant ESI. Say goodbye to, "Gee whiz, Judge, it never crossed our minds they'd want us to save e-mail."

Corporate counsel vilify preservation: "The plaintiffs demand that we preserve everything, and we're spending millions doing so." If plaintiffs' settlement demands don't establish the value of their claims, *why should plaintiffs' preservation demands set the bar for preservation*?

They shouldn't. What defense counsel label as extortionate tactics to force settlement through disproportionate preservation are rarely tactical moves. More typically, plaintiffs' actions mirror the same fear and unsophistication that spur defendants to over-preserve. Uncertain where relevant evidence resides, plaintiffs demand preservation of every place it might be. Equally uncertain and irrationally afraid of the outlier jurist, defendants say "okay" when they should say "no way."

Cooperation in e-discovery doesn't mean bowing to your opponent's demands for over-preservation. Instead, cooperation entails communicating relevant, reliable and specific information about systems, sources and forms to enable the other side to make responsible preservation demands…*even if they won't do so*. It's vital to quickly convey what you *won't* preserve, ideally affording an opponent an opportunity to take the issue to the court before disputed sources are lost. Above all, it's most crucial to understand your client's or your company's data landscape well enough to make *defensible* choices. *Defensible*, not *unassailable*. The repetition of sensible, right-sized preservation decisions based on expertise and diligence will save far more than unwarranted deference to opinions out of step with the mainstream of ESI jurisprudence; what I call "the tyranny of the outlier."

Saints Preserve Us

Per John W. O'Tuel III, Assistant General Counsel for GlaxoSmithKline: "Cases are being settled, discontinued or not brought in the first place because the cost of preservation is too high, the risk of spoliation sanctions is too great, and the impact of ancillary litigation proceedings on discovery disputes is too debilitating." It's a sentiment echoed by many in-house counsel and outside lawyers representing big business, but are there metrics to support the claim?

There's no way to measure cases improvidently "settled, discontinued or not brought," but the solid metrics we have on the risk of spoliation sanctions prove that the sanctions risk for negligent non-preservation is miniscule (.00675% per a 2011 report from the Federal Judicial Center). You're more likely to be hit by lightning. As to Fortune 500 companies being "debilitated" by ancillary litigation on discovery disputes, the dockets don't bear out the hardship, and may I be so bold as to suggest that perhaps the participation of six associates and two partners aren't essential to *every* motion and hearing?

Unfortunately, corporations are swapping the remote threat of a court-imposed sanction for the certainty of self-inflicted monetary sanctions through over-preservation. It's an outcome prompted by irrational fear of the outliers.

No Whine before Its Time

That fear is palpable. During a confab about proposed rules amendments dealing with preservation and sanctions, a colleague leaned over and whispered that she'd "never heard so many grown men *whining* at one time." She wanted to scream, "Grow up, and find your cojones!"

Irrational fears should be cured, not coddled. Yet, many lawyers expect to be excused from acquiring the skills required to make good decisions about ESI preservation and to be insulated from the consequences of their bad decisions. The lawyers' role is to manage risk, not avoid it altogether. That's why they pay us the big bucks.

In e-discovery, it's popular to blame outlier opinions on judges so oblivious to enterprise complexity they view any shortcoming in preservation through the rosy lens of 20/20 hindsight. Yet every area of law has "those opinions" that are "markedly different in value from the others of the sample," as Merriam-Webster would say. They're usually well-reasoned on their peculiar facts, but contain pronouncements that are problematic when exported to other situations and circumstances. Beyond the courtrooms from which they came, outliers must be weighed with care and followed with discretion. Have other courts embraced the decisions? Have they been distinguished as “limited to their unique facts?”

Respect the Best. Forget the Rest

“Outliers” aren’t bad things; the law doesn’t evolve without them. But don’t let outliers force you into a race to the bottom. Just because you do business everywhere, your preservation standards need not reflect every outlier, everywhere. Many forward-thinking and savvy decisions have naturally come out of courts in New York, and districts surrounding the nation’s capital. There have been some outliers, too. There are good and bad e-discovery opinions from places far from Wall Street and the White House.

It brings to mind a bumper sticker common in Texas in the mid-1970s when multitudes from the ailing Rust Belt decamped for the prosperous Sun Belt. It read, “*We Don’t Care How You Do It up North*.” Sometimes it’s *okay* to say, “*We don’t care how you do it up north (or out west or down south).*” If your preservation practices are reasonable, diligent and pursued in good faith, you can *and should* resist the tyranny of the outlier.

Takeaways

- Resist the tyranny of the outlier. Acquire the technical skills—and muster the courage--to make sound decisions.
- Value metrics over anecdotes. Sanctions are a remote risk, and there is no rational reason to expect that a diligent, good faith effort gone awry will prompt punishment.
- Don’t let the inability to achieve “perfect” serve as a disincentive to achieving “good.” A defensible, proportionate and cost-effective preservation is within reach, but only by those with sufficient grasp of both information technology and law to support sound choices. If you aren’t driven by knowledge, you’ll be hounded by fear.

Weighing the Risks of E-Mail Preservation

by Craig Ball

[Originally published in Law Technology News, February 2012]

It's hard to persuade lawyers to accept leaner, less costly preservation protocols. Irrational fear of sanctions and spotty familiarity with information technology have so conditioned lawyers to over-preserve that when advised there's no need to keep something, they reply, "Let's keep it anyway…just to be safe." Proportionality in preservation isn't something you get down at the courthouse. Proportionality begins at home.

It begins by understanding the mechanics of preservation, enabling you to select the most cost-effective approaches and manage risk. For e-mail, there are several options:

Do Nothing

The cheapest, easiest and most common approach to e-mail preservation is to do nothing and hope that messages will be around when obliged to produce them. At first blush, doing nothing to preserve e-mail seems the e-discovery equivalent of Russian roulette, and some jurists call it grossly negligent because messages will inevitably winnow away as employees purge folders and change jobs. Most parties can't do nothing!

But doing nothing is a defensible choice for those whose e-mail systems are configured to automatically save items that fall within the scope of preservation. Litigants with message journaling or archiving systems are examples of entities that can safely do nothing once it's established that the systems hold the relevant messages and the messages aren't going anywhere. Even then, IT personnel should be made aware of the obligation to preserve the archives.

Message journaling grabs messages in transit rather than retrieving them from custodial accounts. Accordingly, journaled messages won't reflect user actions such as foldering, flagging or deletion. In rare instances where such user actions are relevant and material, you'll need to employ another approach.

Custodial-Directed Hold

The second most common approach is "custodial-directed hold," where individual employees are made responsible for finding and preserving relevant e-mail.

There's a gridiron maxim that three things can happen when you throw a football, and two of them are bad. The same can be said of making custodians responsible for preserving their e-mail. Some will do it well, some will do it badly and the rest will do nothing at all. Among those who preserve badly will be a few craven souls who decide that the best defense is a good offense and try to destroy what they were asked to hold.

Despite flaws, custodial-directed holds profit from custodians being the foremost experts on the significance and organization of their own mail. If you're going to preserve subjectively, it’s just good sense to secure the items the custodians themselves identify as relevant.

The effectiveness and defensibility of custodial-directed hold hinge on several factors, foremost among them the ability to educate and motivate custodians without triggering efforts to eradicate incriminating or embarrassing material. The notice needs to be strong enough to prompt action and to fix in every recipient’s mind that they were put under hold, the reasons for the hold and the things they were supposed to do. It’s painful when a deponent testifies, “I got something from legal; but I get those all the time, and I never know what they want me to do. I think it said don’t delete stuff.” Hold notices should be coupled with a method for acknowledgement and backed by steps to insure custodians both understand their obligations and meet them.

Grabbing All Mail

A third approach to e-mail preservation is the wholesale copying of mail accounts, which entails a troubling tradeoff. It’s deceptively easy and cheap to copy the mail server data or pull a tape backup set from the rotation. Both quickly lock down e-mail at the nominal cost of a replacement set of backup tapes or a dollop of storage. If e-discovery never comes to pass and no new holds arise, you’re the genius who gambled and won.

What’s the risk? Grabbing a big slug of data is a gamble because it doesn't eliminate the need to isolate relevant information; it simply defers the effort, perhaps to a time when layoffs or fading memories make it harder to find relevant items. Banking big data is also a gamble because the moment it’s set aside, the dataset put on hold begins to look less-and-less like the data on the server. Instead, it becomes a time tunnel to *all* mail for *all* users at a point in the past. As new claims arise, new legal holds attach to both the current mail and the "time tunnel" collection. Soon, you’ve got to process and search. The data set aside’s become a cancerous cell: aberrant, hard to eradicate and scary as hell. Like a cancer, that time tunnel dataset is expensive to manage and lethal to ignore.

Grabbing Selected Accounts

A more restrained alternative to copying all accounts is copying just the mail of custodians connected to the dispute. This doesn’t eliminate the potential for new legal holds to attach, but it makes for a more manageable volume when the preserved data must be searched. Newer e-mail applications like Microsoft Exchange Server 2010 make preservation of key player data easy to implement. Other mail systems and older versions of Exchange Server require a labor-intensive carving out of individual accounts for preservation.

Targeted Collection

A further refinement is to preserve just the accounts of key custodians and apply searches and filters against those accounts to carve out relevant material. This fifth approach is called "targeted collection," and its upside is that the volume of data preserved is likely to be so finite and focused that it won't come under legal holds involving different issues or parties. Plus, sensibly slashing volume pays dividends in every subsequent phase of e-discovery.

The downside of targeted collection is that it requires that large volumes of information be processed up front, so you surrender some of the benefit of suits never filed or early settlements. Moreover, targeted collections are typically preserved before a party has the full picture of the issues and players in the case and often without benefit of the key custodians' input. Users may be overlooked, and keyword searches are certain to miss relevant material—material that may be gone forever by the time it's missed.

Other Mail

Mail can reside off the server within mail containers (like .OST and .PST files) on custodians' local hard drives, in handheld devices and on the web. Whether this mail must be preserved depends on whether it's relevant and unique with respect to other preserved sources. Container files can be copied or searched and then preserved in container file or single message file formats. Web sources that cannot be secured in situ to guard against deletion may be downloaded, in whole or in part, to a local container file using POP3 or IMAP protocols.

Prospective Preservation

The approaches discussed are all geared to retrospective preservation of e-mail. Where the obligation to preserve is ongoing, they must be supplemented by steps to capture future relevant messaging. This may be accomplished by, e.g., reminders to custodians, periodic re-collection from servers or changing mail server settings to journal or archive messaging.

Preserve Carefully

It's time we change the maxim from "preserve broadly" to "preserve carefully." Many preservation efforts are thoughtless and mechanical, designed by those loath to turn a discriminating eye to the task. While no approach to e-mail preservation is wholly without risk or cost, knowing your options helps you to "right size" the approach.

Take aways:

- There are multiple ways to balance risk and cost when preserving e-mail.
- Proportionality begins at home; that is, proportional preservation is up to you.
- Saving everything eliminates one risk, but introduces others.
- Three things can happen when using custodial-directed hold…and two of them are bad.
- Custodial-directed hold should be a part of most legal holds, but not the only part.
- Microsoft Exchange Server 2010 better supports mail preservation for litigation.
- "Preserve broadly" is safe, but expensive. "Preserve carefully" is safe and cost-effective.

A Quality Assurance Tip for Privileged ESI

by Craig Ball

[Originally published on the Ball in Your Court blog, February 22, 2012]

There have lately been a boatload of good stories written about Google's so-called 'Billion Dollar E-Discovery Blunder.' Yes, it was a blunder, and, though the damages are dwindling, maybe it will end up costing Google a billion bucks; but, I'm not so sure it's all that much an e-discovery issue. That said, I'm going to wind this post up with a suggestion of a simple technique for QA/QC in e-discovery you can use to keep your client or company from the same predicament.

First, the Blunder: Oracle sued Google claiming that Google's Android smartphone platform infringes Oracle's Java programming language patents. With almost $27 billion in revenue and $6 billion in profits, Oracle is #96 on the Fortune 500 list of companies that suck at e-discovery. Google is #92, with $29 billion in revenues and $8.5 billion in profits. So, it's a well-matched, Goliath vs. Goliath fight (and even Goliath is going, "Damn, they're big and rich").

Plus, it's got to be personal for many folks in Silicon Valley. Redwood City and Mountain View, California are just minutes apart, so you can imagine that when the two Larrys (Oracle's Ellison and Google's Page) bump into each other at Fry's or waiting in line at the DMV, it's can't be all guy hugs and fist bumps.

In the cozy bosom of the Valley, folks don't just haul off and sue their (very rich and powerful) neighbors. Instead, they have meetings where, fueled by green tea and organic vegan muffins, the lawyers present their case to avoid all the muss, fuss and bother of a lawsuit. Something like that happened on July 20, 2010, when Oracle's lawyers presented contemplated claims to Google's senior Counsel, Ben Lee. Later, Lee, Google's GC and an engineer named Tim Lindholm put their heads together to strategize about the coming onslaught.

Here, it makes sense to simply quote from the Court of Appeals' 2/6/12 order in*In re: Google*, Inc.:

> *At 11:05 a.m. on August 6, 2010, Lindholm sent an email to the attention of Andy Rubin, Google'sVice President in charge of its Android operating platform. Lindholm also included Lee, himself, and another Google engineer, Dan Grove, on the email. The body of the email provided as follows:*
>
> *Attorney Work Product*
> *Google Confidential*

> *Hi Andy,*
> *This is a short pre-read for the call at 12:30. In Dan's earlier email we didn't give you a lot of context, looking for the visceral reaction that we got.*
>
> *What we've actually been asked to do (by Larry and Sergei) is to investigate what technical alternatives exist to Java for Android and Chrome. We've been over a bunch of these, and think they all suck. We conclude that we need to negotiate a license for Java under the terms we need.*
>
> *That said, Alan Eustace said that the threat of moving off Java hit Safra Katz hard. We think there is a value in the negotiation to put forward our most credible alternative, the goal being to get better terms and price for Java.*
>
> *It looks to us that Obj-C provides the most credible alternative in this context, which should not be confused with us thinking we should make the change. What we're looking for from you is the reasons why you hate this idea, whether you think there's anything we've missed in our understanding of the option.*
>
> *–Tim and Dan*

Before you get your knickers in a twist worrying about how hard this will be on "Safra Katz," know that they were surely talking about Safra Catz, Co-President of Oracle and named one of the Most Powerful (and Highly Paid) women in the world. I'd guess she'd have more to worry about being alone in a room with her Co-President, the overcharged and overcharging Mark Hurd.

This e-mail lodged in several locations around the Googleplex and, after suit, Google produced several copies of this message and withheld other copies, claiming they were privileged and listing them in its privilege log (see below). Ten months later, Google tripped to its error and demanded return of the copies. Inadvertent production? Pretty clearly. Timely action or waiver by use without objection? Non issues, as it turns out.

The Court found that the e-mail was not privileged as either a confidential attorney-client communication or as attorney-work product. The reason I don't think the opinions are important e-discovery rulings is because the rationale and result would have been exactly the same if these had been old-timey paper memos. The medium had nothing to do with the issues or outcome.

The e-discovery nexus arises from the way the copies were retained (reportedly as autosaved backups) and from the failure to intercept the copies before production. There are all sorts of bright ideas emerging from smart folks who have groundbreaking tools they could sell to that search naif, Google, to help it avoid this ever happening again. Reading some of these missives made me think of a post I wrote two years ago for the EDDUpdate blog called, "A Quality Assurance Tip for Privileged ESI."

In that long-ago post, I called for producing parties to run a last-ditch quality assurance linear keyword search against the final production set before it goes out the door. Linear (i.e., across the actual documents in the set, not against an index) because such a belt-and-suspenders approach compensates for inherent and unforeseen omissions and corruptions in the index. The keywords and phrases searched would be unique selections from the highly sensitive privileged documents you'd already found and listed on the privilege log. Yes, that's right, you're searching for things *you already know are privileged and that you believe have already been culled from the set.* Missing these was Google's gaffe.

Not all inadvertent production of privileged material stems from what you fail to find. Often, you let slip the very thing that you know about and are most determined to protect. It's like trying <u>not</u> to think of a purple hippopotamus.

Again, this is a QA/QC technique applied to the production set on the eve of production. It's a simple, effective, quick and cheap way to protect against the most damaging privileged communications slipping through because, as Google learned, they do slip through.

Here, the terms searched might have been, e.g.,:

- "technical alternatives exist to Java"
- "think they all suck"
- "hit Safra Katz hard"
- "Obj-C provides the most credible alternative"
- the subject line of the message

The chance of hitting an unrelated document is nil, but the likelihood of catching a draft or autosaved version is high. You want the terms searched to be unique (bad grammar and misspellings help) and extracted from various points across the body of the document or message (in case it's an early draft or a corrupted or truncated version). You're not looking for unknowns, and don't waste time on the trivial, i.e., on all that silly chaff in most privilege logs. You want to do this for the "if this message makes it to the other side, we're hosed" stuff.

Did the lawyers know this message was a bombshell? You'd think so, considering the license they took in the privilege log when describing the copies withheld; to wit, "Email reflecting advice of counsel in preparation of litigation re alternatives for Java for Android and Chrome" and "Email seeking advice of counsel re technical alternatives for Java." I didn't see any advice of counsel in there, did you? And I just bet the engineers at Google run to the legal department for technical advice when they're stumped. In…my…dreams!

You can stop here, or you can go on to read what I published two years ago (while you imagine me maturely going "*Nyah*, *nyah*, na-*nyah nyah*" in that smug way that makes even me want to punch me):

We squander so much money in e-discovery searching for confidential attorney-client communications. "Squander" because it's an outsize expense that could have been largely eliminated with minimal effort at the time fingers met keyboard. It's not as though counsel are wholly unaware of the sensitivity of privileged communications when made. If it had been a face-to-face conversation, we'd have had the presence of mind to shut the door or ask those outside the ambit of privilege to leave. Lawyers really aren't as stupid as we sound in the reported decisions.

If we have the presence of mind to recognize and protect a confidential attorney-client communication when made face-to-face–if we're savvy enough to say, "*Wait a second while I take this off speakerphone*,"–why are we incapable of bringing the same cautious mien to our electronic conversations? And, why-oh-why do we forget the most important component of quality assurance before producing material posing a risk of inadvertent production of privileged communications?

Ask a judge who's done an *in camera* review of privileged ESI what percentage of the material submitted was truly privileged, and you're likely to hear numbers hovering way below fifty percent. An average assessment of 20% or less wouldn't surprise me. Does anyone *ever* review the definition of a confidential attorney-client communication anymore? Is it not in the Nutshells today?

The favored technique to cull privileged material from ESI entails looking for any material that includes lawyers' names, firm names, lawyer and firm e-mail addresses and words like "privileged." It's a criterion geared to grab *everything*sent to or from an attorney or firm. If this surfeit of material isn't just lazily set aside and forgotten, it's painstakingly reviewed page-by-hourly-charged-page.

Heaven forbid that the profession should ever be forced to surrender this bountiful boondoggle. Imagine, we seed the client's ESI with privileged communications, then bill the client to segregate it. If BP were a law firm, it'd be charging for cleaning up the crude.

…

I post now to suggest something you *will* embrace because it's an essential step in QA/QC that's so obvious; I'm amazed at how rarely it's done.

Assume you've done your privilege review, and you're ready to make production. Suddenly, you hear a little voice in the back of your mind. *Why, it's Judge Paul Grimm*, whispering, "*QA…QC…QA…QC…Victor Stanley."* You know you're supposed to do *something* to check the quality of your production to be assured that you haven't inadvertently allowed privileged communications to slip through your net. But what? You've searched the index for lawyers' names, firm names, e-mail addresses and words like “law,” “advise,” “liable,” “criminal” and “attorney.” You’ve double-checked random samples. What more can you do?

One thing you should absolutely do is search the material about to be produced for examples of confidential attorney-client communications you ***know*** exist. That is, the stuff you most fear the other side seeing. Examples are probably right there in your file and e-mail. You should have a set of unique searches composed to ferret out these bombshells in anything you send to the other side. It’s the stuff for which you most need quality assurance and control, because it’s the stuff that would be most prejudicial if it crossed over.

I also suggest that you search for these core privileged communications *linearly* across the contemplated production, not just within the indices, because–let’s face it–indexed search is fast, but it misses stuff that linear search picks up. I get why you don’t search the collection linearly at the outset–life is short–but when it’s culled to a production set, don’t you want the benefit of *both* technologies to protect against inadvertent production of the most sensitive, privileged material?

FYI: *Linear search* goes through the documents in the collection seeking keywords matches. *Indexed search* looks solely to indices, i.e., lists of words meeting certain criteria for inclusion and exclusion, such words having been culled from the documents in the collection.

I recently commented on a long, thoughtful post of Ralph Losey’s discussing *Mt. Hawley Ins. Co. v. Felman Production, Inc.*, 2010 WL 1990555 (S.D. W. Va. May 18, 2010). I summed up my sentiments this way:

For all the many challenges there are to isolating privileged material in voluminous ESI, finding the privileged items well known to counsel and appearing in their own files need not be one of them.

Gold Standard

by Craig Ball

[Originally published in Law Technology News, April 2012]

Lawyers are in denial to the point of delusion with respect to the reliability of keyword search and human review. Judge John Facciola put it best when he quipped that lawyers think they're experts at keyword search because they once found a Chinese restaurant on Google.

We trust keyword search because we understand it. We trust manual review of documents because we grossly overestimate reviewers' abilities to make sound, consistent decisions about relevance. "To err is human," the Bar seems to say, "but forgive us if we'd rather not divine just how error-prone reviewers really are."

Better approaches to search are arriving as so-called "predictive coding" or "technology assisted review" (TAR) products. Still, it will be years before the rank and file embraces TAR, if only because those hawking TAR tools remain resolutely uninterested in positioning the technology for use by anyone but big corporations and white shoe law firms. Worse, the fervor among vendors to sell something, *anything* they can label predictive coding insures that tools little different from ordinary keyword search will be given a dab of lipstick and pushed out to market as TAR tools. It's messy down in the TAR pit.

Even those adopting predictive coding tools will need to compile "seed sets" of relevant documents to train their tools. So, clunky-but-comfy keyword search and manual review are likely to remain the means to cull seed sets from samples. Despite serious shortcomings, keyword search and manual review will be with us for a while.

Keyword search is the art of finding documents containing words and phrases that signal relevance followed by page-by-page (linear) review of those documents. It's often called the "gold standard" of electronic discovery.

That's ironic, because extracting and refining gold relies less on finding precious aurum than it does on dispersing all that isn't golden. Prospectors use water and chemicals to flush away all but the gold left behind. So, a true "gold standard" for keyword search would incorporate both precise inclusion (smart queries) <u>and</u> defensible exclusion (smart filters).

To illustrate, in one e-discovery dispute over search, the plaintiff submitted keywords to be run against the defendant's e-mail archive for a three-month interval. Unfortunately, the archive held all e-mail for all custodians, and the defendant adamantly refused to segregate by key custodian or deduplicate before running searches. The interval was narrow, but the collection was vast and redundant.

The defendant tested the agreed-upon keywords but shared only aggregate hit rates for each. Thinking the numbers too high, but unwilling to look at the hits in context, the defendant rejected the search terms. The plaintiff agreed the hit counts were daunting but asked to see examples of hits on irrelevant documents before furnishing exclusionary (AND NOT) modifications to flush away more of what wasn't golden.

The defendant refused, insisting it wasn't necessary to see the noise hits in context to generate more precise queries. The parties were at an impasse, with one side grousing "too many hits" and demanding different search terms and the other side uncertain how to exclude irrelevant documents without knowing what caused the noisy results.

A lawyer who dismisses a search because it yields "too many hits" is as astute as the Emperor Joseph dismissing Mozart's *Il Seraglio* as an opera with "too many notes." Mozart replied, "There are just as many notes as there should be." Indeed, if data is properly processed to be susceptible to text search and the search tool performs appropriately, a keyword search generates just as many hits as there should be. Of course, few lawyers craft queries with the precision Mozart brought to music; so when the terms used seem well chosen for relevance, it's crucial to scrutinize the results to learn what tailings are cropping up with the gilt-edged, relevant documents.

Keyword search is just a crude screen: "Show me items that contain these words, and don't show me items that contain those." High hit counts don't always signal a bad screen. If search terms merely divide the collection into one pile holding relevant documents and one without, you're closer to striking gold. Then, you look at what you can reliably exclude with the next screen, and the next; drawing ever closer to that elusive quarry, *documentum relevantus*.

But you must see hits in context to refine queries by exclusion. That seems so manifestly obvious, it's astounding how often it's not done.

When lawyers delegate keyword search, they often get back only aggregate hit counts and mistakenly conclude that's enough information to judge searches noisy or not. If, instead, counsel get their hands dirty with the data, as by personally exploring representative samples using desktop or hosted tools, the parties could work quickly, effectively and cooperatively to zero in on relevant material. Good queries are best refined by knowledgeable people testing them against pertinent, small collections. Lousy outcomes spring from lawyers thinking up magic words and running them against everything.

It's not just a theory. Recently, as part of an early case assessment effort, I sought to rapidly isolate relevant documents from a half million e-mail items culled from four key custodians. That's a volume where you'd expect to see bids from service providers and mustering of review teams. It's a project most firms would see as much more than a weekend's work for one lawyer.

We tried something different. To start, the client exported the four key custodians' e-mail messages for the time period of interest from its e-mail archives. Those 50 gigabytes of messaging went into a desktop processing and review tool.

Extracting and indexing the data overnight, I flagged exception items (e.g., images without extractable text and encrypted files) for further processing, then exported spreadsheets reflecting the most used e-mail addresses. I asked the custodians to flag addresses with no connection to the dispute. Meanwhile, I compiled the customary list of search terms and phrases expected to occur in relevant documents and tested these. Documents with false hits were examined for characteristics permitting mechanical exclusion. Testing, re-testing and re-examination soon produced reliable inclusion and exclusion term lists. Weeks of evaluation took just days because the iterations and results were instantaneous.

The discards were tested, too. For example, material excluded by addresses but containing inclusion terms was carefully checked to insure the hits weren't relevant. Defensible exclusion proved as powerful as inclusion, and potentially relevant material that couldn't be excluded as tailings stayed in the collection as ore. A true "gold standard."

Did it produce a perfectly parsed set of material? Certainly not. Keyword search and human review still fall short of expectations. But it was fast, relatively cheap and afforded cautious confidence that the set produced was more relevant and less riddled with junk than what would have emerged from the usual game of blind man's buff. It was fast and cheap because the person creating and testing the inclusive and exclusive filters was elbows deep in the data and hands on with the search tool. Feedback was immediate. Quality checks could be done at once.

Ideally, e-discovery tools don't put distance between the lawyer and the evidence but, instead, extend our reach and help us get our arms around big data. A lawyer who is hands-on with the evidence and who tests and refines his or her choices is a lawyer who can explain and defend those choices. That's the real golden future of e-discovery. Welcome back, counselor.

Ten Bonehead Mistakes in E-Discovery

by Craig Ball

[Originally published in Law Technology News, June 2012]

Spoiled by Google and legal research, lawyers are woefully unprepared for the difficulty of search in e-discovery.

Search fails us in two, non-exclusive ways: our query will not retrieve the information we seek, and our query will retrieve information we didn't seek. Obviously, we want what we're looking for (high recall) and *only* what we are looking for (high precision).

Recall and Precision aren't friends. Every time Recall has a tea party, Precision crashes with his biker buddies and breaks the dishes.

It's easy to achieve a high recall of responsive ESI. You simply grab it all: *100% of the data = 100% recall*. The challenge is achieving precision. If one out of every hundred items returned is what you seek, 99 items are duds—*1% precision stinks.*

Keyword search followed by human review is called "linear search," and for now, it's standard operating procedure in e-discovery, in part because linear search is mistakenly considered the safest course lest a party fail to produce something responsive or turn over something that should have been withheld.

Linear search is time-consuming, so it's expensive. Worse, it doesn't work well. People make search and assessment errors, and making lots of searches and assessments, they make lots of errors!

Mistakes *can* be subtle and hyper technical, but most are not. If we eliminate bonehead errors, we improve the quality of e-discovery and markedly trim its cost. Search will ever be a battle between Recall and Precision, but avoiding bonehead mistakes limits casualties.

Recently, I ran a blog post sharing five bonehead mistakes I'd observed and asking readers to contribute five more.

Mistake 1: Searching for someone's name or e-mail address in their own e-mail

If you run a list of search terms including a custodian's name or e-mail address against their own e-mail, you should expect to get hits on all messages. I know some of you are saying, "Craig, no one's *that* boneheaded!" Actually, plaintiffs do it, defendants do it, and vendors run these searches without flagging the error. Ask yourself: how often are the proposed search term lists exchanged between counsel carefully broken out by particular custodians or forms of ESI to be searched?

Bill Onwusah, Litigation Support Manager at Hogan Lovells in London, commented that he'd seen this mistake take the form of "searching for a term that shows up in the footer of every single document produced by the organisation," such as the firm's name.

Mistake 2: Assuming the Tool can run the Search

Every ESI search tool has features and limitations. You must understand what data has been indexed and what search methods and syntax are supported.

Most e-discovery tools index words, which means you won't retrieve any information that isn't text (including some PDF, TIF and other pictures of words that haven't been OCR'd to searchable text) or that isn't accessible text (like encrypted documents). Plus, most search tools don't index parts of speech called "noise" or "stop" words deemed so common they'll gum up the works. I call this the "To Be or Not to Be" problem, because all of the words in Hamlet's famous phrase tend not to be indexed in e-discovery.

Syntax mistakes occur when you assume the tool can run the search the way you constructed it. Not every search tool supports every common search method, e.g., wildcard characters, Boolean constructs, stemming, proximity searches or regular expressions, and even when two tools support the same search method, tool A may require you to use different search syntax than tool B.

Mistake 3: Not Testing Searches

Much of what distinguishes a mistake as boneheaded is the ease with which it could have been avoided. When a party to a lawsuit once proposed the letter "S" as a search term, I didn't need to test it to know it was a bonehead choice. But what about all those noisy terms that pop up in file paths or are invariably encountered within ESI yet have nothing to do with the case? Even search terms that appear bulletproof can surprise you. Test your searches to be sure they perform as expected.

Mistake 4: Not Looking at the Data!

Don't just natter on about the *quantity* of hits to evaluate your search; check the *quality* of the hits. *Look* at the data! Minutes spent looking at the data can eliminate weeks or months of reviewing crappy results and a zillion dollars spent in motion practice.

Mistake 5: Ignoring the Exceptions List

It's the rare e-discovery effort where everything processes without exception. Typically, the exception list will reflect hundreds or thousands of items that are encrypted, corrupt, unrecognized or unreadable. You may take a calculated risk to ignore certain exceptional items; but too often, exceptions are misclassified as benign or dismissed altogether. That's boneheaded.

Ed Fiducia, Regional Vice President for EDD vendor Inventus, offered a sixth and seventh for the bonehead mistakes list:

Mistake 6: Assuming That Deduplication Solves My Problem

Ed pointed to the limits of using hashing to identify truly duplicative files. “The rub is the definition of a truly duplicative file.”

For example, e-mail messages sent to multiple addressees won’t deduplicate across custodians because each message reflects its unique message ID and delivery path. Word and PDF versions of the same document won’t hash deduplicate because they’re different file formats.

Hashing leaves “thousands upon thousands of near duplicates that must be identified and reviewed. This leads to not only a dramatic increase in review costs, but a dramatic increase in the probability that documents will be coded inconsistently. Spend more money, get worse results. Not a good combination.”

Mistake 7: Reviewing Fifty Custodians When Five Will Do

Ed Fiducia: “*Preserve* everything? You bet! *Review* everything? Not in my book.”

“The knee jerk reaction is to blame plaintiffs’ attorneys who ask for everything. Equal responsibility goes to defense attorneys who don’t negotiate the process from the start in meet and confer. As a service provider, you’d think I’d push to process and review everything; but over the past 18 years, I’ve seen case after case prove that if the scope of e-discovery is limited from the start--with caveats to allow for additional discovery when warranted--everybody wins.”

Dave Swider, Senior Discovery Consultant for Evolve Discovery, contributed:

Mistake 8: Failing to Search for Common Name Variations

“Here’s one we see pretty often: Searching for names without anticipating variations. We’ll see a search for ‘Robert Smith’ with no variations specified; no Rob, Bob, Bobby, Robby, not even an email address.”

“Similarly, we’ll be asked to search for a complete law firm name: all five names as an exact string, with no domain or proximity search.”

Too, “we see use of wildcards and terms that are far too expansive…. I worked on a case that involved laying one material on top of another in a process called ‘deposition.’ Guess what term appeared on the potential privileged terms list? A common offender in groundwater cases is ‘well.’”

Marc Hirschfeld, President of Precision Legal Services, added:

Mistake 9: Neglecting to Run Searches Against File and Folder Names

“Here is one that I never see attorneys talk about…. I often find a treasure trove of information when the name of a folder holding relevant information includes a search word but the documents inside do not. It’s as if the user pre-identified these documents as relevant but, because the file and folder names weren’t indexed or searched, the treasure is missed.”

Ann Marie Gibbs, National Director of Consulting at Daegis, offered:

10: Failing to Rapidly React to the Problems You Encounter

“Another review oversight we see is a failure to ‘update’ the review set when a ‘false hit’ is running up the review bill. This relates to the mistake where a client declines to accept excellent advice on search selection criteria. If you can’t get them to understand the problem on the front end, you have a second bite at the apple on the back-end.”

Dave Swider sums it up: “The number one boneheaded move by legal staff is simply not bothering to understand how data works and how they can best apply tools that will make their outcomes better. ***Our best clients are those that treat data not like documents, but like data.”***

Imagining the Evidence

by Craig Ball

[Originally published in Law Technology News, August 2012]

As a young lawyer in Houston, I had the good fortune to sip whiskey with veteran trial attorneys who never ran short of stories. One told of the country lawyer who journeyed to the big city to argue before the court of appeals. The case was going well until a judge asked, "Counsel, are you aware of the maxim, *'volenti non fit injuria?'"*

"Why, Your Honor," he answered in a voice like melted butter, "In the piney woods of East Texas, we speak of little else."

Lately, in the piney woods of e-discovery, the topic is technology-assisted review (TAR *aka* predictive coding), and we speak of little else. The talk centers on that sudsy soap opera, *Da Silva Moore v. Publicis Groupe, and* whether Magistrate Judge Andrew Peck of the Southern District of New York will be the first judge to anoint TAR as being "court approved" and a suitable replacement for manual processes now employed to segregate ESI.

TAR is the use of computers to identify responsive or privileged documents by sophisticated comparison of a host of features shared by the documents. It's characterized by methods whereby the computer trains itself to segregate responsive material through examination of the data under scrutiny or is trained using exemplar documents ("seed sets") and/or by interrogating knowledgeable human reviewers as to the responsiveness or non-responsiveness of items sampled from the document population.

Let's put this "court approved" notion in perspective. Dunking witches was court approved and doubtlessly engendered significant cost savings. Trial by fire was also court approved and supported by precise metrics (*"M'Lord, guilt is established in that the accused walked nine feet over red-hot ploughshares and his incinerated soles festered within three days").* Whether a court smiles on a methodology may not be the best way to conclude it's the better mousetrap. Keyword search and linear review enjoy *de facto* court approval; yet both are deeply flawed and brutally inefficient.

The imprimatur that matters most is "opponent approved." Motion practice and false starts are expensive. The most cost-effective method is one the other side accepts without a fight, i.e., the least expensive method that affords opponents superior confidence that responsive and non-privileged material will be identified and produced.
Don't confuse that with an obligation to kowtow to the opposition simply to avoid conflict. The scenario I'm describing is a true win-win:

- Producing parties have an incentive to embrace TAR because, when it works, TAR attenuates the most ***expensive*** component of e-discovery: *attorney search and review*.

- Requesting parties have an incentive to embrace TAR because, when it works, TAR attenuates the most ***obstructive*** component of e-discovery: *attorney search and review*.

Producing parties don't just obstruct discovery by the rare and reprehensible act of intentionally suppressing probative evidence. It occurs more often with a pure heart and empty head as a consequence of lawyers using approaches to search and review that miss more responsive material than they find.

It's something of a miracle that documentary discovery works at all. Discovery charges those who reject the theory and merits of a claim to identify supporting evidence. More, it assigns responsibility to find and turn over damaging information to those damaged, trusting they won't rationalize that incriminating material must have had some benign, non-responsive character and so need not be produced. Discovery, in short, is anathema to human nature.

A well-trained machine doesn't care who wins, and its "mind" doesn't wander, worrying about whether it's on track for partnership. From the standpoint of a requesting party, an alternative that is both objective and more effective in identifying relevant documents is a great leap forward in fostering the integrity and efficacy of e-discovery. Crucially, a requesting party is more likely to accept the genuine absence of supportive ESI if the requesting party had a meaningful hand in training the machine.

Until now, the requesting party's role in "training" an opponent's machines has been limited to proffering keywords or Boolean queries. The results have been uniformly awful.

But the emerging ability to train machines to "find more documents like this one" will revolutionize requests for production in e-discovery. Because we can train the tools to find similar ESI using *any* documents, we won't be relegated to using seed sets derived from *actual* documents. We can train the tools with contrived examples–fabrications of documents like the genuine counterparts we hope to find.

I call this "imagining the evidence," and it's not nearly as crazy as it sounds.

If courts permit the submission of keywords to locate documents, why not entire documents to more precisely and efficiently locate other documents? Instead of demanding "any and all documents touching or concerning" some amorphous litany of topics, we will serve a sheaf of dreams—freely forged smoking guns—and direct, "show me more like these."

Predictive coding is not as linguistically fussy as keyword search. If an opponent submits contrived examples of the sorts of documents they seek, it's far more likely a similar document will surface than if keywords alone were used. As importantly, it's less likely that a responsive document will be lost in a blizzard of false hits. This allows us to rely less on our opponents to artfully construct queries. Instead, we need only trust them to produce the non-privileged, responsive results the machine finds.

There's more to documents that just the words they contain, so mocking up contrived exemplars entails more than fashioning a well-turned phrase. Effective exemplars will employ contrived letterheads and realistic structure, dates and distribution lists to insure that all useful contextual indicia are present. And, of course, care must be taken and processes employed to ensure that no contrived exemplars are mistaken for genuine evidence.

The use of contrived examples may ruffle some feathers. I can almost hear a chorus of, "How dare they draft such a vile thing!" But the methodology is sound, and how we will go about "imagining the evidence" is likely to be a topic of discussion in the negotiation of search protocols once use of technology assisted review is commonplace.

Another "not as nutty as it sounds" change in discovery practice wrought by TAR will be affording requesting parties a role in training TAR systems. The requesting party's counsel would be presented with candidate documents from the collection that the machine has identified as potentially responsive. The requester will then decide whether the sample is or is not responsive, helping the machine hone its capacity to find what the requester seeks. After all, the party seeking the evidence is better situated to teach the machine how to discriminate.

For this to work, the samples must first be vetted by the responding party's counsel for privilege and privacy concerns, and the requesting party must be willing to undertake the effort without fretting about revealing privileged mental impressions. It's going to take some getting used to; but the reward will be productions that cost less and that requesting parties trust more.

*Volenti non fit injuria means "*to a willing person, injury is not done." When we fail to embrace demonstrably better ways of searching and reviewing ESI, we assume the risk that probative evidence won't see the light of day and voluntarily pay too high a price for e-discovery.

Homo Electronicus

by Craig Ball

[Originally published in Law Technology News, October 2012]

We are the transitional generation in terms of the shift from discovery in a world geared to information on paper to one where paper is largely an afterthought. An airline boarding pass is a screen shot of a bar code, gate, time and seat number. We print it in case TSA can't scan our phone, then trash it when we touch down.

Growing up, the organization of information on paper was so ingrained in our education that we take our "paper skills" for granted even as paper has all-but-disappeared. We learned to color inside the lines. Put our name and the date at the top of our papers. Organize alphabetically. Staple and paper clip.

We learned the structure of a "business letter." Date and subject go here, salutation there, and don't forget the CC: and BCC: addressees at the bottom.

All of it marched more-or-less seamlessly into a common culture of paper records management. Correspondence flowed into files, folders, drawers, cabinets and file rooms. Everything had a place, and everything depended upon information being in its place. That is, everything depended upon organizing information from its creation and all along its path until it found its semi-permanent place in the storage and retrieval system.

As information went digital, we clung to metaphors of records management. The screen icons remained files, folders and even envelopes. But while we pretended digital information was still like paper, our culture of records management collapsed. The fleeting phone call and the enduring business letter or "memo to file" all morphed into e-mail. Subject lines ceased to reliably describe contents. File clerks became baristas and file rooms became server rooms. Everyone was left to their own devices—literally—in terms of information management. Computerized search, they promised, would do away with all that pesky management of documents.

And, in many ways, the promise was kept. We draw on vast reservoirs of information using search tools of such instantaneous ingenuity and complexity that we rarely reflect on what transpired for us to find that Chinese restaurant in San Francisco or convert U.S. Dollars to Brazilian Reais at market rates. We've been content to leave it to the geeks.

And there's the nub of the problem in e-discovery. As information stopped being like paper records and everything became databases, lawyers were content to leave organization to the geeks. We can't imagine a competent lawyer *not* knowing how to find a document in a file folder or cabinet; yet, oddly, we can't imagine a lawyer knowing how to fashion a competent ESI search protocol or query a database. We barely expect lawyers to know what ESI protocols and databases *are*. We've set the bar too low for the Bar, and clients and judges are suffering as a consequence.

Part of the problem is that the practical education of lawyers has long depended upon veteran partners handing down the lore of lawyering to associates. But when it comes to e-discovery, veteran lawyers have nothing to share. “Back in the day” war stories about bankers’ boxes in sweltering warehouses aren’t much help when you’re standing in an icy server room.

And when we *do* try to teach e-discovery, we elide over what makes e-discovery challenging: *the technology*. Most e-discovery courses teach the law of e-discovery and give short shrift to the “e.” Well, guess what? The *law* of e-discovery isn’t all that hard to master! You can learn to spout “not reasonably accessible” or “meet and confer” all the livelong day, and you’ll still be as useless as teats on a boar hog when it comes to bringing off an e-discovery effort that works without waste.

The transitional generation lawyer responds, “I’ll hire someone who knows that stuff.”

Okay. That’ll work…for a while.

But someday soon, it will be clear that lawyers *can* learn two things, and sooner still, clients will tire of paying for their lawyer’s to pass the heavy lifting on to e-sherpas.

I say, let’s start learning to carry our own briefcases when it comes to digital evidence. Let’s stop kidding ourselves that this isn’t something we need to understand, and stop being so damned afraid to get our hands dirty with data or look like we might not be the smartest person in the room because we don’t know what goes on under the hood!

I recently asked a speaker on technology-assisted review for his thoughts about the respective strengths and weakness of the various techniques used to cluster documents in predictive coding. He replied that he didn’t know and didn’t need to know. He said, “I don’t need to understand how a jet engine works to fly on an airplane.” I think he forgot that, as lawyers, we are the pilots, not the passengers. We are ultimately responsible for the integrity of our craft.

It breaks my heart when law students question why they need to learn about hashing or unallocated clusters. *“The lawyers I talk to say this is stuff they hire people to handle.”* How am I to respond? *The lawyers you talk to choose to believe that what they don’t know can’t be a measure of their competence?*

Each of us in the transitional generation has to make up our own minds about what we need to know. We can choose to be Eloi or Morlocks. But let’s not kid the next generation of lawyers that they have that choice. They will little know or need our paper-centric skills, and we do them grievous injury when we assure them it’s someone else’s job to understand information technology. We cannot be their mentors on these things, and their easy fluency with consumer technology is insufficient, by itself, to manage e-discovery. They need to learn more than we did, and the best help we can give them is to make sure they understand that.

And we can't stop there. We have an entrenched leadership of lawyers with twenty- to thirty years' experience who cannot simply be idled as we wait for them to shuffle off this mortal coil. We have to re-educate our lawyers—even the gray hairs. It's a task made harder by the reluctance of lawyers of all ages to admit there's a gaping hole in their skill sets that they are patching with the green poultice of wasted client money.

Hard, but not impossible.

I imagine a world where lawyers can and do learn where information resides, the forms it takes, the useful metadata that surrounds it and effective ways to search, manage and present modern evidence without spending so much that no one can afford to turn to the courts to resolve disputes. I see lawyers who roll up their sleeves, use good tools and get their hands dirty with data. These lawyers have evolved from *Homo Erectus* to *Homo Electronicus.* And they will thrive.

Are They Trying to Screw Me?

by Craig Ball

[Originally published on the Ball in Your Court blog, October 9, 2012]

The title of this post is the question posed by a plaintiffs' lawyer who called because he didn't know what to make of a proposal from opposing counsel. The lawyer explained that he'd attended a Rule 26(f) "Meet 'n Confer" where he'd tried to manifest the right grunts and signs to convey that he wanted electronically-searchable production. As neither of the lawyers conferring knew how they might achieve such a miracle, they shared a deer-in-headlights moment, followed by the usual "let me ask my client and get back to you" feint. Some years back, I defined a Rule 26(f) conference as "*Two lawyers who don't trust each other negotiating matters neither understand*." That definition seems to have withstood the test of time.

Before my high-handed cynicism turns you off completely, let me explain that I appreciate that many fine lawyers didn't grow up with this "computer stuff." They earned their stripes with paper and, like me, leapt to law from the liberal arts. They're crazy busy with the constant demands of a trial practice, and ESI is just not a topic that excites their interest. Some are still recovering from the *last* time they tried to pick up pointers from a tech-savvy person and nearly drowned in a sea of acronyms and geek speak.

I feel your pain. I do. Now, let's ease that pain:

The other side proposed:

Documents will be produced as single page TIFF files with multi-page extracted text or OCR. We will furnish delimited IPRO or Opticon load files and will later identify fielded information we plan to exchange.

Are they trying to screw you? Probably not.
Are you screwing yourself by accepting the proposed form of production? Yes, probably.

First, let's translate what they said to plain English.

"Documents will be produced as single page TIFF files…."
They are not offering you the evidence in anything like the form in which they created and used the evidence. Instead, they propose to print everything to a kind of electronic paper, turning searchable, metadata-rich evidence into non-searchable pictures of much (but not all) of the source document. These pictures are called TIFFs, an acronym for Tagged Image File Format. "Single page TIFF" means that each page of a document will occupy its own TIFF image, so reading the document will require loading and reviewing multiple images (as compared to, *e.g.*, a PDF document where the custom is for the entire document to be contained within one multipage image).

If you ever pithed a frog in high school biology, you know what it's like to TIFF a native document. *Converting a native document to TIFF images is lobotomizing the document.* By "native," I mean that the file that contains the document is in the same electronic format as when used by the software application that created the file. For example, the native form of Microsoft Word document is typically a file with the extension .DOC or .DOCX. For a Microsoft Excel spreadsheet, it's a file with the extension .XLS or .XLSX. For PowerPoints, the file extensions are .PPT or .PPTX. Native file formats contain the full complement of content and application metadata available to those who created and used the document. Unlike TIFF images, native files are *functional* files, in that they can be loaded into a copy of the software application that created them to replicate what a prior user saw, as well as affording a comparable ability to manipulate the data and access content that's made inaccessible when presented in non-native formats.

Think of a TIFF as a PDF's retarded little brother. I mean no offense by that, but TIFFs are not just differently abled; they are severely handicapped. Not born that way, but lamed and maimed on purpose. The other side downgrades what they give you, making it harder to use and stripping it of potentially-probative content.

Do they do this because they are trying to screw you? Probably not.
Does it screw you just the same? Well, yeah.

"[W]ith multi-page extracted text or OCR."

A native file isn't just a picture of the evidence*. It's the original electronic evidence*. As such, it contains <u>all</u> of the content of the document in an electronic form. Because it's designed to be electronically usable, it tends to be inherently electronically searchable; that is, whatever data it holds is encoded into the native electronic file, including certain data *about* the data, called **application metadata.** When an electronic document is converted to an image—TIFF—it loses its ability to be searched electronically and its application metadata and utility is lost. It's like photographing a steak. You can *see* it, but you can't smell, taste or touch it; you can't hear the sizzle, and you surely can't eat it.

Because converting to TIFF takes so much away, parties producing TIFF images deploy cumbersome techniques to restore some of the lost functionality and metadata. To restore a measure of electronic searchability, they extract text from the electronic document and supply it in a file accompanying the TIFF images. It's called "multi-page extracted text" because, although the single-page TIFFs capture an image of each page, the text extraction spans *all*of the pages in the document. A recipient runs searches against the extracted text file and then seeks to correlate the hits in the text to the corresponding page image.

If the source documents are scans of paper document, there's no electronic text to extract from the paper. Instead, the scans are subjected to a process called **optical character recognition** (OCR) that serves to pair the images of letters with their electronic counterparts and impart a rough approximation of searchability. OCR sucks, but it beats the alternative (no electronic searchability whatsoever).

"We will furnish delimited IPRO or Opticon load files…."
Whether extracted from an electronic source or cobbled together by OCR, the text corresponding to the images or scans is transferred in so-called "load files" that may also contain metadata about the source documents. Collectively, the load file(s) and document images are correlated in a database tool called a "review platform" or "review tool" that facilitates searching the text and viewing the corresponding image. Common review tools include Concordance, Summation and Relativity. There are many review tools out there, some you load on your own machines ('**behind the firewall**") and some you access via the Internet as **hosted** tools.

To insure that the images properly match up with extracted text and metadata, the data in the load files is "delimited," meaning that each item of information corresponding to each page image is furnished in a sequence separated by delimiters–just a fancy word for characters like commas, tabs or semicolons used to separate each item in the sequence. The delimiting scheme employed in the load files can follow any of several published standards for load file layout, including the most common schemes known as IPRO or Opticon.

"[A]nd will later identify fielded information we plan to exchange."
Much of the information in electronic records is fielded, meaning that is not lumped together with all the other parts of the record but is afforded its own place or space. When we fill out paper forms that include separate blanks for our first and last name, we are dividing data (our name) into fields: (first), (last). A wide array of information in and around electronic files tends to be stored as fields, e.g., e-mail messages separately field information like From, To, Date and Subject. If fielded information is not exchanged in discovery as fielded information, you lose the ability to filter information by, for example, Date or Sender in the case of an e-mail message or by a host of properties and metadata describing other forms of electronically stored information.

Additionally, the discovery process may necessitate the linking of various fields of information with electronic documents, such as Bates numbers, hash values, document file paths, extracted text or associated TIFF image numbers. There may be hundreds of fields of metadata and other data from which to select, though not all of it has any evidentiary significance or practical utility. Accordingly, the proposal to "later identify fielded information we plan to exchange" defers the identification of fielded information to later in the discovery process when presumably the parties will have a better idea what types of ESI are implicated and what complement of fields will prove useful or relevant.

Are they trying to screw you by not identifying fielded information?
No. They're just buying time

Does their delay screw you?
Maybe.

Re-collecting fielded information you didn't expect your opponent would ask for can be burdensome and costly. Waiting too long to seek fielded information from an opponent may prompt the opponent to refuse to belatedly collect and produce it.

So, are they trying to screw you by this proposal? I doubt it. Chances are they are giving you the dumbed down data because that's what they *always* give the other side, most of whom accept it neither knowing nor caring what they're missing. It may be the form of production their own lawyers prefer because their lawyers are reluctant to invest in modern review tools. It probably doesn't hurt that the old ways take longer and throw off more billable hours.

You may accept the screwed up proposal because, even if the data is less useful and incomplete, *you won't have to evolve*. You'll pull the TIFF images into your browser and painstakingly read them one-by-one, just like good ol' paper; all-the-while telling yourself that what you didn't get probably wasn't that important and promising yourself that next time, you'll hold out for the good stuff—the native stuff. Yeah, next time for sure. Definitely. Definitely.

———

Note: Readers should be careful not to confuse the production of ESI in native forms with the use of native applications to open and review the data. ***You don't use native apps to search and review native productions any more than you use a screwdriver as a hammer. Instead, you use review tools tailored to the task.*** While we're at it, you shouldn't let the redaction tail wag the production dog. Go ahead and use TIFFs and OCR for redaction if you wish; but, don't screw up the entire production because you want to use TIFFs to redact a handful of documents! As far as using documents in proceedings, go ahead and print out the few you'll use; but here again, don't get screwed by a TIFF production just so you can print something out. Last I checked, native documents printed out very nicely, too.

The Streetlight Effect in E-Discovery

by Craig Ball

[Originally published in Law Technology News, December 2012]

In the wee hours, a beat cop sees a drunken lawyer crawling around under a streetlight searching for something. The cop asks, "What's this now?" The lawyer looks up and says, "I've lost my keys." They both search for a while, until the cop asks, "Are you sure you lost them here?" "No, I lost them in the park," the tipsy lawyer explains, "but the light's better over here."

I told that groaner in court, trying to explain why opposing counsel's insistence that we blindly supply keywords to be run against the e-mail archive of a Fortune 50 insurance company wasn't a reasonable or cost-effective approach e-discovery. The "Streetlight Effect," described by David H. Freedman in his 2010 book *Wrong,* is a species of observational bias where people tend to look for things in the easiest ways. It neatly describes how lawyers approach electronic discovery. We look for responsive ESI only where and how it's easiest, with little consideration of whether our approaches are calculated to find it.

Easy is wonderful when it works; but looking where it's easy *when failure is assured* is something no sober-minded counsel should accept and no sensible judge should allow.

Consider *The Myth of the Enterprise Search*. Counsel within and without companies and lawyers on both sides of the docket believe that companies have the ability to run keyword searches against their myriad siloes of data: mail systems, archives, local drives, network shares, portable devices, removable media and databases. They imagine that finding responsive ESI hinges on the ability to incant magic keywords like Harry Potter. *Documentum Relevantus!*

Though data repositories may share common networks, they rarely share common search capabilities or syntax. Repositories that offer keyword search may not support Boolean constructs (queries using "AND," "OR" and "NOT"), proximity searches (Word1 near Word2), stemming (finding "adjuster," "adjusting," "adjusted" and "adjustable") or fielded searches (restricted to just addressees, subjects, dates or message bodies). Searching databases entails specialized query languages or user privileges. Moreover, different tools extract text and index such extractions in quite different ways, with the upshot being that a document found on one system will not be found on another using the same query.

But the Streetlight Effect is nowhere more insidious than when litigants use keyword searches against archives, e-mail collections and other sources of indexed ESI,

That Fortune 50 company—call it All City Indemnity—collected a gargantuan volume of e-mail messages and attachments in a process called "message journaling." Journaling copies every message traversing the system into an archive where the messages are indexed for search. Keyword searches only look at the index, not the messages or attachments; so, if you don't find it in the index, you won't find it at all.

All City gets sued every day. When a request for production arrives, they run keyword searches against their massive mail archive using a tool we'll call *Truthiness*. Hundreds of big companies use *Truthiness* or software just like it, and blithely expect their systems will find all documents containing the keywords.

They're wrong…or in denial.

If requesting parties don't force opponents like All City to face facts, All City and its ilk will keep pretending their tools work better than they do, and requesting parties will keep getting incomplete productions. To force the epiphany, consider an interrogatory like this:

For each electronic system or index that will be searched to respond to discovery, please state:

a. **The rules employed by the system to tokenize data so as to make it searchable;**
b. **The stop words used when documents, communications or ESI were added to the system or index;**
c. **The number and nature of documents or communications in the system or index which are not searchable as a consequence of the system or index being unable to extract their full text or metadata; and**
d. **Any limitation in the system or index, or in the search syntax to be employed, tending to limit or impair the effectiveness of keyword, Boolean or proximity search in identifying documents or communications that a reasonable person would understand to be responsive to the search.**

A court will permit "discovery about discovery" like this when a party demonstrates why an inadequate index is a genuine problem. So, let's explore the rationale behind each inquiry:

a. Tokenization Rules - When machines search collections of documents for keywords, they rarely search the documents for matches; instead, they consult an index of words extracted from the documents. Machines cannot read, so the characters in the documents are identified as "words" because their appearance meets certain rules in a process called "tokenization." Tokenization rules aren't uniform across systems or software. Many indices simply don't index short words (e.g., acronyms). None index single letters or numbers.

Tokenization rules also govern such things as the handling of punctuated terms (as in a compound word like "wind-driven"), case (will a search for "roof" also find "Roof?"), diacriticals (will a search for Rene also find René?) and numbers (will a search for "Clause 4.3" work?). Most people simply *assume* these searches will work. Yet, in many search tools and archives, they don't work as expected, or don't work at all, unless steps are taken to ensure that they will work.

b. Stop Words – Some common "stop words" or "noise words" are simply excluded from an index when it's compiled. Searches for stop words fail because the words never appear in the index. Stop words aren't always trivial omissions. For example, "all" and "city" were stop words; so, a search for "All City" will fail to turn up documents containing the company's own name! Words like side, down, part, problem, necessary, general, goods, needing, opening, possible, well, years and state are examples of common stop words. Computer systems typically employ dozens or hundreds of stop words when they compile indices.

Because users aren't warned that searches containing stop words fail, they mistakenly assume that there are no responsive documents when there may be thousands. A search for "All City" would miss *millions* of documents at All City Indemnity (though it's folly to search a company's files for the company's name).

c. Non-searchable Documents - A great many documents are not amenable to text search without special handling. Common examples of non-searchable documents are faxes and scans, as well as TIFF images and some Adobe PDF documents. While no system will be flawless in this regard, it's important to determine *how much* of a collection isn't text searchable, *what's* not searchable and whether the portions of the collection that aren't searchable are of *particular importance* to the case. If All City's adjusters attached scanned receipts and bids to e-mail messages, the attachments aren't keyword searchable absent optical character recognition (OCR).

Other documents may be inherently text searchable but not made a part of the index because they're password protected (i.e., encrypted) or otherwise encoded or compressed in ways that frustrate indexing of their contents. Important documents are often password protected.

d. Other Limitations - If a party or counsel knows that the systems or searches used in e-discovery will fail to perform as expected, they should be obliged to affirmatively disclose such shortcomings. If a party or counsel is uncertain whether systems or searches work as expected, they should be obliged to find out by, e.g., running tests to be reasonably certain.

No system is perfect, and perfect isn't the e-discovery standard. Often, we must adapt to the limitations of systems or software. But you have to know what a system *can't* do before you can find ways to work around its limitations or set expectations consistent with actual capabilities, not magical thinking and unfounded expectations.

Master Bates Numbers in E-Discovery

by Craig Ball

[Originally published on the Ball in Your Court blog, December 28, 2012]

The sophomoric title of this post strives to underscore the trial bar's proclivity to self-abuse when it comes to the petulant insistence on Bates numbers embossed on each "page" of ESI produced in discovery. Comments to a recent post on native production and Bates numbers prompted my recall of a famous psychology experiment, where McGill University's James Olds and Peter Milner set up an apparatus electronically stimulating the pleasure (i.e., orgasm) center of rats' brains. Given control of their stimulus, the rats began virtually tossing one off about *2,000 times an hour,* ignoring food, water and hockey. This is why men must rest between orgasms; else, we would die of dehydration in adolescence.

The first comment on Bates numbers came from Mike McBride, who is an accomplished blogger and a veteran of years in IT and litigation support. Mike wrote:

"One of the real challenges in moving firms to native production is the Bates Label, so I'll be interested in seeing how you deal with that. Discussions I've had about native production typically start and end with "how would I keep track of which document is which if it's not labeled?" Unfortunately, that question exposes the larger issue behind it, namely that attorneys are still printing their documents and reviewing them in that format, even when they have a sophisticated tool available to them to do review. When you print out a copy of a document, removing all of the available metadata, the label becomes the only way to truly tie it back to its electronic version. Until we get them to stop printing, I'm afraid native production is still a step too far."

Mike wisely notes that the TIFF wars often rage over Bates labels. Lawyers adore Bates labels. It comforts us to shrink all those documents we reluctantly hand over to our adversaries and emboss a number and cautionary message in their new margins. "ABC00123-PRODUCED SUBJECT TO PROTECTIVE ORDER," they shout, but mean: "*DON'T EVEN THINK* about giving this to another plaintiffs' lawyer because we know you greedy scum would violate the Court's order in a New York minute but for this warning."

Lawyers who revere Bates labeling as a citadel against violation of protective orders ignore the accomplishments of two inventors: Betty Nesmith Graham and Chester Carlson.

Late in the 1950's, Dallas secretary, Bette Nesmith Graham, invented Liquid Paper in her kitchen blender. (She also 'invented' son Mike Nesmith, who would go on to fame as one of the four Monkees in the invented 1960s TV rock band of the same name). Were one so inclined, correction fluids like Liquid Paper make it child's play to remove or alter Bates numbers.

Chester Carlson invented photocopiers (i.e., Xerox machines), so the magic that enables producing parties to shrink documents and emboss Bates labels makes it simple to mask those Bates labels and resize document to their true dimensions.

And don't get me started on Photoshop and other image editing applications!

In short, Bates labels are an illusory safeguard against malfeasance—and have been since Eisenhower was President and Don and Betty Draper were a couple. Lawyers respect court orders because *we value our licenses and reputations*, not because some damned fool embosses a cautionary legend.

So, getting back to Mike's comment that lawyers want Bates numbers to relate printouts to native source files, it couldn't be simpler: *you produce in native forms and simply emboss Bates numbers when the evidence is downgraded to paper or TIFF for use in proceedings.* That is, you add Bates numbers only where and when they're worth the trouble.

Here's one way to skin that cat, excerpted from the exemplar production protocol made an appendix to my paper, *Beyond Data About Data: The Litigator's Guide to Metadata*):

Unique Production Identifier (UPI)

a) Other than paper originals, images of paper documents and redacted ESI, no ESI produced in discovery need be converted to a paginated format nor embossed with a Bates number.

b) Each item of ESI (e.g., native file, document image or e-mail message) shall be identified by naming the item to correspond to a Unique Production Identifier according to the following protocol:

   i. The first four (4) characters of the filename will reflect a unique alphanumeric designation identifying the party making production;

   ii. The next nine (9) characters will be a unique, sequential numeric value assigned to the item by the producing party. This value shall be padded with leading zeroes as needed to preserve its length;

   iii. The final five (5) characters are reserved to a sequence beginning with a dash (-) followed by a four digit number reflecting pagination of the item when printed to paper or embossed when converted to an image format for use in proceedings or when attached as exhibits to pleadings.

   iv. By way of example, a Microsoft Word document produced by Acme in its native format might be named: ACME000000123.doc. Were the document printed out for use in deposition, page six of the printed item must be embossed with the unique identifier ACME000000123-0006.

The original name of the file is furnished in a load file, along with other relevant system metadata.

The logic behind this is simple. Parties tend to produce many more items in discovery than are used in proceedings (when the page-by-page identification function of Bates numbers is advantageous). So, the party who changes the form of evidence is obliged to tie the altered form to the original and append the pagination.

Those who object that this renaming is burdensome or expensive haven't done their homework. Filenames have been used for decades as a means to assign Bates numbers. It's standard operating procedure in TIFF and load file productions. As to cost, good tools to do the job cost nothing. An excellent free file renaming tool is Bulk Rename Utility, available at http://www.bulkrenameutility.co.uk. It allows you to add custom incrementing numbers and a protective legend like "Subject to Protective Order" in the name. And no, renaming a file this way does *not* alter its content, hash value or last modified date.

I make that last point because I lately saw an affidavit from a Director at a prominent national e-discovery consultancy where the muckety-muck swore that "adding bates number [sic] to the file name would alter the metadata file name, date and time last edited and other fields."

Renaming a file absolutely *does not* alter its "date and time last edited." To do so, you must open the file and save it, neither action being necessary to rename a file. Test it, if you have doubts.

Of course, the same affiant swore that an MD5 hash algorithm "counts all of the bits and bytes of an electronic file. The total number of bits and bytes is referred to the hash value." [Sic]…and makes me sick. Like anyone who understands this stuff reasonably well, I think of the total number of bits and bytes as being the file's *size,* not its MD5 hash.

Shall we chalk up this false swearing to "verbal masturbation," in keeping with our theme? Fine, but beware. This sort of "say-it-irrespective-of-the-truth" strategy is straight out of the Case against Native playbook. Going native takes bread from the mouths of many riding the "TIFF everything" gravy train. They won't go down easy, and they don't fight clean.

So, get hold of yourself, and master Bates numbers to insure litigants get all the satisfaction that comes from native production. If the sexually-tinged joshing in this post offends, please know that I've only ribbed for your pleasure.

Busted! How Happy Accidents Prove Data Theft

by Craig Ball

[Originally published on the Ball in Your Court blog, January 26, 2013]

A big part of my practice is assisting courts and lawyers in cases where it's alleged that a departing employee has walked off with proprietary data. There's quite a lot of that. Studies in the U.S. and abroad suggest that some two-thirds of departing white collar employees leave with proprietary data. So, it seems data theft is the norm.

Of course, not all data leaves with the requisite *scienter* ("evil intent") to be called theft. In this wired world, who doesn't have data on thumb drives, phones, tablets, backup drives, webmail accounts, legacy devices, media cards, CDs, DVDs, floppy disks and good ol' paper? You work for a company a while and you're going to end up with their stuff strewn all over your devices and repositories. But, few data theft lawsuits stem from stale data on forgotten media.

The "classic" data theft scenario is the after-hours mass movement of copious quantities of closely-guarded internal documents to an external USB hard drive or capacious thumb drive. While such actions look dastardly at first blush, a few dimmer bulbs may actually act with a pure heart, intending to take only their personal data (like family photos or music), but dragging entire folder families that also hold corporate ESI.

I tend to be skeptical of such claims unless the usage patterns that follow and other forensic evidence bear out the "I really thought it was just my stuff" defense. It's not hard to tell the difference, so long as devices aren't lost or corrupted.

But you may be wondering: How do forensic examiners determine data was taken, and how do they identify and track storage devices used to carry away ESI?

This post is offered as a *general* introduction to *selected* aspects of Windows Registry and artifact analysis and peculiarities of Windows MAC dates and times. The goal is to introduce you to same, not equip you to conduct forensic exams or march into court assuming this is all you need to know. With that fainthearted disclaimer behind us….

Computer Forensics: A Confluence of Happy Accidents

You can roughly divide the evidence in a computer forensic examination between evidence generated or collected by a user (*e.g.,* an Excel spreadsheet or downloaded photo) and evidence created by the system which serves to supply the context required for authenticating and weighing user-generated evidence. User-generated or -collected evidence tends to speak for itself without need of expert interpretation. In contrast, artifacts created by the system require expert interpretation, in part because such artifacts exists to serve purposes having nothing to do with logging a user's behavior for use as evidence in court. Most forensic artifacts arise as a consequence of a software developer's effort to supply a better user experience and improve system performance. Their probative value in court is a happy accident.

For example, on Microsoft Windows systems, a forensic examiner may look to machine-generated artifacts called LNK files, prefetch records and Registry keys to determine what files and applications a user accessed and what storage devices a user attached to the system.

LNK files (pronounced "link" and named for their file extension) serve as pointers or "shortcuts" to other files. They are similar to shortcuts users create to conveniently launch files and applications; but, these LNK files aren't user-created. Instead, the computer's file system routinely creates them to facilitate access to recently used files and stores them in the user's RECENT folder. Each LNK file contains information about its target file that endures even when the target file is deleted, including times, size, location and an identifier for the target file's storage medium. I'm sure Microsoft didn't intend that Windows retain information about deleted files in orphaned shortcuts; but, there's the happy accident–or maybe not so happy, if you are the one caught in a lie because your computer was trying to better serve you.

Similarly, Windows seeks to improve system performance by tracking the recency and frequency with which applications are run. If the system knows what applications are most likely to be run, it can "fetch" the programming code those applications need in advance and pre-load them into memory, speeding the execution of the program. Thus, records of the last 128 programs run are stored in series of so-called "prefetch" files. Because the metadata values for these prefetch files coincide with use of the associated program, by another happy accident, forensic examiners may attest to, say, the time and date a file wiping application was used to destroy evidence of data theft.

Two final examples of how much forensically-significant evidence derives from happy accidents are the USBSTOR and DeviceClasses records found in the Windows System Registry hive. The Windows Registry is the central database that stores configuration information for the system and installed applications—it's essentially everything the operating system needs to "remember" to set itself up and manage hardware and software. The Windows Registry is huge and complex. Each time a user boots a Windows machine, the registry is assembled from a group of files called "hives." Most hives are stored on the boot drive as discrete files and one—the Hardware hive—is created anew each time the machine inventories the hardware it sees on boot.

When a user connects an external mass storage device like a portable hard drive or flash drive to a USB port, the system must load the proper device drivers to enable the system and device to communicate. To eliminate the need to manually configure drivers, devices have evolved to support so-called Plug and Play capabilities. Thus, when a user connects a USB storage device to a Windows system, Windows interrogates the device, determines what driver to use and—importantly—*records information about the device and driver pairing* within a series of keys stored in the

ENUM/USBSTOR and the DeviceClasses "keys" of the System Registry hive. In this process, Windows tends to store the date and time of both the earliest and latest attachments of the USB storage device.
Windows is not recording the attachment of flash drives and external hard drives to enable forensic examiners to determine when employees attached storage devices to steal data! Presumably, the programmer's goal was to speed selection of the right drivers the next time the USB devices were attached; but, the happy accident is that the data retained for a non-forensic purpose carries enormous probative value when properly interpreted and validated by a qualified examiner.

The Scene of the "Crime"

Imagine you're an employee who has grown disenchanted with your employer. Perhaps the company you helped build has changed management, and you feel marginalized. Maybe you were RIF'ed or passed over for promotion or your latest bonus was a disappointment. For whatever reason, your eye is on the door. Then, opportunity knocks. Maybe a competitor or former co-worker calls with a job offer, or you decide to launch a competing venture. Perhaps you're just a sentimental soul and want to keep a complete record of all the fine-but-under-appreciated work you slaved to produce for your soon-to-be-former employer.

See, most people aren't data thieves by nature. A fair amount of self-serving rationalization goes into getting them to the point of persuading themselves it's not *really* stealing…not exactly. It's more like protecting the fruits of their labor. Of course, if they didn't think it was questionable behavior, they could have sought permission, and they sure wouldn't have come in late at night to make the copies or gone to other lengths to cover their tracks.
Do you get the feeling I've investigated a bunch of these cases?

So, my dear data pirate, now that you've decided to liberate your data, how will you package it to go? Will you e-mail selected items to your personal web mail account? Will you burn a CD? Or will you copy selected folders to an external USB device like a thumb drive or hard drive? The last method is the one most often seen.

But before you start moving data, you've got to figure out what you want to take with you. You'll probably start opening folders and checking the contents of files. You may begin amassing the files you want to take in a folder or Zip file or perhaps you'll drag your entire Documents folder to an external drive. In any event, examiners often see evidence of particularized interest in proprietary files manifested as LNK files in the user's RECENT subfolder. These LNK files offer clues to what was attached, what was accessed and when. Examiners may also see MRU (for **M**ost **R**ecently **U**sed) file records in the Registry for various file types of interest.

Another clue that company data was copied to an external storage device is the timeline of events that can be gleaned from examining the system metadata values of files and folders. Years ago, such a timeline in a data theft case might have been constructed primarily from Last Accessed dates, reflecting the last time a file was “touched” by the user or operating system. That’s less feasible today; here’s why:

All computer operating systems employ a set of routines called the File System that serve as the prosaic plumbing handling routine file management tasks. The Windows family of PC operating systems has long employed a file system called **NTFS** (for New Technology File System). NTFS tracks and records several date and time values for each file it manages. These have customarily been called **MAC dates**, for Last **M**odified, Last **A**ccessed and **C**reated dates, but could as easily be termed MACE dates, acknowledging another time value called the **E**ntry Modified date.

MAC Dates are a frequent source of confusion in computer forensics and e-discovery because the meaning accorded Created and Accessed in common parlance is off a bit from their meanings in Windows World. In NTFS, a document’s Created date *could* coincide with the date the document was authored by the user, but it could also signify when a template used to create a document was authored by someone else or when the document was copied from another disk or drive (as “Created” means created *on that storage volume).*

Copy a file you authored on Monday to an external hard drive on Friday and the file’s Created Date on the external hard drive will likely be Friday (unless you used a copying tool that changes the date back to Monday). See how flaky this can get if you’re not sure what you’re seeing?

The flakiest of the MAC dates is the Last Accessed date. In the past, the Last Accessed date signified when a user last opened a file or the last time an antivirus program examined the file for malware, or even the last time the file was previewed while exploring a list of files in Windows Explorer. Last Accessed dates were being updated all the time for a host of reasons, and all that updating consumed computing resources, slowing things down.

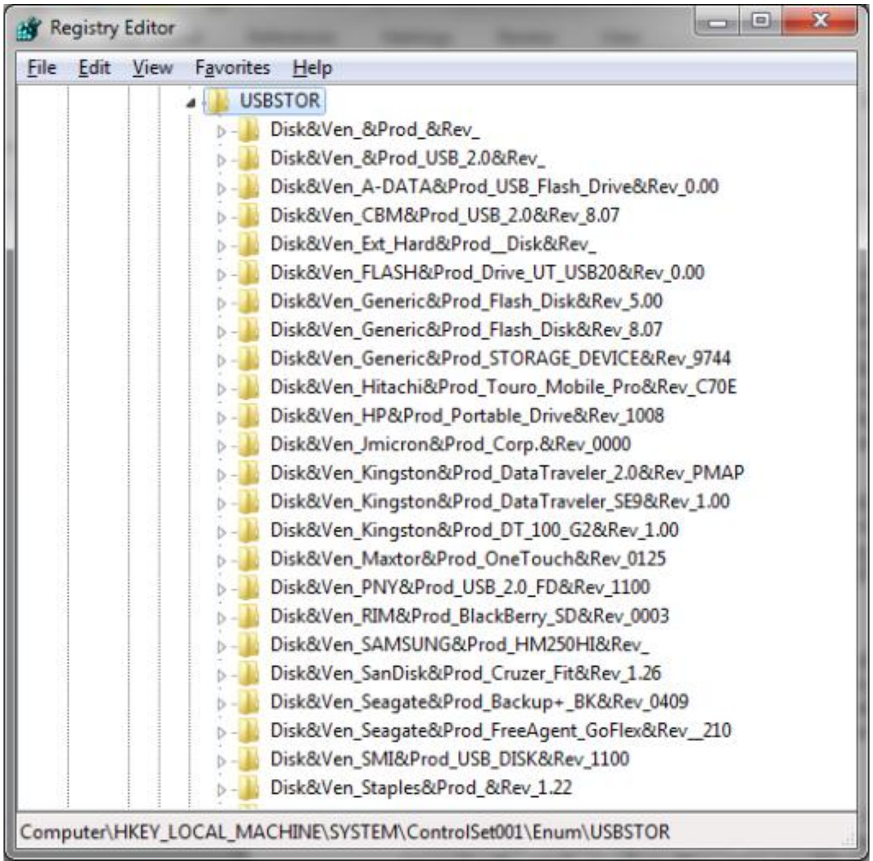
Slow doesn’t sell computers, so Microsoft hated all that automatic updating. Accordingly, when Microsoft introduced its Vista operating system about six years ago, it unceremoniously turned off the routine updating of Last Accessed dates. A user could tinker with the Registry and turn it back on; but if the user doesn’t know updating is turned off, what are the odds a user will edit the Registry to turn it on? Who but forensic examiners would know or care?

Thus, an unreliable metadata value got even more unreliable in Vista, Windows 7 and Windows 8. Windows XP updated Last Accessed dates constantly, and Windows 7 updates them sporadically, such as when files are modified. If an external drive is attached to multiple machines running different versions of Windows, the metadata values will update inconsistently. In sum, updated Last Accessed dates may reliably confirm that a file was touched by a user or machine, but the absence of updated values can't reliably rule it out.

Okay, Byte Bandit, you've assembled what you want to take and unpacked that brand new Seagate Free Agent GoFlex drive you got at Best Buy for ninety bucks, marveling that one terabyte of data can fit onto something the size of a deck of cards. You plug the GoFlex into the USB port of your company laptop and wait for Windows to tell you that it sees the drive and ask you what you want to do with it.

In the brief time that the drive spun up, Windows said "Who goes there?" and the drive responded with its name, rank and serial number; that is, with enough information that Windows could locate and load the right driver to allow the laptop to communicate with this new USB mass STORage device. Armed with this information, Windows dutifully makes a record of the attachment in a Registry key aptly called "USBSTOR" as well as several other Registry keys and in a log called C:\Windows\inf\setupapi.dev.log. Windows also records the date and time of this first attachment in Universal Coordinated Time (UTC), which is to say *almost* in Greenwich Mean Time. Next time someone asks, "What time was it in Leicester Square when you first used your SanDisk Cruzer thumb drive?" you're all set!" Finally, if Windows cannot determine the drive's serial number, it just makes one up. Really. You just *gotta* love Windows! To be fair, it makes up a serial number for a perfectly valid reason, but it's made up all the same.

If you'd like to see what the list of connected USB devices looks like on your machine, simply click the Start button on your Windows machine (unless you're using Windows 8 and can't find the Start button anymore) and enter "Regedit." When the Registry Editor launches, drill down through HKEY_LOCAL_MACHINE to SYSTEM and select any numbered ControlSet, then drill down further through ENUM to USBSTOR. You should



```
USBSTOR
    Disk&Ven_&Prod_&Rev_
    Disk&Ven_&Prod_USB_2.0&Rev_
    Disk&Ven_A-DATA&Prod_USB_Flash_Drive&Rev_0.00
    Disk&Ven_CBM&Prod_USB_2.0&Rev_8.07
    Disk&Ven_Ext_Hard&Prod__Disk&Rev_
    Disk&Ven_FLASH&Prod_Drive_UT_USB20&Rev_0.00
    Disk&Ven_Generic&Prod_Flash_Disk&Rev_5.00
    Disk&Ven_Generic&Prod_Flash_Disk&Rev_8.07
    Disk&Ven_Generic&Prod_STORAGE_DEVICE&Rev_9744
    Disk&Ven_Hitachi&Prod_Touro_Mobile_Pro&Rev_C70E
    Disk&Ven_HP&Prod_Portable_Drive&Rev_1008
    Disk&Ven_Jmicron&Prod_Corp.&Rev_0000
    Disk&Ven_Kingston&Prod_DataTraveler_2.0&Rev_PMAP
    Disk&Ven_Kingston&Prod_DataTraveler_SE9&Rev_1.00
    Disk&Ven_Kingston&Prod_DT_100_G2&Rev_1.00
    Disk&Ven_Maxtor&Prod_OneTouch&Rev_0125
    Disk&Ven_PNY&Prod_USB_2.0_FD&Rev_1100
    Disk&Ven_RIM&Prod_BlackBerry_SD&Rev_0003
    Disk&Ven_SAMSUNG&Prod_HM250HI&Rev_
    Disk&Ven_SanDisk&Prod_Cruzer_Fit&Rev_1.26
    Disk&Ven_Seagate&Prod_Backup+_BK&Rev_0409
    Disk&Ven_Seagate&Prod_FreeAgent_GoFlex&Rev__210
    Disk&Ven_SMI&Prod_USB_DISK&Rev_1100
    Disk&Ven_Staples&Prod_&Rev_1.22

Computer\HKEY_LOCAL_MACHINE\SYSTEM\ControlSet001\Enum\USBSTOR
```

see something that looks like the image below. (If it looks *exactly* like this, GET THE HECK OUT OF MY HOUSE!)

Your Registry Editor won't list the attachment times in USBSTOR; however, tools used by forensic examiners parse the various keys to assemble the data generally permitting a reliable determination of the first and last attachment dates and times (along with other data) for specific USB mass storage devices. *[DO NOT edit your Registry unless you know EXACTLY what you're doing. Just close the window.]*

There you have it, a simplified explanation of some of the methods forensic examiners use to track and trace data theft in Windows systems. When seeking to preserve ESI in cases of suspected data theft, remember that *the lion's share of the forensically revealing data is not stored on the media used to transfer the data* but principally resides on the systems to which the storage medium attached. Endeavor to secure <u>all</u> such evidence items and have them forensically imaged by a qualified examiner before exploring contents.

Lawyers interested in more information on this topic might enjoy my (free) *First Responder's Guide to Employee Data Theft.* Examiners interested in more about Windows Registry analysis should see Harlan Carvey's excellent *Windows Registry Forensics: Advanced Digital Forensic Analysis of the Windows Registry* and download a copy of the also excellent (free) *SANS Windows Artifact Analysis poster*

The Case for Native

by Craig Ball

[Originally published in Law Technology News, February 2013]

Is there any form of ESI production worse than .tiffs and load files? If you've experienced the ease of e-discovery with tools purpose-built for native review, you know what I'm talking about. Once you "go native," you'll never go back!

By native, I mean data in the original electronic formats the producing party uses for, e.g., email, word processing, spreadsheets, and presentations.

A native file is inherently electronically searchable and functional until it's converted to .tiff images, when it loses both searchability and functionality. It's like photographing a steak. You can see it, but you can't smell, taste, or touch it, you can't hear the sizzle, and you surely can't eat it.

Because converting to .tiff takes so much away, parties producing .tiff images attempt to restore a measure of electronic searchability by extracting text from the electronic document and supplying it in a load file accompanying the .tiff images. A recipient must then run searches against the extracted text file and seek to correlate the hits in the text to the corresponding page image. It's clunky, costly, and incomplete.

The irony of .tiff and load file productions is that what was once a cutting-edge technology has become an albatross around the neck of electronic data discovery. To understand how we got to this unenviable place requires a brief history lesson.

Before the turn of the century, when most items sought in discovery were paper documents, .tiff and load file productions made lawyers' lives easier by grafting rudimentary electronic searchability onto unsearchable paper documents. Documents were scanned to .tiff images and coded by reviewers, and their text was extracted via optical character recognition (OCR) software. It was expensive and crude, but speedier than poring over thousands or millions of pieces of paper.

The coding and text had to be stored in separate files because .tiff images are just pictures of pages, incapable of carrying added content. So, in "single page .tiff" productions, each page of a document became its own image file, another file held aggregate extracted OCR text, and yet another held the coded data about the data, i.e., its metadata.

The metadata would include information about the content and origin of the paper evidence, along with names and locations of the various images and files on the media (i.e., CD or DVD) used to transmit same. Thus, adding a measure of searchability yielded a dozen or more electronic files to carry the pieces of a 10-page document.

To put Humpty Dumpty back together again demanded a database and picture viewer capable of correlating the extracted text to its respective page image and running word searches. Thus was born a new category of document management software called "review platforms." Because the files holding the document's OCR'ed text and metadata were destined to be loaded onto a review platform, they came to be called "load files."

Different review platforms used different load file formats to order and separate information according to guidelines called load file specifications. Load files are plain text files employing characters called delimiters to separate the various information items in the load file. Thus, a load file specification might require that information about a document be transmitted in the order: Box No., Beginning Bates No., Ending Bates No., Date, and Custodian. The resulting single line of text delimited by, e.g., commas, would appear: 57,ABC0003123,ABC0003134,19570901,Ball C.

Load files were a headache. But we put up with the pain because adding searchability to unsearchable paper documents was worth it. A stone ax is better than no ax at all.

Because large document cases and attorney review pyramids were integral to law firm growth and profitability, lawyers invested in .tiff review platforms, and service providers emerged to compete for lucrative scanning and coding work. The electronic data discovery industry was born, circa 1987.

Fast forward to 2013, and hardly any documents are born on paper. Today, we seek electronically stored information, viz., email, word-processed documents, spreadsheets, presentations, and databases. With paper, what you see is what you get.

By contrast, ESI divides its digital goodness between information readily seen and information requiring a mouse click or two to view. Documents are layered, multi-media and multi-dimensional, and much ESI defies characterization as a document. Replete with embedded formulae, appended comments, tracked changes, and animated text, ESI thumbs its nose at the printed page.

Despite a sea change in what we seek to discover, lawyers resolutely refuse to embrace modern forms of production. They cling to .tiff imaging and load files, downgrading ESI's inherent searchability and eviscerating the multi-dimensional character of ESI. Thus, an obsolete technology that once made evidence easier to find now deep sixes probative content.

Producing parties dismiss this lost content as "just metadata," as if calling it metadata makes it something you'd scrape off your shoe. In fact, they fear such "metadata" may reveal privileged attorney-client communications (which should clue you in that it's more than just machine-generated minutiae). Producing parties have blithely and blindly been erasing this content for years without legal justification or disclosure in privilege logs.

When a producing party insists on converting ESI to .tiff images over a requesting party's objection, they often rely on *Federal Rules of Civil Procedure* 34(b)(2)(E)(ii), which obliges parties to produce ESI in "the form or forms in which it is ordinarily maintained or in a reasonably usable form or forms." Courts have struggled with the notion of "reasonably usable," but haven't keyed into the fact that .tiff imaging destroys user-generated content. Producing parties are happy to expunge content that may hurt their position and to postpone purchasing software supporting native review, so they've gotten good at making the case against native production.

Requesting parties seeking native production back down too easily because they're desperate to get moving and uncertain how to make the case for native production. Courts tend to be swayed by the argument, "We've always done it this way," without considering why .tiff imaging came into wide use and why its use over objection has become unfair, unwise, and wasteful.

The case against native usually hinges on four claims:

1. You can't Bates label native files.
2. Opponents will alter the evidence.
3. Native production requires broader review.
4. Redacting native files changes them.

Each claim carries a grain of truth swaddled in bunk. Let's debunk them:

1. *You can't Bates label native files*. Nonsense! It's simple and cheap to replace, prepend, or append an incrementing Bates-style identifier to the filename of all items natively produced. An excellent free file renaming tool is Bulk Rename Utility, available at www.bulkrenameutility.co.uk. You can include a protective legend, such as "Subject to Protective Order" in the name; and, no, renaming a file this way does not alter its content, hash value, or last modified date. If the other side grouses that it's burdensome to change file names to Bates numbers, remind them they've long used Bates numbers as file names in .tiff image productions.

It's indeed difficult to emboss Bates numbers on every page of a native file until it's printed or imaged. Yet many forms of ESI (e.g., email, spreadsheets, social networking content, video, and sound files) don't lend themselves to paged formats and will never be Bates-labeled.

We don't put exhibit labels on every item produced in discovery because only a tiny fraction of production will be introduced into evidence. Likewise, little ESI produced in discovery is used in proceedings. When it is, simply agree that file names and page numbers will be embossed on images or printouts.

Sure, file names can be altered, but changing a Bates number or removing a protective legend from a .tiff image or printout is child's play using software found on any computer. Demanding that Bates labeling for ESI be tamperproof is demanding more than was required of .tiff or paper productions.

2. *Opponents will alter the evidence*. Alteration of evidence is not a new hazard, nor one unique to ESI. We never objected to production of photocopies because paper is so easy to forge, rip, and shuffle. Tiffs are just pictures, principally of black and white text. What could be easier to manipulate in the Photoshop era?

Though any form of production is prey to unscrupulous opponents, native productions support quick, reliable ways to prevent and detect alteration. Simply producing native files on read-only media (e.g., CDs and DVDs) guards against inadvertent alteration, and alterations are easily detected by comparing digital fingerprints of suspect files to the files produced.

Counsel savvy enough to seek native production should be savvy enough to refrain from poor evidence handling practices like reviewing native files using native applications that tend to alter the evidence.

3. *Native production requires broader review*. Native forms hold content (such as animated text in presentations and formulae in spreadsheets) added by users but not visible via .tiff. But animated text and formulae aren't what concern your opponent.

The other side worries most about embedded commentary in documents — those candid communications between users and collaborators that are quietly stripped away when imaged. From an evidentiary standpoint, these aren't different from Post-It notes or email between key custodians.

It's crucial to help the court understand that the information stripped away is user-contributed content, and that a form of production isn't reasonably usable if it destroys the information. If opposing counsel argues they put some of the excised content into load files, that's disingenuous: If you cannot see a comment or alteration in context, its meaning is often impossible to divine.

Your opponents may also be reluctant to concede their obsolete tools don't show contemporary content. Fearful that your tools might show content their tools miss, they jettison content rather than upgrade tools.

4. *Redacting native files changes them*. Indeed, that's the whole idea. So the argument that the integrity of native productions will be compromised by removing privileged or protected content is silly! Instead, the form of production for items requiring redaction should be that form or forms best suited to efficient removal of privileged or protected content without rendering the remaining content wholly unusable.

Some native file formats support redaction brilliantly; others do not. In the final analysis, the volume of items redacted tends to be insignificant. Accordingly, the form selected for redaction shouldn't dictate the broader forms of production when native forms have such distinct advantages.

Don't let the redaction tail wag the production dog. If they want to redact in .tiff or PDF, let them, but only for the redacted items and only when they restore searchability after redaction.

Cast Off the Albatross! Tiff production had its day. Now, .tiff dumbs ESI down to the level of paper just so we can use old, familiar tools and workflows. Native production isn't simply better, it's cheaper, too. Why pay to convert native forms to .tiff and load files? Smaller native file sizes also trim the cost of ingestion and storage.

Tiff: You get less, pay more and destroy evidence to boot. Isn't it time to go native?

How to Bates Label Native Production

Nowhere on "La Joconde" — that most famous of all Leonardo da Vinci master-works — does it say "Mona Lisa." Yet, despite theft and 100 years of efforts to "redact" her, using everything from acid to teacups, the world knows Mrs. Giocondo without stamping "Mona Lisa" across her enigmatic smile. Neither must we downgrade or deface electronic evidence produced in its native forms to retain the benefits once derived from Bates-stamping paper documents.

A native file can be given almost any name without altering its contents or changing its hash value (aka its digital "fingerprint"). So, it's fast, free, and easy to rename an electronic file to carry any Bates-style identifier — even a legend like "Produced Subject to Protective Order" — so long as the length of the name stays under 255 characters.

Tips for Bates labeling native production:

► When replacing a file's name (versus prepending or appending an identifier in the name), preserve and produce a record of the original and substitute names.

► Establish an identification naming protocol where, e.g., the first four characters identify the producing party; the next nine are reserved to a unique, sequential numeric value (padded with leading zeroes); and the final five include a separator (i.e., hyphen) and a four-digit number reflecting pagination that is required to be embossed only when the file must be printed to paper or reduced to an image format for use in proceedings or as exhibits.

► If you include a truncated hash value in the filename (e.g., the first and last four digits of the file's MD5 hash value), all parties gain a portable, reliable means to confirm the electronic file is authentic, unchanged, and properly paired with the right name cum Bates identifier. You can't do *that* with printed Bates numbers!

Eight Tips to Quash the Cost of E-Discovery

by Craig Ball

[Originally published on the Ball in Your Court blog, March 21, 2013]

This really happened:
Opposing counsel supplied an affidavit stating it would take thirteen years to review 33 months of e-mail traffic for thirteen people. Counsel averred there would be about 950,000 messages and attachments after keyword filtering. Working all day, every day reviewing 40 documents per hour, they expected first level review to wrap up in 23,750 hours. A more deliberate second level review of 10-15% of the items would require an additional two years. Finally, counsel projected another year to prepare a privilege log.
Cost: millions of dollars.

The arithmetic was unassailable, and a partner in a prestigious law firm swore to its truth under oath.

This could have happened:
On Monday afternoon, an associate attached a hard drive holding 33 months of e-mail for thirteen custodians to the USB port of her computer and headed home. Overnight, e-discovery review software churned through the messages and attachments indexing their contents for search and de-duplicating redundant data. The next morning, the associate identified responsive documents using keywords and concept clustering. She learned the lingo, mastered the acronyms and identified common misspellings. She found large swaths of irrelevant data that could be safely eliminated from the collection and began segregating responsive and non-responsive items. By lunchtime on Wednesday, the software started asking whether particular items were responsive. Before she called it a day, the associate ceded much of the heavy lifting to the program's technology-assisted review capabilities and shifted her attention to searching for lawyers' names and e-mail domains to flag privileged communications. She spent Thursday afternoon sampling items the computer identified as non-responsive to be assured of the quality of review. Before she called it a day, the associate tasked the software to generate a production set and a privilege log for partner review on Friday and wondered if it might be a good weekend to head to the beach.
Cost: 40 associate hours.

These two scenarios contrast the gross disparity in review costs and time between lawyers who approach e-discovery in ignorance and those who do so with skill. The Luddite lawyer who knows nothing of modern methods misleads the court and cheats the client. The adept associate proves that e-discovery is fast and affordable when the right tools and talents are brought to bear. Electronically stored information (ESI) serves us in all our day-to-day endeavors. ESI can and should serve us just as well in our search for probative evidence and in the resolution of disputes.

You Must Make It Happen

Finding efficiencies and avoiding dumb decisions in electronic discovery isn't someone else's responsibility. It's yours. If someone else must perennially whisper in your ear, articulating the issues and answering the questions you should be competent to address, you aren't serving your client.

ESI isn't going away, nor will it wane in quantity, variety or importance as evidence. Each day you fail to hone your e-discovery skills is a day closer to losing a case or losing a client. Each day you learn something new about ESI and better appreciate how to request, find, cull, review and produce it at lowest cost is a day that cements your worth to your clients and makes you a more effective counselor and advocate.

Eight Tips to Quash the Cost of E-Discovery

The following tips are offered to help you slash the outsize cost of e-discovery:

1. **Eliminate Waste**
2. **Reduce Redundancy and Fragmentation**
3. **Don't Convert ESI**
4. **Review Rationally**
5. **Test your Methods and Know your ESI**
6. **Use good tools**
7. **Communicate and cooperate**
8. **Price is what the seller accepts**

1. Eliminate Waste

I once polled thought leaders in electronic discovery about costs. They uniformly agreed that about half of every e-discovery dollar is expended unnecessarily as a consequence of counsel lacking competence with respect to ESI. Half was kind.

Every time you over-preserve or over-collect ESI, every time you convert native data to alternate forms or fail to deduplicate ESI before review and every time you otherwise review information that didn't warrant "eyes on," you add cost without benefitting your client. It's money wasted. Poor e-discovery choices tend to be driven by irrational fears, and irrational fears flow from lack of familiarity with systems, tools and techniques that achieve better outcomes at lower cost. The consequences of poor e-discovery decisions prompt motions to compel or for sanctions, further ratcheting up the cost of incompetence.

2. Reduce Redundancy and Fragmentation

Many complain that electronic discovery has made litigation more costly because there is so much more information available today. Certainly, there are more channels of information available today, allowing an enlightened advocate more probative evidence. Much of what evaporated as a phone conversation now endures as a writing. There is more temporal, photographic and geolocation data to draw on, and more "persons with knowledge of relevant facts" who are privy to revealing information.

Despite there being *more*, the increase doesn't reflect the dire logarithmic leap in data volume some suggest. Much of the growth is attributable to replication and fragmentation. Put simply, human beings don't create that much more *unique* information; they mostly make more *copies* of the *same* information and break it into *smaller pieces*. Yesterday's memo sent to three people is today's 30- message thread sent to the whole department and retrieved on multiple devices. These iterations add a lot to the quantity of ESI, but little in the way of truly unique evidence. Thus, the burden and cost of e-discovery is inversely proportional to a litigant's ability to reduce redundancy and fragmentation. There are many ways to minimize redundancy and fragmentation. Some entail sensible choices during identification and collection; others involve the application of tools and techniques geared to eliminating replication and organizing fragmented information for efficient review.

Anyone who has done a document review can attest to the tedium of seeing the same documents over and over again. Messages repeat within threads or across recipients, and attachments to messages mirror documents from file servers. Some of this can be readily eliminated by simple hash-based de-duplication that costs very little and reliably eliminates documents that are duplicates in all respects. Hash-based deduplication calculates a "digital fingerprint" value (variously called an MD5 or SHA1 value) for each document, allowing redundant documents to be excluded from review.

Nothing offers a more cost-effective means to reduce the cost of document review than deduplication; consequently, no one should undertake a document review without minimally running a simple hash-based deduplication to eliminate replication.

Unfortunately, simple hash-based deduplication doesn't work for e-mail messages (which necessarily reflect different routing information for different recipients) or for documents with minor variations that don't signify material differences in content. For these items, more advanced near-deduplication techniques are needed to eliminate redundancy without increasing the risk that unique documents will be overlooked.

Deduplication is a mechanical process requiring little, if any, human intervention or costly programming. Accordingly, its cost should always be a nominal component of an e-discovery effort. If a service provider attempts to charge princely sums for deduplication, consider it a sign that it's time to find a new vendor. When the volume of information to be deduplicated is modest (e.g., less than 10-15 GB), low cost tools are available to deduplicate without the need to engage a service provider.[1]

3. Don't Convert ESI

It's criminal how much money is wasted converting electronic information into paper-like forms just so lawyers don't have to update workflows or adopt contemporary review tools. Clients work with native forms of ESI because native forms are the most utile,

[1] One of the finest tools for deduplicating collections less than 15GB is called Prooffinder (www.prooffinder.com). It costs $100.00 for an annual license, and all proceeds from its sale go to support child literacy.

complete and efficient forms in which to store and access data. Clients don't print their e-mail before reading it. Clients don't emboss a document's name on every page. Clients communicate and collaborate using tracked changes and embedded comments, yet many lawyers intentionally or unwittingly purge these changes and comments in e-discovery and fail to disclose such redaction. They do it by converting native forms to images, like TIFF.

Converting a client's ESI from its natural state as kept "in its ordinary course of business" to TIFF images injects needless expense in at least half a dozen ways. First, you must pay someone to convert native forms to TIFF images and emboss Bates numbers. Second, you must pay someone to generate load files containing extracted text and application metadata from the native ESI. Third, you must produce multiple copies of certain documents (like spreadsheets) that are virtually incapable of being produced as TIFF images. Fourth, because TIFF images paired with load files are much "fatter" files than their native counterparts, you pay much more for vendors to ingest and host them by the gigabyte. Fifth, it's very difficult to reliably deduplicate documents once they have been converted to TIFF images. Sixth, you may have to reproduce everything when your opponent wises up to the fact that you've substituted cumbersome TIFF images and load files for the genuine, efficient evidence.

4. Review Rationally

Recently, an opponent advised the Court that their projected cost of review encompassed the obligation to look at every e-mail attachment when the body of the e-mail message contained a keyword hit, *even when none of the attachments contained a hit*. They made this representation knowing that the majority of the hits would prove to be noise hits, that is, keywords in a context that doesn't denote responsiveness. Why would a party incur the expense to review the attachments to a message they'd determined was non-responsive when the attachments contained no keywords? It turned out they had separated attachments from e-mail transmittals, surrendering the ability to know which attachments could be eliminated from review because the transmitting message was eliminated from review. That's not a rational approach to review.

A common irrational approach to review is to treat information in any form from any source as requiring privilege review when even a dollop of thought would make clear that not all forms or sources of ESI are created equal when it comes to their potential to hold privileged content. The cost of review accounts for anywhere from 60-90% of the total cost of e-discovery; so, anything that defensibly narrows the scope of review prompts maximum savings. Almost anytime you can use technology to isolate privileged content and prudently employ a clawback agreement or Federal Rule of Evidence 502 to guard against inadvertent disclosure, you can slash the cost of privilege review.

5. Test your Methods and Know your ESI

Staggering sums are spent in e-discovery to collect and review data that would never have been collected if only someone had run a small scale test before deploying an enterprise search. It's easy and inexpensive to test proposed searches against representative samples of data (e.g., one key custodian's mailbox) so as to identify outcomes that will unduly drive up the cost of ingestion, hosting and review. This entails more than simply eliminating queries with large numbers of hits; it requires modifying them to balance the incidence of noise hits against hits on responsive data.

A lot of money gets wasted in e-discovery over disputes that could be quickly resolved if someone simply knew more about the ESI i.e., if someone simply *looked*. Here again, knowing the software and file types used, the nature and configuration of the e-mail system, the retention scheme for backup media or whether a key custodian used a home system for business are all examples of information that can serve to facilitate decisions that will narrow the scope of collection and review with consequent cost savings.

6. Use Good Tools

If you needed to dig a big hole, you wouldn't use a teaspoon, nor would you hire a hundred people with teaspoons. You'd use the right power tool and a skilled operator.
You can't efficiently collect or review ESI without using good tools. Anyone engaging in e-discovery should be able to answer the question, "What's your review platform?" They should be able to articulate why they use one review platform over another, and "because we already owned a copy" is not the best reason.

A review platform is the software tool used to index, sort, search, view, organize and tag ESI. Choosing the right review platform for your practice requires understanding your workflow, personnel, search needs and forms in which ESI will be ingested and produced. Review platforms can be cost-prohibitive for some practitioners, but it's untenable to manage ESI in discovery without a capable review platform.

There are many review platforms on the market, including familiar names like Relativity, Concordance and Summation. There are also Internet-accessible "hosted" review environments and many proprietary review tools touting more bells and whistles than a Mardi Gras parade. Among the most important consideration in selecting a review platform is its ability to accept data in forms that do not to require costly conversion to TIFF images. Additionally, you may want the platform you select to support the most advanced forms of technology-assisted search and review that your budget allows, including predictive coding capabilities.

7. Communicate and Cooperate

Poor communication and lack of cooperation between parties on e-discovery issues contribute markedly to increased cost. The incentives driving transparency and cooperation in e-discovery are often misunderstood. You don't communicate or cooperate with an opponent to help them win their case on the merits; you do it to permit the case to be resolved on its merits and not be derailed or made more expensive by e-discovery disputes.

Much of the waste in e-discovery grows out of apprehension and uncertainty. Litigants often over-collect and over-review, preferring to spend more than necessary instead of giving the transparency needed to secure a crucial concession on scope or methodology.

Communication and cooperation in e-discovery are not signs of weakness but of strength. Cooperation is a means to demonstrate that your client understands its e-discovery obligations and is meeting them. More, it's a means to build trust in the scope and methods of discovery so as to forestall challenges that may prove disruptive to the case and the client's operations. It's even possible that your opponent understands e-discovery or your client's systems better than you do and can propose more efficient ways to scope and complete the effort. What an opponent will accept in a cooperative give-and-take is often less onerous than what you were planning to produce.

Put simply: the more you seek to hide the ball, the more likely a savvy opponent will dig deeper and find something your side missed. Because there are no perfect e-discovery efforts, there are none that can withstand the heightened scrutiny invited by shortsighted stonewalling.

Hubris doesn't help. Most flaws in e-discovery processes can be rectified quickly and cheaply when they surface early. An overlooked variant on a keyword or a missed file type is easy to fix at the outset, but can prove costly or irreparable when discovered months or years later. Moreover, disclosure tends to shift the burden to act. Courts are loathe to entertain belated objections from parties who'd been supplied sufficient information to act promptly.

8. Price is What the Seller Accepts

I've haggled in bazaars and markets from Cairo to Kowloon; but, I've never seen more pliant pricing than among those hawking e-discovery tools and services in the United States.

A famous/infamous e-discovery vendor once quoted $43.5 million for a six-week engagement processing a very large volume of data on an expedited basis. The customer was desperate, but not insane. Rebuffed, the vendor re-quoted the job the next day for several million dollars less. They "sharpened their pencil" again the next day…and the next. Before the week was out, the vendor was proposing to do the job for $3.5 million. They didn't get the work.

Service providers have to pay staff and keep the lights on. So, almost any work beats no work at all. Many will accept work that isn't profitable, if it keeps a competitor from getting the business. Shop around. Make an offer. Only a sucker pays rack rate.

Make yourself sheep and the wolves will eat you. Benjamin Franklin

Are Documents Containing Agreed-Upon Keywords Responsive *Per Se*?

by Craig Ball

[Originally published on the Ball in Your Court blog, March 22, 2013]

More than once, I've faced disputes stemming from diametrically different expectations concerning the use of keywords as a means to identify responsive ESI. I don't recall seeing a case on this; but, it wouldn't surprise me if there was one. If not, there soon will be because the issue is more common than one might imagine.

When requesting parties hammer out agreements on search terms to be run against the producing party's ESI, sometimes the requesting party's expectation is that *any* item responsive to the agreed-upon keywords (that is, any item that's "hit") must be produced unless withheld as privileged. Put another way, the requesting party believes that, by agreeing to the use of a set of keywords as a proxy for attorney review of the entire potentially-responsive collection, and thereby relieving the producing party of the broader obligation to look at everything that may be responsive, those keywords define responsiveness *per se,* requiring production if not privileged.

Now I appreciate that some are reading that and getting hot under the collar. You're saying things like:

- "We *always* have the right to review items hit for responsiveness!"
- "It's the *Request for Production* not the keyword hits that define the scope of e-discovery!"
- "Nothing in the Rules or the law obliges a party to produce non-responsive items!"
  [Expletives omitted]

Perhaps; but, there's sufficient ambiguity surrounding the issue to prompt prudent counsel to address the point explicitly when negotiating keyword search protocols, and especially when drafting agreed orders memorializing search protocols.

To appreciate why expectations should be plainly stated, one need only look at the differing incentives that may prompt disparate expectations.

What is a producing party's incentive to limit the scope of search to only a handful of queries and keywords? Federal law requires a producing party to search all reasonably accessible sources of information that may hold responsive information and to identify those potentially responsive sources that won't be searched. That's a pretty broad mandate; so, it's no wonder producing parties seek to narrow the scope by securing agreements to use keyword queries. Producing parties have tons of incentive to limit the scope of review to only items with keyword hits. It eases their burden, trims their cost and affords requesting parties cover from later complaints about scope and methodology.

What is the requesting party's incentive to limit an opponent's scope of search to only those items with keyword hits? Requesting parties might respond that their incentive is to insure that they get to <u>see</u> the items with hits so long as they are not privileged. By swapping keyword culling for human review, requesting parties need not rely upon an untrusted opponent's self-interested assessment of the material. Instead, if it's hit by the agreed-upon keywords, the item will be produced unless it's claimed to be privileged; in which case the requesting party gets to see its privilege log entry. That's often the contemplated *quid pro quo*.

Both arguments have considerable merit; and, yes, you *can* be compelled to produce non-responsive items, *if the agreement entered into between the parties is construed to create that obligation*. Some might argue that the agreement to use queries is an agreement to treat those queries as requests for production. You don't have to agree, dear reader; but, you'd be wise to plan for opponents (and judges) who think this way.

These are issues we need to pay attention to as we move closer to broader adoption of technology-assisted review. We may be gravitating to a place where counsel's countermanding a machine's "objective" characterization of a document as responsive will be viewed with suspicion. Responding parties see electronic culling as just an extension of counsel's judgment; but, requesting parties often see electronic culling as an objective arbiter of responsiveness. Face it: requesting parties believe that opponents hide documents. TAR and keyword search may be embraced by requesting parties as a means to get hold of helpful documents that would not otherwise see the light of day.

Practice Tip: If you enter into an agreement with the other side to use keywords and queries for search, be clear about expectations with respect to the disposition of items hit by queries. Assuming the items aren't privileged, are they deemed responsive because they met the criteria used for search or is the producing party permitted or obliged to further cull for responsiveness based on the operative Requests for Production? You may think this is clear to the other side; but, don't count on it. Likewise, don't assume the Court shares your interpretation of the protocol. Just settling upon an agreed-upon list of queries may not be sufficient to insure a meeting of the minds.

Cleary Counsels Clarity

by Craig Ball

[Originally published on the Ball in Your Court blog, March 25, 2013]

A colleague flagged an opinion from the Northern District of Oklahoma she'd seen blogged on a big firm website. In the decision, the judge spoke dismissively of "apps" that must be responsible for those hairballs so common in discovery: vague requests and objection-obscured replies. The blogger took the judge's mention of apps too-literally, even noting that the judge failed to name the offending software.

I suspect those literary devices called irony and satire may have been lost on the blogger. The court in *Howard v. Segway, Inc.* wasn't saying both sides had *actually* used bad apps; instead, Magistrate Judge Paul J. Cleary was referencing the current fervor for apps *metaphorically,* as a means to convey that the more things change, the more they stay the same. That is, whether you slavishly draw boilerplate from a paper form or program the same mindless twaddle into a document assembly app on your iPad, drivel remains drivel, and obstruction invites sanctions.

Because we are both witness to discovery before and after the advent of the computer, I felt a kinship with Judge Cleary as he took both sides to task for using discovery devices in ways most lawyers regard as a standard—even enviable–practice. Plaintiff Howard used a thesaurus-like definition of document, larded every request with "any" and "all" and used omnibus phrases like "concerning" or "referencing." Defendant Segway, Inc. objected to each request with that unholy quartet, 'vague, overbroad, unduly burdensome and not reasonably calculated to lead to discovery of admissible evidence,' then "without waiving and subject to said objections, Segway, Inc. gestured in the general direction of something it produced.

This brought me back 30+ years, to my days as a newly-minted associate at a venerable firm in Houston. Associates were tasked to meet periodically to review the firm's standard discovery requests to insure they were state-of-the-art. That meant striking references to Dictabelts and adding to an ever-growing litany of synonyms for "document" in the definitions. Then as now, discovery was a perverse game of Simon Says: if you didn't phrase it *just so*, the other side would find a way to read your request so as not to encompass what you sought.

The irony of my being seen as an expert in discovery is that I have always despised documentary discovery. Before the advent of electronically stored information, I rarely obtained much of value through Requests for Production. Anything helpful tended to be hidden beneath the skirts of some specious objection. Cases didn't settle for value until I got a favorable ruling on my Fifth or Sixth Motion to Compel; then productions were routinely shuffled, sanitized or incomplete. In retrospect, if it hadn't been for depositions and subpoenas, I'd have been perennially in the dark.

Happily, the times they are a-changin'; though less in practice than in the minds of jurists like Judge Cleary who, after 30+ years of the same malarkey, seems mad as hell and unwilling to take it anymore. These jurists are putting some sting into Rule 26(g). The $5,466.50 jointly and severally assessed against Defendant Segway and its counsel, Holden & Carr, isn't going to rock anyone's world; but, it sends a message, warning others who litigate opposite that firm to be wary (though in fairness, apart from the bizarre delaying behavior described, Holden & Carr's use of boilerplate objections to obfuscate isn't new or novel among responding parties). Also in fairness, the criticism visited upon the plaintiff and his counsel could be leveled at most requesting parties.

That's what's refreshing about the *Howard v. Segway, Inc.* decision. It calls out bad behavior that's become so commonplace, we're inured to it.

Save for changing the name, what follows is a verbatim request and response from a case in which I'm consulting:

REQUEST: *All non-privileged communications, whether hand written, typed, computer generated, or emails in John Smith's possession regarding any civil litigation against John Smith. This request is limited to the last 5 years.*

RESPONSE: *Defendant objects to this request as vague and/or ambiguous, as it requires Defendant to speculate as to the meaning of the request, and being overly broad and not reasonably limited as to time and scope. Defendant further objects to this request because it is not reasonably calculated to lead to the discovery of admissible evidence, particularly because it seeks information irrelevant to this litigation. Subject to said objections, Defendant agrees to produce non-privileged documents responsive to this request, if any.*

Perhaps the Request is not a model of limpidity; but, how might a requesting party target it more tightly? Excising "all" and the "any" might help; they add nothing. "Regarding" isn't great, but is there any word (About? Addressing? Discussing?) that's much more precise? Would "*Written communications after March 25, 2008 discussing civil litigation against John Smith*" truly be less likely to draw the same litany of objections?

But the Response is so much worse. It's precisely the sort of block-and-copy boilerplate that offends Judge Cleary, and it's the same response opponents sent me thirty-odd years ago. Then as now, opponents didn't specify what was objectionable about the request. Production was promised, but seldom attached, and I couldn't tell if what was later produced was all that was responsive because, though my opponent's logged documents withheld as privileged, they gave no indication of what was withheld "subject to said objections."

So, how do we fix this mess?

Let’s begin by recognizing that the problem is of longstanding and fueled by ignorance and fear. Requesting parties are scared that something important will be missed by narrowly-drafted requests, and responding parties are scared that, if they don’t object to everything on every conceivable ground, they will surrender their ability to assert an objection when it might actually adhere. Further—and let’s not mince words—both sides behave this way because they are incentivized to be lazy. There is no profit in taking more time to do quality work when the insurance carrier or corporate client will only pay so much and no more. Lazy reliance on forms and boilerplate is shoddy lawyering; but, it’s the prevailing way we train young lawyers to become old lawyers.

Why should we be surprised that this generation of attorneys is no better than the last? We made them in our own image.

I write this essay to float a few steps that judges, law schools and counsel might take to make things better. Someone had the courage to eat the first oyster; someone stepped in front of that Chinese tank. I trust you, dear reader, to muster the courage to say, “Enough!”

Step One for the Bench: Follow Judge Cleary’s lead (or for you east-coasters, Judge Paul Grimm’s lead in *Mancia v. Mayflower*) and put a price tag on absurdly overbroad requests and obfuscating boilerplate objections. Get lawyer skin in the game. Make it a pain-in-the-pocket to get sloppy. Old dogs *can* learn new tricks, but not if we reward them for peeing on the rug.

Step Two for the Bench: Your *values* may be timeless, but your methods not so much. Don’t gauge whether something is acceptable or unacceptable based on how much it mirrors what you did when you practiced law. At least where electronically stored information is concerned, you must hold practitioners to a higher standard of competence than you might have been able to meet “back in the day.”

Step One for Law Schools: Get off your high horses and dedicate third year to compulsory skills training for anyone who plans to do something other than teach the law after graduation. Then, teach students laudable ways to draft and respond to discovery—ways that emphasize thinking over parroting forms and precision over, well, the zero sum game we play now.

Step One for Lawyers Making Requests for Production: Consider incorporating a preamble setting out the scope of information sought and/or a short explanation to accompany each request so as to clarify its scope. Trying to employ universal definitions that apply to every request (or every matter) promotes confusion. If the request isn’t a model of clarity standing on its own, is it better when you have to plug in six defined terms just to get the gist of it?

Step Two for Lawyers Making Requests for Production: Stop trying to fashion a Grand Unified Theory of Discovery manifested as requests so sweeping and convoluted as to be incomprehensible. Less is more. If what you really want is, “the complete conversational thread of e-mail and texting between Jane Pillar and Bob Post during June and July of 2012 where either discuss Paula Plaintiff,” can’t you just ask for it that simply? And please don’t demand “the metadata” unless you can articulate *which* metadata values you seek and have some notion what they signify.

Step One for Lawyers Responding to Requests for Production: Specify the portion of the request you find truly objectionable and *recast* the request to conform to what you <u>are</u> producing. No, it’s really *not* doing your opponent’s job for them. It’s saving your client from costly motions to compel and for sanctions. Plus, it’s meeting your duties under FRCP 26(g) and its state law counterparts.

A Positive Step for Everyone: Take a look at Judge Cleary’s 2011 article in the Oklahoma Bar Journal entitled, *<u>Some Thoughts on Discovery and Legal Writing</u>* *(http://www.craigball.com/Cleary_Discovery_and_Legal_Writing.pdf).* It’s a quick and read that serves to flesh out some of the thinking behind the *Howard v. Segway, Inc*. decision. It’s also a cogent reminder that, if you’re going to litigate in a judge’s court, it’s wise to read what he or she has published on the issue. Anyone reading His Honor’s thoughts in the OK Bar Journal should have had better sense than to respond with a barrage of boilerplate blather.

When Do You Buy T-Shirts?

by Craig Ball

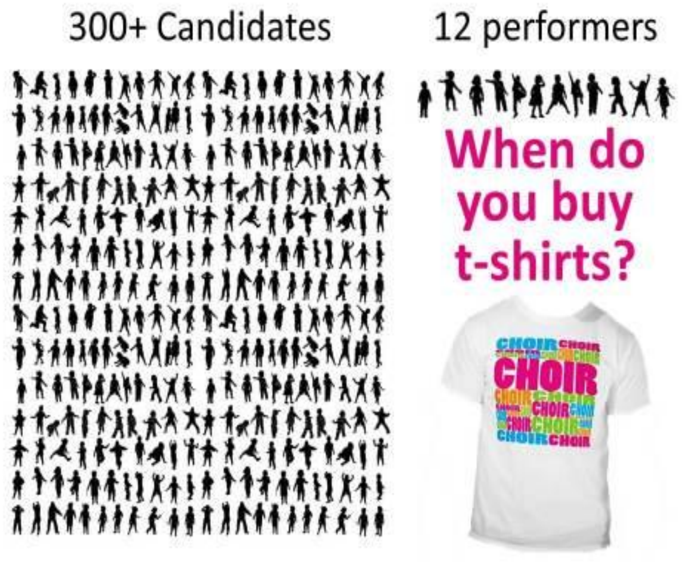
[Originally published on the Ball in Your Court blog, March 27, 2013]

I did an ILTA webcast this morning called "Going Native." Steven Clark ably moderated the panel including Denise Talbert and Maureen Holland. D4 sponsored. Going Native did not mean we spoke in loincloths (although I can really only account for my own attire). We addressed the pros and cons of producing ESI in native and near native forms versus conversion from native forms to TIFF images and load files. I expected agressive pushback as I sang the praises *(Just! Speedy! Inexpensive!)* of native productions; but, steeled for debate, I was instead treated to fine dialog. No one trotted out the usual hackneyed objections to native productions. Advantages and disadvantages were thoughtfully addressed and everyone proved open to flexibility in forms of productions when to do so serves to meet a genuine need or solve a problem.

When polled, roughly half of those attending stated that they weren't making production in native and near-native forms simply because the requesting parties hadn't sought same. Around 16% said they resisted native production out of concern that native productions were harder to track. My sense is that the attendees were open–even eager–to embrace native production. I wasn't surprised by this because there are few audiences for e-discovery education as sophisticated and rational as ILTA audiences. ILTA members tend to be hands on with ESI, affording them a better appreciation of the downsides of image and load file productions. They're typically the ones tasked with cleaning up the messes caused by malformed load files and TIFF creation errors.

That 16% missing out on the advantages of native productions out of concern that native files aren't Bates stamped on each page distresses me because I'm sure they correspond to a much larger percentage of lawyers who can't conceive of litigating without Bates numbers (and protective legends) on every page. It seems a lot of people don't realize that **you don't have to give up Bates numbers and protective legends when you make native productions**. If you approach native productions the right way, the Bates numbers will still be there *when you need them*. I'll explain how that works, but first please indulge me in a little mental exercise.

Imagine that you are the choir director for a big middle school. You're going to assemble twelve students to perform for the town's centennial celebration. The performers will wear matching choir t-shirts you supply at your cost. More than 300 students have signed up to audition for the 12 slots. Will you spend the money and undertake the administrative challenge to buy and fit a t-shirt for every child who auditions, or will you wait and outfit just the twelve performers selected? **When do you buy t-shirts?**



T-shirts are cheap, but you won't want to take on the expense and effort to outfit hundreds of aspiring performers. You'll wait until you've picked the twelve you'll put on stage before you shell out for shirts.

I suggest that this is also how we should sensibly approach Bates numbering native productions.

When do we really want Bates numbers? Isn't it when the items produced are going to perform…as exhibits to pleadings, in depositions and on that rare occasion called "trial?" That's when being able to identify particular documents for the record and be on the same page matters. So, my rule is simple: When you produce natively, every file produced carries its unique Bates number in or as its file name (appended, prepended or replacing the original file name altogether). A load file correlates the source file name to its Bates number. If you wish, other language (e.g., PRODUCED SUBJECT TO PROTECTIVE ORDER) can also be included in the new file name. It's simple. *Really.* Chances are you're *already* using Bates numbers as file names in TIFF productions. We know how to do this easily and at almost no cost.

Now, you secure a court order or agreement requiring any party who prints or images an item produced natively *to include the file name and a page number on the face of the printed or imaged document*. This is also child's play. MS Word can do it. Adobe Acrobat can do it. E-discovery production tools can certainly do it. It's essentially what we've always done when preparing TIFF productions, except this way is a whole lot cheaper and faster.

The Bates numbers will be consistent, because they're the file names. The protective legend will appear on the face of each document for the same reason. Every page will be paginated for ease in reference. When you use printed documents in a proceeding, you furnish copies.

Here's how such a protocol might read:

Unique Production Identifier (UPI)

> a) Other than paper originals, images of paper documents and redacted ESI, no ESI produced in discovery need be converted to a paginated format nor embossed with a Bates number.

> b) Each item of ESI (e.g., native file, document image or e-mail message) shall be identified by naming the item to correspond to a Unique Production Identifier according to the following protocol:
>
> i. The first four (4) characters of the filename will reflect a unique alphanumeric designation identifying the party making production;
>
> ii. The next nine (9) characters will be a unique, sequential numeric value assigned to the item by the producing party. This value shall be padded with leading zeroes as needed to preserve its length;
>
> iii. The final five (5) characters are reserved to a sequence beginning with a dash (-) followed by a four digit number reflecting pagination of the item when printed to paper or embossed when converted to an image format for use in proceedings or when attached as exhibits to pleadings.
>
> iv. By way of example, a Microsoft Word document produced by Acme in its native format might be named: ACME000000123.doc. Were the document printed out for use in deposition, page six of the printed item must be embossed with the unique identifier ACME000000123-0006.
>
> v. The original name of the file will be furnished in a load file, along with other relevant system metadata as agreed upon by the parties or ordered by the Court.

The most common objection this draws is that someone might distribute a document without a protective legend. Yes, the world has its share of untrustworthy people. What would lawyers do were it otherwise? But, do you really think that the protective legend you place in the margins of a TIFF image serves as genuine deterrent to bad guys? Let's recall that, in order to avoid obscuring content, Bates numbers and protective legends are embossed in a gutter of white space created by shrinking the source document on the production page. The same technology that makes it easy to shrink the content makes it easy to enlarge the content to push the Bates number and protective legend off the page. Your copier can do it. Your scanner software can do it, and Adobe Acrobat can do it, too. So, let's not kid ourselves. TIFFing is illusory protection. Fortunately, the risk of abuse is speculative and remote. It can be punished. The punishment in cost and burden that flows from TIFFing for Bates numbers is real and immediate.

Why not stop buying t-shirts for the whole school?

Prooffinder: Touch the Monolith

by Craig Ball

[Originally published on the Ball in Your Court blog, May 9, 2013]

In the spring of 1968, my sixth grade class from suburban Eastchester went to the Loews Capitol Theatre at 51st and Broadway in New York City to see *2001: A Space Odyssey*. It was an unforgettable event. Though much of the movie went over our ten-year-old heads, we got the message about tools and evolution when our hairy forebear flung his bone “hammer” aloft and it became a sleek spaceship. We evolve to use tools, and the tools we use drive our evolution.

We can’t deal with electronic evidence without tools. The more adept we are with those tools, the more adept we become with electronic evidence. Tools that let us touch data—hold it up to the light and turn it this way and that—change the way we look at data. Tools change us.

I’m always preaching that lawyers must get their hands dirty with data and get back in touch with the evidence. It’s a metaphor, but it’s also a manifesto. A master builder needn’t swing every hammer; but, a master builder knows how a hammer feels in the hand.

So, I prodded and pushed e-discovery vendors to bring out tools powerful enough to do real work while inexpensive enough for anyone to own. In 2009, I issued the EDna Challenge. The goal was to equip a small firm lawyer to competently process and review less than 10 gigabytes of run-of-the-mill ESI.

The processing and review tool or service meeting the Edna Challenge must:

1. Preserve relevant metadata;
2. Incorporate de-duplication;
3. Support robust search;
4. Run well on late-model PCs; and, most importantly,
5. Cost less than $1,000.00 over the two-to-three year life of a lawsuit.

Several vendors claimed victory, yet “met” the Challenge by ignoring a crucial need, e.g., requiring the data to miraculously morph into TIFFs and load files, or by eliding over fees for ingestion and storage. A pair of developers got very close to true victory. Splendid products like Vound Software’s *Intella* and GGO’s *Digital WarRoom* could do the work, but alas, couldn’t stay under budget for the life of the case. Still, they deserve kudos for delivering powerful, economical tools. The real “winners” of the EDna Challenge have been all of us who consume e-discovery tools and services: we have benefitted from falling prices and increased competition.

I was sure that if someone built an amazing, affordable e-discovery desktop tool, the world would beat a path to their door. In fact, someone did build a tool that could do virtually everything required to professionally complete a small-scale e-discovery effort.

It was my Holy Grail. The tool was easy to obtain, install and use on a desktop or laptop and cost only **$100.00** for an annual license. And to make even that piddling price painless, the seller committed to donate all proceeds to child literacy!

Of, course, I'm talking about **Prooffinder** by Nuix (www.prooffinder.com), the Sydney-based concern that produces the ESI processing technology used by big corporations, service providers and government agencies all over the world. Finally, here was a tool that any lawyer, paralegal or IT person could afford to put on their Windows machine. Prooffinder has Nuix' superior DNA, but is limited to processing up to 15 gigabytes per matter. That's a minor handicap considering you can open as many "matters" as you wish, and 15GB tends to be ample for, say, the collections of several key custodians.

Here is just a sampling of what Prooffinder can do:

1. Allows fast, eyes-on access to all common forms of electronic data without altering content

Prooffinder opens, parses and supports instant viewing of the contents of hundreds of common and esoteric file formats, including all common data containers (like Outlook PST and OST mail files and compressed archives). You weren't really going to open all those spreadsheets in Excel or risk using Outlook to view e-mail, were you?

2. Sorts and filters on virtually any aspect of the information

Want to see just attached Excel spreadsheets and Word documents containing specified terms from April 3rd through 9th? Click. Click. Click. *No problem*.

3. Deduplicates the data by custodian or across custodians, including e-mail

Seeing just one copy of each item dramatically increases the efficiency of review. Prooffinder even supports configurable near-deduplication.

4. Flexible, powerful and easy-to-use search

If there is a search you want to run, Prooffinder can run it: Boolean, proximity, stemming, you name it. Plus, you can confine your search to just those parts of the data (like file paths or metadata) that interest you.

5. Configurable item tagging

Prooffinder allows you to create the tags you want to use in your review, including assigning tags to keyboard shortcuts.

6. Instantly tweak and test searches

You no longer need to make crucial keyword selections based on mysterious hit counts. Prooffinder allows you to test searches on representative data and quickly determine why searches expected to perform didn't and how over-inclusive searches can be tweaked to perform as expected.

7. Word lists
By compiling all the words into sortable lists, Prooffinder permits identification of common misspellings and acronyms.

8. Presents the complete metadata picture
Prooffinder extracts all the metadata you could want and presents it with just the level of detail you choose. No need to fear that an opponent will see something in the production you couldn't see.

9. Detailed exceptions reporting
Prooffinder flags, quantifies and segregates data that can't be processed for text extraction or requires special handling, e.g., decryption.

10. Impressive Export Capabilities
Though it won't generate conventional load files, Prooffinder builds highly-detailed and customizable production reports, assembles PSTs for production, converts selected items to PDF and generates HTML reports with full text extraction for each item chosen for export.

Use Case
Though I think of Prooffinder as the ideal entry-level tool for solo and small firm practitioners, I recently came to see how hamstrung big firm lawyers are without desktop e-discovery tools.

I was assisting the requesting party in an e-discovery effort where a big firm lawyer rejected proposed search terms as overbroad and couldn't answer my questions about the capabilities and limitations of the search tools his client's e-discovery vendor used. The collection slated for search consisted of a huge volume of e-mail messages and attachments exported to PST container formats for each custodian. The other side couldn't produce any coherent metrics to support its objections. We accepted that hit counts were high, but were confident that noise hits could be quickly and reliably excluded if we could simply see examples of the noisy hits in context so as to identify Boolean constructs that would eliminate them. We also suspected that the hit counts were grossly inflated by inadequate deduplication across custodians holding the same responsive messages.

Additionally, we were concerned that many of the relevant documents we sought were attached to the messaging in non-textual formats (like faxes and scans) and thus weren't responsive to keyword search. The other side had no idea how much of their data comprised such non-searchable documents.

It became clear that the other side couldn't address the problem because their tools didn't allow them to get the answers we needed. The proposed searches were slated to be run against a lot of data, so it was crucial that both sides settle on efficient, effective methods of culling and search.

At an impasse, I proposed that opposing counsel buy a copy of Prooffinder and run it against the PSTs of a couple of agreed-upon key custodians—merely a sensible sample. As the $100 Prooffinder cost the same as just minutes of his billable time, he couldn't object on a cost basis. To his credit, he agreed.

When we sat down together with the Prooffinder-processed sample, it took no time at all to tweak the contentious searches, assess the level of duplication and determine what complement of the data needed to undergo optical character recognition. At first, he worked with the data as I talked him through the queries, so no privileged content was compromised. Very soon, my opponent picked up the intuitive Prooffinder interface and needed no guidance.

We learned that our worst fears about a lack of duplication were warranted but that our concerns about faxes and scans were largely unfounded. Through testing, and by virtue of seeing results instantly, the proposed search terms were crafted to work with little noise. Thanks to Prooffinder (and thanks to a lawyer who wasn't afraid to get his hands a little dirty with the data), the other side saved hundreds of thousands of dollars that would have been wasted poring over multiple copies of irrelevant documents. We got more of the documents we sought and far less junk.

This experience taught me that big firms, too, needed copies on the machines of all their litigators and legal assistants. Then, anyone could quickly look at exemplar digital evidence, run test searches, tweak queries, assess duplication, weigh accessibility and undertake meaningful early case assessment. *Every ape could pick up a bone and start pounding on data.* In so doing, we would evolve to better understand ESI and make smarter, more cost-effective decisions in e-discovery.

Tools Change Us

Having a portal to ESI right on our desktops would change us in much the same way that web browsers gave us the Internet and changed us. Because it's so inexpensive, everyone can afford a copy and everyone can gain a comparable insight into electronic evidence. Opponents can then work together, refining searches by testing instead of guessing. Everyone can see the same metadata and realize that metadata isn't something to fear.

Most of all, it would help restore an aspect of lawyering that's fading away. Once, a prospective client carried in a sheaf of papers—records, correspondence, receipts, bills, pleadings. A lawyer read them and asked questions. That lawyer had a "peek" at the evidence and assessed the matter quickly because he or she had to accept the engagement or turn it away.

Today, the evidence is digital. A client may supply printouts of what they deem important, but it's no longer just a signed contract or a couple of business letters. Much remains unseen, in the e-mail or the network share contents. The evidence lies among hundreds or thousands of threaded messages or exists as revealing comments embedded within drafts of e-documents. Now, it takes weeks or months and costs tens of thousands of dollars before lawyers start to sift through the evidence and assess exposure in earnest—assuming that the untested keyword searches employed serve to return relevant data.

Lawyers need tools that allow them to once more "peek" at the relevant evidence. Prooffinder is ideal for that. I thought it would sell out in days; but I guess not everyone is ready to touch the monolith. Amazingly, there are still a few licenses for sale at www.prooffinder.com. That's great news for you because, in its latest release, Prooffinder is better than ever, and every dollar from license sales still goes to build schools and support the charity, Room to Read.

Just imagine how the savings and efficiencies I saw might be replicated in other cases if more lawyers had the ability to peek into the evidence, secure solid metrics and quickly test and refine searches. If anyone can gain this extraordinary capability for $100 dedicated to child literacy, why hasn't everybody already done so? *Why haven't you?*

Amendments Should Safeguard Meta-Discovery

by Craig Ball

[Originally published on the Ball in Your Court blog, June 9, 2013]

"American laws and American policy view the content of communications as the most private and the most valuable, but that is backwards today," said Marc Rotenberg, the executive director of the Electronic Privacy Information Center, a Washington group. "The information associated with communications today is often more significant than the communications itself, and the people who do the data mining know that." *How the U.S. Uses Technology to Mine More Data More Quickly*, New York Times, June 8, 2013

Marc Rotenberg was commenting on the recent revelation that the U.S. National Security Agency gathers a staggering volume of information about domestic and international telephone calls. When he states, "The information associated with communications today is often more significant than the communications itself…," he doesn't expressly label that "more significant" information as being "metadata," but that's what it is.

Rotenberg's right: *metadata matters*.

At the just-completed, weeklong Georgetown E-Discovery Training Academy, we do something not done at continuing education programs: we test the attendees coming in and test them heading out. Most of our attendees haven't sat for an exam since the bar. The tests include questions like this:

Which statement, if any, below is true:
A) System metadata is probative of some issue in every case.
B) Some files have more metadata than data.
C) Databases are unique in their ability to dispense with metadata.
D) None, all of the statements are false.

The answer sought is **B**.

It's worth adding that some metadata is *more probative* than the contents of the file to which it pertains.

Consider an e-mail that says, simply, "OK." That two letter message—a mere 14 bits of ASCII-encoded text—climbs aboard the internet express toting a big satchel of metadata, then picks up considerably more metadata as it traverses networks and servers. The "who," "when," "where" and "how" of the message's metadata comprises considerably more information than the "what" of its content and often conveys more of what we need to know.

The telephone call data gathered by the NSA may not hold the words exchanged—indeed many calls may not have been answered—but the metadata about numbers called, numbers calling and the time and location of attempts speaks volumes and is, as Mr. Rotenberg aptly puts it, “often more significant than the communication itself.”

No one is up in arms about the NSA’s actions on the ground that it wastes government resources to collect metadata. It’s tacitly acknowledged on all sides of the debate that the information gathered is significantly revealing and probative. It’s evidence, and the government grabs and uses it.

Yet, metadata has enormous utility apart from its innate value as evidence. Metadata enables us organize, manage and make sense of digital evidence. Metadata is the “glue” that holds certain evidence together and the labels on the cans that keep us from grabbing dog food when we want soup. This metadata isn’t evidence, but it’s the electronic equivalent of the evidence bag, the date on the evidence label or the entry in the police property room log that helps us find the evidence and assure ourselves of its integrity.

Say what one will about issues of privacy and governmental intrusion, there should be no debate that metadata is powerfully important information. Unfortunately, a different arm of government, the Judicial Conference of the United States’ Standing Committee on Rules of Practice and Procedure, doesn’t give metadata its due.

By way of background, many believe that the American system of civil discovery is broken. E-discovery makes CEOs worry, and we don’t pay CEOs enough to worry. To help CEOs relax, companies throw gobs of misdirected money at e-discovery. Accordingly, e-discovery is too expensive.

One way to rein in the cost of e-discovery is by narrowing the scope of discovery. Under existing Federal Rule of Civil Procedure 26(b), parties “may obtain discovery regarding any nonprivileged matter that is relevant to any party’s claim or defense—including the existence, description, nature, custody, condition, and location of any documents or other tangible things and the identity and location of persons who know of any discoverable matter. … Relevant information need not be admissible at the trial if the discovery appears reasonably calculated to lead to the discovery of admissible evidence.”

The proposed Rules amendments now wending their way through the public comment process would eliminate all of the language after “claim or defense” and graft proportionality language that’s already a part of the rule onto the first sentence. The latter change accommodates lawyers whose attention deficits or narcolepsy prevent them from reading past the first few lines of the current rule. 😉

Whether you support the proposed amendment or not, you should be concerned that it makes no provision for discovery of metadata or for discovery about information systems. Such efforts are not always easy to characterize as “relevant to any party’s claim or defense." Yet, discovery of metadata and information systems is essential in any case involving ESI (i.e., in each and every case).

Digital is Different

Discovery of ESI presents a need for contextual information rarely encountered in paper discovery. There is an aspect to ESI that reflects both a layer and a penumbra of information that, while not going to the substance of claims or defenses, bears mightily on the integrity and utility of the evidence. These are **application and system metadata**. Courts must condition their thinking to appreciate that, *in the world of ESI, metadata are as important as dates, page numbers and CC:s were for paper correspondence.* If we continue to treat metadata evidence as something apart and optional with respect to the evidence it describes, the proposed change in scope may prove problematic.

Accordingly, discovery of metadata supporting the utility and integrity of ESI requires discrete recognition in the Rules, even though such data may not directly relate to claims or defenses. The Advisory Committee should make clear in the Committee Notes that they do not seek to limit the discovery of meta-information that is either a part of the evidence or which materially bears on its integrity or utility.

Reasonable Discovery *about* Discovery Needs Protection

In a similar vein, the proposed amendment seeks to restrict discovery to matters relevant to any party's claim or defense" at a moment in history when a growing focus of discovery is the exploration of where and how such information can be found in an efficient, effective and affordable manner. No one likes the notion of "discovery about discovery;" however, measured and managed meta-discovery about ESI is a necessity to reining in overbroad discovery and for attacking the guessing game attendant to e-discovery today. The gatekeepers to evidence are no longer lawyers. The gatekeepers are end users and IT. Asking IT about how and where they store data may not be "relevant to a party's claim or defense," but it's highly relevant to a just, speedy and inexpensive process.

The proponents of these amendments may protest that they expect such meta-discovery to continue unabated; but, weigh such protestations against the language they seek to eliminate: *"including the existence, description, nature, custody, condition, and location of any documents or other tangible things and the identity and location of persons who know of any discoverable matter."* The language sought to be struck defines meta-information as relevant. Absent clarification in the Committee Notes or further amendment, cutting such language from the Rule will doubtlessly be raised as a bar to discovery about ESI forms and sources.

Proposed Changes to Rule 30 Shouldn't Limit Meta-Discovery

There's also a pending proposal to amend FRCP Rule 30 to halve to five the number of depositions that may be taken without leave of court and to limit each deposition to six hours. I don;t know if five depositions will suffice to shed light on claims and defenses;

but, I'm sure five won't be enough if you have to use them to elicit essential information about information systems and databases from IT personnel and administrators.

To be more efficient in e-discovery and better able to narrowly craft discovery, there must be disclosure of meta-information about information systems and ESI. These are disclosures which many parties refuse to provide unless compelled to do so through discovery. Consequently, the proposed amended Rule 30 limits should not apply to meta-discovery, such as depositions of IT personnel who have no knowledge of claims and defenses but possess crucial knowledge about the storage and forms of ESI. If a company has a different administrator for each database and a party must learn about each database to carefully tailor discovery to each, a party shouldn't have to choose between merits discovery and meta-discovery. Certainly, parties can seek leave for additional discovery; but, requiring same serves to compound the cost and delay of gaining access to information that forsters efficiency and cost-savings.

We should tailor rules to support the needs of the many diligent litigants who seek discovery in good faith; even when such rules could be exploited by a few who might abuse them. Judges are adept at dealing with abuses, and wield considerable power to rectify actions taken in bad faith. We should trust and encourage judges to do so, just as we should trust counsel not to abuse discovery of meta-information until and unless they have shown they cannot be trusted.

The proposed amendments have much to commend them, and the Committee that drafted them deserves our thanks for fine work. They need our input as well. As I write this it is not clear how one comments on the proposed amendments at this moment (beware of the Rules Committee website as none who venture in emerge unscathed). Though more than one public hearing is likely, the one currently scheduled will take place in Washington, D.C. in November 2013. In the interim, you might send comments by e-mail to Rules_Comments@ao.uscourts.gov and by snail mail to:

Committee on Rules of Practice and Procedure
Administrative Office of the United States Courts
One Columbus Circle, NE
Washington, D.C. 20544

The ‘Not Me’ Factor

by Craig Ball

[Originally published on the Ball in Your Court blog, June 17, 2013]

I’ve been skeptical of predictive coding for years, even before I wrote my first column on it back in 2005. Like most, I was reluctant to accept that a lifeless mass of chips and wires could replicate the deep insight, the nuanced understanding, the *sheer freaking brilliance* that my massive lawyer brain brings to discovery. Wasn’t I the guy who could pull down that one dusty box in a cavernous records repository and find the smoking gun everyone else overlooked? Wasn’t it my rarefied ability to discern the meaning lurking beneath the bare words that helped win all those verdicts?

Well, no, not really. But, I still didn’t trust software to make the sort of fine distinctions I thought assessing relevance required.

So, as others leapt aboard the predictive coding bandwagon, I hung back, uncertain. I felt not enough objective study had been done to demonstrate the reliability and superiority of predictive coding. I well knew the deep flaws of mechanized search, and worried that predictive coding would be just another search tool tarted up in the frills and finery of statistics and math. So, as Herb and Ralph, Maura and Gordon and Karl and Tom sung Hosannas to TAR and CAR from Brooklyn Heights to Zanzibar, I was measured in my enthusiasm. With so many smart folks in thrall, there had to be something to it, right? Yet, I couldn’t fathom how the machine could be better at the fine points of judging responsiveness than I am.

Then, I figured it out: The machine’s <u>not</u> better at fine judgment. I’m better at it, and so are you.

So why, then, have I now drunk the predictive coding Kool-Aid and find myself telling anyone who will listen that predictive coding is the Way and the Light?

It’s because I finally grasped that, although predictive coding isn’t better at dealing with the swath of documents that demand careful judgment, it’s every bit as good (and actually much, much better) at dealing with the overwhelming majority of documents that *don’t* require careful judgment—the very ones where keyword search and human reviewers fail miserably.

Let me explain.

For the most part, it’s not hard to characterize documents in a collection as responsive or not responsive. The vast majority of documents in review are either pretty obviously responsive or pretty obviously not. Smoking guns and hot docs are responsive because their relevance jumps out at you. Most irrelevant documents get coded quickly because one can tell at a glance that they’re irrelevant. There are close calls, but overall, not a lot of them.

If you don’t accept that proposition, you might as well not read further; but if you don’t, I question whether you’ve done much document review.

It turns out that well-designed and –trained software also has little difficulty distinguishing the obviously relevant from the obviously irrelevant. And, again, there are many, many more of these clear cut cases in a collection than ones requiring judgment calls.

So, for the vast majority of documents in a collection, the machines are every bit as capable as human reviewers. A tie. But giving the extra point to humans as better at the judgment call documents, HUMANS WIN! Yeah! GO HUMANS! Except….

Except, the machines work *much faster* and *much cheaper* than humans, and it turns out that there really is something humans do much, much better than machines: *they screw up.*

The biggest problem with human reviewers isn’t that they *can’*t tell the difference between relevant and irrelevant documents; it’s that they often *don’t.* Human reviewers make inexplicable choices and transient, unwarranted assumptions. Their minds wander. Brains go on autopilot. They lose their place. They check the wrong box. There are many ways for human reviewers to err and just one way to perform correctly.

The incidence of error and inconsistent assessments among human reviewers is mind boggling. It’s unbelievable. And therein lays the problem: *it’s unbelievable*. People I talk to about reviewer error might accept that some nameless, faceless contract reviewer blows the call with regularity, but they can’t accept that potential in themselves. “Not me,” they think, “If I were doing the review, I’d be as good as or better than the machines.” It’s the “Not Me” Factor.

Indeed, there is some cause to believe that the best trained reviewers on the best managed review teams get very close to the performance of technology-assisted review. A chess grand master has been known to beat a supercomputer (though not in quite some time).

But so what? Even if you are that good, you can only achieve the same result by reviewing *all* of the documents in the collection, instead of the 2%-5% of the collection needed to be reviewed using predictive coding. Thus, even the most inept, ill-managed reviewers cost more than predictive coding; and the best trained and best managed reviewers cost *much* more than predictive coding. If human review isn’t better (and it appears to generally be far worse) and predictive coding costs much less and takes less time, where’s the rational argument for human review?

What’s that? “My client wants to wear a belt AND suspenders?” Oh, PLEASE.

What about that chestnut that human judgment is superior on the close calls? That doesn’t wash either. First–and being brutally honest–quality is a peripheral consideration in e-discovery. I haven’t met the producing party who loses sleep worrying about whether their production will meet their opponent’s needs. Quality is a means to avoid sanctions, and nothing more.

Moreover, predictive coding doesn’t try to replace human judgment when it comes to the close calls. Good machine learning systems keep learning. When they run into one of those close call documents, they seek guidance from human reviewers. It’s the best of both worlds.

So why isn’t everyone using predictive coding? One reason is that the pricing has not yet shifted from exploitive to rational. It shouldn’t cost substantially more to expose a collection to a predictive coding tool than to expose it to a keyword search tool; yet, it does. That will change and the artificial economic barriers to realizing the benefits of predictive coding will soon play only a minor role in the decision to use the technology.

Another reason predictive coding hasn’t gotten much traction is that Not Me Factor. To that I say this: Believe what you will about your superior performance, tenacity and attention span (or that of your team or law firm), but remember that you’re spending someone else’s money on your fantasy. When the judge, the other side or (shudder) the client comes to grips with the exceedingly poor value proposition that is large-scale human review, things are going to change…and, Lucy, there’s gonna be some ‘splainin to do!

Dogged Pursuit of Direct Access Poses Risks to Counsel

by Craig Ball

[Originally published on the Ball in Your Court blog, June 22, 2013]

In any plaintiff's case, the claimant is Exhibit A. A claimant must be credible because, where the number of lies a jury allows a defendant varies from case to case; the person suing for money gets none. One reason I liked trying wrongful death cases was that the victim couldn't testify.

A common way to prove a claimant isn't credible is by proving the claimant tells different stories about matters made the basis of the suit. Such "prior inconsistent statements" are excluded from the rule against using hearsay testimony, not just as an exception to the rule but by being defined as "not hearsay.[2] So, if a defendant can lay hands on such statements, the statements are coming into evidence and may really hurt.

Nowadays, many prior inconsistent statements are found on social networking sites like Facebook, LinkedIn and Twitter. Facebook posts and tweets with tales of actions and attitudes at odds with claims in court are splendid fodder for impeachment. Even a Facebook photo of a claimant with a smile may serve as ammo for impeachment when mental anguish damages are sought.

Social networking evidence is dynamite, and recent court decisions underscore the difficulty that judges have in deciding whether and how to permit it to be discovered. Facebook pages are part public and part private. Much like one's home, friends can come in, but strangers need an invitation to get past the front door. Also like a private home, the contents of social networking sites are not exempt from discovery; but neither should opposing counsel get to root around the site, hoping to spy something useful.

Still, some courts *are* granting unfettered access to social networking sites. Worse, they are doing it in a manner that promises to blow up in the faces of those *gaining* access through the compelled turnover of log in credentials. This is not a post about *whether* to grant such access—many articles and published decisions do a fine job addressing the pros and cons of same. I write here of the form in which it should (and shouldn't) be done when it must be done.

Twenty years ago, New Yorker cartoonist Peter Steiner created the single most reproduced cartoon from a magazine long renowned for its marvelous cartoons. It featured a canine in front of a computer telling another pooch, "On the Internet, nobody knows you're a dog." Or as Notre Dame footballer Manti Te'o might say it, "On the Internet, nobody knows you're a fraud."

[2] Federal Rules of Evidence, Rule 801(d)(2).

In the world of social networking, you are anyone you say you are. But, if you gained access using credentials (user ID and password), then, you are presumed to be the owner of those credentials, with all the powers and privileges such authentication confers. If opposing counsel logs on as an account owner, opposing counsel invests the genuine account owner with the power to claim that actions attributed to the owner are really counsel's actions. If you are opposing counsel and something gets deleted, how do you prove you didn't delete it? If something gets posted adverse to the account owner, how do you answer the claim that you put it there? After all, you "borrowed" the owner's identity when you logged in as the owner (albeit with the blessing of the court).

Think about that next time you demand an opponent's log in credentials to social networking sites.

In cases where parties fight over access to an opponent's social networking content, the parties are often so caught up with privacy and relevance issues they give short shrift to the mechanics of access. They're like dogs chasing a car; giving no thought to what to do when they catch it.

The approaches courts have taken to forms of production are all over the map:

- **Zimmerman v. Weis Markets, Inc., No. CV-09-1535, 2011 Pa. Dist. & Cnty. Dec. Lexis 187 (Pa. C.P. Northumberland May 19, 2011)** (Court ordered Plaintiff to tender Facebook and MySpace usernames and passwords to Defendant. Court denied request for *in camera* review as an "unfair burden to place on the Court");
- **McMillen v. Hummingbird Speedway, Inc., No. 113-2010, 2010 Pa. Dist. & Cnty. Dec. Lexis 270 (Pa. C.P. Jefferson September 9, 2010)** (Court ordered Plaintiff to surrender Facebook and MySpace usernames and passwords to defense counsel and directed that access be "read only," without specifying *how* such limited access could be accomplished beyond reliance on counsels' rectitude);
- **Romano v. Steelcase, Inc., 907 N.Y.S.2d 650 (N.Y. Sup. Ct. 2010)** (Court required account owner to give access).
- **Offenback v. L.M. Bowman, Inc., No. 1:10-CV-1789, 2011 WL 2491371 (M.D. Pa. June 22, 2011)** (Court looked at social networking content *in camera* and determined that responding party should have made production without necessity of same);
- **Barnes v. CUS Nashville, LLC, 2010 WL 2265668 (M.D. Tenn. June 3, 2010)**(Court volunteered to undertake *in camera* review by "friending" account holder);
- **Thompson v. Autoliv ASP, Inc., No. 2:09-cv-01375-PMP-VCF, 2012 WL 2342928 (D. Nev. June 20, 2012)** (Court ordered upload of all information from Facebook and MySpace accounts onto an external storage device);
- **Giacchetto v. Patchogue-Medford Union Free School Dist., No. CV 11-6323(ADS)(AKT), 2013 WL 2897054 (E.D.N.Y. May 6, 2013)** (Court defined

scope and ordered Plaintiff's counsel, "not Plaintiff," to conduct review and make production);

- **Mailhoit v. Home Depot U.S.A., Inc., ___F.R.D.___, 2012 WL 3939063 (C.D. Cal. Sept. 7, 2012)** (Court defined scope of discoverable content but left mechanism of production to producing party);
- **Robinson v. Jones Lang LaSalle Americas, Inc., No. 3:12-cv-00127-PK (D. Or. Aug. 29, 2012)** (Court defined scope of discoverable content but left mechanism of production to producing party);

If you're thinking the best way to gain access to social networking content is by serving a subpoena on the service provider, think again. The Stored Communications Act (SCA) prohibits internet operators from divulging the contents carried or maintained on the service. 18 U.S.C. § 2702(a)(1) *et seq*. To go this route, you will need either an express, written authorization from the account holder or a court order. Even then, what form of production do you expect to receive?

Ideally, social networking sites would allow users to issue and revoke credentials affording read-only access to site content, or to selected content (like wall posts but not messaging); however, none of the major social networking services do so. Social networking providers not only don't facilitate direct access to user content, some affirmatively prohibit same. As a term of service, Facebook requires users agree that they "will not solicit login information or access an account belonging to someone else."[3] Facebook's terms of service further state, "[y]ou will not share your password [or] let anyone else access your account…."

Accordingly, each court that orders a Facebook user to surrender credentials, also orders the user to violate Facebook's terms of service–a trivial concern perhaps, but one that may give some courts pause.

Apart from gaining access via the service provider or using an opponent's credentials, I made a non-exclusive list of six approaches courts might adopt in dealing with the production of social networking content:

1. **Define scope and charge producing party and/or counsel to examine content and make production.**
2. **Court conducts *in camera* examination.**[4]
3. **Neutral third party expert or special master examines content and makes production within scope.**[5]

[3] https://www.facebook.com/legal/terms (as revised December 11, 2012).

[4] Few courts have the time or inclination to comb through a litigant's social networking sites by in camera review, and fewer attorneys for users will want courts forming impressions about a litigant from such a free-wheeling source. The potential for prejudice is too great.

[5] I've been appointed by courts on many occasions for the purpose of collecting and inspecting private online content. It works, but it's not always cost-effective. Plus, apart from the fact that neutrals have no dog in the fight and thus lack incentive to under- or overproduce, we are generally no better situated to judge relevance than counsel in the case.

4. **Afford supervised access to requesting party (in-person or via screen share), and produce designated content within scope.[6]**
5. **Collect content to external device using agreed-upon tool or method, and produce authenticated copy.[7]**
6. **Requesting party temporarily "Friends" the other side.[8]**

Competent Counsel Continues as Conduit

Like almost any other non-privileged ESI, relevant social networking content is fair game for discovery. By way of analogy, if a photograph's relevant and properly sought in discovery, you have to produce a copy whether the photo's posted on Facebook or comes from a locket around your neck.

The primary responsibility for identifying, preserving and producing relevant social media content lies with producing parties and their counsel. It's their duty to be diligent **and competent** with respect to the evidence within their care, custody and control.

Absent misconduct, incompetence or agreement between the parties, I see little justification to hand over credentials to social media accounts to opponents. The exception for misconduct and incompetence merely acknowledges that the right to keep opponents out of the process presupposes the capability and willingness to fulfill one's discovery duties. If producing parties lack the integrity or the capacity to meet those duties, they forfeit their right to be the conduit for discovery, and courts must intervene to protect the evidence and the process.

In order for this approach to work, there must be some bare level of transparency that allows for confidence in an opponent's ability to act diligently and competently. In stark contrast to paper records, it is not prudent to assume that opponents know how to identify, collect and preserve social networking content. Thus, it's only fair that a requesting party be permitted to inquire about an opponent's plan to preserve and search social networking content and to obtain sufficient answers.[9] This isn't discovery about discovery so much as it is discovery about digital fluency at a time when literacy cannot be assumed.

[6] Affording supervised access, either in person or via screen sharing, has much to commend it. It obviates the need to share credentials with an opponent while still giving the opponent (relatively) unfettered access. Of course, someone must log in as the user, and requesting parties may find working through opposing counsel to be invasive or tedious, especially for sites with a lot of content.

[7] Collecting content to local storage is also effective; however, too few firms possess the tools or skill to facilitate such collection in ways that prove reasonably complete, utile and electronically searchable. Probably the best known purpose-built product facilitating collection from Facebook, LinkedIn and Twitter is X1 Social Discovery, ($945 annual license/ $2,000 perpetual license).

[8] Friending another user in Facebook will not afford access to all of the same content available to the user; however, it might be sufficient in some cases where, e.g., access to messaging is not needed.

[9] Such inquiries are uncomfortable for the resentful majority who lack competence, but the alternative is worse than mere embarrassment; it's the loss of the chance to save the evidence and the loss of a chance to forestall sanctions.

What is Native Production for E-Mail?

by Craig Ball

[Originally published on the Ball in Your Court blog, July 2, 2013]

Recently, I've weighed in on disputes where the parties were fighting over whether the e-mail production was sufficiently "native" to comply with the court's orders to produce natively. In one matter, the question was whether Gmail could be produced in a native format, and in another, the parties were at odds about what forms are native to Microsoft Exchange e-mail. In each instance, I saw two answers; the technically correct one and the helpful one.

I am a vocal proponent of native production for e-discovery. Native is complete. Native is functional. Native is inherently searchable. Native costs less. I've explored these advantages in other writings and will spare you that here. But when I speak of "native" production in the context of databases, I am using a generic catchall term to describe electronic forms with superior functionality and completeness, notwithstanding the common need in e-discovery to produce less than all of a collection of ESI.

It's a Database

When we deal with e-mail in e-discovery, we are usually dealing with database content. Microsoft Exchange, an *e-mail server application*, is a database. Microsoft Outlook, an *e-mail client application*, is a database. Gmail, a *SaaS webmail application*, is a database. Lotus Domino, Lotus Notes, Yahoo! Mail, Hotmail and Novell GroupWise—they're all *databases*. It's important to understand this at the outset because if you think of e-mail as a collection of discrete objects (like paper letters in a manila folder), you're going to have trouble understanding why defining the "native" form of production for e-mail isn't as simple as many imagine.

Native in Transit: Text per a Protocol

E-mail is one of the oldest computer networking applications. Before people were sharing printers, and long before the internet was a household word, people were sending e-mail across networks. That early e-mail was plain text, also called ASCII text or 7-bit (because you need just seven bits of data, one less than a byte, to represent each ASCII character). In those days, there were no attachments, no pictures, not even simple enhancements like **bold**, *italic* or underline.

Early e-mail was something of a free-for-all, implemented differently by different systems. So the fledgling internet community circulated proposals seeking a standard. They stuck with plain text in order that older messaging systems could talk to newer systems. These proposals were called Requests for Comment or RFCs, and they came into widespread use as much by convention as by adoption (the internet being a largely anarchic realm). The RFCs lay out the form an e-mail should adhere to in order to be compatible with e-mail systems.

The RFCs concerning e-mail have gone through several major revisions since the first one circulated in 1973. The latest protocol revision is called RFC 5322 (2008), which made obsolete RFC 2822 (2001) and its predecessor, RFC 822 (1982). Another series of RFCs (RFC 2045-47, RFC 4288-89 and RFC 2049), collectively called Multipurpose Internet Mail Extensions or MIME, address ways to graft text enhancements, foreign language character sets and multimedia content onto plain text emails. These RFCs establish the form of the billions upon billions of e-mail messages that cross the internet.

So, if you asked me to state the native form of an e-mail *as it traversed the Internet between mail servers*, I'd likely answer, "plain text (7-bit ASCII) adhering to RFC 5322 and MIME." In my experience, this is the same as saying ".EML format;" and, it can be functionally the same as the MHT format, but only if the content of each message adheres strictly to the RFC and MIME protocols listed above. You can even change the file extension of a properly formatted message from EML to MHT and back in order to open the file in a browser or in a mail client like Outlook 2010. Try it. If you want to see what the native "plain text in transit" format looks like, change the extension from .EML to .TXT and open the file in Windows Notepad.

The appealing feature of producing e-mail in exactly the same format in which the message traversed the internet is that it's a form that holds the entire content of the message (header, message bodies and encoded attachments), and it's a form that's about as compatible as it gets in the e-mail universe.[10]

Unfortunately, the form of an e-mail *in transit* is often incomplete in terms of metadata it acquires upon receipt that may have probative or practical value; and the format in transit isn't native to the most commonly-used e-mail server and client applications, like Microsoft Exchange and Outlook. It's from these applications–*these databases*–that e-mail is collected in e-discovery.

Outlook and Exchange

Microsoft Outlook and Microsoft Exchange are database applications that talk to each other using a protocol (machine language) called MAPI, for *Messaging Application Programming Interface*. Microsoft Exchange is an e-mail server application that supports functions like contact management, calendaring, to do lists and other productivity tools. Microsoft Outlook is an e-mail client application that accesses the contents of a user's account on the Exchange Server and may synchronize such content with local (i.e., retained by the user) container files supporting offline operation. If you can read your Outlook e-mail without a network connection, you have a local storage file.

Practice Tip (and Pet Peeve): *When your client or company runs Exchange Server and someone asks what kind of e-mail system your client or company uses, please don't say "Outlook." That's like saying "iPhone" when asked what cell carrier you use. Outlook can serve as a front-end client to Microsoft Exchange, Lotus Domino and most*

[10] There's even an established format for storing multiple RFC 5322 messages in a container format called mbox. The mbox format was described in 2005 in RFC 4155, and though it reflects a simple, reliable way to group e-mails in a sequence for storage, it lacks the innate ability to memorialize mail features we now take for granted, like message foldering. A common workaround is to create a single mbox file named to correspond to each folder whose contents it holds (e.g., Inbox.mbox)

webmail services; so saying "Outlook" just makes you appear out of your depth (assuming you are someone who's supposed to know something about the evidence in the case).

Outlook: The native format for data stored locally by Outlook is a file or files with the extension PST or OST. Henceforth, I'm going to speak only of PSTs, but know that either variant may be seen. PSTs are container files. They hold collections of e-mail—typically stored in multiple folders—as well as content supporting other Outlook features. The native PST found locally on the hard drive of a custodian's machine will hold all of the Outlook content that the custodian can see when not connected to the e-mail server.

Because Outlook is a database application designed for managing messaging, it goes well beyond simply receiving messages and displaying their content. Outlook begins by taking messages apart and using the constituent information to populate various fields in a database. What we see as an e-mail message using Outlook is actually a report queried from a database. The native form of Outlook e-mail carries these fields and adds metadata not present in the transiting message. The added metadata fields include such information as the name of the folder in which the e-mail resides, whether the e-mail was read or flagged and its date and time of receipt. Moreover, because Outlook is designed to "speak" directly to Exchange using their own MAPI protocol, messages between Exchange and Outlook carry MAPI metadata not present in the "generic" RFC 5322 messaging. Whether this MAPI metadata is superfluous or invaluable depends upon what questions may arise concerning the provenance and integrity of the message. Most of the time, you won't miss it. Now and then, you'll be lost without it.

Because Microsoft Outlook is so widely used, its PST file format is widely supported by applications designed to view, process and search e-mail. Moreover, the complex structure of a PST is so well understood that many commercial applications can parse PSTs into single message formats or assemble single messages into PSTs. Accordingly, it's feasible to produce responsive messaging in a PST format while excluding messages that are non-responsive or privileged. It's also feasible to construct a production PST without calendar content, contacts, to do lists and the like. You'd be hard pressed to find a better form of production for Exchange/Outlook messaging. Here, I'm defining "better" in terms of completeness and functionality, not compatibility with your ESI review tools.

MSGs: There's little room for debate that the PST or OST container files are the native forms of data storage and interchange for a *collection* of messages (and other content) from Microsoft Outlook. But is there a native format for *individual* messages from Outlook, like the RFC 5322 format discussed above? The answer isn't clear cut. On the one hand, if you were to drag a single message from Outlook to your Windows desktop, Outlook would create that message in its proprietary MSG format. The MSG format holds the complete content of its RFC 5322 cousin plus additional metadata; but it lacks information (like foldering data) that's contained within a PST. It's not "native" in the sense that it's not a format that Outlook uses day-to-day; but it's an export format that holds more message metadata unique to Outlook. All we can say is that the MSG file is a highly compatible *near-native* format for individual Outlook messages–more complete than the transiting e-mail and less complete than the native PST. Though it's encoded in a proprietary Microsoft format (i.e., it's *not* plain text), the MSG format is so ubiquitous that, like PSTs, many applications support it as a standard format for moving messages between applications.

Exchange: The native format for data housed in an Exchange server is its database file, prosaically called the Exchange Database and sporting the file extension .EDB. The EDB holds the account content for everyone in the mail domain; so unless the case is the exceedingly rare one that warrants production of all the e-mail, attachments, contacts and calendars for every user, no litigant hands over their EDB.

It may be possible to create an EDB that contains only messaging from selected custodians (and excludes privileged and non-responsive content) such that you could really, truly produce in a native form. But, I've never seen it done that way, and I can't think of anything to commend it over simpler approaches.

So, if you're not going to produce in the "true" native format of EDB, the desirable alternatives left to you are properly called "near-native," meaning that they preserve the requisite content and essential functionality of the native form, but aren't the native form. If an alternate form doesn't preserve content and functionality, you can call it whatever you want. I lean toward "garbage," but to each his own.

E-mail is a species of ESI that doesn't suffer as mightily as, say, Word documents or Excel spreadsheets when produced in non-native forms. If one were meticulous in their text extraction, exacting in their metadata collection and careful in their load file construction, one could produce Exchange content in a way that's sufficiently complete and utile as to make a departure from the native less problematic—assuming, of course, that one produces the attachments in their native forms. That's a lot of "ifs," and what will emerge is sure to be incompatible with e-mail client applications and native review tools.

Litmus Test: Perhaps we have the makings of a litmus test to distinguish functional near-native forms from dysfunctional forms like TIFF images and load files: ***Can the form produced be imported into common e-mail client or server applications?***

You have to admire the simplicity of such a test. If the e-mail produced is so distorted that not even e-mail programs can recognize it as e-mail, that's a fair and objective indication that the form of production has strayed too far from its native origins.

Gmail

The question whether it's feasible to produce Gmail in its native form triggered an order by U.S. Magistrate Judge Mark J. Dinsmore in a case styled, *Keaton v. Hannum,* 2013 U.S. Dist. LEXIS 60519 (S.D. Ind. Apr. 29, 2013). It's a seamy, sad suit brought *pro se* by an attorney named Keaton against both his ex-girlfriend, Christine Zook, and the cops who arrested Keaton for stalking Zook. It got my attention because the court cited a blog post I made three years ago.

The Court wrote:

> Zook has argued that she cannot produce her Gmail files in a .pst format because no native format exists for Gmail (i.e., Google) email accounts. The Court finds this to be incorrect based on Exhibit 2 provided by Zook in her Opposition Brief. [Dkt. 92 at Ex. 2 (Ball, Craig: Latin: To Bring With You Under Penalty of Punishment, EDD Update (Apr. 17, 2010)).] Exhibit 2 explains that, although Gmail does not support a "Save As" feature to generate a single message format or PST, the messages can be downloaded to Outlook and saved as .eml or.msg files, or, as the author did, generate a PDF Portfolio – "a collection of multiple files in varying format that are housed in a single, viewable and searchable container." [Id.] In fact, Zook has already compiled most of her archived Gmail emails between her and Keaton in a .pst format when Victim.pst was created. It is not impossible to create a "native" file for Gmail emails.
> **Id. at 3.**

I'm gratified when a court cites my work, and here, I'm especially pleased that the Court took an enlightened approach to "native" forms in the context of e-mail discovery. Of course, one strictly defining "native" to exclude near-native forms might be aghast at the loose lingo; but the more important takeaway from the decision is the need to strive for the most functional and complete forms when true native is out-of-reach or impractical.

Gmail is a giant database in a Google data center someplace (or in many places). I'm sure I don't know what the native file format for cloud-based Gmail might be. Mere mortals don't get to peek at the guts of Google. But, I'm also sure that it doesn't matter, because even if I *could* name the native file format, I couldn't obtain that format, nor could I faithfully replicate its functionality locally.[11]

Since I can't get "true" native, how can I otherwise mirror the completeness and functionality of native Gmail? After all, a litigant doesn't seek native forms for grins. A

[11] It was once possible to create complete, offline replications of Gmail using a technology called Gears; however, Google discontinued support of Gears some time ago. Gears' successor, called "Gmail Offline for Chrome," limits its offline collection to just a month's worth of Gmail, making it a complete non-starter for e-discovery. Moreover, neither of these approaches employs true native forms as each was designed to support a different computing environment.

litigant seeks native forms to secure the unique benefits native brings, principally functionality and completeness.

There are a range of options for preserving a substantial measure of the functionality and completeness of Gmail. One would be to produce in Gmail.

HUH?!?!

Yes, you could conceivably open a fresh Gmail account for production, populate it with responsive messages and turn over the access credentials for same to the requesting party. That's probably as close to true native as you can get (though some metadata will change), and it flawlessly mirrors the functionality of the source. Still, it's not what most people expect or want. It's certainly not a form they can pull into their favorite e-discovery review tool.

Alternatively, as the Court noted in *Keaton v. Hannum,* an IMAP[12] capture to a PST format (using Microsoft Outlook or a collection tool) is a practical alternative. The resultant PST won't look or work exactly like Gmail (i.e., messages won't thread in the same way and flagging will be different); but it will supply a large measure of the functionality and completeness of the Gmail source. Plus, it's a form that lends itself to many downstream processing options.

So, What's the native form of that e-mail?

Which answer do you want; the technically correct one or the helpful one? No one is a bigger proponent of native production than I am; but I'm finding that litigants can get so caught up in the quest for native that they lose sight of what truly matters.

Where e-mail is concerned, we should be less captivated by the term "native" and more concerned with specifying the actual form or forms that are best suited to supporting what we need and want to do with the data. That means understanding the differences between the forms (e.g., what information they convey and their compatibility with review tools), not just demanding native like it's a brand name.

When I seek "native" for a Word document or an Excel spreadsheet, it's because I recognize that the entire native file—and *only* the native file—supports the level of

[12] IMAP (for Internet Message Access Protocol) is another way that e-mail client and server applications can talk to one another. The latest version of IMAP is described in RFC 3501. IMAP is not a form of e-mail storage; it is a means by which the structure (i.e., foldering) of webmail collections can be replicated in local mail client applications like Microsoft Outlook. Another way that mail clients communicate with mail servers is the Post Office Protocol or POP; however, POP is limited in important ways, including in its inability to collect messages stored outside a user's Inbox. Further, POP does not replicate foldering. Outlook "talks" to Exchange servers using MAPI and to other servers and webmail services using MAPI (or via POP, if MAPI is not supported).

completeness and functionality I need, a level that can't be fairly replicated in any other form. But when I seek native production of e-mail, I don't expect to receive the entire "true" native file. I understand that responsive and privileged messages must be segregated from the broader collection and that there are a variety of near native forms in which the responsive subset can be produced so as to closely mirror the completeness and functionality of the source.

When it comes to e-mail, what matters most is getting all the important information within and about the message in a fielded form that doesn't completely destroy its character as an e-mail message.

So let's not get *too* literal about native forms when it comes to e-mail. Don't seek native to prove a point. Seek native to prove your case.

Postscript: When I publish an article extolling the virtues of native production, I usually get a comment or two saying, "TIFF and load files are good enough." I can't always tell if the commentator means "good enough to fairly serve the legitimate needs of the case" or "good enough for those sleazy bastards on the other side." I suspect they mean both. Either way, it might surprise readers to know that, when it comes to e-mail, I agree with the first assessment…with a few provisos.

First, TIFF and load file productions can be good enough for production of e-mail if no one minds paying more than necessary. It generally costs more to extract text and convert messages to images than it does to leave it in a native or near-native form. But that's only part of the extra expense. TIFF images of messages are MUCH larger files than their native or near native counterparts. With so many service providers charging for ingestion, processing, hosting and storage of ESI on a per-gigabyte basis, those bigger files continue to chew away at both side's bottom lines, month-after-month.

Second, TIFF and load file productions are good enough for those who only have tools to review TIFF and load file productions. There's no point in giving light bulbs to those without electricity. On the other hand, just because you don't pay your light bill, must I sit in the dark?

Third, because e-mails and attachments have the unique ability to be encoded entirely in plain text, a load file can carry the complete contents of a message and its contents as RFC 5322-compliant text accompanied by MAPI metadata fields. It's one of the few instances where it's possible to furnish a load file that simply and genuinely compensates for most of the shortcomings of TIFF productions. Yet, it's not done.

Finally, TIFF and load file productions are good enough for requesting parties who just don't care. A lot of requesting parties fall into that category, and they're not looking to change. They just want to get the e-mail, and they don’t give a flip about cost, completeness, utility, metadata, efficiency, authentication or any of the rest. If both sides and the court are content not to care, TIFF and load files really are good enough.

A Load (File) Off my Mind

by Craig Ball

[Originally published on the Ball in Your Court blog, July 17, 2013]

I got a call from a lawyer I don't know on Sunday evening. He reported that he'd received production of ESI from a financial institution and spent the weekend going through it. He'd found TIFF images of the pages of electronic documents, but couldn't search them. He also found a lot of "Notepad documents." He'd sought native production, so thought it odd that they produced so many pictures of documents and plain text files.

As it's unlikely a bank would rely on Windows Notepad as its word processor, I probed further and learned that that the production included folders of TIFF images, folders of .TXT files (those "Notepad documents") and folders of files with odd extensions like .DAT and .OPT. My caller didn't know what to do with these.

By now, you've doubtlessly figured out that my caller received an imaged production from an opponent who blew off his demand for native forms and simply printed to electronic paper. The producing party expected the requesting party to buy or own an old-fashioned review tool capable of cobbling together page images with extracted text and metadata in load files. Without such a tool, the production would be wholly unsearchable and largely unusable. When my caller protests, the other side will tell him how all those other files represent the very great expense and trouble they've gone to in order to make the page images searchable, as if furnishing load files to add crude searchability to page images of inherently searchable electronic documents constitutes some great favor.

It brings to mind that classic Texas comeback, "Don't piss in my boot and tell me it's raining."

It also reminds me that not everyone knows about load files, those unsung digital sherpas tasked to tote metadata and searchable text otherwise lost when ESI is converted to TIFF images. Grasping the fundamentals of load files is important to fashioning a workable electronic production protocol, whether you're dealing with TIFF images, native file formats or a mix of the two. I've wanted to write about load files for a long time, but avoided it because *I just hate the damn things*! So, this post is a load (file) off my mind.

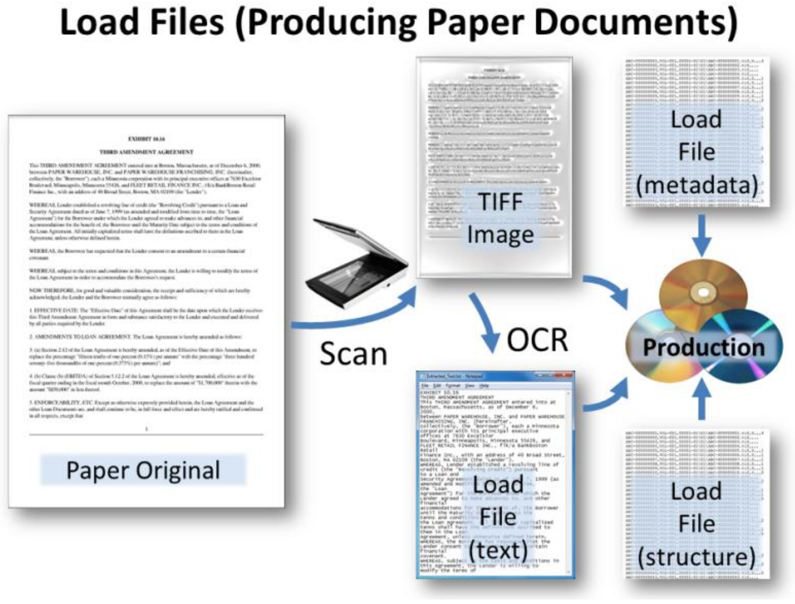
In simplest terms, load files carry data that has nowhere else to go. They are called load files because they are used to load data into, i.e., to "populate" a database. They first appeared in discovery in the 1980s in order to add a crude level of electronic searchability to paper documents. Then as now, paper documents were scanned to TIFF image formats and the images subjected to optical character recognition (OCR). Unlike Adobe PDF images, TIFF images weren't designed to integrate searchable text;

consequently, the text garnered using OCR was stored in simple ASCII[13] text files named with the Bates number of the corresponding page image. Compared to paper documents alone, imaging and OCR added functionality. It was 20th century computer technology improving upon 19th century printing technology, and if you were a lawyer in the Reagan-era, this was Star Wars stuff.

Metadata is “data about data.” While we tend to think of metadata as a feature unique to electronic documents, paper documents have metadata, too. They come from custodians, offices, files, folders, boxes and other physical locations that must be tracked. Still more metadata takes the form of codes, tags and abstracts reflecting reviewers’ assessments of documents. Then as now, all of this metadata needs somewhere to lodge as it accompanies page images on their journey to document review database tools (a/k/a “review platforms”) like Concordance or Summation–venerable products that survive to this day. This data goes into load files.

Finally, we employ load files as a sort of road map and as assembly instructions laying out, *inter alia*, where document images and their various load files are located on disks or other media used to store and deliver productions and how the various pieces relate to one-another.

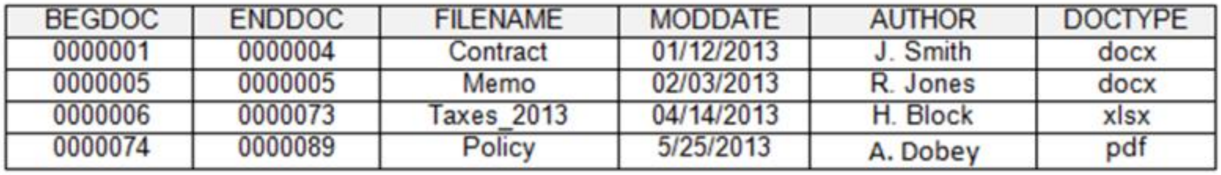
So, to review, some load files carry extracted text to facilitate search, some carry metadata about the documents and some carry information about how the pieces of the production are stored and how they fit together. Load files are used because neither paper nor TIFF images are suited to carrying the same electronic content; and if it weren’t supplied electronically, you couldn’t load it into review tools or search it using computers.



Before we move on, let’s spend a moment on the composition of load files. If you were going to record many different pieces of information about a group of documents, you might create a table for that purpose. Possibly, you’d use the first column of your table to give each document a number, then the next column for the document’s name and then each succeeding column would

[13] ASCII is an acronym for American Standard Code for Information Interchange and describes one of the oldest and simplest standardized ways to use numbers—particularly binary numbers expressed as ones and zeroes–to denote a basic set of English language alphanumeric and punctuation characters.

carry particular pieces of information about the document. You might make it easier to tell one column form the next by drawing lines to delineate the rows and columns, like so:



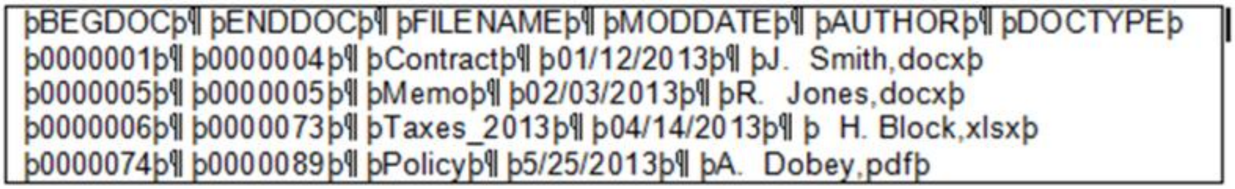
BEGDOC	ENDDOC	FILENAME	MODDATE	AUTHOR	DOCTYPE
0000001	0000004	Contract	01/12/2013	J. Smith	docx
0000005	0000005	Memo	02/03/2013	R. Jones	docx
0000006	0000073	Taxes_2013	04/14/2013	H. Block	xlsx
0000074	0000089	Policy	5/25/2013	A. Dobey	pdf

Those lines separating rows and columns serve as delimiters; that is, as a means to (literally) delineate one item of data from the next. Vertical and horizontal lines serve as excellent visual delimiters for humans, where computers work well with characters like commas, tabs and such. So, if the data from the table were contained in a load file, it might appear as follows:

```
BEGDOC,ENDDOC,FILENAME,MODDATE,AUTHOR,DOCTYPE
0000001,0000004,Contract,01/12/2013,J. Smith,docx
0000005,0000005,Memo,02/03/2013,R. Jones,docx
0000006,0000073,Taxes_2013,04/14/2013,H. Block,xlsx
0000074,0000089,Policy,5/25/2013,A. Dobey,pdf
```

Note how each comma replaces a column divider and each line signifies another row. Note also that the first or “header” row is used to define the type of data that will follow and the manner in which it is delimited. When commas are used to separate values in a load file, it’s called (not surprisingly) a “comma separated values” or CSV file. CSV files are just one example of standard forms used for load files. More commonly, load files adhere to formats compatible with the Concordance and Summation review tools. Concordance load files typically use the file extension DAT and the þ¶ þ characters as delimiters, e.g.:

Concordance Load File



```
þBEGDOCþ¶ þENDDOCþ¶ þFILENAMEþ¶ þMODDATEþ¶ þAUTHORþ¶ þDOCTYPEþ
þ0000001þ¶ þ0000004þ¶ þContractþ¶ þ01/12/2013þ¶ þJ.  Smith,docxþ
þ0000005þ¶ þ0000005þ¶ þMemoþ¶ þ02/03/2013þ¶ þR.  Jones,docxþ
þ0000006þ¶ þ0000073þ¶ þTaxes_2013þ¶ þ04/14/2013þ¶ þ  H. Block,xlsxþ
þ0000074þ¶ þ0000089þ¶ þPolicyþ¶ þ5/25/2013þ¶ þA.  Dobey,pdfþ
```

Summation load files typically use the file extension DII, but do not structure content in the same way as Concordance load files; instead, Summation load files separate each record like so:

Summation Load File

```
; Record 1
@T 0000001
@DOCID 0000001
@MEDIA eDoc
@C ENDDOC 0000004
@C PGCOUNT 4
@C AUTHOR J. Smith
@DATESAVED 01/12/2013
@EDOC \NATIVE\Contract.docx
; Record 2
@T 0000005
@DOCID 0000005
@MEDIA eDoc
@C ENDDOC 0000005
@C PGCOUNT 1
@C AUTHOR R. Jones
@DATESAVED 02/03/2013
@EDOC \NATIVE\Memo.docx
; Record 3
@T 0000006
@DOCID 0000006
@MEDIA eDoc
@C ENDDOC 0000073
@C PGCOUNT 68
@C AUTHOR H. Block
@DATESAVED 04/14/2013
@EDOC \NATIVE\Taxes_2013.xlsx
; Record 4
@T 0000074
@DOCID 0000074
@MEDIA eDoc
@C ENDDOC 0000089
@C PGCOUNT 15
@C AUTHOR A. Dobey
@DATESAVED 05/25/2013
@EDOC \NATIVE\Policy.pdf
```

Just as placing data in the wrong row or column of a table renders the table unreliable and potentially unusable, errors in load files render the load file unreliable, and any database it populates is potentially unusable. Just a single absent, misplaced or malformed delimiter can result in numerous data fields being incorrectly populated. Load files have always been an irritant and a hazard; but, the upside was they supplied a measure of searchability to unsearchable paper documents.

Fast forward to a post-personal computer, post-Internet era

The overwhelming majority of documents and communications are created and stored electronically, and only the tiniest fraction of these will ever be printed. Electronic documents are inherently searchable and do things that paper documents can't, like dynamically apply formulas to numbers (spreadsheets), animate text and images (presentations) or carry messages and tracked changes made visible or invisible at will (word processed documents). Electronic documents also have complements of information within and without called metadata that tend to be lost when electronic documents are printed or imaged. Some of this metadata has evidentiary value (e.g., date and time information) and some has organizational value (e.g., file names).

Because electronic documents are inherently electronically searchable, there's no need to image them or use optical character recognition to extract searchable text. Moreover, there's less need for error-prone load files to populate review tools. Despite these advantages, many lawyers prefer to approach electronic documents in the same way they handled paper documents. That is, they convert searchable electronic documents to non-searchable, non-functional TIFF images and then attempt to graft on electronic searchability by extracting text and metadata to load files.

So, why is an old, error-prone method of data transfer still used in electronic discovery? Good question; because it's not cheaper, and it's certainly not better. Mostly, it's just familiar, and they have a sunk cost in outmoded tools and techniques. Why do some people still use thermal fax paper (for that matter, why do they still use fax machines)?

To be fair, there's a lingering need for load files in e-discovery, too. Even native electronic documents have outside-the-file or "system" metadata that must be loaded into review tools; plus, there's still a need to keep track of such things as the original monikers of renamed native files and the layout of the production set on the production media. In e-discovery, load files—and the headaches they bring–will be with us for a while; *understanding* load files helps ease the pain.

Acrobat to the Rescue: Searching Unsearchable Productions

by Craig Ball

[Originally published on the Ball in Your Court blog, July 21, 2013]

In a perverse irony, lawyers often ‘brag’ about how little they know about information technology; but in situations where admitting confusion could help them, they clam up. Abraham Lincoln said, “Better to remain silent and be thought a fool than to speak out and remove all doubt.” But with respect to problems in electronic discovery, it’s foolish to stay silent.

Sadly, many requesting parties are flummoxed by what’s produced to them. Rather than confess their confusion, they suffer in silence, opening or printing TIFF images one page at a time with nary a clue how to search what they’ve received. And when a production arrives broken—lacking some essential element required for completeness or functionality—the silent majority often doesn’t know what they’re missing. Instead, they laboriously flail away at the evidence, hoping to turn up something useful. It’s a painful and unnecessary ordeal.

Case in point: a client received a production of about 5,000 documents; mostly e-mail messages, all produced as Adobe Portable Document Files or PDFs. Though the documents derived from inherently searchable electronic originals, all the PDFs were created without a searchable text layer, and no extracted text or any fielded data were furnished in accompanying load files. *Ouch!*

E-discovery denizens reading this will grasp the deviousness of the production. It ruthlessly destroys any ability to search or sort the documents electronically and runs afoul of the Federal mandate stating, “If the responding party ordinarily maintains the information it is producing in a way that makes it searchable by electronic means, the information should not be produced in a form that removes or significantly degrades this feature.” *Comments to Rule 34(b) of the Federal Rules of Civil Procedure.*

Innocent mistake? Hardly. The producing party is a Fortune 50 corporation with a storied history of discovery abuse. It’s not their first rodeo.

The producing party surely knows that it will have to supply a replacement production, if sought; but, it also knows that most requesting parties won’t raise a ruckus for fear that an objection will prompt a humiliating, “apparently you don’t understand how to use what we gave you.” With the lack of e-discovery competence extant, most opponents will let it pass unaware. Ignorance is bliss, more so when you can take advantage of the ignorant.

But stripping out searchability and holding back load files is advantageous even when sprung on a savvy opponent like my client. *It buys time.* Depositions must be put off and discovery deadlines or trial dates moved. Opponents squander resources fiddling with the broken production, drafting motions and hiring experts. It’s a tactic that rarely engenders sanctions or cost-shifting because few judges are going to punish a producing party who agrees to promptly supplement—leastwise not on the first go-round. Every dog gets one bite…per lawsuit.

So, if you’ve received a production like the brain dead PDFs mentioned, how do you muddle through and deny your opponent the benefit of such delaying tactics? There’s no pat answer, but I’ll describe the quick-and-dirty approach I took to assist a lawyer who, on the eve of depositions, said, “I’ve just got to go forward with what I’ve got.”

If you’re stuck with unsearchable document images, there are three things you can do to add electronic searchability:

1. You can obtain the native source document or a near-native counterpart;
2. You can obtain extracted text and the requisite load file data that pairs the text with the images; or,
3. You can run optical character recognition (OCR) against the images to extract text.

The third option is the only one you can undertake without obtaining further production from the other side, so it was the only option here.[14][1]

For the most part, the PDFs produced held clean text. That is, because they derived from electronic originals, there were few handwritten annotations, skewed scans, funky fonts or other characteristics to confound OCR. OCR is error-prone at its best; but, it performs abysmally on anything but clean text images.

Once they had the extracted text of the documents in an electronic format, my clients would need a means to pair the extracted text with the correct page image and to search the text. If the mechanism employed indexed the text so as to speed search and supported Boolean and proximity searching, even better.

So, I turned to Adobe Acrobat. The old version 9 Pro edition of Acrobat on my machine is up-to-date enough to create Acrobat Portfolios, run OCR against the contents and even optimize the index for speedier search. It also supports Boolean and proximity searching in a simple-to-use interface that includes a preview mechanism and a basic way to annotate notable documents.

While you need Adobe Acrobat versions 9, 10 or 11 to create a portfolio the recipient of the portfolio just needs the ubiquitous, free Acrobat Reader application to open, view and search it. A PDF Portfolio supports a simple browser-style viewer format in Acrobat Reader, so the documents are very quick to peruse.

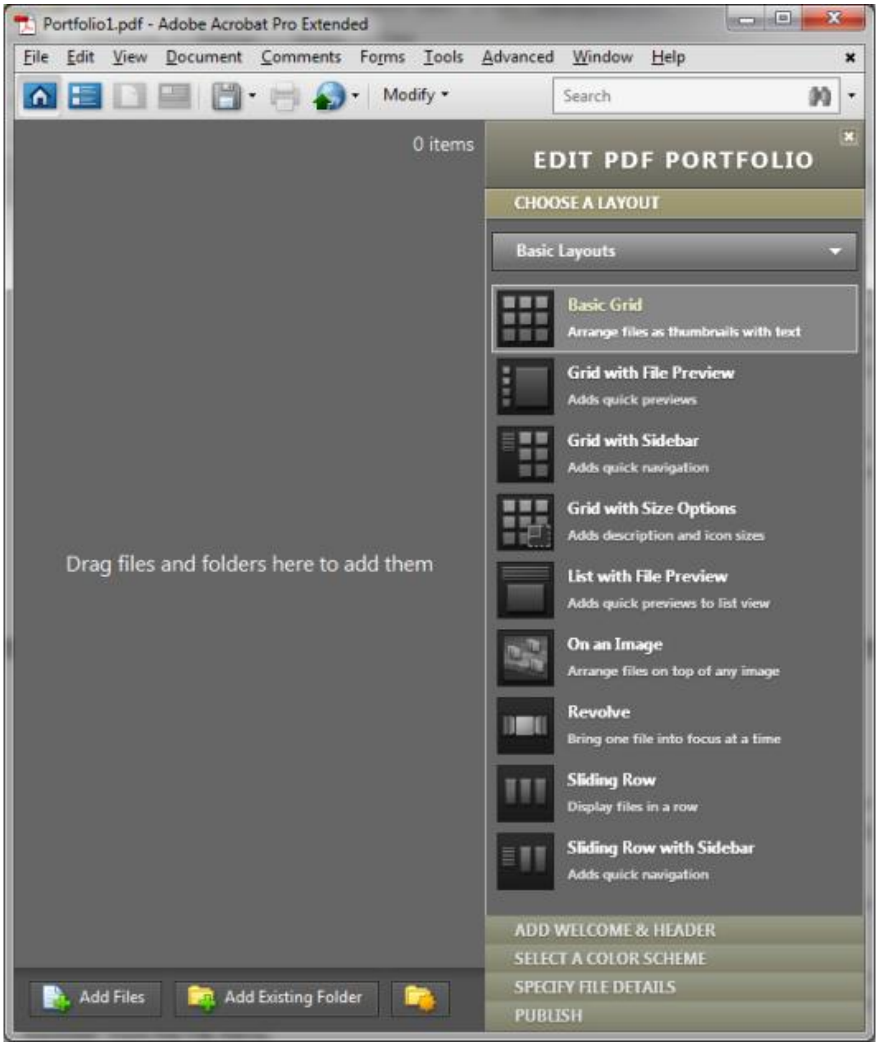
[14] I suppose you could have typists recreate all of the text in the documents manually; but, I shudder to think what that would cost.

Here, I need to reiterate the key difference between Adobe Acrobat products that just seems to stymie so many. Adobe gives away a program called Adobe Reader. It reads PDF formats, but it doesn't create them. Repeat: it doesn't create PDFs or Portfolios. It just reads them. It's called "Reader." Why? Because IT DOESN'T CREATE PDFs. It's free, so enjoy what it does, which is read PDFs. Only.

Adobe sells products called Acrobat (so named because you have to perform gymnastics to get people to understand that the Reader product just reads PDFs). The Acrobat products *create* PDFs, including Portfolios, from Version 9 forward. This is how Adobe makes money: free reader, $350 writer.

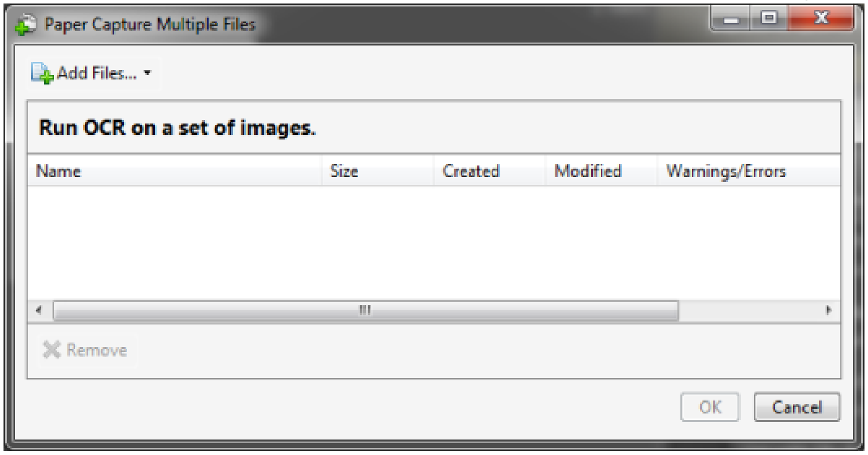
But like most law offices, you already have a copy of the Adobe Acrobat program. The writer, not the…oh, never mind.

To create the searchable Portfolio from almost 5,000 non-searchable PDFs comprising 1.7GB of data, I began by copying the PDFs I wanted to make searchable into a separate folder. Next, I ran Adobe Acrobat and selected "Create PDF Portfolio" from the File menu. The Edit PDF Portfolio window seen below opened.



I then selected "Add Existing Folder" from the bottom of the window and pointed the program to the folder I'd filled with unsearchable PDFs. Acrobat began assembling the Portfolio from the files. It took only a few minutes to 'bind' the documents into a virtual notebook; however, what I had wasn't yet searchable.

The next step was to run optical character recognition against all the documents in the Portfolio. Adobe Acrobat has a built-in basic OCR capability. From the Document menu, I selected OCR Text Recognition>Recognize Text in Multiple Files Using OCR. The dialog box that appeared allowed me to Add Files > Add Open Files. As I'd not yet saved it, the portfolio in progress was called "Portfolio1.pdf" by default. I selected it and my Output Options; then, I left for dinner because it would take hours for Acrobat to extract text from an estimated 30-40 thousand page images using optical character recognition.



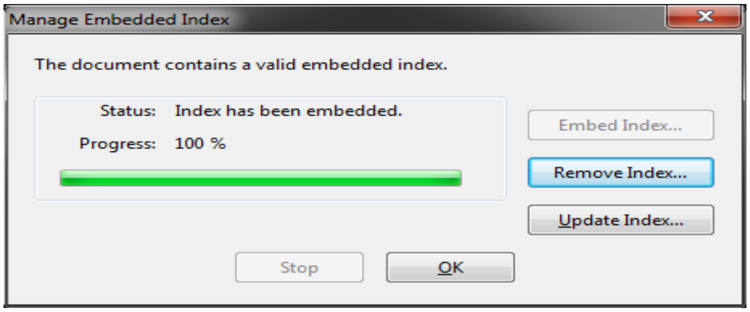
Run OCR on a set of images.

Name	Size	Created	Modified	Warnings/Errors

Before you vendors reading this add, "Our tool would be better for this," please remember that the goal here was fast and cheap. Your wares cost more than free and carry a steeper learning curve than an application law firms already have and use. Adobe Acrobat doesn't deliver the benefits of applications purpose-built for e-discovery; but, it's the butter knife that serves as a decent screwdriver in a pinch.

When the OCR engine completed its work, all of the documents in the collection were now text searchable…sort of. Text in uncommon typefaces or unclear to the OCR engine was rendered incorrectly or not at all. Gray scale content remained largely unsearchable. What emerged was far more utile than what was produced, but fell short of what *should* be exchanged in e-discovery.

Searching was slow because each PDF in the portfolio had to be searched one-by-one. To speed search, the next step was to generate an index for the contents of the portfolio. From the Advanced menu, I selected Document Processing, set my parameters and generated an index.[15] I let this run for a few hours more until completion.



The document contains a valid embedded index.

Status: Index has been embedded.

Progress: 100 %

Now, I had something I could give my client to enable his team to run text and proximity searches against the collection,[16] even if the only tool they had to use was a free copy

[15] In Acrobat 10 and 11, look for this option in the Tools menu.

[16] For Boolean and proximity searches, use the Advanced Search dialogue box. If you have trouble getting the Advanced Search box to appear (as I did with Acrobat 9), try this: Open Acrobat, then open the Advanced Search dialogue box and only then open the Portfolio file. The window stays open and supports the advanced search options.

of Acrobat Reader. It’s even feasible for reviewers using Acrobat to add tags in each document’s description field (or in a custom field added by the reviewer) and sort by those fields and tags. A lagniappe of the process is that, by consolidating the PDFs into a Portfolio, they’re compressed and stored more efficiently. Even with the added text, the searchable Portfolio is one-third the collective size of the documents it holds.

My client can go prepare for the depositions. Acrobat rode to the rescue; yet, the Portfolio workaround detailed here is far from optimum. It’s triage: quick, low cost and preferable to having no review platform and no ability to search the production, but no substitute for a proper production.