Metadata, Metadata, it's Everywhere!

By Peter Coons, Senior Vice President, Digital Forensics Expert, D4, LLC

Metadata, Metadata, it's Everywhere!

By Peter Coons

Metadata - That term is used a lot in the eDiscovery world and many people use it but are unsure of its meaning. So what is metadata?

Let's look to a reliable eDiscovery resource for an answer. The Electronic Discovery Reference Model glossary[1] is a compilation of definitions on various eDiscovery related terms. Here is one definition for metadata that I think is clear and descriptive:

> "Metadata is information about a particular data set which may describe, for example, how, when, and by whom it was received, created, accessed, and/or modified and how it is formatted. Some metadata, such as file dates and sizes, can easily be seen by users; other metadata can be hidden or embedded and unavailable to computer users who are not technically adept. Metadata is generally not reproduced in full form when a document is printed. (Typically referred to by the less informative shorthand phrase "data about data," it describes the content, quality, condition, history, and other characteristics of the data.)"

Now that we have the definition of metadata we can delve deeper in to how it is created and stored.

All Metadata is Not Created Equal

When attorneys and those in the litigation support field discuss preservation of metadata they are usually referring to the dates and times or MAC information (Modified, Accessed, Create dates). Of course, we want to preserve important information like Subject, To, From, Author but for the most part those metadata elements, which are commonly referred to as application level metadata, can be less volatile than file system level metadata.

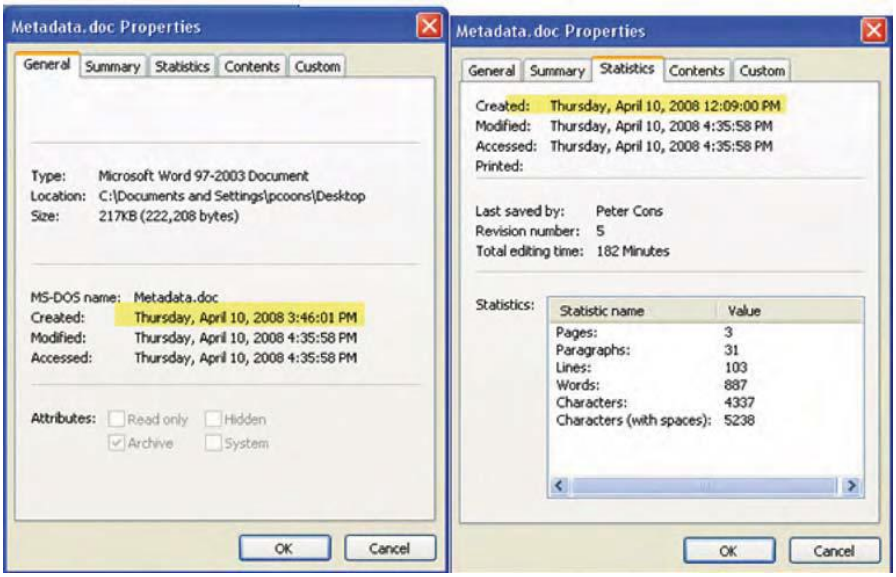
What is application and file system metadata you ask? If you look at a file in Windows[2] explorer and view its properties you may see something like this:

In the example above the Date Modified, Date Created and Date Accessed are all file system level metadata. The Author metadata field is actually an application level metadata element. File system level metadata may contain many elements including file security settings, read/write attributes, ownership information, and, of course, dates and times. These elements can be easily modified by a variety of simple acts, such as moving a file from one drive to another, right clicking on a file within Windows Explorer, emailing the file, saving it in a different format, opening the file, or moving the file to a disk with a different file system. These are generally the easiest metadata elements to change inadvertently.

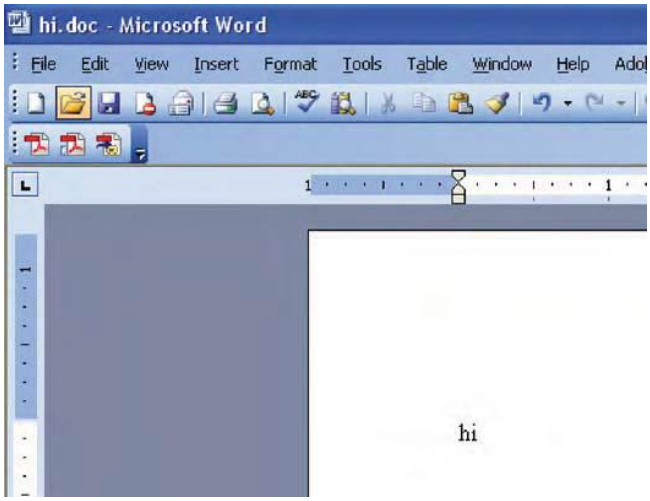
Unwittingly altering file system metadata can also play an important role in the eDiscovery process. Let's say two parties are involved in litigation and the pertinent time frames are between 1/1/01 and 1/1/05. If Party A decides to collect the data and in the process alters dates, then it may be difficult to find all the files that are responsive to the other party's request.

Application level metadata is contained within the data portion of a file and is typically separate from file system metadata. Metadata at the application level is typically buried within the document and some elements can be viewed easily within application. If one opens a Word document and chooses File>Properties, one can see many application metadata items including Author, Dates, Company, Edit time, Version, etc.

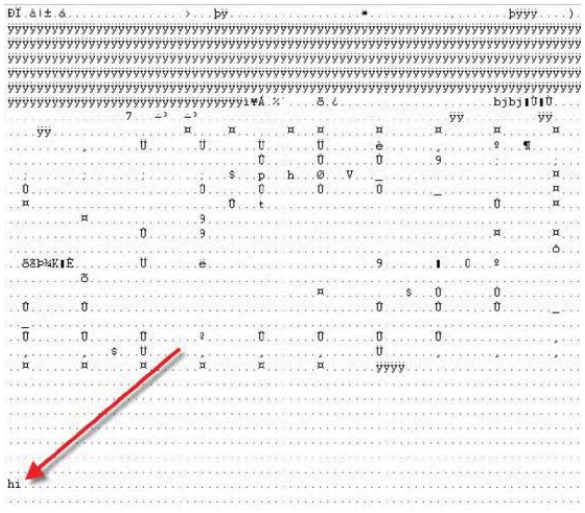
Below are two images that provide a great example of how one document can have different metadata elements – created date. The General tab is actually providing the File System metadata and the Statistics tab is the Application metadata. In this scenario I opened up Microsoft Word at 12:09 to begin writing this article, but I did not save it until 3:46. So as far as the File System is concerned this file did not come to life until 3:46 PM. Prior to being saved it existed as a temporary file — but that explanation is for another article.



Earlier I stated that the application metadata was stored in the data area of the file. If one were to create a Microsoft Word document and type hi and save it, the file would contain much more data than "hi". By viewing the contents through a hex editor[3] one can see into the belly of the document:



(Image A) **Microsoft Word View**



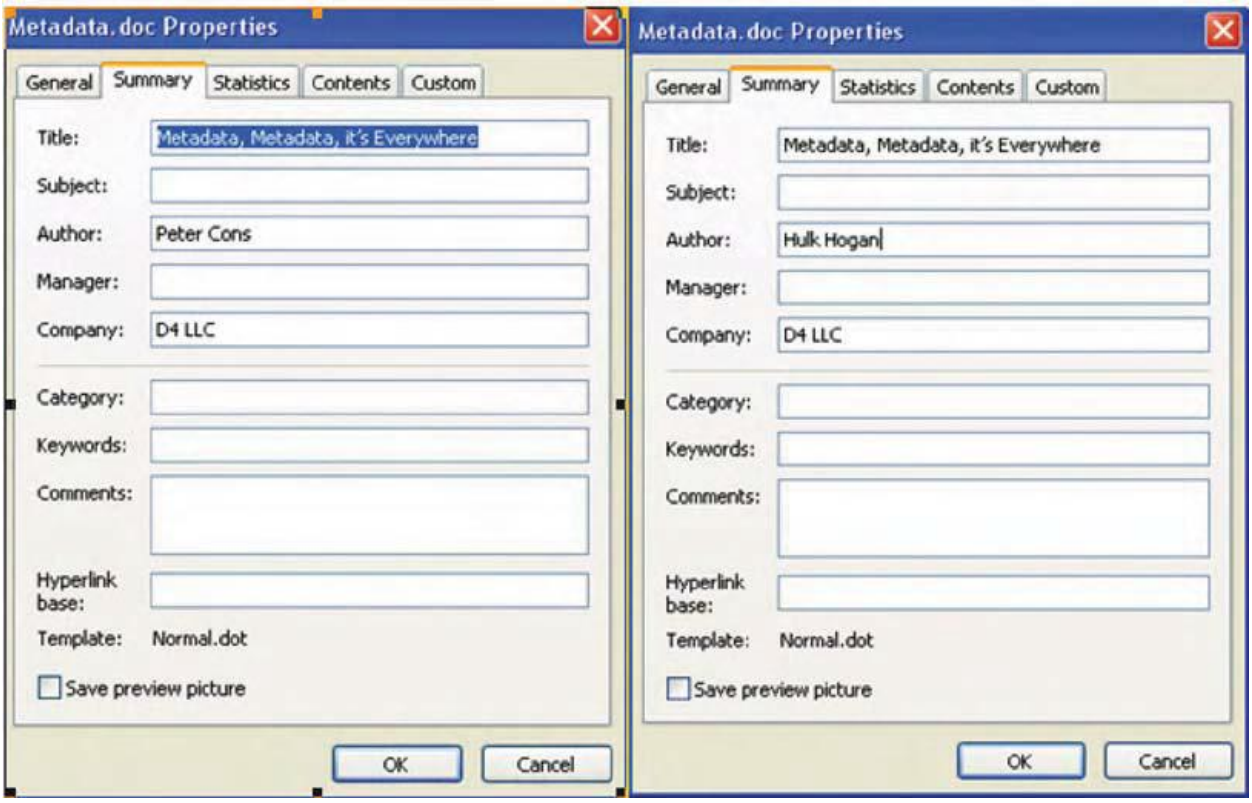
(Image B) **Hex View**

All of the other jumbled mess above before the "hi" is where the application metadata is stored including author, fonts, edit time, dates, etc. It may appear illegible, but trust me, the metadata is in there. This is why it can be a bit more difficult to modify application level metadata unless you are poking around inside the file. Generally, a user cannot modify the author of a document without opening the file in an appropriate application[4]. However, you can see in the example below that is quite easy to modify the author field of a document within Microsoft Word.

"I want it all."

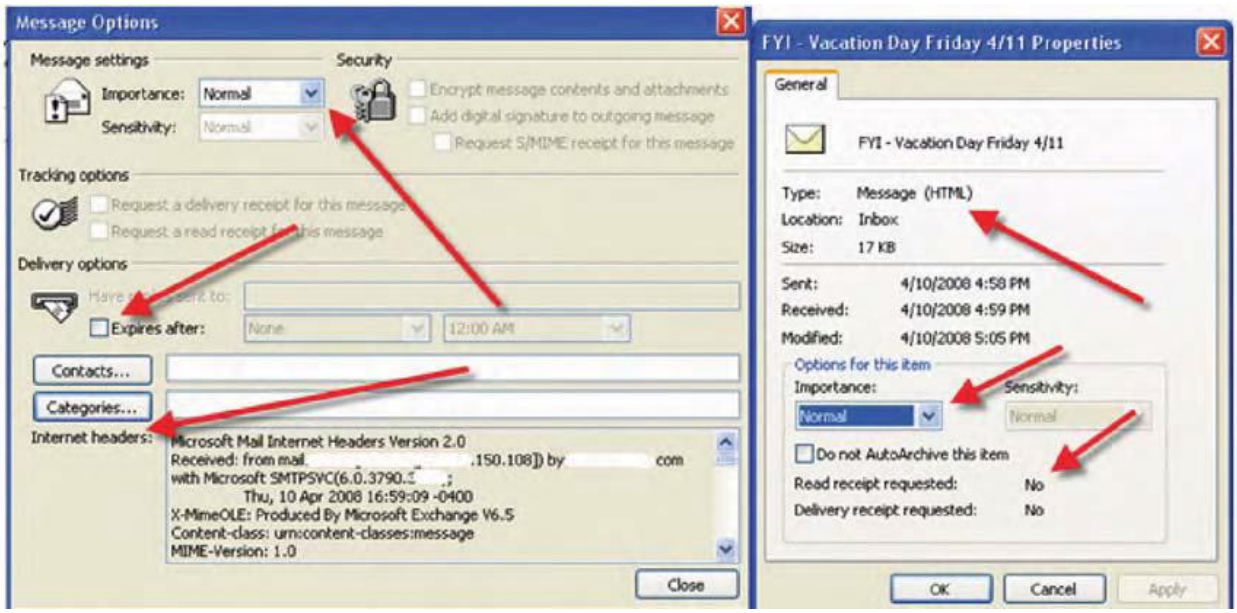
This section refers to the cries from attorneys and litigation support professionals that demand that "all the metadata be preserved and produced."

Let's stick with Microsoft Word. The fact that a document has Times New Roman font is most likely not the lynch pin for winning the case, but it is actually metadata. A Microsoft Word document has many other metadata elements, including custom fields such as "group," "destination," "disposition," as well as custom fields you can create from scratch. These metadata fields may or may not be relevant to any particular case, but in most situations, they will be completely irrelevant.



However, if it is important, I am confident that most vendors will be more than happy to produce all the metadata elements for any and all documents for an additional price.

Let's switch our focus to another common application, Microsoft Outlook. Email messages in Outlook have dozens if not hundreds of potential metadata elements. Below is an example of some of the metadata elements found in one message:



Above we can see the Email Headers[5], Type, Location, and Read Receipt Requested. Not shown, but of likely greater importance is, To, From, CC, BCC, Subject, Sent Date, and Received Date. Do you need all that information loaded into your review database? Maybe yes. Maybe no. But you should be sure what metadata you want before asking for it all!

While we are on the topic of email, how is email different than a typical Office file in terms of metadata?

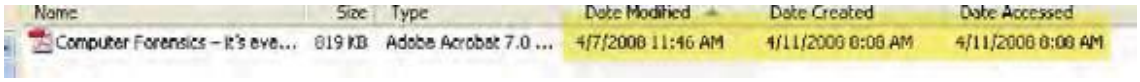
Many email messages are typically contained within email container files. You may be familiar with PST (Outlook) or NSF (Lotus Notes). When messages are contained within mail stores 97% (I have learned never to say 100%) of the important metadata is contained within the container file. In the eDiscovery world it usually does not matter what a PST's file system metadata is. If one moved a PST file to a different drive it would change the MAC dates, but it would typically not affect any of the contents inside the PST or the metadata of the messages and other objects (tasks, to-do, calendar, etc.). Of course the file system date could come into play in certain situations, but that is the 3% buffer above. This fact does not mean that great care should not be taken to preserve the original file system dates and times.

Is Change Easy?

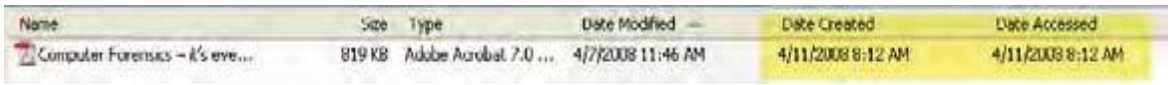
Earlier I stated File System metadata was more volatile than Application System metadata. While I have already provided some insight into that assertion, let's see some examples for MAC dates and how easily they can be altered.

Test #1: Adobe Acrobat File – Copy and Paste within the same volume:

Below are the file system MAC dates for an Adobe Acrobat file.

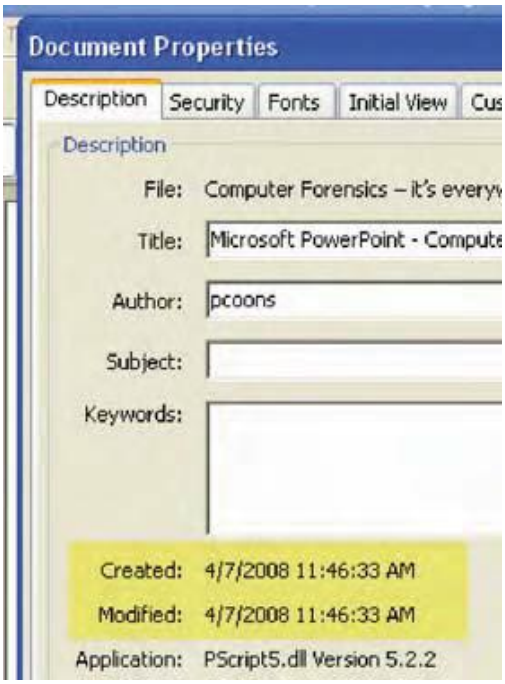


In this example, when we moved the file to a different folder on the same volume[6] using Windows copy and paste through Explorer the results were:



Both the Date Created and Date Accessed were changed, but not the Date Modified. So the File System metadata is reflecting when the file was moved, not actually created. Technically that is not true since it was created in that area (file system) for the first time at 8:12 AM.

When I open up the file to examine the application level metadata the following appears:

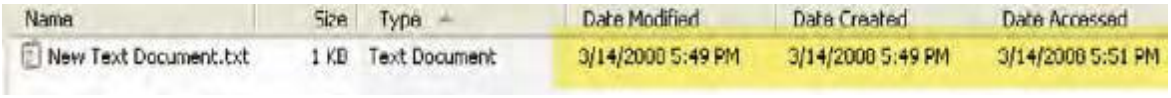


Both the Created and Modified dates in the Document Properties are the same as the Modified date in the File System. The Document Properties screen shown above is identical for the original Acrobat file as well as the copy I made.

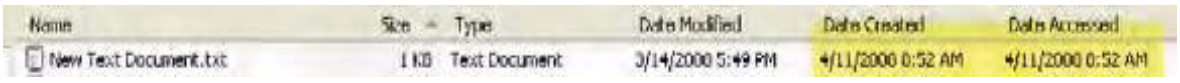
What you can take away from this test is that the file system Modified date may be the most reliable date for analysis as file system Accessed and Created dates are volatile and thus potentially unreliable.

Test #2: Moving a text file from the C drive to an external USB drive:

Below is screenshot of the Text file and its file system metadata on the C drive.



If we copy and paste the file using Windows Explorer to an external USB drive the results are the following:



The results are the same as in Test #1. The conclusion we come to is that moving or copying a file from one volume to another, or within the same volume, without taking proper

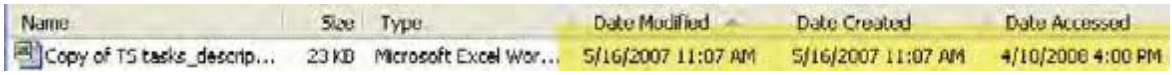
steps to protect the metadata, will result changes to the file system metadata for the Date Accessed and Date Created fields.

What about the Application level metadata for this file? Has it changed or remained the same? A text document has no application metadata for the MAC dates. In fact a text document has no application level metadata at all. You can change the font of a text file, but that change alters the underlying application, and every text document you open from that point forward with have the last assigned font. What you see is what you get.

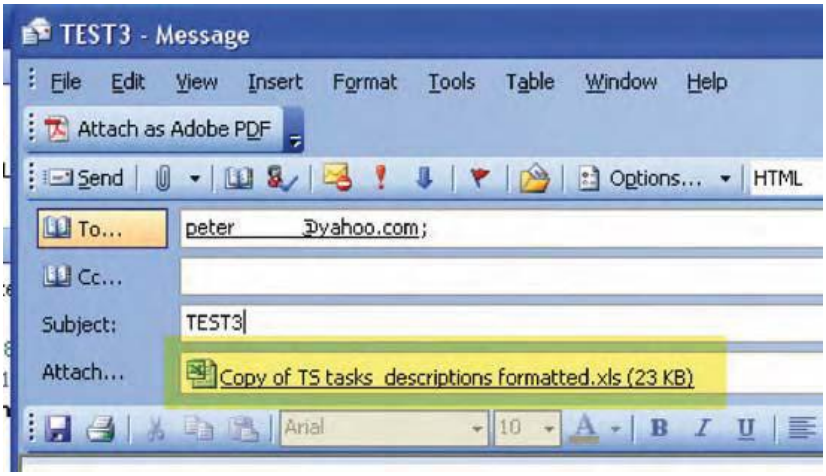
Test #3: Emailing a document from one account to another:

People often send important documents back and forth to one another via email. This simple act can also modify file system metadata. Just as moving files from one volume to another changes metadata, does sending a file by email from one person to another have a similar affect? Why don't I email a Microsoft Excel file and see what occurs.

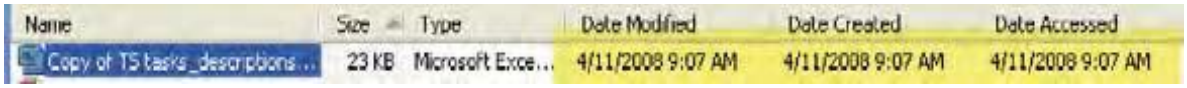
Below we can see the file system metadata for the Microsoft Excel file in Location 1.



When we attach that Microsoft Excel file to an email and send it on its way to Location 2, we assume that the metadata may be changed.

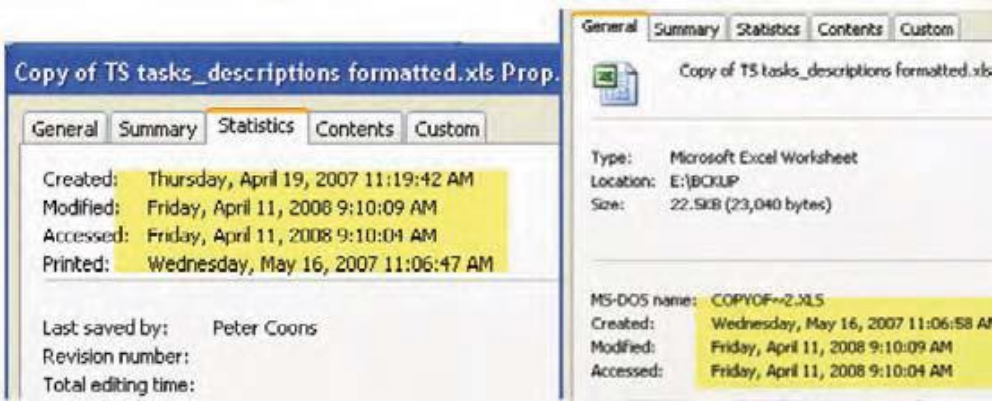


When the Microsoft Excel file is received and saved to a new computer at Location 2 and we look at the File System metadata we see the following:

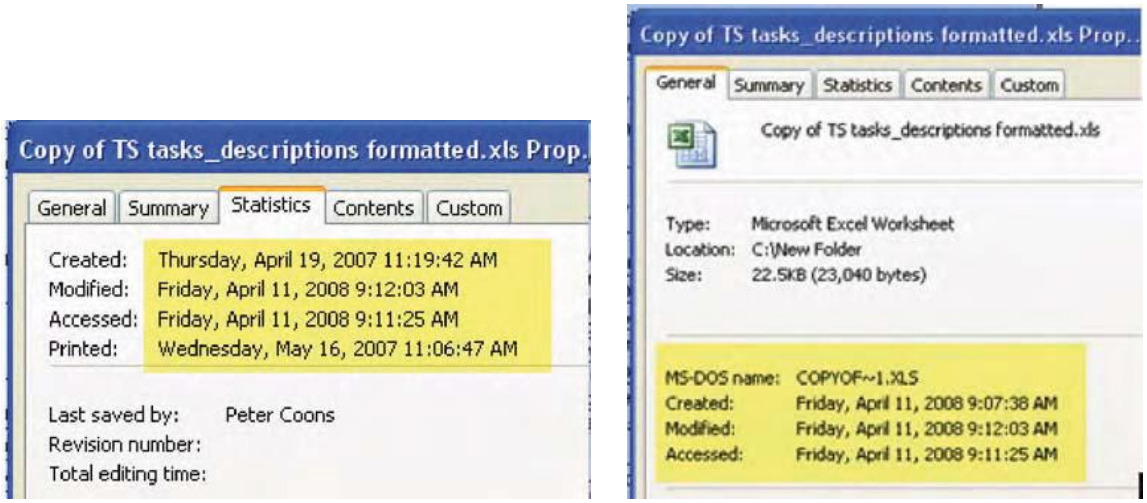


This result in Test #3 is different from the results in tests #1 and #2 as the Date Created has now also changed. What about the application level metadata?

Below is a screenshot of the application metadata for the file that we emailed from Location 1.



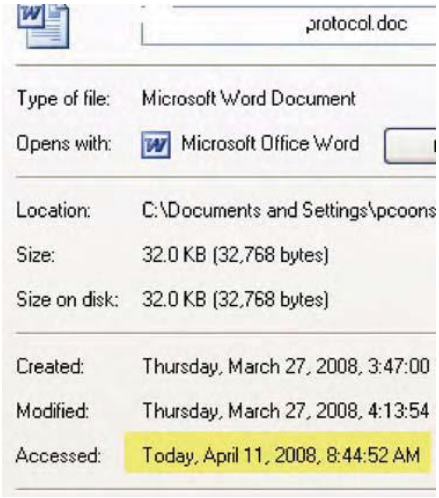
If you recall from earlier in the article, the Statistics tab displays the Application level metadata and the General tab displays the File System metadata. It appears the simple act of opening an XLS spreadsheet and checking its properties will alter both the Modified Date and Accessed Date when in reality nothing has been modified. However, the Created date remains the same. From the information above we can surmise the original file may have been created on April 19, 2007 and moved or copied to a different location on May 16, 2007. Now let's look at the Application metadata after the emailing, saving and opening (to Location 2).



We see that we have retained the Created date in the Application metadata of 4/19/07, but the File System metadata has all changed to our download and open date. Again we see that the simple act of opening the file and checking the internal properties will change the Modified and Accessed dates.

Test #4: Right clicking on a document and checking its properties:

For this test you have to conjure some faith and believe that I did not do anything more than check the file property and that action alone modified the Accessed Date. File System Accessed date is the most volatile of all the dates. It can be changed by a virus scanner or performing a Windows search across multiple files or right clicking on a file.



There are other ways to change metadata. When I Googled "change file dates", as you can see below, I received 11,200,000 hits.

There are programs available that allow a user to open a file and change the metadata (file dates and other information). Below is a file I created in May 1980, or so it appears.

Dates and other metadata (File System or Application) are information stored within the file or file system and if you know how to manipulate it, whether manually or programmatically, it's easy to fool someone.

Have you ever received SPAM from the year 2038? I have, and the spammer does that so it appears at the top of your Inbox. That should be proof enough not to believe everything you see and that it's easy to manipulate metadata.

Fish Out of Water

You can tell a lot about a file by its metadata, but you can decipher even more about a file if viewed within its native environment. By native environment, I am referring to where the document or file was created or existed originally. How can you be sure who created a document if it is so easy to change the Author and other metadata? It is because that uncertainty it may be important to leave files in their native environment.

The fact is that you can tell a lot about a file by looking at the other files and activity surrounding the dates and times in question. This is when a computer forensic expert comes in handy. A forensic expert can identify what other activity was occurring on the same dates or times of the alleged document creation. If other documents were created on the same date and around the same time, who is listed as the author of those documents? Was the alleged author logged into any websites that required personal knowledge of a password that only he or she would be privy to? Was there other activity at that time that can be tied to a specific individual?

Uncovering such clues may help to determine the authenticity and validity of the metadata, if the metadata is in question.

Who wrote that?

I have already shown how easy it is to change the Author in a Microsoft Word document. What about documents which have Authors that don't match up with actual employees of an organization? This can happen very easily.

Let's assume that John Smith receives a Microsoft Word document via email containing a speech from Bill Gates, and John likes a certain quote in the speech. However, instead of starting a new document from scratch, John erases the entire speech from the document, keeping only the quote. John then saves the document and stores it on a public network drive. Later, someone is mining for metadata (because it's fun) and they find a document authored by Bill Gates, per the application level metadata. However, the contents state the document was authored by John Smith. Who is the real author? We know the author

is John Smith, but in the eDiscovery world if we were relying on application level metadata to identify all documents authored by John Smith we would clearly miss this document.

Now let's assume, for sake of argument, that John Smith is a University student and turns in this document for an assignment. When the Teacher's Assistant mines the metadata, Poor John is likely to end up accused of plagiarism.

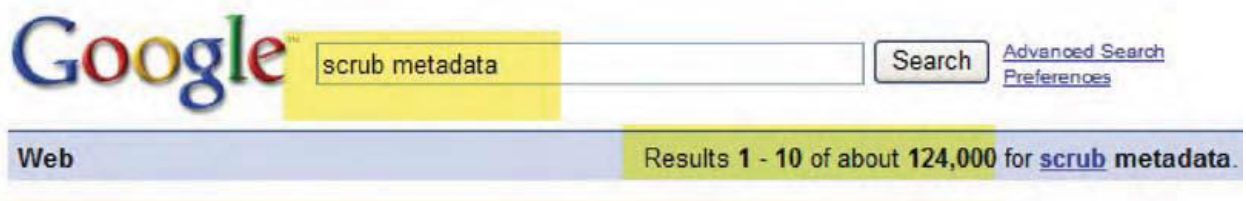
On the political side of the spectrum, former UK Prime Minister, Tony Blair got a byte in the butt from hidden metadata. In February 2003, 10 Downing Street published a dossier on Iraq's security and intelligence organizations. This dossier was cited by Colin Powell in his address to the United Nations the same month. Dr. Glen Rangwala, a lecturer in politics at Cambridge University, quickly discovered that much of the material in the dossier was actually plagiarized from a U.S. researcher on Iraq. Blair's government made the mistake of publishing the dossier as a Microsoft Word file on their Web site.

Shake Off That Metadata

We have seen how metadata can be changed but how do we prevent the next "Blair" incident? There are a number of ways to get rid of metadata in a document.

Arguably, the most popular is to convert a file to a PDF document. A PDF is an Adobe Acrobat file and it can be read by a number of applications including the freely available Adobe Acrobat Reader[7]. The PDF file format is portable (thus the full name: Portable Document Format), and flexible like an acrobat — or the malleable adobe building material used by the ancient Aztecs. It will also wipe away author information, track changes, dates, and other potentially sensitive or confidential data. However converting to PDF does not purge all of the metadata. For example, Acrobat records the name of the original document it was converted from.

Another method to rid yourself of that darn metadata is using specialized metadata removal software. Microsoft offers a software tool for free that integrates into Microsoft Office XP and 2003[8]. From a menu item one can flush all the sensitive metadata prior to sending the document to its next recipient. There are other third party products that profess to scrub all the metadata from myriad applications. Searching Google again, I found about 124,000 hits for the search "scrub metadata."



How To Preserve Metadata

When it's absolutely necessary to preserve metadata, and I assume if you are reading this it most likely is, then one should take care. We have shown that moving a file, emailing a file and even opening a file has the potential to change both File System and Application level metadata. We have also show that it may take some effort to alter the author but also how the author can appear as George Bush. The bottom line is that tools do exist to assist with the preservation of metadata. It's also true that if you know the pitfalls and what not to do or when to do it you are far ahead of the game. Is it necessary to call in a forensic or collection expert when litigation rears its ugly head? That answer depends on a lot of factors, but knowing the factors that go into making that decision is of utmost importance.

Forensic and collection experts frequently create image files to protect the metadata, both File System and Application. Think of an image file as a cocoon or a chrysalis. They also use hardware devices called write blockers to ensure no data can be changed when accessing media. Simply connecting a drive to a computer or booting a computer with an evidence drive attached can have an irreversible effect and modify hundreds of files, if not all the files, on a drive.

Metadata is fragile and sometimes difficult to interpret. Create date does not always equate to the date the file was actually created. Moreover, the create date at the File System level may not match the create date in the Application itself. The author field can be changed with a few clicks of the mouse. So how can one be sure when a document was created and who created it? For the most part we can lean towards believing what we see until there is a reason to question that belief. We also have the benefit of forensic investigative techniques that help determine actual dates and other information if the validity is ever questioned. Just don't believe everything you see and if you do I have a bridge in Brooklyn to sell you.

Consequences of Spoliating Metadata

Bad things happen to good lawyers. When it comes to metadata, one of the worst things that could happen is the failure to properly preserve that metadata when it really matters, such as: (1) when metadata is required to be produced during discovery, (2) when metadata contains potentially exculpatory evidence, regardless of whether or not it is required to be produced, and the authenticity of evidentiary ESI is called into question.

As we've discussed, metadata may contain specific dates, times, and other information about a specific file or events occurring within an operating system. But how does this become relevant to litigation, and how can the spoliation of metadata lead to sanctions, or worse?

We can break this issue down to two core categories: The first category is where metadata itself may be responsive to a discovery request; and the second category is where metadata is not responsive to a discovery request, but may be used in connection with determining authenticity of ESI.

Under the first category, we have metadata that is kept in the usual course of business and is responsive to a discovery demand. Let's start with hypothetical CASE "A" before a Federal Court in which the facts at issue include determining who wrote a series of documents, and on what date and time those documents were created, modified, printed, saved, etc. Let's assume that the Plaintiff in CASE "A" was to issue a discovery demand to Defendant for "all ESI in Defendant's custody or control related to _____." And furthermore, let's say that Plaintiff's demand included instructions, pursuant to Federal Rules of Civil procedure 34(b)(1)(c), that all responsive ESI was to be produced by Defendant "in native format, with metadata attached." However, let us assume that when Defendant sent its IT staff to collect the responsive ESI from its computer systems, the IT staff did not use proper procedures and technology to ensure that both the file system and application metadata was unaltered during the collection. As a result, the metadata has been spoliated, Defendant is unable to meet its production obligations, and Plaintiff is likely to seek monetary and non-monetary sanctions against Defendant.

In the second category, we have metadata that is not necessarily responsive to a discovery demand, but for which the authenticity of the ESI is key. Let's jump into hypothetical CASE "B" in which the facts are fairly straight forward, and the real issue is whether certain ESI propounded by each party as evidence, such as an email message is authentic. Plaintiff claims to have a print out of a memo containing allegedly derogatory comments that was reportedly printed out by a former co-worker and given to Plaintiff. Defendant asserts that Plaintiff falsified the memo, and that Defendant has the original memo which did not contain any of the allegedly derogatory comments. However, let's assume, as happened in CASE "A", that when Defendant sent its IT staff to collect the memo from its computer systems, the IT staff did not use proper procedures and technology to ensure that both the metadata for the memo was unaltered during the collection. Again, the metadata has been spoliated. However, in CASE "B" when Defendant attempts to introduce the supposed original memo into evidence, Plaintiff's objects and provides an expert to show that the metadata associated with that original memo does not match the dates of that email, calling into question the authenticity of Defendant's key evidence. As a result, Defendant is unable to introduce its otherwise exculpatory evidence, and Plaintiff, who has the former co-worker as a supporting witness, wins.

Metadata can help a party tell its story to the jury. However, metadata can also tell a story you don't want the jury to hear, such as how dates and times of evidence do not match up with witness statements. At the end of the day, even inadvertent spoliation of metadata can ruin your entire case. So handle your metadata with care.

Endnotes

1 http://www.edrm.net/wiki/index.php/Glossary

2 Windows XP Service Pack 2

3 http://en.wikipedia.org/wiki/Hex_editor

4 It is possible to modify any metadata element without opening the file in an appropriate application through programming or manipulating in a tool like a hex editor.

5 http://en.wikipedia.org/wiki/E-mail#Header

6 Volume refers to a partition on a drive. Typically one hard drive will have one partition (C:) and therefore one volume. However, a hard drive can be split into multiple partitions and therefore have more than one volume.

7 See: http://www.adobe.com/products/acrobat/readstep2_allversions.html

8 See: http://support.microsoft.com/kb/834427