E-Discovery: The Virtues of Fielding

By Craig Ball

I am a member of the Typewriter Generation. With pencil and ink, we stored information on paper. My generation tends to think of stored information as tangible *things* we persist in calling "documents." But unlike calling a data directory a "folder" (despite the absence of any folded thing) or simulating the sound of a shutter click when taking a digital photo (despite the absence of a shutter), couching requests for production as demands for documents is not harmless skeuomorphism. The mindset that electronically stored information items are just electronic paper documents makes e-discovery more difficult and costly, and it hampers legal professionals as they strive toward competence in e-discovery.

Does clinging to the notion of "document" really hold us back? I think so, because continuing to define what we seek in discovery as "documents" ties us to a two-dimensional view of four-dimensional information. The first two dimensions of a "document" are its flat content--what emerges when you print it to paper or static image format like TIFF. But electronically stored information (ESI) always implicates a third dimension, *metadata and embedded content*, and sometimes a fourth, *temporal* dimension, as we often discover different versions of information items *over time*.

The distinction becomes crucial when considering suitable forms of production and prompts a need to understand the concept of *Fielding* and *Fielded Data.* We must recognize that preserving the fielded character of data is essential to preserving its utility and searchability.

When I say data is "fielded," I mean that information is stored in locations dedicated to holding only particular information (e.g., date, author, zip code, record number and price). Fielding data serves to separate and identify information so you can search, sort and cull it using just certain fields. It's a capability we take for granted in digital applications but that is often crippled or eradicated when data is produced in e-discovery.

Fielding data isn't new. We did it back when data was stored as paper documents. Take a typical law firm letter: the letterhead identifies the firm, the date below the letterhead is understood to be the date sent. A *Re:* line follows, denoting matter or subject, then the addressee, salutation, etc. The recipient is understood to be named at the start of the letter and the sender at the bottom. These conventions governing where to place information are vital to our ability to understand and organize conventional correspondence.

Similarly, all of the common productivity file types encountered in e-discovery (Microsoft Office formats, PDF and e-mail) employ fielding to abet utility and functionality. Native "documents" are natively fielded; that is, a file's content is structured to insure that particular pieces of information reside in defined locations within the file. This structure is understood and exploited by the native application and by tools designed to avail themselves of the file architecture.

We act inconsistently, inefficiently and irrationally when we deal with fielded information in e-discovery. Just a few years ago, lawyers resisted production of spreadsheets in native, fielded formats. Now, only the most Neanderthal counsel challenges the need to produce the native fielding of spreadsheet data. Accordingly, production of spreadsheets in native forms has evolved to become routine and (largely) uncontentious. To reach this point, workflows were modified, Bates numbering procedures were tweaked, and despite dire predictions, none of it made the sky fall. We must now bring the same intelligence to PowerPoint presentations, Word documents and, above all, to discovery from databases.

Take e-mail. All e-mail is natively fielded data, and the architecture of e-mail messages is established by published standards called RFCs—structural conventions that e-mail applications and systems must embrace to insure that messages can traverse any network. The RFCs define placement and labeling of the sender, recipients, subject, date, attachments, routing, message body and other components of every e-mail that transits the Internet.

But when we produce e-mail in discovery, the "standard" practice is to deconstruct each message and produce it in a crudely fielded format that's incompatible with the RFCs and unrecognizable to any e-mail tool or system. Too, the production is almost always incomplete compared to the native messaging.

The deconstruction of fielded data is accomplished by a process called **Field Mapping**. The contents of particular fields within the native source are extracted and inserted into a matrix that may assign the same name to the field as accorded by the native application or rename it to something else altogether. Thus, the source data is "mapped' to a new name and location. At all events, the mapped fields
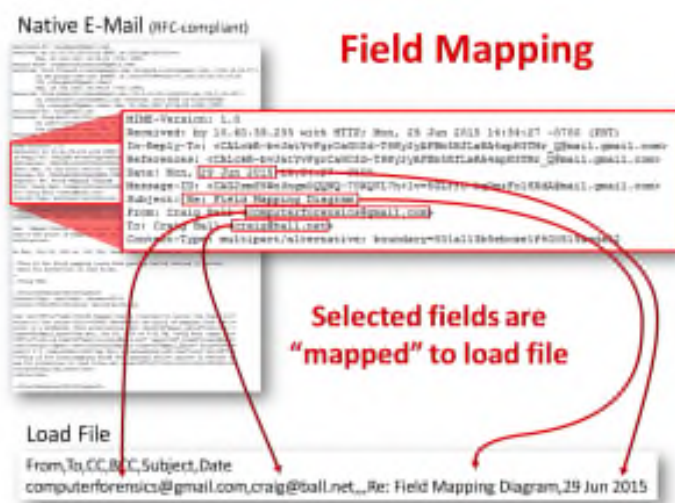
never mirror the field structure of the source file.

The jumbled fielding doesn't entirely destroy the ability to search within fields or cull and sort by fielded content; but, it requires lawyers to rent or buy tools that can re-assemble and read the restructured data in order to search, sort and review the content. And again, information in the original is often omitted, not because it's privileged or sensitive, but the producing party simply elects not to supply it.

The omitted content is not trivial.   In fact, the omitted information significantly aids our ability to make sense of the production, such as the fielded data that allows messages to be organized into conversational threads (*e.g.*, In-Reply-To, References and Message-ID fields) and the fielded data that enables messages to be correctly ordered across time zones and daylight savings time (*e.g.,* UTC offsets).

"Why do producing parties get to recast and omit this useful information," you ask? Not enough lawyers or judges are asking that question.

The answer is that counsel, and especially requesting counsel, are asleep at the wheel. Producing parties have not been challenged on this conduct and, when challenged, have fallen back on crusty claims that it's an industry standard.

E-discovery standards have evolved to acknowledge that e-mail must be supplied with some fielding preserved; but, there is no sound reason to produce e-mail with shuffled or omitted fields. It doesn't cost more to be faithful to the native or near-native architecture or be complete in supplying fielded content; in fact, producing parties pay *more* to degrade the production, and what emerges costs more to review.

Perhaps the hardest thing for lawyers and judges to appreciate is the importance fielding plays in culling, sorting and search.

- It's efficient to be able to cull and sort files *only* by certain dates.
- It's efficient to be able to search *only* within e-mail recipients.
- It's efficient to be able to distinguish Speaker Notes within a PowerPoint or filter by the Author field in a Word document.

Preserving the fielded character of data makes these actions possible and much more. Preserving the fielded data *and* the native file architecture allows use of a broad array of

tools against the data, where restructuring fielded data limits its use to only a handful of pricey tools that understand peculiar and proprietary production formats.

It's not enough for producing parties to respond, "*But, you can reassemble the kit of data we produce to make it work somewhat like the original evidence.*" In truth, you often can't, and you shouldn't have to try.

It ties back to the Typewriter Generation mentality that keeps us defining everything we seek as "documents." Most information sought in discovery today is not a purposeful precursor to something that will be printed. Most modern evidence is data; *fielded* data. Modern productivity files aren't blobs of text, they're ingenious little *databases*. *Powerful, rich, databases.* Their native content and architecture are key to their utility and efficient searchability in discovery. Get the fielding right, and functionality follows.

About the Author

Craig Ball of Austin is a Board-certified trial lawyer who limits his practice to service as a court-appointed Special Master and consultant in computer forensics and electronic discovery. A founder of the Georgetown University Law Center E-Discovery Training Academy, Craig serves on the Academy's faculty and also teaches Electronic Discovery and Digital Evidence at the University of Texas School of Law. For nine years, Craig penned the award-winning *Ball in Your Court* column on electronic discovery for American Lawyer Media and now writes for several national news outlets. Craig has published and presented on forensic technology more than 1,650 times, all over the world. For his articles on electronic discovery and computer forensics, please visit www.craigball.com or his blog, www.ballinyourcourt.com.