

Accidental publicist

Agencies risk unwitting release of sensitive information using popular office software

BY [Michael Arnone](#)

Published on Apr. 10, 2006

A new front line of national and corporate security is emerging, and some of the most common document applications, including Microsoft Word documents and PDFs, are putting people on it without their knowledge. In the past several years, federal agencies and private-sector companies have released documents on the Internet that they thought did not contain sensitive content, but they actually did. That has led to embarrassment, scandals, firings and national security breaches when unintended readers discovered the hidden data.

At least 20 press reports between October 2000 and December 2005 show that the release of hidden, sensitive data is a serious and pervasive problem. For example, a July 2005 Pentagon report on cyberattackers, saved as a PDF, included data hidden in the structure of the documents that listed the IP addresses of attacked Defense Department computers, making them vulnerable to future assaults.

The New England Journal of Medicine revealed last December that Merck had deleted information connecting its Vioxx painkiller to an increased risk of heart attack from a major study on the drug that the company submitted in 2000. The authors wrote the study in Microsoft Word, which retained the deleted text as part of the application's Track Changes function. Merck stopped selling Vioxx in 2004 and is paying hundreds of millions of dollars on thousands of lawsuits based on health problems and deaths linked to the drug.

A March 2004 study by the Institute of Electrical and Electronics Engineers found that of 100,000 documents surveyed, half contained 10 to 50 hidden words, one-third had 50 to 500 hidden words, and 10 percent had more than 500 hidden words.

"Our society spends millions of dollars protecting information from hackers and malicious insiders while spending almost nothing to prevent sensitive information from leaking out in legitimate and routine electronic document exchanges," said Ronald Hackett, program manager at SRS Technologies' Systems Solutions Division. The company sells software that finds and removes hidden data.

7 types of files to watch for

These seven common document formats often contain hidden data problems.

- Microsoft Word (.doc)
- Microsoft Excel (.xls)
- Microsoft PowerPoint (.ppt)
- Microsoft Rich Text Format (.rtf)
- Hypertext Markup Language (.html)
- Extensible Markup Language (.xml)
- PDF (.pdf)

Microsoft has created the Remove Hidden Data (RHD) plug-in for Office 2003, said Gray Knowlton, a senior product manager on Microsoft's development team for the Office application suite. RHD removes hidden data saved by the Track Changes function or in comments in Word, Excel and PowerPoint files. Another tool, Word Redactor, helps users redact information.

Adobe Systems has a PDF Optimizer tool that can examine hidden data, said Gregory Pisocky, a business development manager at Adobe.

Analysts disagree whether the tools work. Andrew Jaquith, a senior analyst at the Yankee Group, said they do.

Stacey Quandt, research director for security solutions and services at the Aberdeen Group, a research firm, said RHD finds what it looks for but can't find all the potential hidden data in a document.

SRS Technologies has created Document Detective, which extracts all hidden data for users to review and remove as they wish, said Ronald Hackett, a program manager at the company. Document Detective provides automatic warnings based on the organization's security policies for sharing information. The company's product also includes a redaction tool.

— *Michael Arnone*

“Ironically, the biggest threat to sensitive information may be honest users just doing their jobs,” Hackett said.

A lot of work must be done to educate users and vendors about document security, said Paul Stamp, a senior analyst at Forrester Research. Government and industry users need tools to access and deal with hidden data. He said document management vendors are beginning to recognize the problem of accidentally releasing sensitive data.

“It’s something we’re aware of,” said Gray Knowlton, a senior product manager on Microsoft’s development team for the Office application suite. “It’s something we spend a lot of time thinking about.”

But is enough being done to prevent avoidable losses? With few tools and little training available to teach people how to remove hidden information, Hackett and other analysts said, the problem is likely to continue.

The WYSIWYG problem

The causes of much of the hidden data problem are users’ ignorance of how digital documents work and software companies’ tendency to give customers too much of what they want — ease of use and flexibility. The core of the issue is the “what you see is what you get” (WYSIWYG) concept, a driving force behind the evolution of application user interfaces for the past 30 years. The idea is to conceal software’s inner workings from users so that the documents on their screens seemingly mirror how the final versions will appear to others.

Whereas paper documents have only two sides separated by a fraction of an inch of pressed wood pulp, digital documents are small file systems within their larger applications. They can contain reams of material — including metadata, older versions and deleted items — in multiple layers that don’t appear on the screen or in printouts.

“Paper is WYSIWYG,” said Andrew Jaquith, a senior analyst at the Yankee Group. “What you see in an electronic document is not necessarily what you get. It’s everything ever done with the document that may still be in it.”

WYSIWYG interfaces encourage users to act as they would in the real world, which provides a false sense of security, Jaquith said. Removing data from a digital document is not the same as using an eraser or a permanent marker on a paper one.

Another common problem with hidden data comes when application vendors make it too easy to use new software features that have unforeseen consequences, Hackett said. He cites the Ad Hoc Review function, a document-sharing tool, in the Microsoft Windows XP operating system.

Without alerting the user, Windows XP automatically starts the Ad Hoc Review with Tracked Changes function when someone using the Outlook e-mail client sends or replies to a Word, Excel or PowerPoint document, Hackett said. This function stores complete copies of every version of the document, even though only the final version is immediately visible. He and Knowlton disagree whether the feature is easily disabled to prevent inadvertent data release.

Hackett said he reviewed 101 federal documents last December and found that the Ad Hoc Review option was enabled on 30 percent of them.

Outlook automatically turns on Track Changes because the program presumes that the user wants to compare changes others make to the original document, Knowlton said. Duyen Truong, a Microsoft spokeswoman, disagreed with Hackett’s claim and said people other than the document creator can turn off Track Changes by accessing the Reviewing toolbar under the Tools menu.

Microsoft should remove the automatic feature from Outlook and warn users more about how the Ad Hoc Review function works, Hackett said.

Who's responsible?

Most people don't know or forget that applications track changes, Jaquith said. That underscores the common opinion that the problem is the software's use, not its development. "It's a classic case of folks not necessarily reading the owner's manual for these things," he said.

No one agrees, however, who is ultimately responsible for training people in how their software works and how to remove hidden data so they don't unwittingly release sensitive information. Software companies offer enough tools, training and information for users to adequately protect sensitive information, said Knowlton and Gregory Pisocky, a business development manager at Adobe Systems.

Hackett agrees that improper use is the issue but added that companies are not blameless. "They make the user responsible and wash their hands of it," he said. Pisocky said software companies are not responsible for warning users about hidden data because the document software has no way to determine whether the unseen information is sensitive.

Users should buy third-party tools that find hidden data, enable human review and remove what needs to go, Hackett said.

Until software companies improve their products, users must ensure they don't reveal hidden data, Jaquith said. They should be aware of whether documents track changes and when they redact information, he said.

Users shouldn't be blamed for releasing hidden data that they didn't know was there, said Stacey Quandt, research director for security solutions and services at the Aberdeen Group, a research firm. The responsibility falls on organizations, which must establish policies that account for the risks of the technology they use, she said.

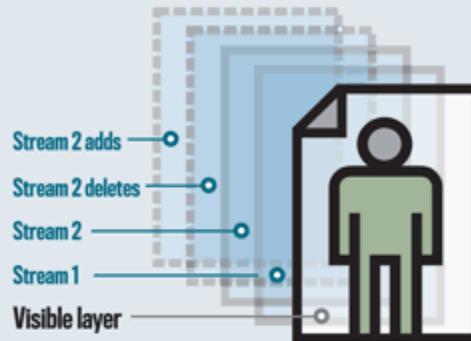
Microsoft has added a number of relevant features to its Office 2007 suite, due in January 2007, Knowlton said. Office 2007 contains an upgraded version of its Remove Hidden Data plug-in tool, called Document Inspector, which detects and removes hidden text, document properties, headers and footers, and all kinds of annotations and changes. The suite will not contain the Send as Attachment for Review function, which also enables Ad Hoc Review.

Pisocky declined to comment on whether future versions of Adobe software will do anything to help users find and control sensitive information.

Dissecting digital documents

Many popular electronic document formats rely on the following techniques to capture information and user activity, which can pose risks for information security.

Layers: Digital documents consist of multiple layers of data, including many that users typically don't see. Layers can contain deleted information, tracked changes and metadata, such as information about previous users, earlier versions and file names.



Embedded objects: When files such as photos and spreadsheets are embedded as objects in electronic documents, the entire file is included, not just the part visible on the page. Cropping or resizing images does not get rid of the full file's data, which could include sensitive information. Even if users place objects off to the side of the page, those objects are still in the file. Furthermore, embedded objects can contain other embedded objects.



Redaction: People sometimes try to redact text by drawing black boxes over it. Although it makes the text unreadable in normal viewing mode, the box does not remove the text. It is simply another object or layer that others can easily remove or sidestep to reveal the underlying information.

