Diving Deeper to Catch Bigger Fish

By: William P. Butterfield, Conor R. Crowley, Melinda R. Coolidge[1]

I. Introduction

The debate over whether (and which) metadata fields should be included in electronic document production has continued unabated for years. When producing parties grumble at producing documents in their native format out of fear of accidentally divulging secrets through hidden metadata fields, receiving parties often compromise by accepting a searchable image file with limited metadata fields instead.[2] However, ignoring the powerful functionality and information provided by metadata can seriously limit the value of the documents produced to the receiver.

Metadata is "information associated with and made part of an electronic document that is not visible in the normal viewing or printing of that document."[3] Often described as "data about data," metadata is "information that documents different aspects of a set of data, such as its attributes, deleted text, comments, field lengths, locations, associations, ownership, quality, etc."[4]

---

[1] William P. Butterfield is a partner in the Washington, D.C. office of Hausfeld, LLP. Conor R. Crowley is a partner in the McLean, VA office of Daley Crowley LLP. Melinda R. Coolidge is an associate in the Washington, D.C. office of Hausfeld, LLP.

[2] There are also conflicting views in the United States as to whether lawyers violate their ethical obligations when they search an adversary's metadata to uncover confidential information or other attributes about the document. *See, e.g.*, Norman C. Simon, *Electronic Discovery: The Great Metadata Debate*, 14 (The Metropolitan Corporate Counsel, May 2008), available at http://www.metrocorpcounsel.com/pdf/2008/May/14.pdf. This paper does not address ethical issues associated with metadata production and review.

[3] *See* Dennis Kennedy, *FRCP and Metadata: Avoiding the Lurking e-Discovery Disaster* (Workshare Whitepaper 2007) ("Metadata can include descriptive, historical, technical or other information and is often generated automatically. Classic examples of metadata are the information contained in 'document properties' of Microsoft Word documents and comments or tracked changes in Word documents.").

[4] *See* E-Discovery Updates: *Native Production is a Risky Business, Not a Cost Saving Measure or Required by the New E-Discovery Rules* (Feb. 18, 2008), http://ediscoverylaw.us; David Haynes, *The Five Purposes of Metadata* (CILIP Aug. 11, 2004) (citing G. Tozer, *Metadata Management for Information Control and Business Success.* Artech House, 1999) (Metadata is "the means by which the structure and behaviour of data is recorded, controlled, and published across an organization.")).

Metadata is not unique to electronic data; paper production also has forms of metadata not immediately available to reviewers (paper size, whether a paper was printed, typed or copied, paper type, highlighting color, etc.)[5]

There are several categories of metadata. System metadata is automatically added by the computer system, such as file creation date and time stamps, file location, and author information.[6] Application metadata is automatically added by the application being used, like word and character counts, total editing time, etc.[7] Document metadata is user-added, like comments, highlighting and track changes, which are usually the most revelatory if the three types of metadata.[8]

Native file format is the "actual format in which the document was created in the application in which it was originally created," (e.g., a Microsoft Word document would be produced in the .doc format).[9] Inclusion of metadata in native format files can exponentially increase the efficiency and efficacy of a document review; however, the utility of certain metadata is often not apparent when the parties agree on the parameters of the production. By agreeing to non-native format documents with limited metadata, the receiving party may be multiplying the amount of time it will take (or potentially eliminating the ability) to pinpoint the cast of characters, key events, timeframe, location, and motives involved in the lawsuit, which could be discovered through electronic manipulation of metadata.

Because conversion from native format into an image or pdf can limit the number or utility of metadata fields, receiving parties will receive the most useful productions by insisting on native format. Such a production will increase the reviewer's flexibility, their variety of search tools, and permit improved information management, allowing the reviewer to more quickly and thoroughly unlock the secrets found just below the surface of (typically) the millions of documents produced.

---

[5] *See* Kennedy, *supra* n.3.

[6] *See id.*; W. Lawrence Westcott II, *The Increasing Importance of Metadata in Electronic Discovery*, 14 Rich. J.L. & Tech. 10, http://law.richmond.edu/jolt/v14i3/article10.pdf.

[7] *See* Kennedy, *supra* n.3.

[8] *See id.*

[9] *See id.*

II. Format Options

When negotiating a document production involving electronically stored information, parties typically choose one of four production format options: (1) paper only, (2) image files without metadata, or (3) image files with metadata, and (4) electronic native format production.

The question of the format in which documents will be produced is still subject to vast disagreement. In the United States, while many experts agree that Fed. R. Civ. P. 26(b) and 34(a) and (b) "eliminat[e] any doubt that metadata is subject to discovery,"[10] parties will still "routinely oppose[] plaintiffs' requests for metadata, calling them 'overly broad and burdensome requests' and 'completely irrelevant…. In some ways, it's a fishing expedition.'"[11] However, one e-discovery practitioner noted that "[t]here is a pattern that has developed in these cases that suggests that, when a party requests metadata early on in the e-discovery process, they're usually going to get it."[12] Another commentator recently observed that:

> … there is a growing tendency to request electronic documents in native file format, in no small part because of the availability of the metadata associated with these documents. There is a growing consensus that judges will be sympathetic to requests for production in native file formats and that the trend will be toward native file format becoming the default form of production under the new rules.[13]

Some producing parties that fear accidental production of privileged metadata may still prefer to print all of their electronically stored documents and do their production on paper. Clearly, all the uses of metadata will be lost through this type of production. Furthermore, when documents are converted from electronic to paper format, they can lose their authenticity.[14] While the receiving party can scan and OCR paper documents to make them keyword-searchable, taking this extra step increases costs, while converting the files from electronic format to paper, and then back to electronic format, is an inefficient and unnecessary litigation

---

[10] *See id*.

[11] *See* Tresa Baldas, *Metadata Grows in Legal Import*, (National Law Journal, Jan. 26, 2009) (quoting Melissa Geist, Reed Smith's director of complex litigation electronic discovery from the Princeton, N.J. office).

[12] *See id.* (quoting Wayne Matus, co-national head of Pillsbury Winthrop Shaw Pittman's electronic discovery practices).

[13] *See* Kennedy, *supra* n.3.

[14] *See* Philip J. Favro, *A New Frontier in Electronic Discovery: Preserving and Obtaining Metadata*, B.U. J. Sci. & Tech. L. (Winter 2007).

practice.[15] Productions of non-searchable image files (essentially a scanned copy of the underlying printed document[16]) can contain the same limitations as paper productions. Furthermore, even when keyword-searching (often vastly more efficient than reviewing paper documents) is possible, its utility has its limits as well, as discussed in section III, *infra*.

While simply producing a searchable image file in the first place can be efficient for both parties (note that the producing party will still have to convert every file from its native format to an image file), it does not maximize the utility of an electronic production, as would a native format file with metadata. As discussed *infra*, while keyword searching is a powerful tool in document review, its value is limited, and should be only one tool in a reviewer's arsenal. Furthermore, some courts have held that by converting electronic files to image files and removing the metadata, the defendant had impermissibly altered the records and "essentially created new documents."[17]

Receiving images with metadata increases the efficiency and efficacy of a document review. With both images and pdfs, metadata can be captured during the rendering process and attached in an accompanying file.[18] However, any metadata fields not specifically requested will not be produced, and sometimes reviewing metadata produced in a separate file can be more challenging than reviewing metadata never removed from a native format document. Missing metadata fields will again limit the reviewer's ability to fully capitalize on data stored naturally in a document.[19]

Native-format documents can be the easiest to produce, and will almost certainly be the most useful for reviewers. Reviewing native-format documents is cost-effective, allowing for control of large volumes of data and ensuring document integrity.[20] As a result, American courts have begun ordering production of native-format documents.[21] Native-format documents arrive

---

[15] *See* Favro, *supra* n.14 (In the electronic age, it is not a question of whether electronic data should be produced, but what form the production should take.)

[16] *See* Kennedy, *supra* n.3.

[17] *See, e.g., Hagenbuch v. 3B6 Sistemi Elettronici Industriali S.R.I.*, No. 04 C 3109, 2006 WL 665005, *2 (N.D. Ill. Mar. 8, 2006) (as discussed in Favro, *supra* n.14).

[18] *See* Favro, *supra* n.14.

[19] *See id.*

[20] *See id.*

[21] *See* Favro, *supra* n.14; Kennedy, *supra* n.3.

authenticated, eliminating the need for extra steps to meet basic evidentiary standards.[22] Furthermore, they arrive with their original metadata fields intact, allowing electronic manipulation of data to maximize the efficacy of sorting and searching.[23] Producing parties can eliminate protected or privileged metadata, while keeping other metadata intact.[24] The next two sections describe the common and enhanced uses of native-format metadata in document review.

III. Common Uses of Metadata

Inclusion of basic metadata fields allows the reviewer to use other tools, on top of keyword-searching, to organize information received in an electronic document production. Keyword-searching is an important tool in the arsenal of a document reviewer; keywords can permit reviewers to search and sort documents quickly – as long as the relevant documents contain the right keyword. However, savvy defendants may have invented their own lexicon or used code words to avoid detection by harried document reviewers. Furthermore, reviewers can be hampered when keyword searches end up being over-inclusive (turning up documents not relevant to the search), or under-inclusive (leaving out relevant documents that should be located).[25] Thus, both the precision and recall of searches for relevant information can be adversely affected when searches rely on keywords. In *United States v. O'Keefe*, Judge Facciola of the D.C. District Court stated: "for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than [other] search terms… is truly to go where angels fear to tread."[26]

Because keyword-searching does not produce perfect results, adding another tool to the reviewer's arsenal – searching and sorting using metadata – can reveal information and relationships that are not apparent from pure keyword-searching. Metadata fields allow reviewers to immediately sort documents by relevance, and pinpoint key actors, relevant time

---

[22] *See* Favro, *supra* n.14 ("Because the inherent traits of metadata may enable a party to establish such things as when a document was created, the identity of the person who prepared the document, the purpose for doing so, and where the document was subsequently maintained, a party can more easily meet the authentication requirements under the Federal Rules of Evidence and applicable state law.")

[23] Indeed, some U.S. courts have ordered production of metadata. *See Aguilar v. U.S. Immigration and Customs Enforcement Div.*, 255 F.R.D. 350 (S.D.N.Y. 2008); *In re Seroquel Prods. Liab. Litig.*, 2007 WL 219989 (M.D. Fla. Jan. 26, 2007); *see also Williams v. Sprint/United Management Co.*, 230 F.R.D. 640 (D.Kan. 2005).

[24] *See id.*

[25] *See The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery* (August 2007 Public Comment Version) (citing *Alexander v. FBI*, 194 F.R.D. 316 (D.D.C. 2000); *Quinby v. WestLB, AG*, 2006 WL 2597900 (S.D.N.Y. Sept. 5, 2006)).

[26] *U.S. v. O'Keefe*, 537 F.Supp.2d 14, 24 (D.D.C. 2008).

frames, and major events. For instance, metadata fields such as the dates documents were created and edited, and the authors of those documents, allow for quick searching and sorting to create a chronology or a cast of characters.[27] When the ultimate goal is winnowing a large set of documents down to the most relevant, technologies that improve precision and recall provide a clear advantage over cruder keyword-based search techniques.

a. Metadata Enhances Searching Techniques

Metadata enhances retrieval performance by establishing a context for individual descriptors.[28] Search results become much more precise when the reviewer can search the author and subject fields separately from the body of the document, retrieving only documents that are from Smith (by searching only the author field), or only documents that are about Smith (by searching only the subject field).[29]

b. Automatic Authentication of Documents

Metadata can help authenticate documents and data by demonstrating that the "electronic document has been kept securely, is a complete record, and has not been tampered with."[30] In fact, the authentication provided in native-format document review can even undercover fraud.[31] In one case, a party alleged that a forwarded email could have been forged by the forwarder and therefore could not be authenticated.[32] Metadata analysis traced the unique identifier value of the Outlook user in the forwarded email message back to the original computer from which it

---

[27] *See, e.g., Boyapati, et al., v. Rockfeller Mgmt. Corp., et al.*, 2008 WL 2690051 (FCA), [2009] ALMD 1495, (Fed. Ct. Australia, 2 July 2008) (Court found that cross-claimant's testimony was contradicted by metadata identifying the author and dates of creation of certain electronic files).

[28] *See* Haynes, *supra* n. 4 ("Appropriate metadata tags around the different data elements allow search engines to seek information in a more discriminating way.")

[29] *See id.*

[30] *See id.*

[31] *See* Nolan M. Goldberg and Scott M. Cohen, *Turning Obscure Bits of Data into Hard Evidence: A Proposal for the Unorthodox Use of a Document Request to Capture System Metadata*, 8 (The Metropolitan Corporate Counsel, Sept. 2008) available at http://www.proskauer.com/news_publications/itn/data/003359/_res/id=sa_PDF/08.pdf ("It is often overlooked that applications, operating systems, and file systems also general number 'system' files to which the metadata may related, which, despite any concealment efforts, can reveal the existence of a file that has not been produced, or even that file's contents.").

[32] *See* Conrad J. Jacoby, *E-Discovery Metadata in the Real World*, available at http://my.advisor.com/doc/17368.

was forwarded, proving that the email had not been tampered with.[33] In another case, metadata analysis revealed that promissory notes had been fraudulently backdated.[34]

c. Information Management

In the same way that metadata can be used to sort documents based on creation and edit dates, and the authors and recipients of emails, metadata fields can be used to automate the management of information. Even the most basic fields, such as author or recipient, can facilitate the implementation of rules-based auto-classification software that appropriately, albeit crudely, tags electronic files with the appropriate confidentiality designations or retention periods.

IV. Enhanced Uses of Metadata

In addition to the standard uses of metadata, e-discovery vendors are now able to render information obtained from metadata into charts, graphs, or otherwise useful formats to help the reviewing party visualize everything in a case from who calls whom on Saturdays to the change in a supervisor's emotive tone over time.

a. Communication Analysis

By examining metadata fields in a variety of electronic communications (emails, instant messages, website posts, etc.), reviewers can sense the pulse of organizations and individuals. Reviewers can chart who is in regular communication, who the regular actors are, and who is always consulted prior to final decision making, from the basic metadata "to" and "from" fields of emails, or from the "author" or "comments" fields in Word documents. By examining regular patterns of communication, reviewers can spot whether decisions routinely made through electronic communication uncharacteristically start taking place at in-person meetings without written documentation.

b. Proximity Analysis

Reviewers can manipulate metadata fields to identify individuals' proximity to each other. From various metadata fields, one can discover how close certain people are to each other, whether they work on projects together, and whether they have a personal as well as professional relationship. Reviewers can look for different modes of communications (*e.g.*,

[33] *See id.*

[34] *Australian Securities & Investments Comm., v. West, et al.*, 2008 WL 2059662 (SASC), 100 SASR 496 (Sup. Ct. So. Australia, 9 May 2008).

multiple email accounts, instant messages, and phone calls) as well as communication during non-work hours. This information will help the reviewing party determine who they should depose or interview by identifying who would be most likely to know information about another employee and their work or habits.

c. Behavioral Analysis

Information about individuals' behavior can help develop the background of a case, or can be crucial to proving the basic elements of a case. From metadata, reviewers can determine which tasks a person has prioritized, and how long he spends on those tasks versus other tasks. Computer programs can measure a supervisor's emotive tone over time to determine if and when his opinions about a project seemed to change.

d. Workflow Analysis

By examining a variety of metadata fields, reviewers can assess the consistency of an organization's processes, such as hiring and due diligence, and pinpoint periods of significant variation from that consistency for further review. The reviewing party can get a feel for the organization's document creation and editing process, and determine whether a usual decision maker was avoided in certain instances.

e. Deletion Analysis

Metadata is a particularly useful tool to determine whether data has been tampered with. For instance, by examining time information on emails and documents, periods of no communication when there is usually communication, or conflicting time stamps, may indicate improperly withheld or altered information. E-discovery vendors can calculate whether data is missing because it was routinely deleted, or whether information was tampered with in anticipation of litigation. For example, in *Krumwiede v. Brighton Associates, L.L.C.*, metadata revealed that on or shortly after the date the plaintiff was ordered to return his laptop to the defendant, thousands of files were accessed, moved, or deleted.[35]

[35] *Krumwiede v. Brighton Associates, L.L.C.*, No. 05 C 3003, 2006 U.S. Dist. LEXIS 31669 (N.D. Ill. May 8, 2006) (as discussed in Westcott, *supra* n.6).

V. Complexity Associated with Metadata Production

a. Data Mapping

Native-format documents can have more than ten thousand associated metadata fields. When production is made in a format other than native, e.g., pdf, the image file must be accompanied by a load file that associates the extracted metadata fields with the image file. Although only a small subset of the available metadata fields is normally produced in non-native productions, the complexity of the task increases as the number of different native formats increases. When converting documents created by different applications, which may use different metadata fields or different names for the same metadata fields, to a common format, the conflicts among metadata field naming conventions must be resolved. The need to employ generic metadata fields that are common across various formats will inevitably result in less precise, and ultimately, less accurate metadata, than that available with native-format documents.

b. Challenges of Native File Format Document Production

Since native file production is the method by which most parties will receive all of the metadata, it is useful to conduct a brief review of the complexity associated with native production. While reviewing native-format documents can have many advantages, there are also several challenges associated with native-format production and review.

First, current technology has not made Bates numbering a native-format document particularly easy.[36] Consider that if the producing party were to open the document to add an electronic Bates number, the document's metadata would inadvertently be changed as well.[37] Similarly, when the reviewer opens the native-format document, she will also be changing that document's metadata (*e.g.*, date document last opened).[38] Reviewing parties can solve this latter problem by copying all of the documents produced, keeping one pristine version, and allowing reviewers to access the copied version.

Second, native-format documents can also present additional challenges to the producing party, because these documents cannot be effectively redacted. Until a tool exists that will make it possible to redact portions of a native file, some parties have designed a special protocol to address that problem. Under that protocol, the parties agree to exempt partially redacted

[36] *See* Mary Mack, *Native File Review: Simplifying Electronic Discovery?* (Legal Tech Newsletter, Dec. 1, 2005).

[37] *See id.*

[38] *See id.*

documents from a production that is otherwise in native format. The partially redacted documents are produced in paper or in a static image like TIFF or pdf.

Finally, native production sometimes adds complexity and expense to the producing party in conducting a privilege review. The "Track Changes" feature to Microsoft Word documents and similar functions in other applications sometimes reflect comments or edits made by attorneys, or reveal questions asked to attorneys. Thus, where a native document utilizes the Track Changes feature, its inclusion in a production set potentially requires lawyers to review all of the comments and edits to make sure they are not subject to attorney-client, work product or other privileges. This review can add significant additional expense to the producing party. The Track Changes feature also adds complexity, because there are potentially several versions of a document, each containing separate edits. Thus, parties must address this issue when addressing document authenticity.

These challenges aside, courts have, in some instances, overruled a producing party's objections and required production of native-format documents. In *In re Verisign Inc. Securities Litigation*, the magistrate found that production of image files was insufficient, and that the production "must include metadata as well as be searchable."[39] The court agreed, stating that it:

> understands that it may be difficult for Defendants to incorporate their redactions and [B]ates numbers into the .pst format, but it is not convinced that the responsive documents are so replete with privilege redactions that such a task would transcend all reasonableness.[40]

VI. Conclusion

This paper has only touched on the potential uses of metadata. Of critical importance is the fact that the reviewing party usually cannot know which metadata fields will be critical for its review until it has the documents in hand.

Producing parties can protect any privileged metadata while still permitting a reviewing party to gain innumerable insights into the case by having access to a variety of metadata fields in their electronic document review. Producing documents in their native format, where the maximum number of metadata fields is preserved, will result in the most efficient and efficacious document production and review for both parties.

---

[39] *In re Verisign Inc. Securities Litigation*, No. C 02-02270 JW, 2004 U.S. Dist. LEXIS 22467, at *7 (N.D. Cal. Mar. 10, 2004) (as discussed in Westcott, *supra* n.6).

[40] *Id.* at *14.