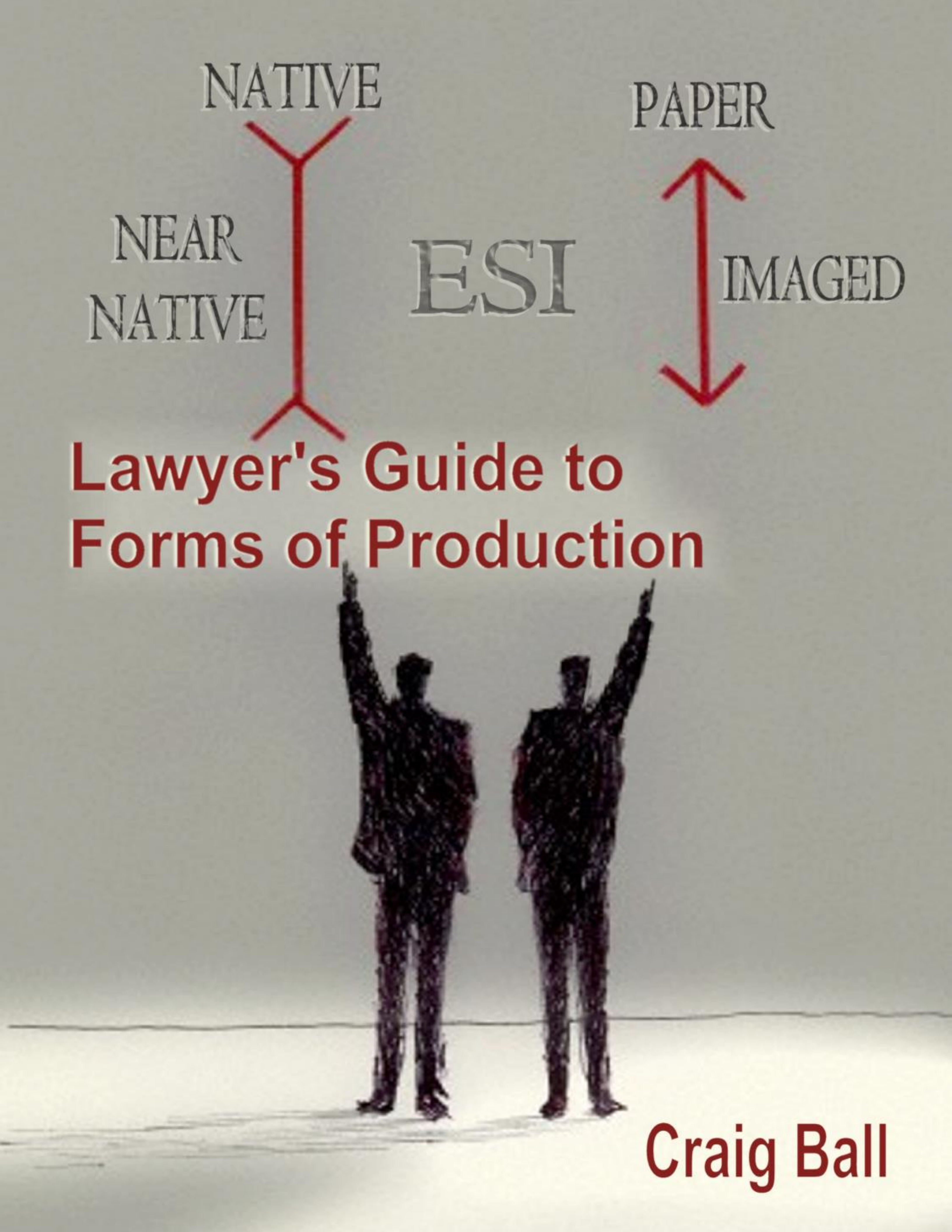
Lawyer's Guide to Forms of Production



Craig Ball

Lawyer's Guide to Forms of Production

Craig Ball

We are awash in digital information, and the forms in which it's produced in discovery dictate whether it is complete and functional or just a pale imitation of the original. Forms of production matter more than most lawyers appreciate, perhaps because it's difficult for lawyers to assess the full content and capability of ESI without knowing what they're missing.

This article explains the significance of forms of production and lays out options to guide the reader in making sensible selections. It seeks to help lawyers eschew the wasteful and outmoded practice of downgrading digital information to paper-like forms and, instead, embrace *forms that function*—that is, forms of production that preserve the integrity, efficiency and functionality of digital evidence.

Contents

Background

The currency of discovery is information, once embodied over many centuries as oral testimony and paper records. Historically in litigation, oral testimony was memorialized by official transcript and paper was faithfully copied to other paper. The law little concerned itself with “forms” of production because options were few and the precept that evidence be faithful to its source and complete in its content was deeply rooted in our jurisprudence. To alter, destroy or degrade evidence was abhorrent—it rarely occurred, and when it did, spoliation was punished severely.

Then, something extraordinary happened. In barely the space of a generation, information became digital, created, communicated and recorded as a sequence of binary data termed “ones” and “zeroes.” These, in turn, manifested as encoded forms capable of reflecting the full spectrum of human creativity: pictures, sounds, text messages, e-mail, spreadsheets, presentations, databases, documents and more. What were once just “flat” forms of information little different than paper acquired new dimensions and depth.

Digital photographs hold EXIF data revealing where they were taken and by what camera. Digital spreadsheets carry formulae supporting complex calculations. Word processed documents store editorial histories and are laced with conversations between collaborators. Presentations feature animated text and rich media, including sound, video and dynamic connections to other data. Databases don’t “store” documents as much as assemble them on demand. Even conversations—once the most ethereal of interactions—now loiter as text messages and voice data packets traversing the internet and cellular networks.

This torrent of data is itself described or bolstered by more data called *metadata* supporting the ability to find, use and trust digital information. All of a sudden, the forms in which information is supplied determine if it is intelligible, functional and complete.

Growing Tension

Initially, lawyers accustomed to the printed page wanted none of the digital deluge and proved adept at keeping it at bay, expending princely sums to print digital data to images approximating the appearance of paper records. Many then printed these out for review and proceedings, little caring about the content and utility surrendered. Lawyers loved that pages could be embossed with Bates numbers and that paper and paper-like forms, lacking the dynamic character of digital information, couldn't be plumbed for embedded content, yield metadata intelligence or be inadvertently altered.

But a growing cadre of lawyers came to appreciate that electronically stored information (ESI) produced as images and load files was incomplete and ungainly. These lawyers questioned why they should settle for production in static forms that lack the innate searchability, functionality and full complement of data as that enjoyed by producing parties. They began demanding more utile and complete forms of production, first for forms like spreadsheets, which are manifestly incapacitated by conversion, and then for more subtly degraded forms, like presentations and word processed documents. Now, they seek e-mail in native and near-native forms, for it's become clear that even e-mail productions lose key capabilities when converted to static images for production.

The 2006 amendments to the Federal Rules of Civil Procedure supported the move toward more functional and complete forms of production. Rule 34(b) was amended to state that requesting parties may specify the form or forms in which electronically stored information is to be produced. The objective was to forestall costly defacement of data in discovery absent demonstrated need and justification by the producing party.

But in a bizarre turnabout, the practice of downgrading electronic evidence to images became so commonplace that litigants routinely argued that requesting parties must show cause to *stop* producing parties from downgrading the evidence! The burden was backwards. More peculiarly, lawyers have seen only image and load file productions for so long that most seem oblivious to—and threatened by—alternatives. They fail to appreciate that their clients *never* work with information in the forms used by lawyers and may barely recognize these forms when confronted with them in depositions or on the witness stand.

Even when attorneys for producing parties use native and near-native forms when reviewing for responsiveness and privilege, the final step before production is often to downgrade the evidence to images. This is the 21st century equivalent of shuffling paper documents before they are produced, though most do it without guile, believing that it has "always" been done thusly and so must be appropriate.

And, indeed it <u>is</u> appropriate when the parties *agree* upon the form of production. Absent Court intervention, parties are free to incur any waste, hardship or delay they jointly consent to tolerate, even to the point of printing everything out. What a producing party is <u>not</u> free to do (absent order of the Court) is downgrade forms of production against the express, timely wishes and objections of a requesting party.

Notwithstanding the attendant waste and inefficiency, the reality is that quite a few lawyers are content with paper-like productions and seek TIFF and PDF images in lieu of native and near-native forms. Accordingly, this paper also addresses the essential elements of imaged productions in the hope that, until they cease altogether, they can be accomplished with as little confusion as possible.

What are the options for forms of production?

Options for forms of production include native forms, near-native forms, imaged production (PDF or, more commonly, TIFF images accompanied by load files containing searchable text and metadata) and paper (printed out) productions. It is not necessary—and rarely advisable—to employ a single form of production for all items exchanged in modern discovery; instead, tailor the forms to the data in a *hybrid* production. TIFF and load files work best for items requiring redaction and for scans of paper records; native forms are best suited to spreadsheets, electronic presentations and word processed documents; and e-mail and database content are best exchanged in near-native forms.

Paper

Converting searchable electronic data to costly and cumbersome paper is usually a step backward, but not always. Paper still has its place. In a case where the items to be produced are paper documents and the volume of same so modest that electronic searchability isn't needed, paper remains a reasonable choice.

Imaged Production

Here, production consists of files that are digital "pictures" of the documents, e-mails and other electronic records. These images are typically furnished in formats like Adobe's Portable Document Format (PDF) or as a series of Tagged Image File Format (TIFF) images. Converting ESI to TIFF images strips the evidence of its electronic searchability and metadata. Accordingly, load files accompany TIFF image productions to hold searchable text and selected metadata. Searchable text is obtained either by extraction from an electronic source or, in the case of scanned paper documents, by use of optical character recognition (OCR). Load files are composed of delimited text, meaning that values in each row of data follow a rigid sequence and are separated by characters like commas, tabs or quotation marks. Using load files entails negotiating their organization or agreeing to employ a structure geared to review software such as AD Summation or LexisNexis Concordance.

Imaged formats are ideal for production of scanned paper records, microfilm and microfiche, especially when OCR serves to add electronic searchability. Even some other types of ESI work reasonably well in imaged formats, albeit at a higher cost than native production, so long as the ESI lends itself to a printed format and remains electronically searchable and complete when printed or imaged. However, when the ESI holds embedded information (such as collaborative content or formulae in spreadsheets) and otherwise extends beyond the confines of printable information (as with voice mail, databases or video), imaged production breaks down.

Native Production

In native production, the defendant duplicates and produces the actual data files containing responsive information. Word documents are supplied in their native .DOC or .DOCX formats. Excel spreadsheets are supplied in .XLS and .XLSX formats. PowerPoint presentations are produced in native .PPT and .PPTX formats. The immediate benefits to the producing party are speed and economy--little or nothing must be spent on image conversion, text extraction or optical character recognition. Because items produced natively are inherently searchable, functional and complete, forms of production are unlikely to be a bone of contention. The benefits to the requesting party are substantial. Using native review tools or the same applications used to create and manipulate the data, requesting parties see the items produced exactly as they appear to the producing party.[1] Embedded commentary and metadata aren't stripped away, deduplication is facilitated, e-mail messages may be threaded into conversations, time zone irregularities can be reconciled and, every step of the way in discovery and trial preparation, costs are reduced and utility is enhanced.

Sounds great, but native production is not without its issues. Specialized native applications needed to view obscure data formats may be prohibitively expensive or difficult to deploy without expertise and pricey tools (*e.g.,* specialized engineering applications or enterprise database software). Additionally, care must be exercised not to corrupt native data while reviewing it. The rule of thumb is that native production is preferable, but it takes a measure of competence to work with it.

Near-Native Production

Some ESI can't be feasibly or prudently handed over in true native formats. Common examples include enterprise e-mail, databases and social networking content. Near-native forms are forms which preserve the essential utility, content and searchability of native forms, but which are not, strictly speaking, native forms. Near-native forms are sufficiently similar in it content and structure to permit the near-native forms to be brought back into the native application with little or no loss of content or utility. It's possible

[1] Don't confuse the production of ESI in native *forms* with the use of native *applications* to open and review data. When someone produces an Excel or Word file, the recipient need not–and should not–use Excel or Word to undertake review of same. In a pinch, maybe; but, it's best to use review tools purpose-built to the task. Competent counsel needs the utility and completeness of native or near native production; but competent counsel does not want to alter the evidence by using native applications for review. Native applications (Word, PowerPoint, Excel, etc.) are not designed to protect the integrity of their files as evidence, nor do they support the workflow of review, tagging and the other common features of a well-crafted review tool. The proper way to undertake litigation review of native ESI is by use of a tool designed for review of native ESI, not the native applications themselves.

because the structure of the data (a/k/a the “fielding” of the data) can be easily mapped back and forth between the native and near-native iterations. This is in contrast to imaged forms which do not preserve fielding or functionality (although load files are often supplied with imaged productions in a feeble effort to restore a diminished level of functionality).

For example, when the contents of a single e-mail account are exported from a corporate Exchange mail database to a PST container, the container is not native to the mail server,, but it replicates the pertinent content as well as the essential functionality of the source. Not “native,” but near enough.

Exports of information from database applications are often produced in so-called “delimited” formats not native to the database, but which nevertheless support the ability to interpret the exported data in ways faithful to the native source. Collection and production from social networking sites like Facebook won’t replicate the precise manner in which the content is stored in the Cloud; yet, it’s possible to tender the information in near-native forms that mirror much of the essential utility, completeness and searchability of the online original.

Hosted Production

Arguably, a fifth alternative is hosted production. This is production *without* production, in that the tendered information resides on a controlled-access website. The requesting party reviews the data through an online application (similar to a web browser) capable of searching and displaying a variety of electronic formats. More commonly, hosted data and online review tools are used internally by a producing party’s counsel to search the data for privileged and responsive items rather than as a means to afford access to the requesting party. The items identified are then duplicated onto transfer media (*e.g.,* optical disks or a hard drive) and produced in one or more of the formats described above.

Forms of Production in the Federal Rules

The drafters of the Federal Rules recognized there would be fewer e-discovery 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 forms disputes would be identified and addressed in the earliest days of a lawsuit, well before the first requests for production were served, obviating costly do-overs. Opponents who couldn’t work out forms disputes were expected to get them in front of the Court quickly.

Rule 26(f)(3)(C) of the Federal Rules of Civil Procedure requires the parties to a lawsuit to submit a discovery plan to the Court prior to the first pretrial conference. The plan must address “any issues about disclosure or discovery of electronically stored information, including the form or forms in which it should be produced.”

In practice, discussions about forms rarely occur in a timely or constructive way. Lawyers still leave meet and confer sessions having agreed that CDs or DVDs will be the form of production.[2]

> **Practice Tip:** A requesting party's first step in getting forms that function is to furnish opponents with a clear and practical written specification of preferred forms *before* the initial Rule 26(f) conference, affording opponents the opportunity to assess feasibility, cost and burden of producing in the preferred forms. At the outset, a requesting party may not know all the various forms in which the data resides natively on opponents' systems; however, a requesting party can usually anticipate the *most common* forms of ESI that will be sought (*e.g.*, e-mail, word processed documents, presentations and spreadsheets) and put forward a preferred production format for those.

Rule 34(b)(1)(C) of the Federal Rules of Civil Procedure allows a requesting party to "specify the form or forms in which electronically stored information is to be produced." Yet, it's common for requests for production to be wholly silent on forms of production, despite pages of absurdly detailed definitions and instructions. It's a squandered opportunity.

The Federal Forms Cha-cha-cha

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).

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 over. 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.[3] Fed. R. Civ. P. 34(b)(2)(D).

> **In a Nutshell**
>
> The 2006 Federal Rules amendments gave requesting parties the right to designate the form or forms in which ESI is to be produced. A responding party may object to producing the designated form or forms, but if the parties don't subsequently agree and the Court doesn't order the use of particular forms, the responding party must produce ESI as it is ordinarily maintained or in a form that is reasonably usable. Moreover, responding parties may not simply dump other forms on the requesting party, but must disclose the other forms before making production so as to afford the requesting party the opportunity to ask the Court to compel production in the designated form or forms.

[2] That *should* be funny because optical disks are production *media,* not production *forms*. If you're not laughing, maybe you're one of those lawyers.

[3] This may qualify as the most ignored obligation in the Federal Rules of Civil Procedure.

Step Four: If the requesting party won't accept the forms the producing party 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 a requesting party will be stuck with TIFFs when 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).[4] They also argue it's unduly expensive and burdensome to start over in order to produce the same ESI in a different form.[5]

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."

Unfortunately, the risk of additional production has proven insufficient to promote good practice.

[4] The Rule states, "A party need not produce the same electronically stored information in more than one form."

[5] Having conducted a review of TIFF images, the producing party may assert that it is unduly burdensome to relate relevance and privilege assessments made of images to native counterparts.

> **Practice Tip:** A requesting party shouldn't wait until the response date to learn if an opponent refuses to furnish the forms sought. Press for a commitment; and if one is not forthcoming, move to compel ahead of the response date. Don't wait to hear the Court say, "Why didn't you raise this before they spent all that money?"

Texas Practice re: Forms of Production

Texas was the first state to enact rules of procedure dealing with ESI. Texas Rules of Civil procedure Rule 196.4 went into effect in 1999, seven years before the Federal e-discovery rules emerged. Accordingly, Texas practice diverges from federal practice when it comes to form of production. Rule 196.4 states,

> To obtain discovery of data or information that exists in electronic or magnetic form, the requesting party must specifically request production of electronic or magnetic data and specify the form in which the requesting party wants it produced. The responding party must produce the electronic or magnetic data that is responsive to the request and is reasonably available to the responding party in its ordinary course of business. If the responding party cannot – through reasonable efforts – retrieve the data or information requested or produce it in the form requested, the responding party must state an objection complying with these rules. If the court orders the responding party to comply with the request, the court must also order that the requesting party pay the reasonable expenses of any extraordinary steps required to retrieve and produce the information.

Under Texas practice, the requesting party's specified form of production is afforded more weight than in federal courts. Production in specified forms that are available to a responding party in the ordinary course of business is mandatory, unless such production cannot be accomplished despite reasonable efforts. The primary relief available to a producing party is not the use of an alternative form of production but, instead, the right to obtain mandatory cost shifting for extraordinary steps required to produce in the requested form.

Thus, under Texas jurisprudence, a request for production in the forms in which the responding party uses the information in the ordinary course of business seems both least likely to be objectionable and least likely to prompt cost shifting. Nevertheless, it can be daunting to help courts (and opponents) whose only exposure to e-discovery has been to imaged productions appreciate that imaged formats require extraordinary steps to generate and produce, where native forms are easiest to retrieve and produce.

Onward through the Fogg: Learning the Language of Forms

E. D. Fogg, a veteran trial lawyer, brought a diversity action in the Western District of Texas against Huevos & Huevos (H&H), manufacturer of a medical device called trans-testicular mesh used to treat scrotal droop. Citing mesh migration causing pain, disfigurement and marked increase in vocal pitch, the suit seeks damages for Fogg's clients and others similarly situated.[6]

Fogg has handled several medical device liability claims over his long career but lacks experience with e-discovery. In the past, he sought paper production or had his staff print things out. But he's gone to some CLE programs that counseled him to get electronically searchable ESI, and he has decided to give it a go.

Recently, Fogg attended a Rule 26(f) "Meet 'n Confer." He made all the right grunts and signs to convey to opposing counsel that he wanted electronically searchable production. But as neither knew how they might achieve such a miracle, they shared a deer-in-headlights moment, followed by the usual "let me check with my client and get back to you" feint.[7]

Some days later, Fogg received a letter from opposing counsel stating:

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.

Fogg sought out Jennifer Xavier, a young associate at the Fogg firm. Jen was hired to help usher the firm into the 21st century and "make sure we do the e-stuff right." Jen is finding that change doesn't come easy.

Fogg handed Jen the document and bluntly asked, "Are they trying to screw me?"

Jen glanced at the proposal and replied, "*Are they trying to screw you? Probably not. But are you screwing yourself by accepting the proposed form of production? Yes, probably."*

She went on to explain the proposal in 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

[6] All of this nonsense is intended to be fictional. Any resemblance to any real person, company, product or scrotum is unintended and coincidental.

[7] 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.

images (as compared to, *e.g.*, a PDF document where the custom is for the entire document to be contained within one multi-page 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 used by the software application that created the file. For example, the native form of a 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 backward little brother. 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, yes.

"[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 all 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 electronically searched, and its application metadata 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 documents, 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 files 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 Summation, Concordance/Opticon and iPro.

"[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 full name) into two 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 fielded 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 want can be burdensome and costly. Waiting too long to specify fielded information you seek 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 without knowing 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 many more billable hours.

"Tell them you want native production," she offered as Fogg headed for the door, but Jen didn't share what she was thinking:

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 or print them out and painstakingly read them one by one. All the while, you'll be 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.

Onward through the Fogg: Load files

One Monday morning, two months later, E.D. Fogg once more dropped in on Jen Xavier. He looked tired. Fogg reported that he'd gotten production from Huevos & Huevos, Inc. and spent the weekend going through it. He found TIFF images of the pages of electronic documents, but couldn’t search them or even copy out any content. He also received a lot of “Notepad documents.” He insisted he’d notified opposing counsel in writing that he wanted “everything in native,” so he wasn’t sure what to make of all those pictures of documents and text files.

Thinking it unlikely that a multinational medical device company would use Windows Notepad as its word processor, Jen 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. Fogg couldn’t make head nor tails of the last.

Jen figured out that Fogg received an imaged production from an opponent who ignored Fogg’s specification of native forms and simply printed everything to electronic paper. The other side expected Fogg 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, Jen knew the production would be wholly unsearchable and largely unusable. Even then, search would be pretty lousy. But she also knew that when Fogg 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 constituted some great favor.

Jen sighed and thought of that classic Texas comeback, “Don’t piss in my boot and tell me it’s raining.”

The Lowdown on Load Files

Load files are the unsung digital sherpas of e-discovery, 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.

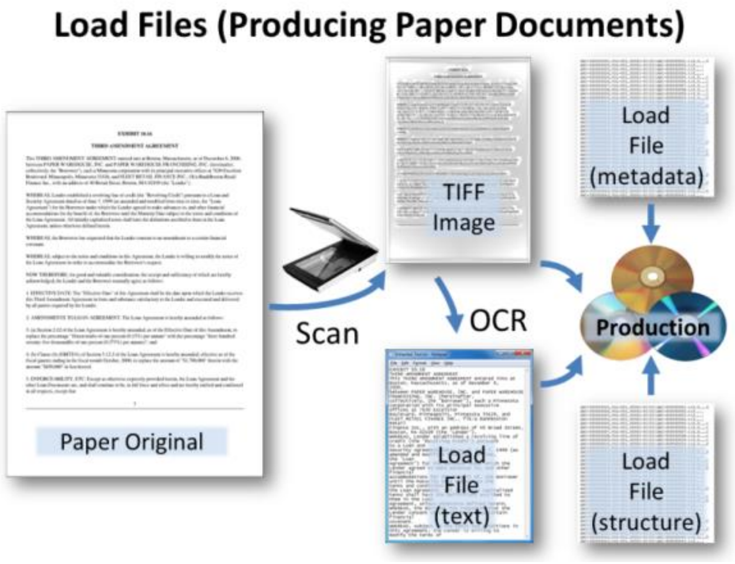
In simplest terms, load files carry data that has nowhere else to go. They are called load files because they are used to load data into, *i.e.,* “populate” a database. They first appeared in discovery in the 1980s in order to add a rough electronic searchability to paper documents.

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. Then, as now, paper documents were scanned to .tiff images and coded by reviewers, and their text was extracted via optical character recognition (OCR) software.

Unlike Adobe PDF images, TIFF images weren’t designed to integrate searchable text; consequently, the text garnered using OCR was stored in simple ASCII[8] text files named with

[8] 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.

the Bates number of the corresponding page image. 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.



While we tend to think of metadata as a feature unique to electronic documents, paper documents have metadata, too. They come in the form of custodian names, office locations, file and folder labels, box numbers and other physical descriptors that must be tracked. Still more metadata takes the form of codes, tags and abstracts reflecting reviewers' assessments of documents and names. Finally, load files serve as a sort of road map laying out, *inter alia*, where document images and their various load files are located on disks or other media used to deliver productions and how the various pieces relate to one another.

Thus, adding a measure of searchability required more than a dozen separate electronic files just to carry the content of one 10-page document.

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. It was expensive and crude, but speedier than poring over thousands or millions of pieces of paper.

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" like Concordance or Summation—venerable products that survive to this day.

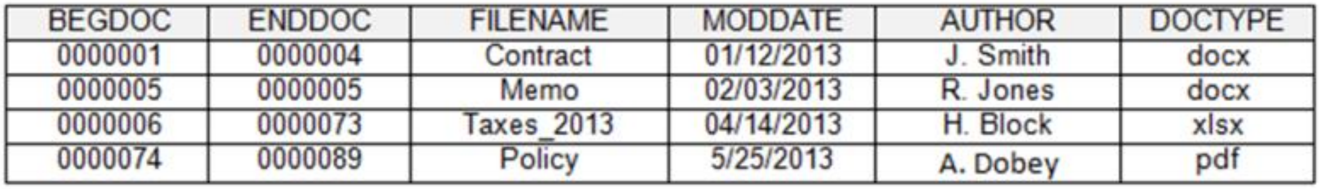
Different review platforms used different load file formats to order and separate information according to guidelines called "load file specifications." Again, load files are plain text files employing characters called delimiters to field (separate) the various information items in the load file. Thus, 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.

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 fit together. Load files are used because neither paper nor TIFF images are suited to carrying the same electronic content; and if it wasn'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 two columns of your table to number the first and last page of each document, then the next column for the document's name and then each succeeding column would carry particular pieces of information about the document. You might make it easier to tell one column from the next by drawing lines to delineate the rows and columns, like so:



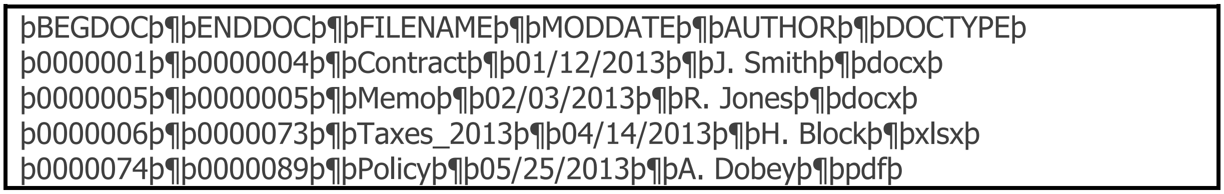
BEGDOC	ENDDOC	FILENAME	MODDATE	AUTHOR	DOCTYPE
0000001	0000004	Contract	01/12/2013	J. Smith	docx
0000005	0000005	Memo	02/03/2013	R. Jones	docx
0000006	0000073	Taxes_2013	04/14/2013	H. Block	xlsx
0000074	0000089	Policy	5/25/2013	A. Dobey	pdf

Those lines separating rows and columns serve as delimiters; they (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 be delimited 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,05/25/2013,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 (thorn, ALT-0254) and ¶ (pilcrow, ALT-0182) characters as delimiters, *e.g.:*

Concordance Load File



```
þBEGDOCþ¶þENDDOCþ¶þFILENAMEþ¶þMODDATEþ¶þAUTHORþ¶þDOCTYPEþ
þ0000001þ¶þ0000004þ¶þContractþ¶þ01/12/2013þ¶þJ. Smithþ¶þdocxþ
þ0000005þ¶þ0000005þ¶þMemoþ¶þ02/03/2013þ¶þR. Jonesþ¶þdocxþ
þ0000006þ¶þ0000073þ¶þTaxes_2013þ¶þ04/14/2013þ¶þH. Blockþ¶þxlsxþ
þ0000074þ¶þ0000089þ¶þPolicyþ¶þ05/25/2013þ¶þA. Dobeyþ¶þpdfþ
```

Summation load files typically use the file extension .DII, and 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 16
@C AUTHOR A. Dobey
@DATESAVED 05/25/2013
@EDOC \NATIVE\Policy.pdf
```

Two more load files are worth mentioning**: Opticon image load files** and **overlay load files**.

Opticon load files employ the file extension .OPT and are used to pair Bates numbered pages with corresponding page images and to define the **unitization** of each document; that is where it begins and ends. A document may be unitized *physically*, as when its constituent pages are physically joined by clips, staples or bindings, or a document may be unitized *logically*, where its constituent pages belong together though they are not physically unitized (as might occur when documents are bulk scanned or a transmittal references an enclosure). Logical unitization requires judgment based on organizational clues in the documents, such as consecutive page numbers, titles and consistent page headers, typefaces and/or document structures. Logical unitization is also a means to track family relationships between container files and contents and e-mail messages and attachments.

Opticon load files typically employ a simple seven-field, comma-delimited structure:

1. Page identifier (e.g., Bates number),
2. Volume label (optional),
3. Path to page image,
4. New document marker (Y),
5. Box identifier (optional),
6. Folder identifier (optional),
7. Page count (optional).

Opticon Load File

```
0000001_0001,,TIFF\001\0000001_0001.tif,Y,,,4
0000002_0002,,TIFF\001\0000002_0002.tif,,,,
0000003_0003,,TIFF\001\0000003_0003.tif,,,,
0000004_0004,,TIFF\001\0000004_0004.tif,,,,
0000005_0001,,TIFF\001\0000005_0001.tif,Y,,,1
0000006_0001,,TIFF\001\0000006_0001.tif,Y,,,68
0000007_0002,,TIFF\001\0000007_0002.tif,,,,
0000008_0003,,TIFF\001\0000008_0003.tif,,,,
0000009_0004,,TIFF\001\0000009_0004.tif,,,,
0000010_0005,,TIFF\001\0000010_0005.tif,,,,
```

Opticon load files are typically used in conjunction with Concordance load files. The Opticon load files carry the image links and unitization, and the Concordance load files supply metadata and anything else. Of course, everything in an Opticon load file could be included in a Concordance load file. Using both is a last-century practice supporting legacy tools.

Overlay load files are not a form of load file as much as a particular usage of same. Overly files are used to update or correct existing database content by replacing data in fields in the order in which the records occur, without matching any specific identifier. Think of overlay load files as blindly replacing the next "X" records in the database. Thus, it's crucial that the order of the data within the overlay file match the order of the data to be replaced. That is, they must be sorted in the same way and the overlay must not add or omit any fields.

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 one 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.

To be fair, there's a lingering need for load files in e-discovery. Even native electronic documents have outside-the-file or "system" metadata that must be loaded into review tools; plus, it's essential 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. U*nderstanding* load files helps ease the pain.

Onward Through the Fogg: The Case against Native

At lunchtime, E.D. Fogg stuck his head into Jen Xavier's office and asked, "Care for a taco? I'm buying." At the local Mexican joint, Fogg described his call to opposing counsel to ask why they hadn't produced the native forms he'd specified. The defense lawyer said, "*E.D., if we'd given you what you asked for, we wouldn't have any Bates numbers on the documents. Gotta have Bates numbers, amigo!*"

"*We talked about whether to give you native stuff, but we couldn't risk your accidentally altering the files. Plus, if we gave you native, we'd have to review the metadata and other stuff inside the files for privilege. Who's going to pay for that?*"

"*And E.D.,*" he added before hanging up, *"we had to redact privileged communications from a few of these files. We can't redact native files without changing them. You don't want us changing the evidence, do you? Adios.*"

Jen laughed. "They ran the case against native on you—right out of the producing party playbook."

Jen saw that E.D. was wounded by her laughter, and sought to console him.

“E.D., they do this to *everybody*—and they get away with it! Don’t worry. I’m going to tell you how to push back. Under the Rules, you get to specify the form or forms of production. You asked for ‘native.’ They didn’t object. They didn’t tell you they were giving you junk. They just blew you off. Huddle up. Here’s what we’re going to do.”

The Case Against Native

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.

Despite the sea change in what we seek to discover, lawyers cling to .tiff imaging and load files, obsolete technology that once made evidence easier to find but which 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-style identifiers, remind them they've long used Bates numbers as file names in .tiff image productions.

Granted, it's indeed difficult to emboss Bates-style identifiers on every page of a native file unless it's printed or imaged. Yet many native forms of ESI (*e.g.,* spreadsheets, social networking content, video, and sound files) don't lend themselves to paged formats and will never be Bates labeled.

But this limitation is no big deal! 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 e-mail 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, overall, 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.

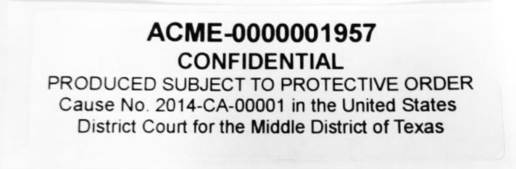
More on Mastering Bates Numbers

Few discussions of better, cheaper production get past the petulant protest, "*but we can't Bates number!*" So, it behooves us to look closely at our options when it comes to Bates-style identifiers.

Bates numbering is the assignment of a unique sequential (typically alphanumeric) identifier to each item produced in discovery. The name pays homage to Edwin G. Bates, a 19th-century inventor of a patented handheld automatic numbering machine. In the paper production era, self-inking Bates stamps were used to imprint an incrementing number on each page of documents produced, facilitating page retrieval, tracking the source of the production, consistent reference to particular pages in proceedings, ordering of pages and identification of missing pages. With the advent of electronic printing, "Bates stamping" became "Bates labeling," though the process lost all ties to Mr. Bates' much-loved stamp. Now, anything that could be printed to an adhesive label became part of Bates numbering, and labels began to feature the names of producing parties and matters. Some included protective legends warning the items were subject to protective order.[9] The move to document imaging brought the practice of embossing Bates identifiers on the face of each page image. To avoid obscuring content, page images are often shrunk to create white space on which to emboss Bates numbers and protective legends.

Bates identifiers are useful organizational tools, and lawyers are quite attached to them. Unfortunately, a lot of lawyers imagine that every item produced in discovery is a document, and every document is comprised of one or more discrete pages. Increasingly, responsive items are data and metadata, not documents; hence, much modern evidence cannot be readily or functionally produced as 8½ x 11 pages or page images. Modern evidence handling demands a transformation in how lawyers deploy Bates identifiers. Fear of change has proven a potent sticking point in the transition to less costly and more utile native and near-native production formats.

Bates-style automatic numbering machine



ACME-0000001957
CONFIDENTIAL
PRODUCED SUBJECT TO PROTECTIVE ORDER
Cause No. 2014-CA-00001 in the United States
District Court for the Middle District of Texas

Where Bates identifiers are concerned, making the transition to modern forms of production requires acceptance of three propositions:

- Printouts and images of ESI are not the same as ESI;
- Most of what's produced in discovery is never used in proceedings; and
- Names of electronic files can be changed without altering contents.

Printouts and images of ESI are not the same as ESI

Everyone understands that a signed document is not the same as one that's unsigned. They contain different information, and they function differently (just try probating an unsigned will). It may be that the differences don't matter to meet a particular need, but no lawyer would argue that the signed- and unsigned documents are one and the same.

[9] 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 1960s-70s 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, so the magic that enables producing parties to shrink documents and emboss Bates labels makes it simple to mask those Bates labels by resizing documents to their true dimensions.

Likewise, a Microsoft Word document in its native .DOC or .DOCX format is not the same thing as a printout or page image of the document. They contain different information, and they function differently. Though the differences may not matter to meet a particular need, no lawyer should equate the two in discovery. The native document carries more information than its printed counterpart, and it is inherently functional, searchable and complete. Moreover, the native document is described by more and different metadata—information often invaluable in identifying, sorting and authenticating evidence.

Once we grasp that imaged and native forms of ESI are not interchangeable, it's easy to appreciate why it's worth the effort to embrace alternate means of Bates numbering items and reap the advantages of native and near-native productions.

Most of what's produced in discovery is never used in proceedings

There's no doubt that tagging items produced in discovery with unique sequential identifiers is valuable. It will be the practice for a long time to come, even as paper all but disappears from the discovery process. But, we will do it differently.

Adding Bates numbers to ESI the old-fashioned way adds cost. True, the per page expense to emboss a Bates identifier is trivial, but processing ESI to make paging and embossing possible is expensive, measured both by the cost to convert ESI to images and the time wasted in dealing with less utile and complete forms.

The better approach is to produce ESI as ESI and rename the information items produced to carry unique, sequential alphanumeric identifiers. It's practical, effective, easy and cheap. In fact, changing file names to conform to Bates numbers is already standard practice in discovery. What do you think they've been doing for decades when naming all those TIFF image files?

Yet, there are times where having an information item in a tangible form sporting a unique page identifier is useful, such as in deposition, as exhibits to pleadings as evidence introduced in hearings and at trial.

So, how do we enjoy the substantial cost savings and efficiencies that flow from producing ESI as ESI while still having the benefit of Bates identifiers in proceedings?

The answer is to emboss Bates numbers only on items that are needed in paged and numbered formats for use in proceedings. That's nearly always a tiny fraction of the whole because *most of what's produced in discovery is never used in proceedings*.

Of course, the producing party can't make that determination at the time of production. So, the solution is to visit responsibility for embossing page identifiers upon the party who converts ESI to paged formats. Those are the only items requiring page references; otherwise, identification of the native file *(i.e.,* the document, not its pages) by Bates identifier is wholly sufficient. In short, *everyone produces in native forms and when anyone downgrades evidence to paper or*

TIFF for use in proceedings, they will emboss Bates numbers on pages pursuant to a common protocol and supply copies of the reformatted and embossed pages to all parties. Parties thus add Bates numbers to pages only where and when such page-by-page granularity is needed and its cost justified.

Names of electronic files can be changed without altering contents

Paper records don't have unique names. Certainly, paper documents can be titled and people can agree to refer to a document in a consistent way; but, not even *Magna Carta* or *Declaration of Independence* uniquely refers to just one tangible thing. Yet within electronic file systems, files must be identified uniquely. The identifier is ultimately the physical or logical address of the data stream(s) on the storage medium; but, that address can then correlate to a user-friendly name as long as the name meets naming conventions (*i.e.,* rules) imposed by the file system. In the past, DOS file names could be nearly any combination of eight alphanumeric characters followed by a dot and up to three characters (usually denoting the type of file). Later, file systems evolved to support file names up to hundreds of characters in length.

If you change the title of a paper document, you're altering the evidence. By contrast, *the names of electronic files can be changed without altering the contents of the ESI*. A file and the friendly name given to it are independent of one another. File names are *system metadata* and reside outside the files they identify.[10] Accordingly, a Bates identifier and even a protective legend can be made part of a file name without effecting any change in the file, so long as the length of the name stays under 255 characters.[11] The original name of the file is furnished in a load file, along with other relevant system metadata. To wit:

> Each information item produced should be identified by naming the item to correspond to a Bates identifier according to the following protocol, supplying the original file name data in a delimited load file as described below:
>
> i. The first four (4) characters of the filename will reflect a unique alphanumeric designation identifying the party making production;
>
> ii. The next six (6) characters of the filename may reflect any printable content you wish to include for your convenience, padded with leading zeroes as needed to preserve its length;[12]

[10] As distinguished from *application metadata,* which always reside within a file and are thus integral to the file.

[11] This is the limit in Microsoft Windows NTFS (*e.g.,* Windows XP, Vista, 7 & 8). When establishing a naming protocol, consider the length of the fully qualified path to root in your calculation, and note that maximum path lengths may be shorter when using delivery media formatted in file systems other than Windows NTFS, *e.g.,* optical media.

[12] 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!

iii. The next nine (9) characters will be a unique, consecutive numeric value (“Bates number) assigned to the item by you. This value shall be padded with leading zeroes as needed to preserve its length;

iv. 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 converted to an image format for use in proceedings or when attached as exhibits to pleadings.

v. By way of example, a Microsoft Word document produced by you in its native format might be named: ACSS000319000000123.docx. Were the document printed out for use in deposition, page six of the printed item must be embossed with the unique identifier ACSS000319000000123-0006. Scans of paper documents and items requiring redaction produced as TIFF or PDF images should be identified on each page in a manner that does not obscure content by embossing the Bates number of the file followed by a dash or underscore and the page number. Respond to each request for documents by listing the Bates numbers of responsive documents produced

Production should include a delimited load file supplying relevant system metadata field values for each document by Bates number. The field values supplied should include (as applicable):

a. Source file name
b. Source file path
c. Last modified date
d. Last modified time
e. Custodian or source
f. Document type
g. MD5 hash value
h. Redacted flag
i. Hash de-duplicated instances (by full path)

The case against imaged production

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. Parties (*i.e.,* clients) work with native forms of ESI because native forms are the most utile, complete and efficient forms in which to store and access data. Parties don’t print their e-mail before reading it. Parties don’t emboss a document’s name on every page. Parties 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:

1. You must pay someone to convert native forms to TIFF images and emboss Bates numbers;
2. You must pay someone to generate load files containing extracted text and application metadata from the native ESI;
3. You must produce multiple copies of certain documents (like spreadsheets) that are virtually incapable of being produced as TIFF images;
4. 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;
5. It's very difficult to reliably deduplicate documents once they have been converted to TIFF images; and
6. 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.

But what if they *want* imaged production?

If you're reading this thinking, "*lots of requesting parties want TIFF or PDF images,"* you'd be right. Many requesting parties are unaware there are alternatives to imaged productions. Some like the simplicity of opening each page image in a browser or in Acrobat Reader. Others are wed to outdated review tools incapable of ingesting modern forms of ESI production. Still others attribute little value to electronic searchability or to content they've never seen. Whatever the reasons, if counsel can agree upon a form of production and their clients are content to bear the consequences of that agreement, let no man put same asunder!

That is, unless the work of the Court is somehow impeded by their agreement. The efficient administration of justice may warrant judicial intervention notwithstanding agreements between the parties. Put another way, just because two lawyers think they can fly doesn't mean the judge should unlock the door to the courthouse roof! The public has an interest in the just, speedy and inexpensive administration of justice, and if evidence will be lost, resolution delayed or substantial waste engendered by the forms of production employed, it's within the ambit of the Court to intercede and reform the process in the public interest.

The fact some lawyers choose not to seek more utile and complete forms of production shouldn't serve to hinder those lawyers who do, especially when production in native and near-native forms serves to lower the cost of discovery and further its utility for all.

Producing parties often adopt a paternalistic view of native and near-native production, as if saying of requesting parties, "If we give them native forms, they will just break them."

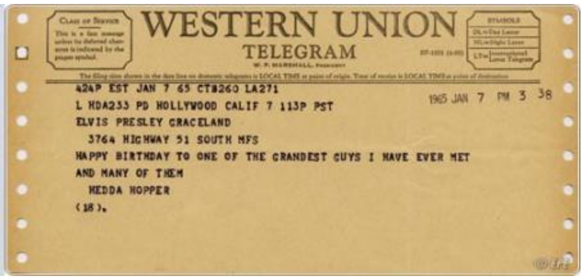
Modern Requests for Modern Evidence

So, how do capable lawyers like E. D. Fogg avoid being taken advantage of in electronic discovery? To start, they need to rethink how they approach requests for production because the boilerplate in common use is antiquated and non-specific. The first step in getting the information you want in the form you want it is to clearly ask for it, making good use of procedural rules.

Consider this typical definition of Document pulled from requests served on one of the world's largest technology companies by a top-notch law firm:

> "Document" or "documents" means all written, typed, or printed matters, and all magnetic, electronic, or other records or documentation of any kind or description (including, without limitation, letters, correspondence, telegrams, memoranda, notes, records, minutes, contacts, agreements, records, or notifications of telephone or personal conversations, conferences, inter-office communications, E-mail, microfilm, bulletins, circulars, pamphlets, photographs, facsimiles, invoices, tape recordings, computer printouts and work sheets), or reproduced, and all compilations of data from which information can be obtained, and any and all writings or recordings of any type or nature, in your actual possession, custody, or control, including those in the possession, custody, or control of any and all present or former directors, officers, employees, consultants, accountants, attorneys or other agents, whether or not prepared by you. A document or tangible item is within your control if you have a right to compel a third party to produce it to you.

Clearly, this meandering missive has been handed down, old lawyer to new, for ages, or at least since the era of telegrams and microfilm. Sure, language was added to reflect "new" technologies (*e.g.,* facsimiles and e-mail), but, no one has lately read the definition with an eye toward modern evidence.[13]



WESTERN UNION
TELEGRAM

424P EST JAN 7 65 CT8260 LA271
L HDA233 PD HOLLYWOOD CALIF 7 113P PST
1965 JAN 7 PM 3 38
ELVIS PRESLEY GRACELAND
3764 HIGHWAY 51 SOUTH MFS
HAPPY BIRTHDAY TO ONE OF THE GRANDEST GUYS I HAVE EVER MET
AND MANY OF THEM
HEDDA HOPPER
(18).

Arguably, the phrasing is so all-encompassing as to make the antiquated examples irrelevant—but then, why define "document" at all, if the definition and the examples supplied don't matter?

Slipshod requests for production of ESI sow the seeds for failed discovery. When requesting parties trot out creaky definitions, couch requests in language of a bygone era and don't specify utile and complete forms of production, they're inviting opponents to produce junk.

What follow are simple steps to cull superfluous verbiage from requests and add back specificity and clarity. Crisp, clear requests are the hardest to sidestep and the easiest to enforce.

[13] The free floating fragment, "or reproduced," in the middle of this mess suggests no one's read the definition for clarity either. It also tacks on important but misplaced language about custody and control that's probably better suited to an instruction addressing scope of discovery than one defining "document."

Dump the Definition of "Document"

Requesting parties must ditch the boilerplate and draft requests as sleek and modern as ESI itself.

You'll frame better requests for modern evidence by observing that much of what we seek in discovery today bears but a passing resemblance to conventional paper documents. Most digital evidence—including e-mail—exists as data within databases. For the last 30 years or so, nearly all paper records are the unsearchable fossil remnants of their original electronic sources. Information stored in the Cloud, on social networking sites as tweets and Facebook pages or texted between handheld devices is information assembled on demand. What we see manifested as a "document" is likely data assembled on demand and presented in a user-friendly format. Often, there was no "document" until it was constructed onscreen just a moment ago.

So, we must get past thinking about discovery as the quest for "documents" and start focusing on what we really seek: *information in utile and complete forms.*

To that end, the definition of "document" must give way to an alternate term like "information." It's not necessary or desirable to employ a thesaurus-like litany of types of information and run the risk of evasion for failure to name a variant.[14] If you absolutely must supply a definition, why wouldn't something like this suffice?

> *Requesting party seeks relevant, non-privileged information in all forms in which it is stored and communicated.*

You'll want to tack on some "including, but not limited to" examples, but resist the temptation.

Cut the Crap

While you're excising all those mentions of "including, but not limited to," why not eliminate the mentions of "any and all?" They don't add clarity, and they're lightning rods for objection. If drafting a request without "any and all" makes you quake, add this to your instructions:

> *Requests for production should be read so as to encompass any and all items responsive to the request.*

If you must incorporate examples of responsive items in a request, just say "including" and add an instruction that says, "*Examples of responsive items set out in any request should not be construed to limit the scope of the request*."

Don't define a term unless you use it

[14] *Expressio unius est exclusio alterius.*

Word processing makes it so easy to recycle requests from other cases. When drafting discovery, who doesn't start with something from something they've borrowed from somewhere else? But when you define a term and either fail to use it or use an undefined variant, your request broadcasts your reliance on forms—it becomes easy to prove you haven't customized your request to the case. Sloppy requests are the easiest to sidestep and hardest to enforce.

Before you serve discovery, check to see that you've defined only terms you've used and that you've used the terms in ways consistent with your definitions. Do you *really* need to define terms like "concerning" or "date?" Is your definition of "you and your" more metaphysical than practical?

Specify the forms you seek

The most common error seen in requests for production is the failure to specify the forms sought for ESI production. Worse, requests often contain legacy boilerplate specifying forms the requesting party *doesn't* want. When the Federal Rules of Civil Procedure were amended in 2006 to grant requesting parties the right to specify the "form or forms of production," requesting parties acquired a powerful mechanism to modernize discovery. Most have failed to use it.

Every request for production should specify forms of production sensibly and precisely. Don't assume that "native format" is clear or sufficient; instead, specify the formats sought for common file types, *e.g.:*

<table>
<tr><td colspan="2">Information that exists in electronic form is specifically requested to be produced in native or near-native formats and should not be converted to imaged formats. Native format requires production in the same format in which the information was customarily created, used and stored by you. The table below supplies examples of the native or near-native forms in which specific types of electronically stored information (ESI) should be produced.</td></tr>
<tr><td>Source ESI</td><td>Native or Near-Native Form or Forms Sought</td></tr>
<tr><td><i>Microsoft Word documents</i></td><td><i>.DOC, .DOCX</i></td></tr>
<tr><td>Microsoft Excel Spreadsheets</td><td>.XLS, .XLSX</td></tr>
<tr><td>Microsoft PowerPoint Presentations</td><td>.PPT, .PPTX</td></tr>
<tr><td>Microsoft Access Databases</td><td>.MDB, .ACCDB</td></tr>
<tr><td>WordPerfect documents</td><td>.WPD</td></tr>
<tr><td>Adobe Acrobat Documents</td><td>.PDF</td></tr>
<tr><td>Images</td><td>.JPG, .JPEG, .PNG</td></tr>
</table>

E-mail	Messages should be produced in a form or forms that readily support import into standard e-mail client programs; that is, the form of production should adhere to the conventions set out in RFC 5322 (the internet e-mail standard). For Microsoft Exchange or Outlook messaging, .PST format will suffice. Single message production formats like .MSG or .EML may be furnished, if source foldering data is preserved and produced. For Lotus Notes mail, furnish .NSF files or convert to .PST. If your workflow requires that attachments be extracted and produced separately from transmitting messages, attachments should be produced in their native forms with parent/child relationships to the message and container(s) preserved and produced in a delimited text file
Databases	Unless the entire contents of a database are responsive, extract responsive content to a fielded and electronically searchable format preserving metadata values, keys and field relationships. If doing so is infeasible, please identify the database and supply information concerning the schemae and query language of the database, along with a detailed description of its export capabilities, so as to facilitate crafting a query to extract and export responsive data.
Documents that do not exist in native electronic formats or which require redaction of privileged content should be produced in searchable .PDF formats or as single page .TIFF images with OCR text furnished and logical unitization and family relationships preserved.	

The language in the specification table above is technical, and so may draw howls from the other side that they don't understand it. It isn't written for the lawyers. Instead, it speaks to the skilled information technology personnel and e-discovery service providers who will be doing the collection and processing of data.

Address Redaction

Because redaction tends to impact just a small part of most productions, it's important not to let it co-opt the forms of production. Restated, *don't let the redaction tail wag the production dog.*

Producing parties generally prefer to redact ESI in the way they once redacted paper documents: by blacking out text. To make that possible, ESI is converted to non-searchable TIFF images in a process that destroys electronic searchability. So after redaction, electronic searchability must be restored by using OCR to extract text from the TIFF image.

This language should suffice:

"Information items that require redaction shall be produced in static image formats, e.g., single page .TIFF or multipage PDF images with logical unitization preserved. The unredacted content of each document should be extracted by optical character recognition (OCR) or other suitable method to a searchable text file produced with the corresponding page image(s) or embedded

within the image file. Redactions should not be accomplished in a manner that serves to downgrade the ability to electronically search the unredacted portions of the item."

A TIFF-OCR redaction method works reasonably well for text documents, but it fails miserably applied to complex and dynamic documents like spreadsheets and databases. Unlike text, you can't spellcheck numbers, so the inevitable errors introduced by OCR make it impossible to have confidence in numeric content or reliably search the data. Moreover, converting a spreadsheet to a TIFF image strips away its essential functionality by jettisoning the underlying formulae that distinguishes a spreadsheet from a table.

For common productivity applications like Adobe Acrobat and Microsoft Office, it's now feasible and cost-effective to redact natively so as to preserve the integrity and searchability of evidence. Consequently, where it's important to preserve the integrity and searchability of redacted documents, you should determine what redaction methods are contemplated and seek to agree upon methods best suited to the task.

Specify the medium of production

A well-crafted request should designate the *medium* of ESI production as well as the *forms* of production. It's important not to confuse the two. The medium of production is the mechanism used to convey the electronic production to the requesting party. If you're receiving 100GB of data, you don't want it tendered on 143 CDs!

In the definitions and instructions section of your Request, set out acceptable options for the medium of production like, "*Productions smaller than 10GB should be made using DVD recordable optical media. Productions larger than 10GB but smaller than 128GB should be made using a flash/thumb drive or portable external hard drive. Productions larger than 128GB should be made using a portable external hard drive."*

Alternatively, *"Production should be made using appropriate electronic media of the producing party's choosing provided that the production media chosen not impose an undue burden or expense upon a recipient."*

Secure the production in transit

Consider the potential for confidential data to be lost or stolen in transit: *"All productions should be encrypted for transmission to the receiving party. The producing party shall, contemporaneously with production, supply decryption credentials and passwords to the receiving party for all items produced in an encrypted or password-protected form."*

Don't leave the form of the load file to chance

Every electronic file has a complement of descriptive information called *system metadata* residing in the file table of the system or device storing the file. Different file types have different metadata. Every e-mail message has *"fields"* of information in the message "*header*" that support better searching, sorting and organization of messages. This may be data probative in its own right or simply advantageous to managing and authenticating electronic evidence. Either way, you want to be certain to request it sensibly and precisely. Simply demanding "the metadata" reveals you don't fully understand what you're seeking.

Develop a comprehensive production protocol tailored to the case and serve same with discovery. Short of that, specify the particular items of metadata and header fields you seek, *e.g.:*

Produce delimited load file(s) supplying relevant system metadata field values for each information item by Bates number. The field values supplied should include:

a. ***Source file name*** *(original name of the item or file when collected from the source custodian or system);*
b. ***Source file path*** *(fully qualified file path from the root of the location from which the item was collected);*
c. ***Last modified date and time*** *(last modified date and time of the item when collected from the source custodian or system);*
d. ***Custodian or source*** *(unique identifier for the original custodian or source);*
e. ***Document type****;*
f. ***Production File Path*** *(file path to the item from the root of the production media);*
g. ***MD5 hash*** *(MD5 hash value of the item as produced);*
h. ***Redacted flag*** *(indication whether the content or metadata of the item has been altered after its collection from the source custodian or system);*
i. ***Embedded Content Flag*** *(indication that the item contains embedded or hidden comments, content or tracked changes);*
j. ***Deduplicated instances*** *(by full path);and*
k. ***UTC Offset*** *(The UTC/GMT offset of the item's modified date and time, e.g., -0500).*

The following additional fields shall accompany production of e-mail messages:

l. ***To*** *(e-mail address(es) of intended recipient(s) of the message);*
m. ***From*** *(e-mail address of the person sending the message);*
n. ***CC*** *(e-mail address(es) of person(s) copied on the message);*
o. ***BCC*** *(e-mail address(es) of person(s)blind copied on the message);*
p. ***Subject*** *(subject line of the message);*
q. ***Date Received*** *(date the message was received);*
r. ***Time Received*** *(time the message was received);*
s. ***Attachments*** *(beginning Bates numbers(s) of attachments, delimited by comma);*
t. ***Mail Folder Path*** *(path of the message from the root of the mail folder);and*
u. ***Message ID*** *(unique message identifier).*

Consider de-duplication

In your Request, you may wish to specify whether the production should or should not be de-duplicated, *e.g.: “Documents should be vertically de-duplicated by custodian using each document’s hash value. Near-deduplication should not be employed so as to suppress different versions of a document, notations, comments, tracked changes or application metadata.”*

Exemplar Production Protocols

Appendices 2 and 3 are examples of production protocols, sometimes called data delivery standards. The protocol in Appendix 2 is geared to civil litigation and represents the lowest cost approach to production of ESI. It seeks native production of common file types and forces parties to incur the cost of conversion to imaged formats only when needed for redaction. This exemplar protocol specifies near-native alternatives for production of native forms when near-native forms are preferable. It’s designed to be lightweight in the sense that it minimizes the need for extensive processing of data for production and accommodates parties lacking specialized e-discovery tools. Crucially, the Protocol in Appendix 2 keeps the focus on the forms of production used by the producing *parties* versus the degraded and incomplete forms generated for counsel’s use.

Appendix 3 contains the U.S. Securities and Exchange Commission’s Data Delivery Standards as of January 17, 2013. Because they seek production of both native and imaged counterparts for every item produced, the SEC Standards fail rather spectacularly on cost-effectiveness; however, the SEC Standards are an impressively complete and well-informed approach to production and correlation of forms suited to both the most and least sophisticated tools to review the production. Because the SEC Standards require production of the same ESI in more than one form, they seem to run afoul of Federal Rule of Civil Procedure 34(b)(2)(E)(iii). Unless the parties wish to bear the cost of such an expansive dual-format production, it’s best to specify a single utile and complete form of production *for each type or category* of ESI sought, as exemplified in Appendix 2.

Onward through the Fogg: Native production of e-mail

As Jen approached the elevator, she glanced up from her phone to see E.D. Fogg holding the door in a courtly way. He was beaming. “I sent them the production protocol you drafted, Jen, and what a difference. I think they’re really nervous.”

“Why do you think so?” Jen asked—pleased to see admiration in the old fellow’s eyes that had nothing to do with her legs.

“They’ve hired some East Coast e-discovery counsel and are begging to meet and confer. I could hardly get them to return phone calls before.”

“That’s great, E.D.”

"*And* they asked for a settlement demand," he added with unconcealed glee.

"Jen?"

"Yes, E.D."

"Thanks."

"Just doing my job, E.D."

"And one more dumbass question, okay?"

"You bet."

"*What in the hell is RFC 5322?!?!*"

They laughed all the way to the 57th floor.

What is Native Production for E-mail?

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, GroupWise mail—all are *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 letter). 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 so that older messaging systems could talk to newer systems. These proposals are 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 take in order to be compatible with e-mail systems and also adhere to RFC protocols.

The RFCs governing e-mail have gone through several major revisions since the first RFC concerning e-mail protocols circulated in 1973. The latest iteration, circa 2008, is called RFC

5322, a revision of RFC 2822 (circa 2001), which superseded RFC 822 (circa 1982). A separate series of RFCs (RFC 2045-47, RFC 4288-89 and RFC 2049) addressed ways to graft text enhancements, foreign language character sets and multimedia content onto plain text e-mails, collectively called Multipurpose Internet Mail Extensions or MIME.

So, if you asked, "What's the native form of an e-mail *as it traversed the internet between mail servers,"* the answer would likely be, "plain (7-bit ASCII) text adhering to RFC 5322 and MIME." This is the same things as saying ".EML format," and, if the content remains faithful to the RFC and MIME protocols, it can be functionally the same as the MHT format. You can even change the file extension of a properly formatted message from EML to MHT and back 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 carries the entire contents of the message (header, message bodies and encoded attachments) at a pertinent point in time, and it's a form that's about as compatible as you can get in the e-mail universe.[15]

Unfortunately, the form of an e-mail in transit is often incomplete in terms of metadata that may have probative or practical value, and the transiting format is rarely native to the database that sends or receives the message (which tends to be where e-mail messages turn up).

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 features. 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 a local (*i.e.,* retained by the user) container file to support offline operation. If you can read your e-mail using Outlook without a network connection, you have a local storage file.

> **Practice Tip:** 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 webmail services; so,

[15] 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)

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, let's 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, *i.e.*, 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 the database. *What we see as an e-mail message in Outlook is actually a report queried from a database.* The native form of Outlook e-mail carries these fields and adds metadata not a part of the transiting message. Such metadata fields may 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 RFC 5322 messaging. Whether this MAPI metadata is superfluous or invaluable depends upon what questions 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 container file format is widely supported by applications designed to view, process and search e-mail. Moreover, the 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 and data-relating features other than e-mail.

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 5822 format? 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 RFC 5822 message components as well as more and different information than its simpler RFC 5322 cousin, but it lacks information (like foldering data) contained within the source PST. It's not "native" in the sense that it's not a format that Outlook uses day-to-day; but MSG is an export format that holds most of the message metadata unique to Outlook. All we can say is that it's 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, prosaically called the Exchange Database and sporting the file extension .EDB. The EDB holds the account content for everyone in the mail domain. Accordingly, unless the case warrants production of all the e-mail, attachments, contacts, and calendars for everyone, 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, it's never done that way, and there's little to commend it over simpler approaches.

So, if you're *not* going to produce in the "true" native format of EDB, everything left to you is termed "near-native," if it preserves the requisite content and essential functionality of the native form. If it doesn't preserve content and utility, you can call it whatever you want. "Garbage" springs to mind; but, to each their own.

More seriously, 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 you are meticulous in text extraction, diligent in metadata collection and careful in load file construction, you *could* produce Exchange content in a way that's sufficiently complete and utile as to make a departure from native less problematic—assuming, of course, that you produce 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. Further, such a mutation is likely to be of use only if you employ an e-discovery review platform supporting workflows based on paper document collections.

Litmus Test for Forms that Function

This affords us a litmus test to distinguish "native" forms from less functional forms: ***Can the form produced be imported into common e-mail client or server applications?***

You have to admire the simplicity. If a message is produced in a form that can be imported into e.g., Microsoft Outlook, it's a form likely to be searchable, sortable, utile and complete. More, it's a form that anyone can assimilate into whatever review platform they wish at lowest cost But, if the e-mail produced is so degraded 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.

The criterion, *"Will the form produced function in an e-mail client?"* enables parties to explore a broad range of functional native and near-native forms, not just PSTs. Such forms retain essential features like **Fielded Data,** allowing users to reliably sort messages by date, sender, recipients and subject, as well as**Message IDs**, supporting the threading of messages into coherent conversations. Functional forms supply the **UTC Offset Data** within e-mails that allows messages originating from different time zones and using different Daylight Savings Time settings to be normalized across an accurate timeline. Forms that Function don't disrupt the **Family Relationships** between messages and attachments. Forms that Function are **inherently electronically searchable**.

Best of all, producing Forms that Function means that all parties receive data in a form that anyone can use in any way they choose, visiting the costs of converting to alternate forms on the parties who want those alternate forms and not saddling parties with forms so degraded that they are functionally fractured and broken.

If you are a requesting party, don't be bamboozled by an alphabet soup of file extensions when it comes to e-mail production (PST, OST, MSG, EML, DBX, NSF, MHTML, TIFF, PDF, RTF, TXT, DAT, XML). Instead, tell the other side, "*I want Forms that Function. If it can be imported into Microsoft Outlook and work, that form will be fine by me.*"

If the other side says, "W*e will pull all that information out of the messages and give it to you in a load file*," say, *"No thanks, leave it where it lays, and give it to me in a Form that Functions!*"

Gmail

Gmail is a giant database in a Google data center someplace (or in many places). Who can guess what the native file format for cloud-based Gmail might be? Mere mortals don't get to peek at the guts of Google. But it little matters because, even if we *could* name the native file format, we can't obtain that format, nor can we faithfully replicate its functionality locally.[16]

Since we can't get "true" native, how can we otherwise mirror the completeness and functionality of native Gmail? When we get down to the lick log, a litigant doesn't seek native forms for grins. A 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?!?!

[16] 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 employed true native forms, as both were designed to support a different computing environment.

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,* 2013 U.S. Dist. LEXIS 60519 (S.D. Ind. Apr. 29, 2013)*,* an IMAP[17] capture to a PST format (using Microsoft Outlook or a collection tool) is a practical alternative. What you get will not 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.

Just Get Forms that Function

So, what's the native form of e-mail? maybe it doesn't matter., We should be less hung up on the term "native" and instead specify the actual form or forms we seek that are best suited to 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 your review tools), not just demanding native like it's a brand name.

When we seek "native" for a Word document or an Excel spreadsheet, it's because we recognize that the entire native file—and *only* the native file—supports the level of completeness and functionality we need, a level that can't be fairly replicated in any other form. But when we seek native production of e-mail, we can't expect to receive the entire "true" native file. We 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. 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 when it comes to native e-mail. Don't seek native to prove a point. Seek native to prove your case.

[17] IMAP (for Internet Message Access Protocol) is another way that e-mail client and server applications can talk to each another. The latest version of IMAP is described in RFC 3501. IMAP is not a form of e-mail storage, but 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).

Production from Databases

Enterprises increasingly rely on complex databases to manage recordkeeping and business processes such that at least some of the evidence in your case likely exists only as a value derived by querying a database.

Requesting parties typically ignore databases altogether in discovery. Else, they often demand entire databases, little thinking what such a demand will entail were it to succeed. If the database is built in Microsoft Access or some other simple tool, it's feasible to acquire the hardware and software licenses required to duplicate the producing party's database environment sufficiently to run the application. But, if the data sets require massive storage resources or are built on enterprise-level database management systems (DBMS) like Oracle or SAP, mirroring the environment is just about out of the question. I say "just about" because the emergence of Infrastructure-as-a-Service Cloud-based computing suggests the potential for mere mortals to deploy enterprise-level computing environments on a pay-as-you-go basis.

A more likely production scenario is to narrow the data set by use of filters and queries, then either export the responsive date to a format that can be analyzed in other applications (*e.g.*, exported as extensible markup language (XML), comma separated values (CSV) or in another delimited file) or run reports (standard or custom) and ensure that the reports emerge in a fielded, delimited format that supports electronic search.

Before negotiating a form of production, investigate the capabilities of the DBMS. The database administrator may not have had occasion to undertake a data export and so may have no clue what an application can do much beyond the confines of what it does every day. It's the rare DBMS that can't export delimited data. Next, have a proposed form of production in mind and, if possible, be prepared to instruct the DBMS administrator how to secure the reporting or export format you seek,

Remember that the resistance you experience in seeking to export to electronic formats may not come from the opposing party or the DBMS administrator. More often, an insistence on reports being produced as printouts or page images is driven by the needs of opposing counsel. In that instance, it helps to establish that the export is feasible as early as possible.

As with other forms of e-discovery, be careful not to accept production in formats you don't want because, like-it-or-not, many Court give just one bite at the production apple. If you accept it on a paper or as TIFF images for the sake of expediency, you often close the door on re-production in more useful forms.

Even if the parties can agree upon an electronic form of production, it's nevertheless a good idea to secure a test export to evaluate before undertaking a high volume export.

Gathering data on databases

Few advocates obtain useful insight into database production capabilities from opposing counsel. Often, opponents lack the grounding in DBMS needed to elicit and convey the information, or the other side so fears your effort to “invade” their databases that they adopt an uncooperative—at times, hostile—bunker mentality.

It’s the exceedingly rare case where discovery entails one party gaining direct access to another’s databases. Unfortunately, it’s nearly as rare for databases to be competently queried for discoverable content and their contents supplied in utile and complete forms.

By rights, database discovery should be one of the easiest and least contentious aspects of e-discovery. In practice, it’s a swamp.

To get database content produced in utile and complete forms requires requesting parties to inquire into the structure of the data and the capabilities of the DBMS, particularly reporting and export capabilities. To that end—and especially where opposing counsel can’t or won’t cooperate—it’s prudent to depose persons knowledgeable about the databases holding potentially responsive information

The following notice of deposition is an example of topics selected to elicit information needed to frame an efficient and effective database discovery effort.

NOTICE OF DEPOSITION(S) PURSUANT TO F.R.C.P.

PLEASE TAKE NOTICE that the deposition(s) of ***ABC Corporation*** pursuant to Fed. R. Civ. P. 30(b)(6) will take place at ***date/time/location***. The deposition, if not completed on the noticed date, shall be continued, if necessary, from day to day thereafter, excluding weekends and holidays, until completed. The deposition(s) will be conducted under the supervision of an officer who is authorized to administer an oath and will be recorded stenographically and on video.

Pursuant to Fed. R. Civ. P. 30(b)(6), ***ABC Corporation*** must designate and produce at the deposition(s) for examination one or more "officers, directors, or managing agents, or other persons who consent to testify" and who possess sufficient knowledge to testify as to the Deposition Topics listed below.

DEFINITIONS AND INSTRUCTIONS

The following definitions and instructions apply to this Notice:

1. The phrase "DATABASE OR SYSTEM" refers to a devices and mechanisms to store, access and retrieve data, including retired or "legacy" devices or systems, employed,

purchased by, leased, accessed, queried, subscribed to, summarized, controlled and/or provided to or obtained by you, including ***[list systems known to be of interest]***, or any other relevant databases that have not specifically identified herein.

2. For each DATABASE or SYSTEM that holds potentially responsive information, we seek to question the designated person(s) who, with reasonable particularity, can testify on your behalf about information known to or reasonably available to you concerning the Deposition Topics listed below.

3. The phrase "SUBJECT MATTER OF THE ACTION" means ***[e.g., the claims made in the operative complaint in this cause or other relevant subject matter]***;

3. Each deponent is instructed to produce at the deposition the requested information items listed in Exhibit A to this Notice ***[i.e., subpoena duces tecum]..***

DEPOSITION TOPICS

1. The standard reporting capabilities of the database or system, including the nature, purpose, structure, appearance, format and electronic searchability of the information conveyed within each standard report or reporting template that can be generated by the database or system or by any overlay (*e.g.,* third-party) reporting application(s);

2. The enhanced reporting capabilities of the database or system, including the nature, purpose structure, appearance, format and electronic searchability of the information conveyed within each enhanced or custom report (or template) that can be generated by the database or system or by any overlay (*e.g.,* third-party) reporting application;

3. The flat file and structured export capabilities of each database or system, particularly the ability to export to fielded/delimited or structured formats in a manner that faithfully reflects the content, integrity and functionality of the source data;

4. Other export and reporting capabilities of each database or system (including any overlay reporting application) and how they may or may not be employed to faithfully reflect the content, integrity and functionality of the source data for use in this litigation;

5. The structure of the database or system to the extent necessary to identify data within potentially responsive fields, records and entities, including field and table names, definitions, constraints and relationships, as well as field codes and field code/value translation or lookup tables.

6. The query language, syntax, capabilities and constraints of the database or system (including any overlay reporting application) as they may bear on the ability to identify, extract and export potentially responsive data from each database or system;

7. The user experience and interface, including datasets, functionality and options available for use by persons involved with the subject matter of the action;

8. The operational history of the database or system to the extent that it may bear on the content, integrity, accuracy, currency or completeness of potentially responsive data;

9. The nature, location and content of any training, user or administrator manuals or guides that address the manner in which the database or system has been administered, queried or its contents reviewed by persons involved with the subject matter of the action;

10. The nature, location and contents of any schema, schema documentation (such as an entity relationship diagram or data dictionary) or the like for any database or system that may reasonably be expected to contain information relating to the subject matter of the action;

11. The capacity and use of any database or system to log reports or exports generated by, or queries run against, the database or system where such reports, exports or queries may bear on the subject matter of the action;

12. The identity and roles of current or former employees or contractors serving as a database or system administrator for databases or systems that may reasonably be expected to contain (or have contained) information relating to the subject matter of the action; and

13. The cost, burden, complexity, facility and ease with which the information within databases and systems holding potentially responsive data relating to the subject matter of the action may be identified, preserved, searched, extracted and produced in a manner that faithfully reflects the content. integrity and functionality of the source data.

Forms of Production Matter

Once, forms of production hardly mattered.
Paper was paper.
Today, forms of production can spell the difference between winning and losing.
Forms of production matter.

Ask parties about the forms of ESI they use daily and it's doubtful you'll hear a peep about TIFF images or load files. *Parties don't use that junk; only lawyers do.* When clients create, communicate and collaborate, they do it using forms geared to native applications with file extensions like .XLSX, .DOCX, .PPTX, .MSG, etc. They choose and use functional and complete native and near-native forms. Those are the forms witnesses consult to reconstruct events and refresh their memories. Those are the forms witnesses recognize at deposition and in trial.

Yet, too often, e-discovery is a bait and switch con game: We request the parties'_modern data, but receive the lawyers' dilapidated junk. Once that inequity dawns on everyone, perhaps we will bid goodbye to wasting millions on senseless downgrading of ESI and ring in a new era of hands-on analytics.

If you are a requesting party, it's time to take a hard look at the language of the definitions and instructions accompanying your requests for production. If you're like most, you didn't draft that language from scratch. You borrowed some boilerplate from someone who borrowed it from someone who drafted it in 1947. That hand-me-down verbiage is long past retirement age; so, retire it and craft modern requests for a modern digital world.

We will never be less digital than we are today. We will never return to a world where paper is the preferred medium of information storage and transfer. Never.

We must move forms of production upstream, from depleted images and load files to functional native and near native forms retaining the content and structure that supports migration into any form. So, isn't it time we demand modern evidence and obtain it in the forms in which it serves us best? Utile forms. Complete forms. *Forms that function.*

About the Author

Craig Ball of Austin is a Board Certified trial lawyer, certified computer forensic examiner, law professor and electronic evidence expert. He limits his practice to serving as a court-appointed special master and consultant in computer forensics and electronic discovery and has served as the Special Master or testifying expert in computer forensics and electronic discovery in some of the most challenging and celebrated cases in the U.S. 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. For Craig's articles on e-discovery and computer forensics, please visit www.craigball.com or his blog, www.ballinyourcourt.com.

Appendices

Appendix 1
Broken Badly: Anderson Living Trust v. WPX Energy Production
Reprinted from *Ball in Your Court*, May 8, 2014

U.S. District Judge James Browning is a fine fellow. There are many reasons to say so; but the first is that, though he sits in New Mexico, he was born in the Great State of Texas. Judge Browning kindly spoke to my E-Discovery class at the law school in September 2012. I'd sought him out because he'd been ably grappling with e-discovery issues in a case styled *S2 v. Micron.* In his remarks to my class, he splendidly recounted some of the challenges faced by judges who ascended to the bench before the Age of Digital Evidence. Judge Browning has one of those C.V.s that could make any lawyer hate him (e.g., Yale, varsity letterman, Law Review editor-in-chief, Coif, Supreme Court clerk); but he's a good judge and a nice guy to boot.

I share my admiration of Judge Browning to underscore that I feel a bit of a rat in expressing misgivings about his recent opinion in *The Anderson Living Trust v. WPX Energy Production, LLC, No. CIV 12-0040 JB/LFG. (D. New Mexico March 6, 2014)*. I think he got it wrong in some respects–not on the peculiar equities of the case before him, but in his broader analysis of Rule 34 of the Federal Rules of Civil Procedure and in conjuring a Hobson's choice for requesting parties.

The *Anderson Living Trust* case is a fight over gas leases; but the merits don't matter. As the Court succinctly put it, the issue, is "whether the Defendants must arrange and label approximately 20,000 pages of documents stored in hard copy form which, at the Plaintiffs' request, were scanned and produced as searchable PDF files…."

The Defendants planned to unilaterally convert the paper into TIFF images with OCR load files, but acceded to Plaintiffs' request that they supply searchable PDFs instead. So, the Defendants converted paper records to TIFFs and then to searchable PDF images. Remember, at the start of the litigation*, the source evidence items were paper records*, not electronically stored information (ESI). Litigation alone prompted their conversion to crudely-searchable electronic formats.

In stark contrast to ESI, paper documents are inherently unsearchable. Thus, paper records are that rare form of evidence that is enhanced, rather than degraded, by conversion to page images and by use of optical character recognition (OCR) to approximate searchable text. As it happens, TIFF images cannot carry the text, but PDF images can. Think pants with pockets versus skirts without pockets. When you use TIFF images for production, text has to go *somewhere* and, since TIFFs have no "pockets," the text goes into a purse called a "load file." Load files are meant to be loaded into a database called a review tool where, paired with corresponding page images, non-searchable hard-copy documents acquire a rudimentary searchability.

Because the searchable text was derived by OCR (as opposed to text extracted from an electronic source), PDF wouldn't outshine TIFF in terms of the accuracy of text searchability. Both would be equally rife with errors, but both still better than paper. Accordingly, the principal distinctions between the two image formats go to convenience—you can search text in a PDF without messing with load files, and PDFs are more compact than TIFFs despite holding both page images and text.

The Plaintiffs' request for PDFs suggests they were seeking a form of production they could manage without review tools; but, considering the source was paper and OCR, they gained little by demanding PDF and, as it turned out, ceded quite a bit.

When Plaintiffs received the PDF production, they concluded they were unable to manage it unless the Defendants either organized the documents as they'd been kept in the usual course of business or indicated which items produced were responsive to which Request.

The defense furnished an index of their production but declined to do more, contending that, because the paper documents were now deemed ESI, the Plaintiffs could no longer secure the benefits of rules governing production of "documents."

The dispute thus turned on how to apply Fed. R. Civ. P. Rule 34(b)(2)(E), which provides:

> **Producing the Documents or Electronically Stored Information**. Unless otherwise stipulated or ordered by the court, these procedures apply to producing documents or electronically stored information:
>
> (i) A party must produce documents as they are kept in the usual course of business or must organize and label them to correspond to the categories in the request;
>
> (ii) If a request does not specify a form for producing electronically stored information, a party must produce it in a form or forms in which it is ordinarily maintained or in a reasonably usable form or forms; and
>
> (iii) A party need not produce the same electronically stored information in more than one form.

The Defendants made a compelling case, insisting they had: (i) produced the documents "as they are kept in the usual course of business"; (ii) "provided information about the particular way in which the documents are ordinarily maintained"; (iii) "provided an index identifying documents by category"; and (iv) "produced the documents so that they are fully searchable." The Defendants added the *coups de grâce* that organizing the production to correlate with the discovery would be a lot of effort that was unlikely to benefit anyone. They well played the proportionality card (although the word "proportionality" appears nowhere in the opinion).

Still, the Court agreed with the Plaintiffs and directed the Defendants to label their responsive documents to correspond to the Plaintiffs' requests. The ruling was a model of *stare* *decisis,* considering the Court cited three prior opinions in which it had reached a similar result ordering a party to tie Bates numbers to specific Requests.

The Court reflected on its reasoning as enunciated from the bench:

The Court said that it understood rule 34(b)(2)(E)(i) — requiring a responding party to produce documents "as they are kept in the usual course of business or . . . organize and label them to correspond to the categories in the request" — to apply both to hard copy documents and to ESI, as both are subsets of the catchall term "documents," and that rule 34(b)(2)(E)(ii) and (iii) are additional provisions related only to the production of ESI. … The Court expressed uncertainty regarding how a party would produce ESI "in the usual course of business," and that, if the party were to go through the documents and remove privileged or unresponsive documents before placing the files on a storage device, this production would not meet the "usual course of business" requirement and the party would have to label the documents to correspond to the categories in the request. … The Court compared hard copy document storage to ESI and noted that it would be difficult to find an analog to allowing opposing counsel access to boxes of information kept in warehouses, because it would require the responding party to give the other party access to the responding party's computer system, or to place all of the files on a storage device without culling out any unresponsive or privileged files….

The Court, it appears, was less than supremely confident in its ruling. Understandably, as these were not your usual, imperious "Let them eat TIFF" Defendants. They approached the Plaintiffs and inquired about preferred forms of production. They supplied native files in native forms *and* paired same to particular requests. They promised that the paper documents were ordered as in the usual course, notwithstanding their digitized format. They graciously added searchability to unsearchable paper documents and furnished an index.

I've seen worse…*much* worse.

A month later, the Court held a hearing on an unrelated matter and the defense asked the Court to reconsider its directive to pair the Bates numbers of the electrified paper with the requests to which they are responsive. It's not clear from the opinion why the work hadn't been done in the intervening month, and I expect Plaintiffs' counsel couldn't have passed a BB when the Court said it "was seriously rethinking its prior ruling" and that "some of the commentators and some of the cases conflate."

I hope it wasn't some blindsided young associate who had to comment and conflate all that back to the partners. [Have you seen *Breaking Bad*? They don't mess around in Albuquerque].

The Court made it easy for the defense, stating that, if the Defendants could prove that the ESI was produced as it was kept in the usual course of business, without litigation-related alteration, then the Court was inclined to rule that no labeling would be required. But the Court added, "once you've set there and you had your paralegals go through it, you've decided that this is relevant, this is not relevant, we're going to go through it for privilege, we're going to Bates stamp it, I don't think that's . . . the usual course of business."

Undaunted, the defense provided a declaration, and a damn fine one, too! The Declaration establishes that the only documents removed from production were privileged ones, and supplies a breakdown of contents by Bates number ranges and sources. The affidavit also confirms my suspicion that the venerable Plaintiffs' firm was trying to navigate e-discovery in an "old school" way, without benefit of basic e-discovery tools.

The Plaintiffs supplied no counter-declaration from anyone with any e-discovery expertise (or from anyone at all, insofar as PACER reveals).

Reminder: *Evidence is good. Judges like evidence more than lawyer talk*.

At this juncture, the Court could have put this matter to bed in three ways, without muss, fuss or dicta:

1. The Court could have found that the material in question derived from hard-copy documents clearly subject to 34(b)(2)(E)(i) at the start of litigation, and the conversion of these paper documents to searchable electronic forms for use by counsel in discovery didn't change their essential character for purposes of requiring that they be organized as they are kept in the usual course of business or organized and labeled to correspond to the categories of a request. Plaintiff prevails.

2. The Court could have found that the electronic counterparts of the paper documents had been produced in substantially the way they were kept in the usual course of business, making reasonable allowances for variations attendant to the parties' agreement to scan and OCR the material and the need to protect privileged content. Documents withheld as privileged would necessarily be identified in a privilege log by Bates number, so, their location within the collection could be readily established. Defendant prevails.

3. The Court could have found (and did find) that the Parties agreement respecting production was a stipulation that altogether removed the issue from the purview of Rule 34(b)(2)(E), and the Court could have fashioned any outcome it deemed proper and proportionate without further need to address the (inapplicable) Rule 34(b)(20(E) in dicta.

Instead, the Court pursued a broad-ranging assessment of Rule 34(b)(2)(E) that stirs an eddy of uncertainty.

1. For example, the Court termed "unavailing" the argument that the source documents weren't ESI because they existed in hard copy form and were only imaged for production. The Court reasoned that the agreement to image the documents was, in fact, the parties stipulating out of the rule's default provisions.

This makes little sense. Certainly, parties can agree to stipulate around Rule 34(b)(2)(E); but such a stipulation should be clear and express. It needn't follow that because one party accedes to another party's desire to make paper records more convenient, organization of the information is optional. Why should a mutually beneficial endeavor come at the risk that an opponent is free to destroy the usual and customary organization of the evidence or at the cost of a requesting party's right to know what's responsive to what?

The parties merely settled upon *forms* of production. They made no bargain respecting the *organization* of production, nor did they agree to restrict the *scope* of production. These are three distinct dimensions of discoverable information.

It was the producing party's avowed intention to convert the hard-copy documents to TIFF images. Had they done so without the agreement of the requesting party, they would nonetheless have been obliged to produce the hard-copy documents as kept in the usual course or organized and labeled to correspond to the request. Judge Browning made quite clear that, "'if at the beginning of the litigation the documents existed in document hard copy form,' then the Defendant could not unilaterally convert the documents into ESI." However, if the requesting party cooperates and allows a producing party to convert hard-copy documents to TIFF or PDF, the requesting party is now agreeing, *sub silentio*, to forego organization of the documents. The producing party is thus free to make an unholy mess of the production from an organizational standpoint, and there's not a thing the requesting party can do about it. That doesn't add up.

2. At the start of the litigation, the 20,000 pages produced were paper records kept in the usual course of the Defendants' business. From the standpoint of the usual course of the Defendants' business, they never changed form. That is, they did not become ESI in conjunction with the customary operation or recordkeeping of the Defendants' business. So, they remained subject to the provisions of Rule 34(b)(2)(E)(i). It's a mistake to equate conversion for the convenience of the lawyers to conversion in the usual course of the litigants' business.

If a lawyer elects to convert ESI to another form like scanned images, the destination form may be the form used in the course of the lawyer's business, but it's not the form used by the producing party

The Court fails to distinguish between the form and organization of information as used by the parties to an action and the (de)form and (dis)organization occasioned by counsel's wish to convert information to something else. We frequently encounter this assumption in e-discovery. That is, producing parties assume that requesting parties can't demand any form more complete or utile than the dumbed-down versions used by producing party's counsel.

That's not the rule.

Requesting parties should be entitled to obtain forms of production that mirror the forms the producing parties use, not compelled to accept the degraded forms preferred by the producing party's lawyer.

3. The *form* of production does not implicate the *organization* of production. They are different, and each presents different opportunities for abuse. A party can produce information items in utile, complete and searchable forms but still disrupt organization or logical unitization (i.e., folder structure) so as to render the production all but useless. The notion that electronic search adequately offsets the risk of shuffling and other organizational mischief is untenable—ask anyone who's ever gotten a malformed load file.

Organizational information, like foldering data and file locations (paths), are as essential to utile production today as they were in the paper era, if not more so.

4. I don't share the Court's view that potentially responsive documents/ESI collected as maintained in the usual course of business lose that character when privileged documents are culled. Granted, the collection is not as complete as kept in the usual course, but the remaining documents are still organized and in the same form as kept in the usual course. In the paper era, it was customary for the boxes from the legal department to be spirited out of the warehouse; and when privileged contents turned up, they, too, were pulled. Production by inspection didn't oblige a producing party to abandon privilege. Parties need not do so with respect to ESI.

5. Most problematic of all is the Court's conclusion that provisions 34(b)(2)(E)(i) and 34(b)(2)(E)(ii) "apply to distinct, mutually exclusive categories of discoverable information," being "Documents," which the Court calls "a term that does not include ESI," governed exclusively by 34(b)(2)(E)(i), and ESI, which the Court says is governed exclusively by 34(c)(E)(ii). The Court relies on the views of a distinguished commentator, John K. Rabiej. With respect to Professor Rabiej–who was closely involved with the amendments process—his disjunctive interpretation of 34(b)(2)(E) is one thoughtful view; but one that seems oddly out of step with the Committee Notes.

The Court's embrace of such a distinction is regrettable because the 2006 Federal Rules amendments and the Committee Notes that accompany them go to some pains to underscore that the term "documents" includes ESI. In fact, defining "Documents" to encompass data compilations has a long and uncontentious history in the Rules.

The Court saw the perils, stating, "There is something to be gained from imposing basic organization requirements onto massive productions of ESI; artifacts of ESI can be jumbled beyond usefulness — by dumping them out of their file directories and onto the requesting party — just as easily as hard copy documents can." Indeed, and *it happens all the time*, though more often as a consequence of carelessness than of bad faith.

The organization of ESI in production can fairly and efficiently be made to mirror its organization in the usual course of business. It typically requires little more than competent handling of system metadata. It doesn't require granting an opponent access to a responding party's computer systems.

Here again, it's useful to distinguish the three dimensions of discoverable ESI: *form, organization* and *scope*. If a party culls privileged content before producing the data for inspection, form and organization remain the same; only scope changes—and it's appropriate that privileged content be outside the general scope of discovery. Any minimal impact on organization is offset by the obligation to log content withheld as privileged.

6. Finally, it's an ill wind which blows no man to good. Judge Browning clarified that his ruling did not apply unless the requesting party sought conversion to an imaged format. "'[I]f at the beginning of the litigation the documents existed in document hard copy form,' then the Defendant could not unilaterally convert the documents into ESI."

By that reasoning, if at the beginning of the litigation the documents existed as ESI, the producing party cannot unilaterally convert the documents into paper or paper-like forms (e.g., images) *unless the requesting party stipulates to same*.

Quoting Professor Rabiej, the Court notes that "while (E)(i) document production gives the producing party the right to choose whether to produce 'in the usual course of business" or "label[ed] … to correspond to the categories in the request,' (E)(ii) puts the ball in the requesting party's court by first giving them the option to 'specify a form for producing' ESI. Fed. R. Civ. P. 34(b)(2)(E)(i)-(ii). It is only if the requesting party declines to specify a form that the producing party is offered a choice between producing in the form 'in which it is ordinary maintained' — native format — or 'in a reasonably useful form or forms.' Fed. R. Civ. P. 34(b)(2)(E)(i)-(ii)."

That's powerful stuff, and dead right. Producing parties have long assumed that they were free to ignore a requesting party's specification of form so long as they produced in a form claimed to be "reasonably usable." Not so. As the Court notes, the "reasonably usable" option applies only when a requesting party fails to specify a form.

The lesson for requesting parties is always, *always*, ALWAYS specify forms for production in your requests. If you want Word documents produced natively, SAY SO! If you want e-mail in functional forms, specify the forms! The Rules afford requesting parties the crucial right to specify form or forms of production, and lawyers who fail to avail themselves of that right are inviting production of less utile and -complete forms. If you wear a "KICK ME" sign on your bootie, don't be surprised by the boot.

So, with apologies to Judge Browning, the result seems right, but the rationale not so much. It's dicta likely to be cited in support of mischief, and I know that's not what the Court wished.

Appendix 2: Exemplar Production Protocol

1. "Information items" as used here encompass individual documents and records (including associated metadata) whether on paper or film, as discrete "files" stored electronically, optically or magnetically or as a record within a database, archive or container file. The term should be read broadly to include e-mail, messaging, word processed documents, digital presentations and spreadsheets.

2. Responsive electronically stored information (ESI) shall be produced in its native form; that is, in the form in which the information was customarily created, used and stored by the native application employed by the producing party in the ordinary course of business.

3. If it is infeasible to produce an item of responsive ESI in its native form, it may be produced in an agreed-upon near-native form; that is, in a form in which the item can be imported into the native application without a material loss of content, structure or functionality as compared to the native form. Static image production formats serve as near-native alternatives only for information items that are natively static images (*i.e.*, photographs and scans of hard-copy documents).

4. The table below supplies examples of agreed-upon native or near-native forms in which specific types of ESI should be produced:

Source ESI	Native or Near-Native Form or Forms Sought
Microsoft Word documents	.DOC, .DOCX
Microsoft Excel Spreadsheets	.XLS, .XLSX
Microsoft PowerPoint Presentations	.PPT, .PPTX
Microsoft Access Databases	.MDB, .ACCDB
WordPerfect documents	.WPD
Adobe Acrobat Documents	.PDF
Photographs	.JPG, .PDF
E-mail	Messages should be produced in a form or forms that readily support import into standard e-mail client programs; that is, the form of production should adhere to the conventions set out in RFC 5322 (the internet e-mail standard). For Microsoft Exchange or Outlook messaging, .PST format will suffice. Single message production formats like .MSG or .EML may be furnished, if source foldering data is preserved and produced. For Lotus Notes mail, furnish .NSF files or convert to .PST. If your workflow requires that attachments be extracted and produced separately from transmitting messages, attachments should be produced in their

	native forms with parent/child relationships to the message and container(s) preserved and produced in a delimited text file.

5. Absent a showing of need, a party shall produce responsive information reports contained in databases through the use of standard reports; that is, reports that can be generated in the ordinary course of business and without specialized programming efforts beyond those necessary to generate standard reports. All such reports shall be produced in a delimited electronic format preserving field and record structures and names. The parties will meet and confer regarding programmatic database productions as necessary.

6. Information items that are paper documents or that require redaction shall be produced in static image formats scanned at 300 dpi e.g., single-page Group IV.TIFF or multipage PDF images. If an information item contains color, the producing party shall not produce the item in a form that does not display color. The full content of each document will be extracted directly from the native source where feasible or, where infeasible, by optical character recognition (OCR) or other suitable method to a searchable text file produced with the corresponding page image(s) or embedded within the image file. Redactions shall be logged along with other information items withheld on claims of privilege.

7. Parties shall take reasonable steps to ensure that text extraction methods produce usable, accurate and complete searchable text.

8. Individual information items requiring redaction shall (as feasible) be redacted natively, produced in .PDF format and redacted using the Adobe Acrobat redaction feature or redacted and produced in a format that does not serve to downgrade the ability to electronically search the unredacted portions of the item. Bates identifiers should be endorsed on the lower right corner of all images, but not so as to obscure content.

9. Upon a showing of need, a producing party shall make a reasonable effort to locate and produce the native counterpart(s) of any unredacted .PDF or .TIF document produced. The parties agree to meet and confer regarding production of any such documents. This provision shall not serve to require a producing party to reveal redacted content.

10. Except as set out in this Protocol, a party need not produce identical information items in more than one form and shall globally de-duplicate identical items across custodians using each document’s unique MD5 hash value. The content, metadata and utility of an information item shall all be considered in determining whether information items are identical, and items reflecting different information shall not be deemed identical.

11. Production should be made on CD, DVD or hard drive(s) using the medium requiring the least number of deliverables. Label all media with the case number, production date, Bates range and disk number (1 of X, if applicable). Organize productions by custodian, unless otherwise instructed. All documents from an individual custodian should be confined to a single load file. All productions should be encrypted for transmission to the receiving party. The producing party shall, contemporaneously with production, supply decryption credentials and passwords to the receiving party for all items produced in an encrypted or password-protected form.

12. Each information item produced shall be identified by naming the item to correspond to a Bates 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 six (6) characters will be a designation reserved to the discretionary use of the party making production for the purpose of, e.g., denoting the case or matter. This value shall be padded with leading zeroes as needed to preserve its length;

    iii. The next nine (9) characters will be a unique, consecutive numeric value assigned to the item by the producing party. This value shall be padded with leading zeroes as needed to preserve its length;

    iv. The final six (6) characters are reserved to a sequence consistently beginning with a dash (-) or underscore (_) followed by a five digit number reflecting pagination of the item when printed to paper or converted to an image format for use in proceedings or when attached as exhibits to pleadings.

    v. By way of example, a Microsoft Word document produced by Acme in its native format might be named: ACMESAMPLE000000123.docx. Were the document printed out for use in deposition, page six of the printed item must be embossed with the unique identifier ACMESAMPLE000000123_00006. Bates identifiers should be endorsed on the lower right corner of all printed pages, but not so as to obscure content.

    vi. This format of the Bates identifier must remain consistent across all productions. The number of digits in the numeric portion and characters in the alphanumeric portion of the identifier should not change in subsequent productions, nor should spaces, hyphens, or other separators be added or deleted except as set out above.

13. Information items designated Confidential may, at the Producing Party's option:

a. Be separately produced on electronic production media prominently labeled to comply with the requirements of the **[DATE]** Protective Order entered in this matter; or, alternatively,

b. Each such designated information item shall have appended to the file's name (immediately following its Bates identifier) the following protective legend:
~CONFIDENTIAL-SUBJ_TO_PROTECTIVE_ORDER

When any item so designated is converted to a printed or imaged format for use in any submission or proceeding, the printout or page image shall bear the protective legend on each page in a clear and conspicuous manner, but not so as to obscure content.

14. Producing party shall furnish a delimited load file supplying the metadata field values listed below for each information item produced (to the extent the values exist and as applicable):

Field Name	**Sample Data**	**Description**
BegBates	ACMESAMPLE000000001	First Bates identifier of item
EndBates	ACMESAMPLE000000123	Last Bates identifier of item
AttRange	ACMESAMPLE000000124 - ACMESAMPLE000000130	Bates identifier of the first page of the parent document to the Bates identifier of the last page of the last attachment "child" document
BegAttach	ACMESAMPLE000000124	First Bates identifier of attachment range
EndAttach	ACMESAMPLE000000130	Last Bates identifier of attachment range
Parent_Bates	ACMESAMPLE000000001	First Bates identifier of parent document/e-mail message. ***This Parent_Bates field should be populated in each record representing an attachment "child" document. ***

Child_Bates	ACMESAMPLE000000004;
	ACMESAMPLE000000012;
	ACMESAMPLE000000027
		First Bates identifier of "child" attachment(s); may be more than one Bates number listed depending on number of attachments.
		***The Child_Bates field should be populated in each record representing a "parent" document. ***

Custodian	Houston, Sam	E-mail: mailbox where the email resided. Native: Individual from whom the document originated
Path	E-mail: \Deleted Items\Battles\ SanJac.msg Native: Z:\TravisWB\Alamo.docx	E-mail: Original location of e-mail including original file name. Native: Path where native file document was stored including original file name.
From	GuerreroJ@hotmail.com; David Crockett [mailto: Davy@Crockett.net]	E-mail: Sender Native: Author(s) of document **semi-colons separate multiple entries **
To	Genl. A.L. de Santa Anna	Recipient(s) **semi-colons separate multiple entries **
CC	Jim.Bowie@bigknife.com	Carbon copy recipient(s) **semi-colons separate multiple entries **
BCC	AustinSF@state.tx.gov	Blind carbon copy recipient(s) **semi-colons separate multiple entries **
Date Sent	03/18/2014	E-mail: Date the email was sent

Time Sent	11:45 AM	E-mail: Time the message was sent
Subject/Title	Remember the Alamo!	E-mail: Subject line of the message
IntMsgID	<A1315BC17ABD4774BF779CB3E3E62B9B@gmail.com>	E-mail: For e-mail in Microsoft Outlook/Exchange, the “Unique Message ID” field; For e-mail in Lotus Notes, the UNID field. Native: empty.
Date_Mod	02/23/1836	E-mail: empty. Native: Last Modified Date
Time_Mod	01:42 PM	E-mail: empty Native: Last Modified Time
File_Type	XLSX	E-mail: empty Native: file type
Redacted	Y	Denotes that item has been redacted as containing privileged content (yes/no).
File_Size	1,836	Size of native file document/email in KB.
HiddenCnt	N	Denotes presence of hidden Content/Embedded Objects in item(s) (yes/no)
Confidential	Y	Denotes that item has been designated as confidential pursuant to protective order (yes/no).
MD5_Hash	eb71a966dcdddb929c1055ff2f1ccd5b	MD5 Hash value of the item.
DeDuped	E-mail: \Inbox\SanJac.msg Native: Z:\CrockettD\Alamo.docx	Full path of other instances de-deduplicated by MD5 hash **semi-colons separate multiple entries **

15. Each production should include a cross-reference load file that correlates the various files, images, metadata field values and searchable text produced.

16. Parties shall respond to each request for production by listing the Bates identifiers/ranges of responsive documents produced, and where an information item responsive to these discovery requests has been withheld or redacted on a claim that it is privileged, the producing party shall furnish a privilege log.

Appendix 3: U.S. S.E.C. Data Delivery Standards

Available from:

http://www.sec.gov/divisions/enforce/datadeliverystandards.pdf

U.S. Securities and Exchange Commission

Data Delivery Standards

The following outlines the technical requirements for producing scanned paper collections, email and electronic document/ native file collections to the Securities and Exchange Commission. The SEC uses *Concordance®* 2007 v9.58 and *Concordance Image®* v4.53 software to search, review and retrieve documents produced to us in electronic format. Any proposed production in a format other than those identified below, the proposed use of *Predictive Coding*, *computer-assisted review* or *technology-assisted review* (TAR), or the use of de-duplication during the processing of documents, must be discussed with and approved by the legal and technical staff of the Division of Enforcement (ENF) and the methodology must be disclosed in the cover letter. We appreciate your efforts in assisting us by preparing data in a format that will enable our staff to use the data efficiently.

General Instructions

1. A cover letter should be included with each production. *This letter MUST be imaged and provided as the first record in the load file.*
   The following information should be included in the letter:
   a. List of each piece of media (hard drive, thumb drive, DVD or CD) included in the production by the unique number assigned to it, and readily apparent on the physical media.
   b. List of custodians, identifying:
      1) The Bates range (and any gaps therein) for each custodian
      2) Total number of records for each custodian
      3) Total number of images for each custodian
      4) Total number of native files for each custodian
   c. List of fields in the order in which they are listed in the data file.
   d. Time zone in which emails were standardized during conversion (email collections only).
2. Documents created or stored electronically MUST be produced in their original electronic format, not printed to paper or PDF.
3. Data can be produced on CD, DVD or hard drive; *use the media requiring the least number of deliverables.*
4. Label all media with the following:
   a. Case number
   b. Production date
   c. Bates range
   d. Disk number (1 of X), if applicable

5. Organize productions by custodian, unless otherwise instructed. All documents from an individual custodian should be confined to a single load file.
6. All productions should be checked and produced free of computer viruses.
7. All produced media should be encrypted.
8. Passwords for documents, files, compressed archives and encrypted media should be provided separately either via email or in a separate cover letter from the data.

Delivery Formats

I. *Concordance®* Production

All scanned paper, email and native file collections should be converted/processed to TIFF files, Bates numbered, and include fully searchable text. Additionally, email and native file collections should include linked native files.

Bates numbering documents:
The Bates number must be a unique, consistently formatted identifier, i.e., an alpha prefix along with a fixed length number for EACH custodian., i.e., ABC0000001. This format MUST remain consistent across all production numbers for each custodian. The number of digits in the numeric portion of the format should not change in subsequent productions, nor should spaces, hyphens, or other separators be added or deleted.

The following describes the specifications for producing image-based productions to the SEC and the load files required for *Concordance®* and *Concordance Image®*.

1. Images

a. Images should be single-page, Group IV TIFF files, scanned at 300 dpi.
b. File names cannot contain embedded spaces.
c. Bates numbers should be endorsed on the lower right corner of all images.
d. The number of TIFF files per folder should not exceed 500 files.
e. Rendering to images PowerPoint, AUTOCAD/ photographs and Excel files:
   1) PowerPoint: All pages of the file should be scanned in full slide image format, with any speaker notes following the appropriate slide image.
   2) AUTOCAD/ photographs: If possible, files should be scanned to single page JPEG (.JPG) file format.
   3) Excel: TIFF images of spreadsheets are not useful for review purposes; because the imaging process can often generate thousands of pages per file, a placeholder image, named by the *IMAGEID* of the file, may be used instead.

2. *Concordance Image®* Cross-Reference File

The image cross-reference file is needed to link the images to the database. It is a comma-delimited file consisting of seven fields per line. There must be a line in the cross-reference file for every image in the database.

The format is as follows:
ImageID,VolumeLabel,ImageFilePath,DocumentBreak,FolderBreak,BoxBreak,PageCount

ImageID: The unique designation that *Concordance®* and *Concordance Image®* use to identify an image. ***Note***: *This imageID key* ***must*** *be a unique and fixed length number. This number will be used in the .DAT file as the ImageID field that links the database to the images. The format of this image key must be consistent across all productions. We recommend that the format be a 7 digit number to allow for the possible increase in the size of a production.*

VolumeLabel: Optional

ImageFilePath: The full path to the image file.

DocumentBreak: The letter "Y" denotes the first page of a document. If this field is blank, then the page is not the first page of a document.

FolderBreak: Leave empty

BoxBreak: Leave empty

PageCount: Optional

Sample

```
IMG0000001,,E:\001\IMG0000001.TIF,Y,,,
IMG0000002,,E:\001\IMG0000002.TIF,,,,
IMG0000003,,E:\001\IMG0000003.TIF,,,,
IMG0000004,,E:\001\IMG0000003.TIF,Y,,,
IMG0000005,,E:\001\IMG0000003.TIF,Y,,,
IMG0000006,,E:\001\IMG0000003.TIF,,,,
```

3. *Concordance®* Data File

The data file (.DAT) contains all of the fielded information that will be loaded into the *Concordance®* database.

a. The first line of the .DAT file must be a header row identifying the field names.

b. The .DAT file must use the following *Concordance®* default delimiters:

Comma	¶	ASCII character (020)
Quote	þ	ASCII character (254)
Newline	®	ASCII character (174)

c. Date fields should be provided in the format: mm/dd/yyyy

d. All attachments should sequentially follow the parent document/email.

e. All metadata associated with email, audio files, and native electronic document collections must be produced (see pages 4-5).

f. The .DAT file for scanned paper collections must contain, at a minimum, the following fields:

1) FIRSTBATES: Beginning Bates number
2) LASTBATES: Ending Bates number
3) IMAGEID: Image Key field
4) CUSTODIAN: Individual from whom the document originated
5) OCRTEXT: Optical Character Recognition (file path, or text)

Sample of .DAT file (when text files are provided separately)

```
þFIRSTBATESþ¶þLASTBATESþ¶þIMAGEIDþ¶þCUSTODIANþ¶þOCRTEXTþ
þPC00000001þ¶þPC00000002þ¶þIMG0000001þ¶þSmith, Johnþ¶þE:\TEXT\PC00000001.TXTþ
þPC00000003þ¶þPC00000003þ¶þIMG0000003þ¶þSmith, Johnþ¶þE:\TEXT\PC00000003.TXTþ
þPC00000004þ¶þPC00000005þ¶þIMG0000004þ¶þSmith, Johnþ¶þE:\TEXT\PC00000004.TXTþ
```

Sample of .DAT file (with text)

```
þFIRSTBATESþ¶þLASTBATESþ¶þIMAGEIDþ¶þCUSTODIANþ¶þOCRTEXTþ
þPC00000001þ¶þPC00000002þ¶þIMG0000001þ¶þSmith, Johnþ¶þ*** IMG0000001 ***®®The world of
investing is fascinating and complex, and it can be very fruitful. But unlike the banking
world, where deposits are guaranteed by the federal government, stocks, bonds and other
securities can lose value. There are no guarantees. That's why investing is not a spectator
sport. By far the best way for investors to protect the money they put into the securities
markets is to do research and ask questions.®® *** IMG0000002 ***®®The laws and rules that
govern the securities industry in the United States derive from a simple and
straightforward concept: all investors, whether large institutions or private individuals,
should have access to certain basic facts about an investment prior to buying it, and so
long as they hold it. To achieve this, the SEC requires public companies to disclose
meaningful financial and other information to the public. This provides a common pool of
knowledge for all investors to use to judge for themselves whether to buy, sell, or hold a
particular security. Only through the steady flow of timely, comprehensive, and accurate
information can people make sound investment decisions.þ
þPC00000003þ¶þPC00000003þ¶þIMG0000003þ¶þSmith, Johnþ¶þ***IMG0000003 ***®®The result of this
information flow is a far more active, efficient, and transparent capital market that
facilitates the capital formation so important to our nation's economy.þ
þPC00000004þ¶þPC00000005þ¶þIMG0000004þ¶þSmith, Johnþ¶þ *** IMG0000004 ***®®To insure that
this objective is always being met, the SEC continually works with all major market
participants, including especially the investors in our securities markets, to listen to
their concerns and to learn from their experience.®® *** IMG0000005 ***®®The SEC oversees
the key participants in the securities world, including securities exchanges, securities
brokers and dealers, investment advisors, and mutual funds. Here the SEC is concerned
primarily with promoting the disclosure of important market-related information,
maintaining fair dealing, and protecting against fraud.þ
```

The text and metadata of Email and the attachments, and native file document collections should be extracted and provided in a .DAT file using the field definition and formatting described below:

Field Name	Sample Data	Description
FIRSTBATES	EDC0000001	First Bates number of native file document/email
LASTBATES	EDC0000001	Last Bates number of native file document/email **The LASTBATES field should be populated for single page documents/emails.
ATTACHRANGE	EDC0000001 - EDC0000015	Bates number of the first page of the parent document to the Bates number of the last page of the last attachment "child" document
BEGATTACH	EDC0000001	First Bates number of attachment range
ENDATTACH	EDC0000015	Last Bates number of attachment range
PARENT_BATES	EDC0000001	First Bates number of parent document/Email **This PARENT_BATES field should be populated in each record representing an attachment "child" document
CHILD_BATES	EDC0000002; EDC0000014	First Bates number of "child" attachment(s); can be more than one Bates number listed depending on the number of attachments **The CHILD_BATES field should be populated in each record representing a "parent" document
CUSTODIAN	Smith, John	Email: mailbox where the email resided Native: Individual from whom the document originated
FROM	John Smith	Email: Sender Native: Author(s) of document **semi-colon should be used to separate multiple entries
TO	Coffman, Janice; LeeW [mailto:LeeW@MSN.com]	Recipient(s) **semi-colon should be used to separate multiple entries
CC	Frank Thompson [mailto: frank_Thompson@cdt.com]	Carbon copy recipient(s) **semi-colon should be used to separate multiple entries
BCC	John Cain	Blind carbon copy recipient(s) **semi-colon should be used to separate multiple entries
SUBJECT	Board Meeting Minutes	Email: Subject line of the email Native: Title of document (if available)
DATE_SENT	10/12/2010	Email: Date the email was sent Native: (empty)
TIME_SENT	07:05 PM	Email: Time the email was sent Native: (empty) **This data must be a separate field and cannot be combined with the DATE_SENT field
LINK	D:\001\ EDC0000001.msg	Hyperlink to the email or native file document **The linked file must be named per the FIRSTBATES number
MIME_TYPE	MSG	The content type of an Email or native file document as identified/extracted from the header
FILE_EXTEN	MSG	The file type extension representing the Email or native file document; will vary depending on the email format
AUTHOR	John Smith	Email: (empty) Native: Author of the document
DATE_CREATED	10/10/2010	Email: (empty) Native: Date the document was created

TIME_CREATED	10:25 AM	Email: (empty) Native: Time the document was created **This data must be a separate field and cannot be combined with the DATE_CREATED field
DATE_MOD	10/12/2010	Email: (empty) Native: Date the document was last modified
TIME_MOD	07:00 PM	Email: (empty) Native: Time the document was last modified **This data must be a separate field and cannot be combined with the DATE_MOD field
DATE_ACCESSD	10/12/2010	Email: (empty) Native: Date the document was last accessed
TIME_ACCESSD	07:00 PM	Email: (empty) Native: Time the document was last accessed **This data must be a separate field and cannot be combined with the DATE_ACCESSD field
PRINTED_DATE	10/12/2010	Email: (empty) Native: Date the document was last printed
FILE_SIZE	5,952	Size of native file document/email in KB
PGCOUNT	1	Number of pages in native file document/email
PATH	J:\Shared\SmithJ\October Agenda.doc	Email: (empty) Native: Path where native file document was stored including original file name.
INTFILEPATH	Personal Folders\Deleted Items\Board Meeting Minutes.msg	Email: original location of email including original file name. Native: (empty)
INTMSGID	<000805c2c71b$75977050$cb8306d1@MSN>	Email: Unique Message ID Native: (empty)
MD5HASH	d131dd02c5e6eec4693d9a0698aff95c 2fcab58712467eab4004583eb8fb7f89	MD5 Hash value of the document.
TEXT	From: Smith, John Sent: Tuesday, October 12, 2010 07:05 PM To: Coffman, Janice Subject: Board Meeting Minutes Janice; Attached is a copy of the September Board Meeting Minutes and the draft agenda for October. Please let me know if you have any questions. John Smith Assistant Director Information Technology Phone: (202) 555-1111 Fax: (202) 555-1112 Email: jsmith@xyz.com	Extracted text of the native file document/email

4. Text

Searchable text of the entire document must be provided for every record, at the document level.

a. Extracted text must be provided for all documents that originated in electronic format. The text files should include page breaks that correspond to the 'pagination' of the image files. Note: Any document in which text cannot be extracted must be OCR'd, particularly in the case of PDFs without embedded text.

b. OCR text must be provided for all documents that originated in hard copy format. A page marker should be placed at the beginning, or end, of each page of text, e.g. *** IMG0000001 *** whenever possible. The data surrounded by asterisks is the *Concordance®* ImageID .

Sample page markers with OCR text:

> *** IMG0000001 ***
>
> The world of investing is fascinating and complex, and it can be very fruitful. But unlike the banking world, where deposits are guaranteed by the federal government, stocks, bonds and other securities can lose value. There are no guarantees. That's why investing is not a spectator sport. By far the best way for investors to protect the money they put into the securities markets is to do research and ask questions.
>
> *** IMG0000002 ***
>
> The laws and rules that govern the securities industry in the United States derive from a simple and straightforward concept: all investors, whether large institutions or private individuals, should have access to certain basic facts about an investment prior to buying it, and so long as they hold it. To achieve this, the SEC requires public companies to disclose meaningful financial and other information to the public. This provides a common pool of knowledge for all investors to use to judge for themselves whether to buy, sell, or hold a particular security. Only through the steady flow of timely, comprehensive, and accurate information can people make sound investment decisions.

c. For redacted documents, provide the full text for the redacted version.

d. Delivery
The text can be delivered two ways:
1) As multi-page ASCII text files with the files named the same as the ImageID field. Text files can be placed in a separate folder or included with the .TIF files. The number of files per folder should be limited to 500 files.
2) Included in the .DAT file.

5. Linked Native Files

Copies of original email and native file documents/attachments must be included for all electronic productions.
a. Native file documents must be named per the FIRSTBATES number.
b. The full path of the native file must be provided in the .DAT file for the LINK field.
c. The number of native files per folder should not exceed 500 files.

II. Audio Files

Audio files from telephone recording systems must be produced in a format that is playable using Microsoft Windows Media Player™. Additionally, the call information (metadata) related to each audio recording MUST be provided. The metadata file must be produced in a delimited text format. Field names must be included in the first row of the text file.

The metadata must include, at a minimum, the following fields:

1)	Caller Name:	Caller's name or account/identification number
2)	Originating Number:	Caller's phone number
3)	Called Party Name:	Called party's name
4)	Terminating Number:	Called party's phone number
5)	Date:	Date of call
6)	Time:	Time of call
7)	Filename:	Filename of audio file

III. Video Files

Video files must be produced in a format that is playable using Microsoft Windows Media Player™.

IV. Electronic Trade and Bank Records

When producing electronic trade and bank records, provide the files in one of the following formats:

1. MS Excel spreadsheet with header information detailing the field structure. If any special codes exist in the dataset, a separate document must be provided that details all such codes. If details of the field structure do not fit in the header, a separate document must be provided that includes such details.

2. Delimited text file with header information detailing the field structure. The preferred delimiter is a vertical bar "|". If any special codes exist in the dataset, a separate document must be provided that details all such codes. If details of the field structure do not fit in the header, a separate document must be provided that includes such details.

V. Electronic Phone Records

When producing electronic phone records, provide the files in one of the following formats:

1. MS Excel spreadsheet with header information detailing the field structure. If any special codes exist in the dataset, a separate document must be provided that details all such codes. If details of the field structure do not fit in the header, a separate document must be provided that includes such details. Data must be formatted in its native format (i.e. dates in a date format, numbers in an appropriate numerical format, and numbers with leading zeros as text).

2. Delimited text file with header information detailing the field structure. The preferred delimiter is a vertical bar "|". If any special codes exist in the dataset, a separate document must be provided that details all such codes. If details of the field structure do not fit in the header, a separate document must be provided that includes such details.

The metadata must include, at a minimum, the following fields in separate columns:

1) Account Number:	Caller's telephone account number
2) Originating Number:	Caller's phone number
3) Terminating Number:	Called party's phone number
4) Connection Date:	Date of call
5) Connection Time:	Start time of call
6) End Time:	End time of call
7) Elapsed Time:	Duration in minutes of the call

Each field of data must be loaded into a separate column. For example, Connection Date and Connection Time must be produced in separate columns and not combined into a single column containing both pieces of information. Any fields of data that are provided in addition to those listed here must also be loaded into separate columns.

VI. Email Native File Production

When approved, Outlook (.PST) and Lotus Notes (.NSF) email files may be produced in native file format. A separate folder should be provided for each custodian.