

By Conrad J. Jacoby,*
Kim Araneo,*
and Mel Goldenberg*

Finding the Right Format of Production for

Electronic Information

Under the amended Federal Rules of Civil Procedure that took effect December 1, 2006, litigants must discuss specifically how they would like to receive electronically stored information—e-discovery materials—that will be exchanged over the course of the lawsuit. In the past, such discussions about the “format of production” tended to be afterthoughts delegated to a litigation support professional or the least experienced member of the legal team. This is a topic that requires active involvement by the top members of a legal team, some of whom may never have had a reason to get involved with what seems to be a minor logistical detail.

Competing issues arise in the context of picking a production format for digital discovery materials in a case. While several formats are used today, each has benefits—and shortcomings—that may make it suitable for some but not other electronically stored information (ESI) productions. Put another way, no one single “best” production format works in all cases, and it is up to the legal team to make a judgment call based on their understanding of the needs of their case.

Today TIFF productions have expanded to include a database load file that contains searchable text that has been extracted from the document at the time of TIFF conversion.

With today's technology, three basic formats are used for producing electronic documents, plus a few minor variations on each. First, of course, people simply print out electronic documents and exchange paper. While paper is universally criticized as old fashioned and fundamentally inadequate, consider the following:

- Paper can be read without any additional technology investment.
- Paper production can easily be divided among multiple reviewers.
- Paper productions can be redacted and bates-numbered with a minimum of technology investment.

Balanced against those considerations, consider the following shortcomings of print production. First, printing electronic documents can create a voluminous physical record that is impossible to review. A single Excel spreadsheet, for example, may print out as over 10,000 pages. Second, printing electronic documents hides some information—the legendary “meta-data”—that could be important in helping authenticate a document or convey information about what the author was thinking. Finally, paper documents are not searchable—at least, not very easily. Every lawyer has worked with a box of discovery documents where key documents were identified by post-it notes and color-coded tape flags. That approach may work for very small document collections where a lawyer can keep a general sense of the entire discovery request in short-term memory, but it breaks down for any collection that exceeds even

half a box of material. This is perhaps the most important reason why judges increasingly disfavor paper productions unless they are made pursuant to a specific request by the party seeking discovery.

A second format for producing electronic materials is through digital images, usually in Group IV TIFF format. A TIFF image is nothing more than a digital photograph of the document. It's much like printing a document to paper, except that 14,000 images can be stored on a single CD, dramatically shrinking the amount of physical storage space required to store a document collection and greatly reducing the cost of duplicating large portions of the collection when required. TIFF images can be used in just about every litigation support tool on the market; they've been an industry standard for over a decade. It's a very popular production format, and even in an era where “native format” productions are receiving significant publicity, many litigants still exchange discovery documents in TIFF image.

TIFF images didn't become so popular just because they could be squeezed onto CD-ROMs and computer hard drives. TIFF productions commonly include log files that break these images into discrete documents, making it much faster and easier to work with discovery documents. Instead of flipping through all 10,000 pages of a voluminous Excel spreadsheet—to find where it ends (if nothing else), that entire document in TIFF form can be classified and moved about as a single entity. In addition, storing that one document won't take

up a quarter of an associate's office. TIFF images can also easily be redacted and electronically numbered using relatively inexpensive tools. Several studies and numerous document review teams have found it is generally much faster to review TIFF images than an identical hard-copy production, particularly if the review team is using flat-panel monitors that don't create as much eyestrain as older CRT displays.

As recently as only a few years ago, many TIFF productions were merely just that—TIFF images with a data file providing document breaks. That was considered a reasonable production format at the time, and this format is still used as a production format in some cases today. However, TIFF images in and of themselves have the same problem as paper documents—they cannot easily be searched. To do that, the recipient of the TIFF images must spend money to have the documents run through an OCR process to create searchable text. For documents that started out in searchable electronic form, this is an extremely inefficient way of extracting information, especially since the process that creates the TIFF images can also harvest the full text of those documents—with no OCR errors—at the same time.

Today TIFF productions have expanded to include a database load file that contains searchable text that has been extracted from the document at the time of TIFF conversion. Extracting the raw text gives you exactly what was typed into the document—there are no OCR transcription errors. In addition, a few courts have found that

some amount of extracted text or objective data should be included as part of a TIFF production in order to make the production format “reasonably useful” under the Rules of Civil Procedure. As one final caveat, though, native text extraction preserves all the spelling errors of the original writer. Materials like Instant Messaging (“IM”) logs that contain many abbreviations and misspellings can still be difficult to search, even if their text is perfectly extracted.

While a combination of TIFF image and extracted text is the most common production format at the moment, continuing limitations—and developments—in technology highlight a number of shortcomings that make this production methodology unsuitable as a universal production format for all cases. First, TIFF conversion doesn’t necessarily guarantee the document will show up exactly as it might have looked when it was printed out on a specific computer. Usually, the substance of the document is more important than its appearance, but disputes can center on how prominently specific contract terms or disclaimers were displayed in a particular document. Second, current text extraction technology does not necessarily pull out every single shred of potentially searchable information from a document. This is done as a matter of policy, since much document metadata has little if any value. Service bureaus and extraction software make educated decisions about which fields are likely to contain potentially interesting information and extract only those, inevitably leaving behind other metadata. Most of the time, this

level of text extraction is sufficient to meet a party’s discovery needs, but it is possible that a particularly esoteric piece of metadata might be required to help make a point about a specific document. If that metadata wasn’t extracted at the time of initial processing, it may not be easily available.

A final concern about producing electronic information in TIFF format is the cost of converting into TIFF image and extracting their searchable text. For voluminous amounts of ESI, it can be expensive to process the collection—money the producing party has to spend just to determine that many of the electronic documents are irrelevant and will not have to be produced in discovery. At some level, this simply feels inefficient.

That line of reasoning has been one driving factor for the rapidly increasing popularity of the third format—the so-called “native file” production format. The idea behind native file production is remarkably simple: Why pay to process voluminous electronic information that’s already searchable? Why not simply review these electronic files in their existing format to find the files that are actually responsive? After completing review, only the responsive files would be turned over, still in the same format that they were originally received. The receiving side then has the option (and cost) of processing these materials into whatever they want—TIFF images, paper, or other appropriate format. On a theoretical level, native file production saves the producing party a significant amount of money and permits the requesting party

access to exactly the same information that the producing party has. Imagine a litigation environment with no more disputes over incomplete production of information!

Unfortunately, nothing is ever simple in litigation. First, it has been difficult to develop technology permitting a legal team to search, review, and categorize large amounts of disparate file types and documents. While amazingly sophisticated search tools have long been available—think of Google and Yahoo! and old-time search engines like Altavista—output from these search engines simply did not fit into the work flow of a litigation document review. You couldn’t tag batches of documents. You couldn’t add notes describing why a given document is important to the case. Tools that integrate solid search with those kinds of review functions had to be built from the ground up—something that continues to this day.

Second, working with electronic files in their native format may actually hinder typical document review efficiency. Most importantly, current technology does not permit a reviewer to redact a native file. One can easily remove or alter text in a native document, but taking these actions changes enough key information used to authenticate ESI that normal automated protocols cannot easily validate the edited document against the original version stored in the ordinary course of business. A second limitation to native file review is that current technology does not permit Bates numbers to be attached to pages of native files. Page-level control numbers offer an easy way to identify

a few courts have found that some amount of extracted text or objective data should be included as part of a TIFF production in order to make the production format “reasonably useful” under the Rules of Civil Procedure.

For law firms that do not have the internal infrastructure to hold all the discovery data—or the paraprofessional specialists to maintain large databases—a key question is whether a legal team has a sufficient litigation budget to hire a third-party hosting service to store the data.

specific passages in key documents, but, as with redaction, adding a running number to native file documents changes document metadata, making it difficult to authenticate the file for admission into evidence. As a consequence, using native files as exhibits can be awkward and inconvenient. Instead of quickly referencing a specific page within the production, the examining attorney may need to reference “screen X of document entitled ‘working notes in preparation for meeting,’ as found on John Smith’s computer on November 8, 2006.”

Between the separate shortcomings of native file and TIFF image plus extracted text productions, decisions about production formats depend greatly on the needs of a specific litigation matter. First, consider the technology available to the legal team for working with incoming electronic discovery materials. Paper is unlikely to be a viable production format, unless discovery is limited to a very small number of documents and e-document metadata will never be of any use in the case. That situation is ever less likely to occur, so legal teams will mostly likely choose between some variation of TIFF production and native file production.

Most legal teams already have ready access to software tools that will work with TIFF images and searchable database text. If not, basic litigation software is a fairly modest expense. However, software may not be the problem. TIFF images can take up a lot of digital storage space. Does a law firm have the empty hard disk space on its computer network to store five million

TIFF images? Ten million TIFF images? Once considered exotic, these document collections are increasingly commonplace in the world of e-discovery.

For law firms that do not have the internal infrastructure to hold all the discovery data—or the paraprofessional specialists to maintain large databases—a key question is whether a legal team has a sufficient litigation budget to hire a third-party hosting service to store the data. External hosting, also known as “online repositories” or “ASPs,” offers outsourced expertise and virtually unlimited storage capacity, albeit for a sometimes steep price. ASPs charge monthly fees for data storage and user access. For litigation lasting a year or less, online repositories may be cheaper than investing in new storage capacity but relatively few cases settle that quickly. Over time, the recurring costs of online repositories can put a significant dent in a litigation budget. On the plus side, outsourcing discovery document hosting also purchases dedicated project and document management expertise, which may not be available from an overbooked internal litigation support staff

Given the cost and storage issues inherent in generating and working exclusively with TIFF images, are native files a better solution? After all, producing these files in their original form avoids substantial processing costs. In addition, multiple tools can index and search native file collections. Shouldn’t this be the less expensive and more efficient option?

Unfortunately, working with native files may require investing in entirely new infrastructure or ASP hosting. The two most common litigation support software systems in use today—Concordance and Summation—aren’t as adept at working with native files as they are with TIFF images and extracted text. Indeed, they can’t deal at all with certain types of native electronic data, such as mainframe computer files, complicated databases, and other data files increasingly exchanged in discovery. Using existing litigation support tools to work with native files may end up working smoothly only after the production has been converted from native format into TIFF image and extracted. That can quickly eliminate any theoretical savings from working with native files.

Second, a receiving party may not care about its ability to redact sensitive information—that’s a problem for the party that produced the native files—but it certainly does care about using these documents as substantive evidence. It can be a tedious process to authenticate native files, and a law firm may need to bring in an e-discovery expert for that purpose. Dealing with the logistical issues of authenticating ESI can be a powerful distraction for a legal team and one that reduces the amount of time and energy available for substantive legal analysis and case preparation.

Given the many competing priorities in litigation, no single production format consistently stands head and shoulders above others at this time. Litigation-specific analysis will continue to point one way in some cases and another at other times. That said, legal teams will usually find their choice of production format becomes clear after they have answered the following questions:

- What tools does the team already have for working with electronically stored information?
- What internal expertise and staff does the team already have for working with electronically stored information?
- What budget does the team have for working with electronically stored information?
- Does the team have a protocol in place for authenticating electronic materials exchanged in discovery?

In addition to helping legal teams understand their priorities for a specific case, these questions also serve as a consistent analytical process that can be used in a broad range of legal matters to identify the best way with which to work with digital information. Over time, working through this analysis in a variety of matters will help attorneys and paraprofessionals develop a “gut feeling” about the most efficient ways to proceed with e-discovery. While such instincts must always be reviewed in light of developments in the law and in technology, they still provide a helpful foundation for working successfully with these materials.

*Conrad J. Jacoby writes and lectures extensively on e-discovery and litigation management. He received his B.A. from Yale University and his J.D. from Georgetown University Law Center.

*Kim Araneo has been working in the field of electronic discovery for nearly 15 years. She earned a B.S. from Louisiana Tech University and J.D. from Mississippi College of Law.

*Mel Goldenberg is president of TechLaw Solutions. He received his B.S. from Boston University. TechLaw Solutions is a pioneer in litigation support and information management.