

LawTechnologyNews

THE PROFESSION'S LARGEST TECH PUBLICATION

Products, Systems & Services for Legal Professionals



Ball in Your Court

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 back up 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 back up tape. We must be vigilant to avoid a reflexive calculus like: All back up tapes are inaccessible > Inaccessible means undue burden is presumed > Good cause must be shown before access is granted > Requesting party pays cost of converting data from inaccessible" 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.

Craig Ball, a member of the LTN Editorial Advisory Board, is a litigator and trial technology consultant based in Montgomery, Texas. E-mail: craigball@gmail.com.

LawTechnologyNews

THE PROFESSION'S LARGEST TECH PUBLICATION

Products, Systems & Services for Legal Professionals



Ball in Your Court

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 in place to double back over what has been seen 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 are applied 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 the third wave of data is analyzed, what about what was missed in waves one and two? Here, 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.

Craig Ball, a member of the LTN Editorial Advisory Board, is a litigator and trial technology consultant based in Montgomery, Texas. E-mail: craigball@gmail.com.

LawTechnologyNews

THE PROFESSION'S LARGEST TECH PUBLICATION

Products, Systems & Services for Legal Professionals



Ball in Your Court

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, back up 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.

Craig Ball, a member of the LTN Editorial Advisory Board, is a litigator and trial technology consultant based in Montgomery, Texas. E-mail: craigball@gmail.com.