



New Document Metadata Changes in Microsoft Office 2007

Esquire Innovations, Inc., a leading provider of Microsoft Office integrated practice management software services and applications for the legal market, counts more than 550 law firm and in-house council clients utilizing its applications. Esquire Innovations has been developing, supporting, and selling document creation, formatting, re-purposing, comparing, and metadata management software applications in the legal industry since 1999. Innovative Software + Astonishing Services = Extraordinary Client Success

ESQUIRE
Innovations, Inc.

This article first appeared in the July 2009 issue of Law
Journal Newsletters – Legal Tech Newsletter

Randall Farrar
Esquire Innovations, Inc.
July, 2009



Microsoft Office documents contain information other than the actual content that is intended for the reviewer to view and edit. This information is called document metadata. In some cases the disclosure of this document metadata may have potential for harm or embarrassment for a law firm.

The document metadata contained in a Word document other than the intended text doesn't necessarily create risk of adverse disclosure, because some document metadata is necessary for formatting or macro automation of the document. However, some document metadata, such as Tracked Changes, may be used to share among cooperators but should not be shared with adversaries or in some instances clients, because it contains author and date metadata.

Metadata Changes Start with the New File Format

Microsoft in its ever increasing effort to make the end user experience easier when using Office has presented users with a different paradigm for the user interface. Along with this change in the user interface is a complete change in the Microsoft Office document format. Prior to Microsoft Office 2007 the file format was binary, which made it difficult to access internal metadata. Now in Microsoft Office 2007 the file format is XML. In other words, the new file format can be accessed without using the intrinsic Microsoft Office application, such as Word. This ease of file access means that document metadata can easily be viewed and changed.

Viewing Metadata in a Microsoft Word 2007 Document

It is easy to view metadata in a Microsoft Word 2007 (".Docx") document without Microsoft Word. A .Docx file is a combination of 10 or more xml files contained within one Zip file (as strange as that sounds). The only difference between a Word 2007 file and a Zip file is the extension. To view metadata in a Microsoft Word 2007 .Docx File you only need to save a Word 2007 document to your desktop and rename the extension by changing the ".docx" to a ".zip" extension. After you rename the extension, double click on it and this will open it in Windows Explorer (or your default Zip application) and display the Word document as XML and at the root level of the XML file.

Each XML file in the Microsoft Office 2007 document is called a "part." Within the document part is the main document part. When inspecting the main document part, you'll notice many elements that you know are in your document, elements like styles and content. Additionally you can view metadata elements such as the document properties and author information. The chart below shows some of these metadata xml parts.

Microsoft Office 2007 Document Metadata XML Parts Summary

Part	Path	Description	Microsoft Application
app.xml	root\docProps	Defines application file properties. These properties include the number of characters, words, lines, paragraphs, and pages in the document	Word Excel PowerPoint
core.xml	root\docProps	Defines core file properties. Includes creator name, creation date, print date, title, and document description.	Word Excel PowerPoint
custom.xml	root\docProps	Defines custom document properties	Word Excel PowerPoint
settings.xml	root\word	Defines document variables	Word
document.xml	root\word	Defines tracked changes and author information.	Word
recipientData.xml	root\word	Contains contact data mail merge operation	Word
revisionHeaders.xml	root\xl\revisions	Defines tracked changes, author and date information.	Excel
comments1.xml	root\xl	Defines comments in the workbook, comment author and comment dates and times	Excel
notesSlide1.xml	root\ppt\notesSlides	Defines slide note information	PowerPoint
commentAuthors.xml	root\ppt	Defines information about each author who has added a comment to the document. That information includes the author's name, initials, a unique author-ID, a last-comment-index-used count, and the author display color.	PowerPoint

Microsoft Office 2007 has introduced three new metadata elements; Custom XML, Content Controls and Mail Merge Recipient List. Although these new metadata elements introduce advanced user and automation power, they introduce completely new metadata risks for law firms.

Custom XML

The inability to place non-Microsoft Office document data or information in the old Microsoft Office file format was a major limitation for developers. For instance, if a developer wanted to insert a small data list in the document that could be referenced for automation, it could not be done. It would have to be placed elsewhere and linked to using elements such as in custom properties and document variables. This was a clumsy way of doing things for a myriad of reasons. Microsoft Office 2007 now allows developers to place small data lists or large databases in a Word, Excel and PowerPoint documents. Although this is great for developers, Microsoft has opened Pandora's Box with this feature.

The risks with Custom XML are that it stays with the document until you remove it using Document Inspector or automation. A user will not be able to see the Custom XML, and has to remember to use the Document Inspector.

Content Controls

In the old days a developer would use bookmarks to place static information that a user types in or selects from a dialog box into a document. In Office 2007 Microsoft provided a much more sophisticated approach (called Content Controls), which can be used for the same purpose, but the text being placed in the document is dynamic. Content controls can be inserted into a document to provide more dynamic user options - for example you may want to allow a user to pick a value from a drop-down list, for different "closings" in a letter.

For the most part Content Controls are benign, but a developer can take Content Controls to the next level by linking them directly to Custom XML. The risks are that Content Control links stay with the document and often the user is completely unaware of this link or even the contents of the Custom XML itself.

Mail Merge Recipient List

Microsoft has greatly improved its Mail Merge feature with Office 2007. You no longer need a PhD in rocket science to get it done. Yet, this improvement adds new metadata risks. The new Mail Merge features allow the user to connect the document to an external data source. The data source could be a Microsoft Outlook contact list, an external database or a file that contains the information to be merged into a document. These lists could be, for example, the names and addresses of recipients for a letter.

The metadata risk is that this link source is stored in the document.

- If this source is on a server then the server path is stored.

- If the source is an Outlook contact list the mailbox address is stored, i.e., Mailbox – Randall.
- If the list is manually created by the user the data source path is stored along with the database structure.

Summary

With the advent of Microsoft Office 2007 users are able to create and work with documents much more efficiently. The new XML file format provides many advantages over the old binary format and opens up a gigantic door for a new generation of software solutions. But, along with these benefits is a new metadata layer that law firms need to be aware of. When firms begin to migrate to Microsoft Office 2007 they will need to address these new risks with their internal document metadata policies and to ensure that the firm's metadata software solution – software that eliminates metadata - addresses them.

Randall Farrar is the president and co-founder of Esquire Innovations, a leading provider of Microsoft Office integrated practice management software services and applications for the legal market. He can be reached at randall.farrar@esqinc.com