

tidBytes

Doing It Right Once: “Data Wellness”

by Larry G. Johnson, Esq.

Digital data glut is undermining major enterprises everywhere. Big companies literally “don’t know what they know,” and as a result all kinds of useful wheat is being lost in ever-larger mountains of chaff. When lawyers have to respond to electronic discovery requests, how can they hope to wade through all their clients’ electronic data, often in the millions of documents (terabytes), when even *they* can’t?

Imagine a sci-fictional, Mayo Clinic-like magical treatment where you’d never get sick again. Even if it took days and cost a lot of money, would you take it? You bet you would. You’d wonder why *anyone* wouldn’t. Yet corporate and government enterprises, drowning in the electronic data they generate every day, have been reluctant to take available “Mayo Clinic cures” to trim fat from their digital data and exploit the useful information available in their data stores. Especially the data called “unstructured data:” all those emails and their attachments; word-processed files; spreadsheets; photos; sound files (including voice mail); PowerPoints and video. In other words, all the stuff you and everybody else create.

It’s the unstructured data that’s causing congestive heart failure in corporate USA. Of course, MIS staff, consultants and software developers have made various efforts to attack the problem. But, as far as I can tell, until recently, most knowledge management software applied to unstructured data has been like herding cats. Top-down categorization systems (“taxonomies”), such as the ancient Dewey Decimal System used in libraries, or the WestLaw Key Number system, are as limiting as they are useful. Rarely do two people precisely agree on how to categorize a set of documents, let alone agree on which documents fit what category. It can be hit and miss. Descriptive HTML or XML metatags in Web documents suffer from the same limitations. And all global taxonomies are subject to the limiting interest (bias) of the categorizer. What may be important to a lawyer in defining a document may not be the same paragraph of interest to an accountant, or to somebody in R&D. Chaotic data can lead to equally chaotic taxonomies for the same unstructured data.

But, to steal a campaign slogan in this election season: “Hope is on the way.” Software solutions are emerging that do a one-time, across-the-board cleanup and organization of the enterprise digital data chaos, with a bonus at the end: You get a kind of 21st Century digital überlibrary for all future electronic data coming into or generated from within the enterprise.

All these solutions use mathematical modeling, called latent semantic profiling, and some of these vendors go further by adding social networking algorithms, linking people to actions they take as reflected in the documents. And both of these technologies promise to merge into something still evolving and enormously exciting: the Semantic Web, or as some call it: “Internet 2.0.”

Letting the documents organize themselves

Latent semantic profiling organizes documents into clusters based on their meaning, arrived at by pattern-matching and indexing of all words in a data set, and the relationship of each word to every other word, i.e. their contexts. Such software can thus distinguish the use of the word “diamond” when used in a set of documents dealing with baseball, as opposed to documents that talk about jewelry. The more documents the software clusters by meaning, the better a job it does in grouping them according to their textual content.

Digital voice files can also respond to semantic profiling, since measurable sound waves parallel the smallest units of recognizable speech, called “phonemes.” This technology can also be speaker-independent and language-independent; it doesn’t matter who is speaking or if it’s in Russian or Arabic. Can you guess what government agencies are already using this stuff?

In the environment of these semantic profiles (imagine a unique analog wave for each document that is linked to other “document waves” of like kind), after all the unstructured, human-generated documents have been extracted, processed and “de-duplicated” (not a small task in itself), you can now experience a quantum leap in your ability to find what you’re looking for. You no longer have to rely on keyword search terms or “fuzzy logic” or thesauruses to hit the bull’s-eye. You have graduated to “content searches” where you use a browser-like application to find any document “about” a topic, say “toxic waste in rivers.” Type those words in, and it may produce on the screen a document that does not contain the words “toxic”, “waste” or “river,” but which does contain language about *arsenic dumping* in the *Mississippi*.

As with all good technology, it will not only be more powerful than what it replaces, it will be easier to use. If you have a key document in hand (or even one that you create that would be your “dream” smoking gun), all you have to do is tell the software to find other documents like it. And it will. An entire document can be, if you will, the search term.

Adding people to the mix

Some companies that offer content search technology stop there. But others go further in offering social network analysis capabilities. For example, one company, Cataphora (www.cataphora.com), isolates “conversations” between people engaged in a common project or action, based on the content of their email exchanges. This capability can, for example, help define who within a vast enterprise is providing valuable talent and leadership on a specific project. It can also assist in uncovering cabals and conspiracies by showing, mapped over time, who did what and when, and whether people who ordinarily shouldn’t have access to certain information suddenly do. Further, an innocuous email with a message like, “Stay alert – tomorrow a certain bird will be singing,” can make a whole lot of sense if the full context of people’s actions and their communications is isolated.

The implications for data security, prevention of sexual harassment and other inappropriate employee activity, IP protection – and electronic discovery of key documents – are self-evident.

The Semantic Web – the Next Big Thing

The guy who invented the World Wide Web, Sir Tim Berners-Lee, is working on the completion of what some call Internet 2.0. In simplest terms, the web will soon evolve into one gigantic Semantic Web, whereby information will be linked in ways similar to the way semantic profiling and social network analysis work. In time, every piece of information on the Web will be linked to every other piece of information in a massive neural matrix so you can manipulate unimaginable amounts of data for any purpose.

Say, for instance, you're a photographer and you want sunset shots of 20 famous landmarks in Europe. You tell the Semantic Web that's what you want to do 90 days from now. What you get back is an itinerary that plots the optimum travel routes according to all the European train schedules, the available hotel rooms, and the weather forecasts and sunset times for the places you're interested in. Click a button and all the necessary hotel and train reservations are made, based on a budget averaged from what you spent on all your prior travels, with everything arranged to be paid from your bank account. Cool, no?

The payoff: the überlibrary

For Sir Tim, creator of the Semantic Web, “the most exciting thing is serendipitous reuse of data: one person puts data up there for one thing, and another person uses it another way.” Using semantic profiling and social network analysis software, you can get similar results. If I am corporate counsel assisting in responding to document production requests, I can find examples of attorney-client privileged documents; I then instruct the software to find all documents like them. I can put those documents into a database of my own private taxonomy called “privileged documents.” I can then set up subcontainers for specific cases (“attorney-client privilege documents in the XYZ case”), so that semantic profiles start to emerge *on my documents that show what I am interested in*.

Accountants in the corporation can look at the same data I have looked at, but for their own purposes, and they, too, will start building their containers. R&D people will do their thing, too.

Over time, it won't make any difference if there are inconsistencies in how users of the master dataset make up categories for their private use. No matter how you or I interpret the data, the relationships between the documents in the master dataset remain unchanged.

So you get to have your cake and eat it too: independent constituency organization of data without disturbing the data. For lawyers, the “constituency” is the case or matter at hand, and the organizing principle is relevance. Issues can shift, change or emerge late in a case, but so can the queries to the data set. And, theoretically, in class action and “pattern litigation” cases (e.g., products liability cases), whole sets of documents can be used as the “search term:” *here are all the documents that were relevant to the issue of failure to warn in the XYZ case; find me everything that is like those documents in this case.*

Why CEOs get paid the big bucks

Right now, most electronic discovery is done in a piecemeal, hodge-podge fashion, involving two worlds that don’t communicate well with each other: IT managers and lawyers. The IT people rarely understand the litigation process and what lawyers need, and the trial lawyers are often clueless about how to find and capture the relevant e-data on hard drives, servers and backup storage media.

Rather than breach that gap effectively with consultants knowledgeable in both IT and legal worlds, attempts are made to minimize the interface between them. For example, courts and think tanks like the Sedona Conference, of which I am a member, promote sampling of data first to see if the most likely data sources reveal anything. That may work if you assume a few emails, for example, will be coherent and complete enough to have evidentiary value. But what about the “bird singing” email cited above? The reference may be code that can only be untangled if a much broader context is examined.

But we haven’t been talking about just e-discovery here. Cutting e-discovery costs and pushing for efficiencies in litigation are just a small subset of the bigger problems facing corporations with digital data glut. What corporations need to do – for all sorts of departmental constituencies, not just corporate counsel – is radical, overall, top-down digital data rationalization and knowledge management using the tools described above. It is a CEO-level problem requiring CEO decisions.

There is an inevitability of “doing it once right” with all corporate data. Once *all* of it is cleaned up and rationalized, channeling new data thereafter into the myriad of private structured environments in the Semantic Web-like überlibrary should be relatively painless – even exciting. Like going to the Mayo Clinic, it can be expensive and intrusive, but it’s the best way to jettison old habits and start on the path to “data wellness.”

Larry Johnson is President of Legal Technology Group, Inc. (www.legaltechnologygroup.com), a company composed of lawyer technologists providing expert litigation support services to trial lawyers and digital data audit and risk management services for corporate counsel.

References and resources:

http://www.dmreview.com/article_sub.cfm?articleId=1009161, for a discussion of structured and unstructured data, and the challenges in integrating them.

Sir Tim Berners Lee quotation from “Sir Tim Berners Lee: He Created the Web, Now He’s Working on Internet 2.0,” *MIT Technology Review*, October 2004, pp. 40-45.

Companies with some or all of the features described in this column include, in alphabetical order: Attenex, Autonomy, Cataphora, DolphinSearch, Engenium, Fios and Syngence.

For integrating content search software into a best-practices Litigation Preparedness Plan, see my article: <http://www.legaltechnologygroup.com/lpp.pdf>.