# The Dangers of Document Metadata:
## The Risks to Corporations

> "What you
> can't see
> can hurt you."
>
> — Gartner Group,
> Research Note on
> Metadata in Office,
> January 2003

## Overview

Corporations today face risks from situations that are either under their direct control or from conditions that they might not even be aware of that puts their corporation at potential harm. An example of the latter is document metadata—hidden information contained in Microsoft® Office documents including Microsoft® Word, Microsoft® Excel, and Microsoft® PowerPoint files. Whenever a document is created, edited, or saved, metadata is automatically added to the document. This information is transmitted every time a document is emailed to parties, both internally and externally to the corporation. Microsoft Word's collaboration features, such as comments and Track Changes, result in a significant amount of metadata being included in documents. Originally conceived to shed light on data, document metadata categorizes information to make it easier to track and find.[1] When used properly, metadata is useful. But when used carelessly, it makes it easy for other people to find out details about the document and other privileged information that could harm corporations.

## Why Remove Document Metadata?

Every time a document is created, metadata is automatically added to it. Some of the information stored in the document may also be confidential—e.g. previous versions or information that may have been rejected or accepted—and may also expose corporations to hidden risks when it is emailed to people outside the company. The problem is not that metadata is added to a document, but rather, it is often more difficult to remove the metadata once it has been added. And because this type of information travels with the document every time it is emailed to others, sensitive or confidential information may be transmitted unknowingly. Within Microsoft Word, the ability to view comments and suggested changes from other people can be useful when collaborating with several parties. In most cases, these collaboration features significantly enhance the user experience. However, changes that are not accepted still remain with the document, even though they appear to be invisible. These changes can easily be displayed by turning on the "Show markup view." This can result in embarrassing situations where external parties are able to see information that was not intended for their eyes.

With the release of Microsoft XP, Microsoft added some metadata removal features. However, it is still painfully short of a complete, automated solution and relies heavily on technical end user intervention. For example, under the Security Tab in the Options menu, Microsoft provides the ability to remove personal information from a file upon save and to warn users before printing, saving, or sending a file that it contains tracked changes or comments. Additionally, users can turn off "fast saves" under the Save tab in the Options menu to ensure that deleted data is really deleted. The only problem is that users must go to several different places within the application to remove different types of metadata. There is no central place to manage the different settings and most users of Microsoft Word are not even aware that these protection options are even available to them. This problem is further compounded when documents are attached from within Microsoft Outlook and sent to outside parties. Used as the de facto way to electronically exchange documents in an enterprise environment, Microsoft Outlook does not offer warnings about metadata in attached documents or zipped files. Thus, the potential to accidentally send Microsoft Word documents containing harmful metadata (and thus expose a corporation to the inadvertent disclosure of sensitive information) is amplified tremendously with every document that is sent back and forth in a collaboration.

In financial reporting documents such as spreadsheets, metadata can be saved with Microsoft Excel files. A quick review of the Fortune 1000 Web sites shows that 33% of these Web sites contain Microsoft Excel documents publicly posted either directly on the company's corporate Web site or linked to a third party site for SEC filings. Accidental posting of Microsoft Excel documents that contain potentially harmful document metadata can be easily viewed by anyone who downloads these documents. Because metadata is not always viewable, document users can unwittingly send confidential information to people outside their organization. In fact, there have been several widely publicized, high profile cases in which document metadata proved to be the culprit. (See side bar on page 2.) Gartner Group estimates that "by year end 2004, fewer than 25% of enterprises will have policies surrounding removal of metadata from Microsoft Office documents."[2] The bottom line: document metadata can get corporations in big trouble—putting the organization at financial risk, a competitive disadvantage, and placing them in an embarrassing situation with costly consequences.

1 Doherty, Sean, The Dark Side of Metadata, Network Computing, 4 September 2003.
2 Silver, Michael, Metadata in Office, Gartner Group Research Note, 23 January 2003.

## Types of Document Metadata and Their Associated Risks

Document metadata comes in many forms. Below is a list of the types of metadata found in Microsoft Office documents and the risks that each type of metadata poses to a corporation.

### ▣ Document Properties

*Microsoft Word, Microsoft Excel, and Microsoft PowerPoint documents*. Document properties are details about a file that help identify it that includes a descriptive title, subject, author, manager, company, category, keywords, comments, and hyperlink base. Document properties display information about a file to help organize the files so that they can be easily found at a later date.

#### ▪ Risks

The names of authors and the name of the company can display sensitive information about a corporation. It is possible that if a document has been sent outside your own corporation, the author name and company name contained in the built-in properties could be a name other than your own. In addition, if documents are re-purposed or used as a template for a new document, information that is specific to a previous client such as pricing, terms, or the client's name can be stored as hidden information within the new document.

### ▣ Document Statistics & File Dates

*Microsoft Word documents only.* Document statistics include information on when the document was created, when it was modified, when it was accessed, and when it was printed. In addition, document statistics display the name of the person it was last saved by, the revision number, and the total editing time. Other statistics include number of pagers, paragraphs, lines, words, and characters.

#### ▪ Risks

Document statistics can create embarrassing situations when the hours billed do not match the total editing time. In addition, the "last saved by" metadata shows the last person who edited the document. This can be risky if it is discovered that the person whose rate and time is billed out is different than the person who actually worked on the document.

### ▣ Document Reviewers

*Microsoft Word documents only.* Document reviewers consist of a list of users that have added or accepted any track changes. When the names of reviewers are removed, but not the Track Changes, the revisions remain with the document. However, the user name associated with each revision will be removed. It is recommended that the names of the document reviewers be removed when removing track changes.

#### ▪ Risks

The risk from the Document Reviewers metadata is that it can expose who has suggested what changes.

### ▣ Custom Properties

*Microsoft Word, Microsoft Excel, Microsoft PowerPoint documents.* Custom Properties includes any property fields added manually to a document or by various programs to help manage and track files.

#### ▪ Risks

Custom Properties are normally things specific to an organization. The potential risk arises because it is easy to see a history of this document.

### ▣ Hidden Text

*Microsoft Word documents only.* Hidden text are text blocks that have been formatted as hidden. Unless specifically selected to be viewed in Microsoft Word, hidden text is not displayed within the document.

#### ▪ Risks

Hidden text can contain notes that are particular to a document. As hidden information that is not cleansed, the hidden text can potentially be viewed by unintentional parties.

### ▣ Comments

*Microsoft Word, Microsoft Excel, Microsoft PowerPoint documents.* Comments are notes and suggestions that are added to a document via the comment feature to help facilitate an online review.

#### ▪ Risks

Comments, like hidden text, unless intentionally removed can display sensitive information to external parties because comment metadata travels with the document. Microsoft Excel and Microsoft PowerPoint documents are especially susceptible to this risk as there is no internal mechanism built into these applications to warn a user that comments are embedded. Gartner Group states that "while Microsoft is aware of potential problems [with comments], it does not have a comprehensive solution to solve this problem."

### ▣ Track Changes and Document Revisions

*Microsoft Word and Microsoft Excel documents*. The Track Changes feature tracks changes (inserted, deleted, and moved text) made to a document during an online review. As changes are made to a document using Track Changes, a new revision of the document is kept by the application. This revision history exists, even after changes to the document have been accepted or rejected.

#### ▪ Risks

Track Changes shows the history of changes to the document. If Track Changes is left on, but the highlight on the

screen is turned off, every change made to the document still remains. This is like recording every single keystroke made to the document that can be viewed by subsequent reviewers. Thus, even though the Track Changes are not visible, it still travels with the document and, in some circumstances, it can be sent to and seen by an unintentional party with potentially disastrous consequences.

## Headers and Footers

*Microsoft Word, Microsoft Excel, Microsoft PowerPoint documents.* Headers and footers are areas in the top and bottom margins of each page in a document. Text or graphics can be inserted in headers and footers—for example, page numbers, the date, a company logo, the document's title or file name, or the author's name—that are printed at the top or bottom of each page in a document.

### Risks

Custom header and footers can contain descriptions such as filename, path, the date and time the document was modified, or other information that is deemed important to make it easy to retrieve and edit a file. Unfortunately, the information contained in footers and headers is often overlooked when the document is shared. Failure to remove this information can result in revealing confidential information.

## Footnotes

*Microsoft Word documents only.* Footnotes attributed to content are embedded as metadata into Microsoft Word documents.

### Risks

Footnotes may expose private, internal document is used in the organization.

## White Text

*Microsoft Word documents only.* White text is blocks of text that have been formatted with a font color of white on a background of white. The text appears invisible when viewed or printed and can be used to hide information in a document.

### Risks

White text is commonly used when documents are posted to the Internet so that can be more readily found by search engines. However, white text can also be viewed by external users. Depending upon what was actually written as white text, the information can be very damaging. White text can also be used for particular field codes such as the "include text" field code, which can point to a file location. If this file location code is embedded in a document, users can unknowingly be updating the code and can potentially expose the document to a hacker.

## Small Text

*Microsoft Word documents only.* Any text block contained in a document that is less that five (5) points is considered small text. The text is so small that it will not be visible when viewed or printed and can be used to hide information in a document.

### Risks

Like white text, small text is commonly used to put information in documents so they can be found by search engines. Small text can also include sensitive information that was not meant to be distributed externally.

## Macros

*Microsoft Word documents only.* If a task is repeated in Microsoft Word, it can be automated using a macro. A macro is a series of commands and instructions that are grouped together as a single command to accomplish a task automatically.

### Risks

There are several reasons to strip out custom macros. For example, macros can be set for templates that may have some amount of pre-populated data. There may be a time when the information contained in these templates should not be seen by external audiences. Another example, macros can be linked to internal databases or intranets. Having access to the internal file naming structure is generally information that most corporations do not want outside their firewall. Lastly, macros are often quite complex and, if developed in-house, may represent the company's intellectual property. If macros are included in the document, the information is freely shared with any outside party.

## Previous Versions

*Microsoft Word documents only.* Previous versions show the number of times that a document has been versioned over its lifetime. This function enables Microsoft Word to save prior versions of a document as a part of the electronic file.

### Risks

The risk associated with previous versions is that a recipient can access any of the previous versions that have been saved. Therefore, the party reviewing the document can go back to any version and see what was changed in the document lifecycle. This metadata, while useful in some instances, can disclose sensitive information.

## Routing Slips

*Microsoft Word and Microsoft Excel documents only.* Routing slips are used to create a distribution list of reviewers in a particular order. Routing slips are manually created by adding in recipients' email addresses. When files are routed, it is sent as an attachment in an email message.

Workshare

## TYPES OF METADATA

Your name

Your initials

Your company or
organization's name

The name of your computer

The name of the network
server or hard disk on which
you saved the document

Other file properties and
summary information

The names of previous
document authors

Track Changes

Document revisions

Document versions

Template information

Hidden text

Comments

Macros

Hyperlinks

Routing information

Nonvisible portions of
embedded Object

Linking and Embedding(OLE)
objects

US OFFICE
208 UTAH STREET, SUITE 350
SAN FRANCISCO, CA 94103
USA
T 888 404 4246
F 415 975 3854
sales.us@workshare.com

UK OFFICE
20 FASHION STREET
LONDON E1 6PX
UNITED KINGDOM
T +44 (0)20 7426 0000
F +44 (0)20 7426 0060
sales.uk@workshare.com

■ **Risks**

Routing slips reveal the names of the people that the document was sent to for review. This may be information that should stay confidential rather than distributed externally. An example of how this information can be used is when email addresses are put in the routing slips. If this document is then published to the Internet, the email address can be displayed for all to see.

◉ **Fast Saves**

*Microsoft Word documents only.* Fast saves is an option in Microsoft Word that saves just the changes that were made to a document, resulting in the history of the changes being saved with the document file. Turning fast saves off and saving the document will remove the changes and store only the final version of the document.

■ **Risks**

Like other metadata, changes saved during a fast save can expose sensitive information to external parties when viewed using a text or hex-editor. Deleted text can still exist in the electronic file. According to the Gartner Group's Research Note on Metadata in Office, "users can easily forget that metadata exists when they send the document to someone else. Some metadata is never visible, such as pieces deleted by users but not really deleted by Microsoft Office when operating with fast save turned on."[2]

◉ **Hidden Slides**

*Microsoft PowerPoint documents only.* Hidden slides are slides that are hidden so that they are not shown during a slide show.

■ **Risks**

A master Microsoft PowerPoint slide deck may contain some slides that are uses as backup or that are for internal use only. To prevent accidental showing of these slides, it is best to strip out any hidden slides before sending the slide deck out externally.

◉ **Hyperlinks**

Microsoft Word and Microsoft Excel documents only. Documents can contain hyperlinks to other documents or Web pages and are displayed as blue underlined text. Hyperlinks in Microsoft Excel files can be seen in: a link to a cell in another Microsoft Excel document, a named link to a named reference in another Microsoft Excel document, a link to another document, an OLE link that inserts another document as an icon, and an OLE link that inserts another document as text.

■ **Risks**

Hyperlinks can maintain a link to a site that corporations may not wish to disseminate such as files that may exist on a computer's local file system, on a corporation's internal database, or on an intranet. Disclosing the file path, or the location of where the files are stored can invite potential hackers to gather sensitive corporate information.

### Conclusion

With the rapid adoption of email systems, documents that previously were printed or faxed to others have evolved into electronic documents that can be zapped around the globe with a click of the mouse. The prevalence of attaching Microsoft Office documents to emails means that those senders of documents can unwittingly share confidential metadata without knowing it and that there is a high degree that the inadvertent disclosure of sensitive information can occur. Corporations who rely on email to transmit documents must take steps to ensure that metadata does not reach unintended parties.[5] Organizations must be enlightened to the importance of understanding the consequences of document metadata, its liability issues, and the risk it poses to corporations.

IT professionals are investigating ways to sufficiently protect their corporations from exposing confidential metadata. They have begun to institute written policies on how to share or distribute documents outside the organization, to train users on the existence and removal of document metadata, and to deploy software applications that effortlessly and automatically remove document metadata from document attachments without any user interaction. Policies such as the new SEC Rulings regarding Corporate Web Site state: "Extra care should be taken to ensure that only the final versions of documents are posted [on the Web site]. Documents should also be cleansed of all metadata or other internal codes before posting to avoid the accidental disclosure of hidden data."[6] According to Gartner Group, through year end 2004, enterprises will have to rely on policy more than technology to ensure metadata is removed from Microsoft Office documents and does not leave the company. [2]

While policies may be difficult to enforce, savvy IT professionals are searching for software solutions that seamlessly solve the removal of metadata with little human interaction and training. This means that the software must not only integrate with Microsoft Office products to remove metadata, but it must also seamlessly work with the email systems that act as the primary distribution vehicle for those documents. Finding the right integrated metadata removal solution that fits effortlessly into the workflow that people have adopted today will put corporations less at risk and more in control by preventing the inadvertent disclosure of sensitive information.

5 Bershad, Robert and Laura Bandrowsky, Metadata: Out of Sight, But Not Out of Mind, Legal Tech Newsletter, May 1, 2002.
6 Hale and Dorr LLP, Spotlight Returns to Corporate Web sites, Find.Law, 2002.