Bloomberg Law Reports® Technology Law

METADATA: WHY THE FUSS? A White Paper on Metadata

Arlen L Tanner, Shook, Hardy & Bacon LLP

Metadata is a term frequently used and often misunderstood. Some metadata can be useful in placing a document into proper context but most metadata is of little or no use in litigation. In certain circumstances, some metadata may reveal too much or be embarrassing. Understanding metadata is the first step in determining its usefulness or liability in any given situation.

Types of Metadata

The common definition of metadata is data about data.[1] It is sometimes hidden statistical information generated by a software program or operating system such as date and time of creation, author's name, date and time last modified.[2] The term metadata is often used in an over-generalized way to encompass application metadata, system or file metadata and embedded or hidden data.[3] From a purist standpoint, embedded or hidden data are data, not metadata, and may at best be referred to as pseudo metadata. Such information, intentionally placed in the document by the user, should more properly be considered data. True metadata is of two general types, either information placed in the document or file by the software application or information about the document placed in a separate file associated with the document.

An example of the latter would be a profile of a document kept in a document management system. The metadata would be the quickly searchable fields of information about the document that would aid in locating and categorizing the actual document. When a document is saved in it, many document management systems require certain database fields of information about the document, such as author, title, subject matter, revision, key words, etc. The information in this profile is metadata, information used by the document management system to describe or categorize the actual document, but the information is not part of the actual document. Electronic library catalogs use metadata to speed searches. Many websites also use metadata tags to lead search engines to their sites.[4]

An example of metadata created by the software application is the typical type of metadata discussed in litigation requests and will be discussed in the most detail. This type of metadata can be divided into file system metadata and application metadata, sometimes called OLE metadata. Application metadata is added by the software program that created the document and file system metadata is added by the operating system when the file is saved or moved. Some metadata exist simply to help the computer software deal with the file. For example, metadata in a compressed video helps the computer achieve a higher compression rate.[5]

Data that is created by the user that is hidden, in comments, track-changes, formulas, speakers' notes or is contained in another files or object that is embedded, are examples of pseudo metadata. This data was generated by the user rather than by the software application or operating system but is not readily apparent in the document. However, in litigation this type of pseudo metadata is consistently cited as an example of potentially critical metadata. In litigation, metadata seems to have been simplified to any data stored in electronic files that is not apparent to the user and might not appear when a file is printed.[6]

Pseudo Metadata—Hidden Data

There are many ways to place hidden or embedded data into a document. Using comments in a document created with Microsoft Word,® Excel® or PowerPoint®[7] is a perfect example. Comments enables the document creator or another user to place hidden commentary about the document without editing the document itself. In the native version a subsequent reader could activate and read the comments. If the document is printed to paper or an image format like TIFF, the comments would not be visible unless special measures are taken to reveal them. Comments can sometimes contain interesting insights on the document but could also be full of banter or rude comments.

Track changes reveal modifications and often show who suggested the change. Providing documents to others without removing the track changes could reveal strategies and weaknesses in bargaining positions. A normal business practice should include the use of scrubbing software, PDF or other formats to transmit documents.[8]

Other examples of hidden data include using the same color of font as the background and using symbol fonts. While using the same color font as the background may appear invisible, it is not invisible to text searches and can be readily extracted. A TIFF image or hardcopy will not always show the "invisible" text however, depending on the print settings. Changing a font to a symbol set may preclude a text search, but if it is a common symbol font, a reader could easily convert the font to a text font and read the "hidden message."

Another type of hidden data is the searchable text underlying an image. Searchable PDF files are a typical example of this. Redacting the image of a searchable PDF will not remove the hidden searchable text. Not understanding this simple fact has created some large blunders, especially when a "redacted" PDF is posted on the internet or public court filing system where anyone can extract the searchable text to reveal the actual redacted information.[9]

A user can embed data into other documents. For example, a complete spreadsheet, presentation or other document can be embedded into another document. When the document is printed to paper or image, there is a small icon indicating the presence of the embedded document but, unless the embedded document is opened and printed separately, its contents are not available. A two page Word document could have multiple other documents embedded in it. When those embedded documents are extracted, those two pages could expand to hundreds or more. Embedded files may also be multi-media files such as audio, video or graphics files. Graphics files may be embedded either as an icon or as the actual, visible graphics item. In the later event, the printed image will reflect the actual graphic. These types of embedded files, while not always easy to process, are generally ascertainable.[10] There are other methods of hiding information in unrelated files that are not easy to detect.

Steganography is hiding files or information in such a way that no one knows there are hidden files. There are several methods of hiding files or messages inside of other documents. For example, a word document, other pictures or even a video clip could all be hidden inside a picture file that appears quite innocent. Opening the picture file shows just the picture with no indication that other files may be embedded.[11] Fortunately, most civil litigations do not have to worry about this type of hidden data.[12]

File System Metadata

The best example of Windows' file system metadata is the information visible with a right-click of the mouse and by selecting "properties." Under the "General" tab is displayed the type of file, the location, the size, the size on disk, the date created at that location, the date last modified, the date last accessed and file attributes. The "Summary" tab will display a basic set of file system metadata. Selecting "Simple" shows fewer items and "Advanced" shows additional file system metadata items. Depending on the file type, settings and application versions, such things as author, revision number, date last printed and other statistics are shown. Some of these items, such as "modified" (a/k/a "date last modified"), are generally reliable. Other items may have been changed or may not exist due to settings. Items available under the "summary" tab will vary based on the file type. Microsoft Office files will generally have more "summary" information than most non-Office files.

One metadata item that is often misunderstood is the "created" date under the "General" properties tab. It may be the same or different than the "date created" item under "summary." The "created" date under the "General" tab is the date that file was stored on that storage device. It may be much later than when the file was originally created and may be later, earlier or the same as the "date last modified." When seeking the most reliable date for a file, the "date last modified" is best. It shows when that file itself was last saved rather than when the file was copied from one location to another, which is captured by the "created" date. What is somewhat misleading is that one could open a file and, without altering anything or even moving the cursor, hit "save" and the "modified" date would change. However, it accurately shows when the file was last saved, whether anything in the document was changed or not and is the most reliable of the date metadata fields.[13]

The "accessed" or "last accessed date" is also a very fickle item. Viewing a file will change the "accessed" date. Even looking at the properties of a file using the right-click, may update the "accessed" date. Copying a file using the Windows "drag and drop" method will change both the "accessed" date and the "create" date, but not the "last modified date." Tools exist for examining the metadata of files and for copying files in a way that preserves both the "create" and "accessed" dates of a file.[14]

Other file system metadata items include file type, file size, size on disk and location. File type is based on the file extension. If the file extension is changed, this "file type" metadata item will change and the computer may not be able to open the file.[15] The "location" shows the folder path where the document is located. This almost always is a useful piece of metadata. It shows where the file was kept in the normal and ordinary course. Location, however, is lost when the file is copied unless measures are taken to capture or preserve the location metadata.[16] "Size" is the file's actual byte size and "size on disk" shows how much space (clusters) is allocated to the file on the disk. Often the "size on disk" is larger because two files cannot share a cluster and clusters are fixed in size. A file size that seems larger than the number of pages in the document may be an indication of significant embedded files, media or graphics usage in the file.

Application Metadata

Different types of files may contain varying types of application metadata. Application metadata is dependant on the file type. PDF documents and JPG documents have different application metadata fields than Microsoft Office documents.

For example, PDF documents have much less application metadata than Microsoft Office documents. However, PDF documents do have application metadata. Viewing the "properties" of a PDF document will reveal a "PDF" tab in addition to the normal system metadata file tabs. The PDF tab will show, in addition to "title," "author," "subject," "keywords," date "created," and date "modified," the application that created the PDF. The information on the PDF tab may differ from the file system metadata on other tabs. For example, there may be an "author" listed on the PDF tab and not one for "author" on the file system tab.

JPG files contain application image metadata such as the pixel width and height, the horizontal and vertical resolution and bit depth. Some jpg photographs will also contain a number of additional metadata items about the camera and camera settings and the date the picture was taken.[17] Video and audio files may also have unique document metadata. A video file may list the pixel width and height as well as the duration, bit rate, audio sample size, audio format, frame rate and data rate. Audio files may list the bit rate, audio sample size, channels, audio sample rate, audio format as well as artist, album, tracks, title, genre, year, and comments.

The email fields, "to," "from," "date sent," and "subject line" are useful fields in litigation. The fields "CC" and "BCC" may also be useful. While they are commonly referred to simply as metadata, technically, in email from Outlook or Lotus Notes, these are email database fields. These, and sometimes other email fields, are commonly included as metadata produced in litigation.[18]

In litigation, most document metadata attention focuses on Microsoft Office documents.[19] In versions of Word prior to Word 2003, there was often metadata available showing the revision history[20] of a document,[21] including the "previous authors" or "previous document authors," often referred to as the "last ten authors" list. Word 2003 (version 11) and later versions do not store the previous authors or revision history metadata.[22] If a document, that was created in an earlier version of Word, is opened in Word 2003 or a later version and saved, the revision log is "corrupted," and the revision information is no longer available. Speakers' notes in PowerPoint presentation may provide understanding and context to the slides. When processing PowerPoint files for litigation, vendors can reveal the speakers' notes in several ways, including rendering to TIFF using the "print notes page" option, extracting the speakers' notice to a metadata field or creating "notes" image pages following the corresponding slide image page.

Depending on the scope of the litigation, track changes or comments may be useful. The last user of the Word or Excel document may have hidden track changes or may have accepted or rejected them and turned track changes off. If the track changes were just hidden, they exist as hidden metadata and can be revealed.[23] If accepted or rejected before turning track changes off and saving the documents, then the track changes should not become viewable metadata.[24] Comments in Word, Excel or PowerPoint may be visible or hidden, but in either case are available as metadata to anyone with the native copy of the file. A user may delete comments or hidden text and save the document. If not removed, the hidden text and comments are easily viewed.

Hidden track changes and hidden comments, together with other hidden data, present review challenges. If the review is in native, html or one of the other pseudo-native review forms, reviewers must check for hidden information. If the review is conducted with TIFF or PDF images, the review portion would go faster with fewer issues, if the production is also in image format. If the review and production are to be in image format, then the instructions to the vendor should define what is to be done with hidden and embedded data. Often, the direction is to expose hidden text and comments, including any track changes, in one of a variety of ways. There are other times when, due to the issues in the case, the document is imaged as it was saved by the user so if the user hid track changes or comments, but did not remove them, they would not be visible on the TIFF or PDF image. Counsel must understand the issues in the litigation when determining the proper treatment of hidden text and comments.

An important review consideration is the possible existence of hidden data in the extracted text when the image does not reflect the hidden data. In such situations, hidden track changes or comments may exist in the extracted text but not on the visible image. If reviewers, in such cases, look only at the images, additional information may be produced without review, some of it may be privileged. Hidden metadata in extracted text is especially a concern when privileged information is in comments or track changes that are not reviewed because they are not visible on the face of the document. Each case should have, as part of the case management or protective order, some language that inadvertent disclosure or production of documents or metadata may not be considered a waiver. The language should include a "clawback" option.[25]

Metadata can often be much ado about very little or it can become a major issue. The best practice in dealing with metadata is to discuss metadata at the beginning of the case (in the Rule 26(f) meetings in federal cases) and agree what should be done with metadata. The issues to discuss include the format of production[26] and the metadata fields to be provided. In certain cases there should be an agreement on how certain documents will be processed to account for existing comments, track changes, PowerPoint speakers' notes and spreadsheet formulas. Different matters may require a variation, but the following table sets forth the basic metadata fields used in litigation.[27]

List of Common Useful Metadata Fields Used in Production

BegDocID	The beginning Bates number of a document. This is not extracted as a true metadata item, but is provided on production to identify a document and to unitize the pages of a document.
EndDocID	The ending Bates number. This is the number of the last page of a document. This is not extracted as a true metadata item, but is provided on production to unitize the pages of a document.
AttachBegID	The Bates number of the first page of an attachment to an email. This is not extracted as a true metadata item, but is provided on production to unitize the pages of a document.
AttachEndID	The Bates number of the last page of an attachment to an email. This is not extracted as a true metadata item, but is provided on production to unitize the pages of a document.
Source	This indicates the custodian or department who had custody of the document prior to collection. This is not generally a metadata item extracted from the documents, but is provided to show custody and control.
DocDate	For email the sent date, for other electronic documents, the date last modified.
To	The "to" recipient of an email

From	The sender of an email
CC	The carbon copy email recipient
BCC	The blind carbon copy email recipient
Subject	The email subject line
Filename	The file name of an electronic document (email from an email folder or inbox would not have a file name). This field is for individual stored files, often called "loose files."
Extension	The file extension of an electronic document, e.g., doc, xls, pdf, etc.
Path	The file path (folder and subfolders) where the electronic documents were stored prior to being collected for litigation.
Title	If available, the optional title used in electronic documents
DocType	Email or electronic documents. Optionally, the DocType field can indicate if a document is an email attachment.
Application	The application that created the electronic document. E.g., Microsoft Word.
Author	If available, the author field in electronic documents.
Text	The searchable text, extracted from searchable documents. This may be provided as separate files or as a field with the other metadata.

Under certain circumstances other metadata fields may be requested. Examples of other metadata may include the create date, the internal MS Office date created, formulas extracted from Excel, email importance, date last accessed, file size, comments, track change authors and dates, speakers' notes and versions. Much of the other metadata is of little use in most litigations.

Metadata in Correspondence (Mining)

An unsettled area in legal ethics is whether counsel may ethically "mine" for metadata from correspondence between counsel. Clearly, metadata mining is appropriate when dealing with information received in native format in discovery, but when correspondence is between counsel, and the sending party leaves metadata in the document, there are different opinions as to whether the receiving party can "mine" that metadata. For example, if track changes or comments were not permanently removed from the document, but were not visible when sent, some jurisdictions would permit counsel to reveal them and other would not. New York, Arizona, Florida, Alabama, Maine, New Hampshire and West Virginia say it is not ethical to mine metadata from correspondence with opposing counsel. The ABA's model opinion, Colorado, District of Columbia, Pennsylvania, Maryland, Vermont and Minnesota say it is acceptable to do so. The District of Columbia makes it a little more complicated by adding, unless the recipient "knows" the metadata was inadvertently sent. Pennsylvania tempers their opinion by saying if the recipient "concluded" it is inadvertent, the recipient must notify the sender. Vermont adds that you are to "disclose" receipt of metadata. Minnesota states that their Rule of Ethics 4.4(b) applies and therefore the sender must be notified.

The best rule of thumb when corresponding with an opposing party or counsel is to send correspondence in PDF form or use a metadata "scrubber" application on the document being sent. When using track changes between opposing sides so a document cannot be scrubbed or produced in PDF, the sender must take care to understand what information is being provided. For example, track changes will indicate who made the changes and when

they were made. There may be those providing edits on one side that should not be revealed to the other side. In such situations, one possible solution may be to use a copy of the document for all contributors on one side to use and then have one person enter the changes into the document to be shared with the other side.

Conclusion

Metadata is a focus for lawyers, courts and vendors of e-discovery software. When data is collected, forensically-sound methods must be used to avoid risk of metadata spoliation. Some metadata is very useful and in certain cases critical. Most metadata, however, is of no real use in litigation. Discussions of reasonable production of metadata should occur early in litigation to prevent an unnecessary dispute later. It is prudent to know what is revealed by the associated metadata when sending electronic documents to others or posting documents on the internet.

Arlen L Tanner is Of Counsel with Shook, Hardy & Bacon, LLP. The views expressed herein are those of the author and do not necessarily represent those of Shook, Hardy & Bacon, LLP or its clients.

---

[1] "Meta" is from the Greek "μετὰ," meaning transcending—as in above or beyond. http://www.answers.com/topic/meta-2. As a prefix it means "about." *Id.* It conveys a concept which is an abstraction from another concept or used to complete or add to the latter. *Id.; See also* http://en.wiktionary.org/wiki/meta- (last visited July 7, 2010).

[2] *Williams v. Sprint/United Mgmt. Comp.*, 230 F.R.D. 640, 646 (D. Kan. 2005).

[3] *Aguilar v. Immigration and Customs Enforcement*, 255 F.R.D. 350, 354 (S.D.N.Y. 2008). Substantive, also called application metadata, is created by the application and reflects substantive changes made by the user. This would include comments, fonts, spacing or prior edits. System metadata is defined as things like author, date and time of creation or modification. Embedded metadata is defined by the *Aguilar* court as information such as formulas, hidden columns, internally linked contents and database information. *Id.* Application metadata is sometimes referred to as document metadata.

[4] Putting "hidden" metadata key words in the website to lead search engines to the site or to make the site appear higher on search engine's "ranking," has generated controversy and litigation, especially when using a competitor's trademark as a metadata tag on the website. *See, e.g.*, *North Am. Medical Corp. v. Axiom Worldwide, Inc.*, 522 F.3d 1211 (11th Cir. 2008).

[5] J.R. Hidalgo, *On the use of indexing metadata to improve the efficiency of video compression*, Circuits and Systems for Video Technology, IEEE Transactions, Vol. 16, Issue 3, March 2006, pp. 410-19.

[6] For example, the Court in *Williams* determined that formulas in spreadsheets were metadata. *Williams* at 647.

[7] These registered trademarks will be used without the ® indication at all other points in this paper.

[8] Scrubbing software must not be used during discovery on documents subject to legal hold obligations. *See Williams* at 644.

[9] Large portions of the transcript of the Facebook and ConnectU settlement were redacted and a PDF of the transcript was made available. Despite preventative measures to keep the settlement confidential, the Associated Press easily read the hidden searchable text to reveal that Facebook's internal valuation of the company was $3.7 billion, $8.88 per share. This was far less than the $15 billion valuation established by the Microsoft investment in 2007. *See* http://www.techcrunch.com/2009/02/11/the-ap-reveals-details-of-facebookconnectu-settlement-with-best-hack-ever/ (last visited July 7, 2010).

[10] In processing and review, the little graphics, often .jpg images, that individuals or companies use for email backgrounds or signature blocks, are rendered as separate attachments. This increases the attachment and page counts with items that weren't intended to be attachments in the normal usage of the term.

[11] *See* http://www.garykessler.net/library/steganography.html (last visited July 7, 2010). Cryptography, the encrypting of files to restrict or prevent access, has the flaw that the file is generally still visible and encrypted so decryption methods may be applied. Hiding an encrypted file within an innocent graphic or media file, provides a very solid method of securing information. There were allegations that terrorists used pictures posted on the internet, containing hidden files, to pass information to cell members. Sensitive information may be stolen from a company using steganograpy.

[12] Digital watermarking is much more likely to be important in a trial. It is a method of hiding information in a file so that it can be proven later to be the intellectual property of the one who watermarked the document.

[13] Date metadata fields can help clarify the history of a document. At times, however, the date metadata fields can create confusion. For example, using existing documents as templates may show a create date far in advance of when the actual document actually came into being. The date last modified would show when it was last saved but would not help with the question of when the document first came into existence.

[14] Robocopy, a free application from Microsoft, has a proven method of preserving this date metadata. WinRar and recent versions of WinZip have options for preserving this date metadata. Earlier versions of WinZip did not preserve metadata. Many commercial collection tools exist, such as SafeCopy 2 from PinPoint Laboratories, that also preserve the date metadata and the path information.

[15] Processing applications that examine the header of a file would still recognize the application that created the file.

[16] Robocopy, SafeCopy 2 and other collection tools can be set to capture the location (path) information.

[17] The date the picture was taken is based on the date setting of the camera and may be different than the file system date last modified. Other photograph metadata could include the make and model of camera, the shutter speed, the lens aperture, flash mode, focal length, and exposure time.

[18] In addition to the basic e-mail metadata fields that both applications share, Lotus Notes and Outlook have fields the other does not.

[19] Common Office metadata fields may include: Last Saved By, Word Count, Page Count, Paragraph Count, Line Count, Character Count, Chars, Byte Count, Presentation Format, Slide Count, Note Count, Hidden Slides, Multimedia Clips, Last 10 Authors, Routing Slip, Track Changes, Fast Saves, Hidden text, Graphics Hyperlinks, Document Variables, Include Fields, File Name, Title, Author, Comments, App Name, Version Date, Created Date, Last Printed, Date Last Saved, Total Edit Time, Template, Shared, Subject, Category, Company, Keywords, and Manager.

[20] British Prime Minister Tony Blair's office posted a report to its website, as a Word document, that analyzed Iraq's weapons of mass destruction. Contrary to the government's assertions, metadata revealed that it had been drafted by civilians. Additional focus on this document revealed the civilians had plagiarized portions from a thesis written more than ten years previously. Richard M. Smith, *Microsoft Word Bytes Tony Blair in the Butt*, http://www.computerbytesman.com/privacy/blair.htm (June 30, 2003) (last visited July 7, 2010).

[21] Prior versions of documents could be recovered and, with "allow fast saves" activated, other data that a user thought was deleted could be discovered. "Fast Save" was discontinued with Service Pack 3 of Office 2003.

[22] It is interesting that Microsoft started to eliminate revision history metadata after it posted its 1999 Annual Report on its website in a Word document. Revision metadata in the document revealed that a portion had been prepared on a Macintosh. *See* Scott Rosenburg, *Microsoft's Annual Report: Made on Macintosh*, Salon.com at http://www.salon.com/technology/log/1999/10/12/microsoft_report (Oct. 12, 1999) (last visited July 7, 2010).

[23] For examples of real life hidden track change examples, *see* http://www.shaunakelly.com/word/trackchanges/PublicExamplesOfTrackChanges.html (last visited July 7, 2010).

[24] Unless an older version of Word was used with "fast save" or revisions history activated. The "Remove Hidden Data" add-in is a tool that removes many hidden data items. However, there are known issues with this tool. http://support.microsoft.com/default.aspx?scid=kb;en-us;834636 (last visited July 7, 2010).

[25] In federal courts, the inadvertent disclosure provision should conform with the requirements for Federal Rule of Evidence 502. Rule 502, if its provisions are followed, will help protect from waiver, especially in other jurisdictions and in other matters pending in state courts.

[26] The format of production impacts metadata. If native documents are provided, then all metadata is available. If images of documents are provided, a production format with many advantages, then the parties should agree on which useful metadata fields should be provided, together with the searchable text extracted from the documents.

[27] Because metadata aids in establishing how an electronic document was kept in the normal and ordinary course of business, care must be taken to ensure that collection, searching/filtering or processing methods neither change the metadata or create additional metadata. For example, an application such as Clearwell, which helps filter potentially responsive ESI from non-responsive ESI, may, if not properly configured, substitute the Clearwell document ID number as the "file name" for an email that was originally located in the custodian's inbox. Such an email did not originally exist as a "loose file" and would not have a file name. It certainly did not have the Clearwell number as it was kept in the normal and ordinary course of business.