

# Hidden Data in Electronic Documents

Deborah Kernan

GIAC GSEC Practical (v.1.4b, Option 1)

July 5, 2004

## Abstract

Document authors may be unaware that their documents contain hidden data and that there is the potential for the inadvertent release of sensitive information when sharing these documents with others. This paper discusses the various types of data that can be contained within the documents, why it could be a problem if the data owner does not know it exists, and steps that can be taken to minimize or eliminate the types of hidden data saved with documents.

## Introduction

The potential for inadvertent release of sensitive or proprietary data is possible given the fact that electronic documents can contain data that the document owner may not realize is there. One case of the inadvertent release of data is found in the story of Tony Blair's Iraq Dossier<sup>1</sup> in which hidden data in a Microsoft Word document revealed who had contributed to the document, which was later discovered to be plagiarized. Another instance involved the SCO Group's lawsuit against DaimlerChrysler. Hidden data, again in a Microsoft Word document, revealed that Bank of America was originally identified as the defendant instead of DaimlerChrysler<sup>2</sup>. Research conducted by Simon Byers<sup>3</sup> found that all Microsoft Word documents that he collected, mainly from the Internet, and tested revealed some hidden data in them.

The inadvertent disclosure of data also occurred when the Washington Post published a PDF version of a handwritten letter intended for the police and left by the Washington, DC, sniper. Details were blacked out of the PDF file, but it was possible to remove the blackened areas to reveal the underlying text they were intended to hide<sup>4</sup>.

Although there are other applications that may store hidden data in electronic documents, Microsoft's Office software dominates 90% of the market<sup>5</sup>. Therefore, this paper will concentrate on the hidden data contained within Microsoft Office electronic documents and measures that can be taken to reduce the potential for inadvertent release of sensitive or proprietary information that may be included in that hidden data.

## What Data is Saved?

According to the Microsoft Knowledge Base Article 223396, "OFF: How to Minimize Metadata in Microsoft Office Documents,"<sup>6</sup> whenever a document is

created, opened, or saved it may contain hidden data, or meta-data (data that describes other data), that includes user information, such as the user's name and initials, company name, the name of the computer, the path to the location where the document was saved, the names of previous document authors, hidden text or cells, comments, and other information about the document itself. Microsoft Knowledge Base Articles discussing methods to minimize the meta-data in Microsoft Office products suggest that in addition to Word, Excel and PowerPoint also save data that can be found in the electronic form of the workbooks and presentations, and which the document owner may not wish to share with others.

## **What's the Problem?**

One problem with the fact that Microsoft Office products save this meta-data is that the document owner may not be, and usually isn't, aware that the electronic form of the document contains anything more than what he/she has written and can see on the computer screen. The data that is saved is not visible when viewing the document, workbook, or presentation in the normal manner. Some of the data is visible when the item is viewed with a hex editor, but typical users are not likely to employ a hex editor to view the electronic documents they create to ensure that there is no unwanted data present.

When a document is shared electronically, it is possible that the person sharing the document is releasing more information than they intended. It is also possible that this information could contain sensitive or proprietary data. As a basic example, one piece of information that accompanies the document is the path to the location where the document was saved. It is possible that this path may contain the name of a proprietary company network, which could be revealed to the recipient of the electronic document if he had the wherewithal to find it.

Microsoft Word is often used as a collaborative tool, in which one document is routed to multiple authors, and the input from each author is duly noted and attributed to them. As part of the meta-data, Microsoft Word saves the names of the last ten people to work on the document. This information is not visible when the document is viewed in the normal manner. It is this feature of Word that caused difficulties for Tony Blair, as noted in the article referenced at the beginning of this paper.

Microsoft Excel worksheets allow users to hide columns and their contents. These columns and the contents are saved with the document. A user may not intend a recipient to view this data. Viewing the worksheet normally does not provide any indication to the recipient that there are hidden columns to be seen. However, viewing the worksheet with a simple hex editor reveals the hidden columns and their contents.

The electronic versions of Microsoft PowerPoint presentations contain similar meta-data as that which is saved with Microsoft Word documents and Microsoft Excel workbooks.

Some of the meta-data saved with these documents is not visible in the printed form of the document or when the documents are viewed using the associated application. When the document is shared electronically, a hex editor can be used to view this information, and utilities that can be used to extract this information from the file are readily available and easily obtained. Although the meta-data may seem innocuous or harmless, the author may reveal information that he did not intend others to see. In the case of Excel, this could be sensitive financial data that the user thought he had concealed. In the case of Word, this could be the list of contributors to the document and changes or comments attributed to them.

Information contained in headers and footers can also contribute to the inadvertent release of information. In Microsoft Word documents, headers and footers are not seen when the document view is "Normal." Headers and footers are seen in the "Print Layout" view. In Excel workbooks, headers and footers are not seen until the worksheet is printed. Although the user may not realize that header or footer information is present, it is saved along with the document and can be seen when the document is viewed with a hex editor. Headers and footers in PowerPoint presentations are visible when viewing the presentation in the normal manner.

It is clear that information the document owner did not intend to share with others could be released when the document is published in electronic form. In some cases, the information may be harmless, and its inadvertent disclosure, unimportant. On the other hand, the information may be sensitive financial or proprietary company data that could have unforeseen consequences for the document owner or for the company.

## **What Can be Done?**

Several Microsoft Knowledge Base Articles<sup>7</sup> explain how to minimize the amount of meta-data that is saved by the Microsoft Office suite of products. To remove the user name from Word documents, Microsoft Knowledge Base Article 290945 instructs the user to "enter non-identifying strings or spaces in the appropriate text boxes...." on the User Information tab found under Tools, Options. Merely deleting the existing text found in those text boxes is not enough to prevent the information from being saved with the document.

Personal summary information can be found on the Summary and Custom tabs under File, and then Properties, in Microsoft Word, Excel, and PowerPoint. Information on these tabs is saved with the electronic document, and the user may not wish to share this information with others. This information can be

deleted, and unlike the case of User Information, once deleted, the summary information will not be saved with the document.

The Microsoft Knowledge Base Articles<sup>7</sup> present a different set of instructions for removing personal summary information when connected to a network. It seems that if a user is connected to a network, the network user name may be saved even after the personal summary information has been deleted. The user is instructed to copy the document to the local hard drive and not to log onto the network in order to prevent this information from being saved with the document.

The “Fast Save” feature of Microsoft Word and Microsoft PowerPoint decreases the amount of time needed to save a document or presentation by saving only the changes that are made to the document or presentation. It is possible for some deleted text to be saved when this feature is enabled. Microsoft Knowledge Base Article 290945<sup>8</sup> and Microsoft Knowledge Base Article 314800<sup>9</sup> instruct the user to disable this feature if there is a concern that documents or presentations may contain text that the user deleted and had intended to remove from the document or presentation.

When macros are created in Microsoft Word, Excel, and PowerPoint, the recorded macro begins with a header that contains the name of the author who created the macro. The Microsoft Knowledge Base articles previously referenced discuss the steps that can be taken to minimize meta-data associated with macros in each of these documents, instructing the user to remove the author name from the recorded macro using the Visual Basic Editor.

Altogether, this collection of Microsoft Knowledge Base Articles cites 53 topics with instructions to be followed by the user to minimize the amount of meta-data saved with the documents. Some of the instructions are common to all of the Microsoft Office applications; others are unique to a specific application. Interviews conducted by this author of experienced users found that most were unaware that any meta-data was saved by these applications, and they did not know that certain configuration settings within the application could prevent this type of information from being saved with their documents.

## **Automated Tools**

It is likely that typical users of Microsoft Office products are unaware that information that they had no intention of sharing is saved with their documents, and the expectation that they will wade through the instructions provided in the Microsoft Knowledge Base articles to minimize the amount and type of information that is saved is, at best, low. It seems more likely that users, once they are aware that this information exists, would rather employ an automated tool to assist them with the elimination of the meta-data information saved with these documents.

An Internet search with search terms “metadata, Microsoft, removal” returned links to several commercially available utilities designed to remove meta-data from Microsoft documents. Most offer free downloads of trial versions of the utility. The list of utilities includes ezClean by KKL Software<sup>10</sup>, iScrub from Esquire Innovations, Inc.<sup>11</sup>, Metadata Assistant for Word, Excel and PowerPoint from Payne Consulting Group<sup>12</sup>, Doc Scrubber from Javacool Software LLC<sup>13</sup>, Out-of-Sight from SoftWise<sup>14</sup>, Workshare Protect from Workshare<sup>15</sup>, and Metadata Scrubber from BEC Legal Systems<sup>16</sup>. In January of 2004, Microsoft published an add-in to Office 2003/XP called rhdtool.exe that can be downloaded from the Microsoft Download Center and can be used to permanently remove hidden data from Microsoft Word, Excel, and PowerPoint files<sup>17</sup>.

Three of these utilities were randomly chosen, and their features and effectiveness compared when applied against a variety of documents created with Microsoft Office applications that contained one or more standard types of meta-data.

**ezClean.** After installation, the ezClean utility can be invoked from within Microsoft Word, Excel, or PowerPoint. When ezClean is invoked from within one of the supported Microsoft applications, the document is checked for meta-data, and the user is presented with a dialog box that shows the types of meta-data that were found and the number of occurrences for each type. Details for each type of meta-data found are also shown. The user can select to delete all, none, or individual types of meta-data shown on the list. The user is given the option to save the clean document with a different filename, thus preserving the document in its original form (with meta-data). The user can choose to “View Report,” which displays a printable report containing the meta-data that was detected in the document. ezClean can also be integrated with Microsoft Outlook. It can be configured to check outgoing attachments for meta-data, automatically remove detected meta-data, and check e-mail sent to addresses external to the company. ezClean configuration settings can be modified through the ezClean.ini file to provide a customized approach to meta-data removal. Although the ezClean Administrator’s manual provided with the tool states that ezClean can be invoked from the command line with a variety of customizable options, this author failed to accomplish this with the trial version of the tool.

The ezClean.ini configuration settings are divided into two categories; those that affect the behavior of the tool and those that affect how meta-data is handled for each of the applications supported by ezClean. Some of the settings that can be modified that affect the behavior of the tool include the showdlg setting, which controls whether or not the dialog box will be displayed when the tool is invoked, the logging setting, which controls whether or not a log is kept of any errors that occur during meta-data removal and any instances of successful meta-data removal, as well as the save setting, which controls whether a document will

automatically be saved after removal of meta-data, and the inspectionformat setting, which controls the format for the inspection report.

The configuration settings that apply to the handling of meta-data for each application are further divided into four categories; those that affect the default settings in the dialog box for each type of meta-data, those that control the default settings when the tool is run in AutoExec (unattended) mode, those that control whether or not the dialog box will appear in a certain document, and those that allow some items within each meta-data type to be exempt from removal. For the standard types of meta-data within each application, the settings can be configured to always remove it, never remove it, or let the user decide whether or not to remove it.

Some meta-data cannot be removed from documents using ezClean. The Administrator's manual provided with the ezClean utility lists the types of meta-data that ezClean cannot remove, including document statistics and the author, network, and printer history in Excel.

**Workshare Protect.** The Workshare Protect utility can be invoked from within Word and can also be run as a standalone application. The Workshare Protect documentation notes that in addition to Microsoft Word, Excel, PowerPoint, and Outlook, Workshare Protect also works with Lotus Notes and Groupwise email systems, and Interwoven and Hummingbird document management systems. When Workshare Protect is invoked from within Microsoft Word, the document is checked for meta-data, and the user is presented with a "Metadata Risk Report." The report displays the potentially damaging meta-data found in the document, if any. The user can choose to Clean the document, Print the report, Close the report, or view the Help files. If the user chooses to clean the document, a dialog box is displayed that shows the types of meta-data found. The user has the option to remove all of the meta-data found, none of it, or he can choose the individual types he would like to remove. The meta-data is then removed from the document. The user is not presented with an option to save the clean document with a different filename. Instead, he must remember to use "Save As" on the File menu if he wishes to save the clean document with a different filename, thereby preserving the original document (with meta-data).

Workshare Protect provides a configuration management tool that is invoked from within the standalone version of the Workshare Protect application, but some of the configuration settings only apply to documents that are sent as email attachments, either internal to or external to an organization. The configuration management tool allows the user to choose several types of meta-data to remove, including all macros and all Microsoft Excel and PowerPoint headers and footers. It also allows the user to exclude certain types of meta-data from removal when the document is cleaned. When Workshare Protect is run from within a Microsoft Word document, the default list of meta-data types that may be

discovered and can then be either reset or removed includes Document Statistics, Comments, Footnotes, Hidden Text, Document Versions, and Smart Tags. This list does not include headers, footers, or macros. Settings that affect the behavior of the Workshare Protect utility can also be modified from within the configuration management tool. The Help files list the types of meta-data that Workshare Protect can discover and remove, but the documentation provided with the application does not list those types that it cannot discover and remove.

After the trial version of Workshare Protect was installed on a test system, an icon to invoke the application was present in Microsoft Word documents, but was missing from Excel workbooks. When an Excel file was chosen from within the standalone version of the application, an error occurred implying that the format chosen was not supported by the tool. The Help file indicates that one of the main features of Workshare Protect is removal of meta-data from Microsoft Word, Excel, and PowerPoint documents, however, the Troubleshooting section of the Help file (and the User Guide) indicates that meta-data discovery is currently only supported for .doc, .dot, and .rtf files, with future versions to include meta-data discovery for .xls and .ppt format files.

**The Remove Hidden Data add-in.** The Remove Hidden Data add-in from Microsoft works with Microsoft Word, Excel, and PowerPoint 2002 and 2003. After the add-in is installed, a File menu option "Remove Hidden Data" is available in Microsoft Word, Excel, and PowerPoint. The user can open individual files and choose this option to remove the meta-data that would normally be saved with the document, workbook, or presentation. The user can also invoke the add-in on the command line to remove hidden data from multiple files. When the tool is started, the user is asked to choose whether to remove hidden data without prompting or to display prompts so that the user can evaluate the types of meta-data found and decide whether or not to remove them. The user is then prompted to save the cleansed document with a different filename, thus preserving the original (with meta-data). The tool then performs a discovery for meta-data and prompts the user to either remove or keep the data for certain meta-data types found. The user is not given the chance to review all types of meta-data found. For example, comments are automatically deleted without user input. A log file is then displayed containing the details of the discovery and removal process.

Documentation provided with the tool lists the types of meta-data that can be removed by the tool. The list includes user name, comments, personal summary information, headers and footers, file paths, hidden text, and hyperlinks. There are no configuration files to permit customization of meta-data types included for or excluded from removal. The documentation states that the tool can be run on the command line with a variety of options. One option, /M, allows the user to remove all macros in the document files chosen to cleanse. When the tool is run from the command line, the user is not prompted to decide whether or not to

remove the meta-data that is found. Instead, an option, /A, can be used to generate a report of the meta-data found without removing any of it. This author found that the /A option did not generate a report showing the meta-data found. Instead, the usual log file seen at the end of the process when the tool is invoked from within the document was produced. No meta-data was removed. The /M option did not result in the removal of macros from the Microsoft Word document used.

As previously noted, when macros are created in Microsoft Word, Excel, and PowerPoint, the recorded macro begins with a header that contains the name of the author who created the macro. None of the automated tools tried by this author removed the macros or their headers from the documents. They are easily viewed with the Visual Basic editor, and include the author's name, unless the author removes this information manually. If the author does not enter non-identifying strings or spaces in the "Name" box on the User Information tab found under Tools, Options, then each time a new macro is created, the header will be created containing the name found in this box.

The automated tools tried by this author do not offer a complete solution to the removal of the meta-data found in the electronic versions of Microsoft Word, Excel, and PowerPoint documents. For example, Workshare Protect and ezClean do not identify headers or footers in Microsoft Word documents, whereas the Remove Hidden Data add-in prompts the user to save or remove headers and footers. The Remove Hidden Data add-in automatically removes comments, with no input from the user, whereas ezClean and Workshare Protect allow the user to decide whether or not to remove comments. None of the tools identified macros and did not remove them or their associated header information. In addition, the resultant "clean" files produced after running the Remove Hidden Data tool and Workshare Protect from within a Microsoft Word document were a few kilobytes *larger* than the original file.

## Conclusion

Whenever a Microsoft Office document is created, opened, or saved it may contain hidden data, or meta-data, that includes user information, such as the user's name and initials, company name, the name of the computer, the path to the location where the document was saved, the names of previous document authors, hidden text or cells, comments, and other information about the document itself. The document owner may not be, and usually isn't, aware that the electronic form of the document contains anything more than what he/she has written and can see on the computer screen. This could lead to the inadvertent disclosure of sensitive or proprietary data when the electronic versions of these documents are shared with others.

Microsoft and other companies have recognized this issue and have developed automated tools that can be employed to find and eliminate some types of meta-

data that is saved with Microsoft Office documents. Microsoft has also published several Knowledge Base Articles that address this issue and provide instructions for the modification of configuration settings within the various Microsoft Office applications to minimize the amount of meta-data that is saved. These tools and instructions do not address document content, and it is up to the author of the document to ensure that the content is free from sensitive or proprietary data prior to release of the document.

The first line of defense in the prevention of inadvertent release of sensitive or proprietary data is user education. If the user is unaware that information can be saved without his knowledge, he cannot begin to take steps to prevent that from happening. Manual modification of certain configuration settings found within the various Microsoft Office applications combined with the use of automated meta-data removal tools can effectively eliminate the possibility that potentially sensitive information contained in the meta-data will unintentionally be shared with others.

© SANS Institute 2004, Author retains

## List of References

1. Smith, Richard M. "Microsoft Word bytes Tony Blair in the butt." June 30, 2003. <<http://www.computerbyesman.com/privacy/blair.htm>>.
2. Shankland, Stephen and Ard, Scott. "Hidden text shows SCO prepped lawsuit against BofA." March 4, 2004. <[http://news.com.com/2102-7344\\_3-5170073.html](http://news.com.com/2102-7344_3-5170073.html)>.
3. Byers, Simon. "Scalable Exploitation of, and Responses to Information Leakage Through Hidden Data in Published Documents." 2003/04/03. <[http://www.user-agent.org/word\\_docs.pdf](http://www.user-agent.org/word_docs.pdf)>.
4. Foss, Kurt. "Washington Post's scanned-to-PDF Sniper Letter More Revealing Than Intended." 26 October 2002. <<http://www.planetpdf.com/mainpage.asp?webpageid=2434>>.
5. CorporateWatch.org. "Microsoft." February 2004. <<http://www.corporatewatch.org.uk/profiles/microsoft/microsoft.pdf>>.
6. Microsoft Corporation. "OFF: How to Minimize Metadata in Microsoft Office Documents." 11/25/2003. <<http://support.microsoft.com/default.aspx?scid=kb;en-us;223396>>.
7. Microsoft Corporation. "HOW TO: Minimize Metadata in Microsoft Word 2002." 10/19/2003. <<http://support.microsoft.com/default.aspx?scid=kb;EN-US;290945>>.  
Microsoft Corporation. "XL: How to Minimize Metadata in Microsoft Excel Workbooks." 5/13/2003. <<http://support.microsoft.com/default.aspx?scid=kb;EN-US;223789>>.  
Microsoft Corporation. "PPT2002: How to Minimize Metadata in Microsoft PowerPoint Presentations." 3/18/2003. <<http://support.microsoft.com/default.aspx?scid=kb;EN-US;314800>>.
8. Microsoft Corporation. "HOW TO: Minimize Metadata in Microsoft Word 2002." 10/19/2003. <<http://support.microsoft.com/default.aspx?scid=kb;EN-US;290945>>.
9. Microsoft Corporation. "PPT2002: How to Minimize Metadata in Microsoft PowerPoint Presentations." 3/18/2003. <<http://support.microsoft.com/default.aspx?scid=kb;EN-US;314800>>.
10. KKL Software. "ezClean – Metadata removal utility for Microsoft Office." 2004. <<http://www.kklsoftware.com/products/ezclean/details.asp>>.

11. Esquire Innovations, Inc. "iScrub – Premier Metadata Removal Utility For Microsoft Office (Word, Excel & PowerPoint)." <[http://www.esqinc.com/printable/products\\_iscrub.htm](http://www.esqinc.com/printable/products_iscrub.htm)>.
12. Payne Consulting Group. "Metadata Assistant for Word, Excel and PowerPoint." 2004. <<http://www.payneconsulting.com/public/products/ProductDetail.asp?nProductID=34>>.
13. Javacool Software, LLC. "Doc Scrubber 1.1." 2003. <<http://www.docscrubber.com>>.
14. SoftWise. "Welcome to a world free of metadata." 2001-2004. <<http://www.softwise.net/prod5.html>>.
15. Workshare. "Workshare Protect Total document protection and control." 2003. <[http://www.workshare.com/products/pr\\_w3protect\\_overview.htm](http://www.workshare.com/products/pr_w3protect_overview.htm)>.
16. BEC Legal Systems. "Metadata Scrubber." <<http://www.beclegal.com/lsy/lsylegspeclegbarmetadata.asp>>.
17. Microsoft Corporation. "Office 2003/XP Add-in: Remove Hidden Data." 1/5/2004. <<http://www.microsoft.com/downloads/details.aspx?FamilyID=144e54ed-d43e-42ca-bc7b-5446d34e5360&displaylang=en>>.