



## Without a trace

02/20/06

By Joab Jackson,  
GCN Staff

**Documents can reveal supposedly deleted data, but fortunately there are ways to make sure that secrets stay secret.**

E-government is all about getting more information in the hands of the citizenry, but maybe the Office of Management and Budget inadvertently carried that idea too far.

In 2002 when Mark Forman, then associate director for IT and e-government at OMB, sent out a document on how to implement the White House's E-Government Strategy, he offered a bit more information than he thought. The final Word document also contained the last few revisions it went through before heading out the door. Oops.

Fortunately for OMB, the revisions added little more to the public record than simple copy editing changes; no deep, dark secrets were revealed. But it wouldn't be the last time data seeped out from the hidden crevices of an electronic document.

### Out of the black

Last April, when the Defense Department's Multi-National Force-Iraq unit issued a report on a shooting investigation, it redacted certain portions that were sensitive. But it wasn't a redaction job well done. An Italian blogger pasted the text of the document into Microsoft Notepad and uncovered the sections that had been blacked out in the published Acrobat Portable Document Format file. Oops.

And late last year, the New York Times pried open a Word document of a presidential speech and discovered that the originator of the White House document (and by extension of the speech

#### The document scrubber's toolbox

Agencies grappling with ways of ensuring their documents are more secure have several resources at their disposal. The tools and guidance below can be found at [www.gcn.com](http://www.gcn.com) by entering the appropriate number in the Quickfind box.

**Microsoft Office 2003/XP: Remove Hidden Data Add-in.** It's only a 270K download and works with Windows 2000 SP4 and Windows XP SP1. Get it now. (Quickfind 532)

**Find and Remove Metadata (Hidden Information) in your Legal Documents.** This Microsoft resource is helpful for anyone who uses Word, Excel and PowerPoint. (Quickfind 533)

**Redacting with Confidence: How to Safely Publish Sanitized Reports Converted from Word to PDF.** The National Security Agency guide should be mandatory reading for all government users who create PDF documents. (Quickfind 534)

**Trace.** Workspace's free tool for uncovering metadata and hidden data in Office documents. (Quickfind 535)

**Document Detective.** The software from SRS Technologies of Hunstville, Ala., reviews Microsoft Office and Adobe PDFs for hidden metadata, showing the results. A related product, the Electronic Document Review System, removes hidden and extraneous data. (Quickfind 536)

itself) was not among Bush's usual cadre of speechwriters. He was a special adviser with an expertise in swaying public opinion. Oops.

The Justice Department, United Nations, United Kingdom and more than one commercial organization have all suffered similar embarrassments. And such information leaks are starting "to have a higher cost" to organizations, said Ken Rutsky, executive vice president of marketing for Workshare Inc. of San Francisco.

Fortunately, there are simple ways to prevent careless information leaking. Last December, the National Security Agency released guidance on how to clean up your documents before sending them out to the world. The document, *Redacting with Confidence: How to Safely Publish Sanitized Reports Converted From Word to PDF*, is a good start, but CSOs, office managers and system administrators should know of other dangers lurking in their office software—and how to root them out.

### **Feature rich, security poor**

All the aforementioned cases show how unsuspecting agencies suffer the effects of feature-rich software, said Ronald Hackett, program manager for SRS Technologies of Huntsville, Ala. SRS makes software that can review and remove the hidden data within PDFs and Office documents.

Software vendors have been eager to add new features to their products to keep customers upgrading. At the same time, they've been making those products as easy to use as possible, meaning more application behavior gets moved into the background, where it goes unnoticed by untrained users. The downside of this approach is that agency workers are usually unaware that applications are performing certain actions, such as tracking changes or collecting data about the user.

As a result, documents tend to collect hidden data, which takes two forms. One is metadata, or data about a document, which is appended by the program, often unnoticed by the user. Then there's data that was part of the original document but somehow has been rendered invisible to the average user. Knowing how to find both types of hidden data is critical to getting rid of it.

### **Metadata**

Last October, Harlan Carvey, a Washington-based security professional and author of the book *Windows Forensics and Incident Recovery*, tried a little experiment. He downloaded a Word document from the Office of Naval Research to find what information he could about the creation of the document from the document itself.

Carvey had written a Perl program that would extract data from the document's file information block, the index usually found at the entryway of the file. When he ran the ONR document through this script he had teased out a variety of information, some of it innocuous. He was able to glean the names of some of the individuals who edited the document, the file path, the version of Word used to compose it and when it was created. He posted the results on his blog.

Metadata is often considered complex technology necessary for sharing applications and data at an enterprise level. But in this case metadata is just the basic information a program may collect about the origins of a document. It could be useful in searching or other advanced features, but it's not necessarily the type of information that must be kept secret. It's the rare instance, such as revealing who created the White House speech, that document metadata contains details that agencies don't want to disclose.

Still, to help get rid of all that metadata, Microsoft Corp. offers a removal tool, which is available both as an option within Office XP, and in the form of a plug-in for batch jobs (see sidebar, below). But NSA warns that these applications are unstable and do not remove all unwanted data. "Reliance on these tools may give a false sense of security," the report concluded. (Microsoft Corp. declined to participate in this story and NSA did not further elaborate on the instabilities.)

PDF files also carry metadata. Open a PDF with Adobe Reader, then click on the Document Properties option in the File menu. If it were specified, this is where you'd find the document's title, author, subject, creation date, what program created the PDF and other assorted bits of data.

Much of this metadata gets pulled into the PDF when it's created; programs such as Adobe Acrobat Distiller and PDFMaker Add-in grab it from the source document. The good news is that Adobe's software, like most other PDF converters, can be configured not to bring this metadata over from the source document.

### **Hidden data**

But it wasn't exposed metadata that brought shame to the Multi-National Force-Iraq. It was hidden data, or data that the user thought was removed but still existed.

The lesson was that simply blacking out text does not remove it, said John Landwehr, director of security for Adobe Systems Inc. To redact, the employee had set the text background color to black, making it appear blacked out. When the document was converted to PDF, all of the original text was carried over, along with the black-on-black formatting.

Even when you've successfully expunged all the metadata from a file, you still have to deal with data that may be invisible to the naked eye. Some is as obvious as black lettering on black text, but the application software can also create more than its fair share of hidden data.

Change tracking, for instance, creates a lot of potentially hidden data. A fundamentally useful idea, change tracking allows documents, as they get passed around the office, to keep track of which user made which changes.

In untrained hands, however, the Track Changes feature in Office can lead to trouble. A document's originator may make the potential mistake of turning on the tracking feature but not choosing the option of highlighting the changes on screen. In which case a user may not realize his or her changes are being logged. Nor may the final editor of a document realize that the

change tracking must be turned off and the changes must be merged into a final version of the document.

In many cases, however, end users aren't to blame for having their changes captured. Hackett described a quirk of the Microsoft Windows/Office environment where Microsoft Outlook, Microsoft's e-mail client, surreptitiously starts the change tracking in a document, even when the user hasn't turned on the feature.

The upshot is when you e-mail a PowerPoint presentation, Excel spreadsheet or Word document to another party, the change tracking is automatically on and, as the file makes it rounds, you know who works on it. To turn off this feature in Outlook, go to Tools>Options>Preferences>E-Mail Options>Advanced E-mail Options and unclick the box next to "Add properties to attachments to enable Reply with Changes."

The fact that this feature is enabled by default is problematic. But it's potentially compounded by the fact that once Outlook starts a document's change tracking, it can only be turned off by the owner of that document, Hackett said.

Another potential weak point of office products is their ability to act as a container for other types of files. Today's Office documents, spreadsheets and presentation slides can hold movies, audio recordings, images, sections of data from other Office documents. While this feature is great for, say, assembling training material, it can also harbor untold amounts of data that may not be visible to someone inspecting the document.

Take images, for instance. When someone embeds an image in a document, instead of cropping it down to an appropriate size, the author may instead just shrink the frame of the image so that it only shows the relevant part of that image. The trouble with this approach is that the entire image is still accessible to other users, not just the visible portion, Hackett said.

Documents can also contain other files that are not visible simply because the author shrank them down in the document to a small size, or they blend in with the background colors. A video that starts off with a color that is identical to the background of a document might not be noticed, Landwehr noted.

PDFs also can contain older document pages that are not visible to most users but can be retrieved with computer forensics or hacking tools, Hackett said. These include images that may sit underneath images that were subsequently placed in the document.

Microsoft Office itself has a feature that allows data from one Office product to be embedded in another. It's easy, for example, to display a pie chart created in Excel inside a Word document, Hackett said. The downside that few users realize is that when they drag data from Excel to Word they're not just embedding a graph; they're including the entire Excel spreadsheet inside the Word file, Hackett said. Another user could open the file and extract the entire worksheet.

In fact, in some cases you can change the .doc extension of a Word file with an embedded spreadsheet to .xls and open up the document in Excel with only the spreadsheet displayed,

Carvey said. “It provides an interesting means of getting data out of an organization, or passing around illicit data.”

### **NSA’s approach**

All this extraneous information is enough to give a system administrator a headache, but NSA provides some good advice. Although written as a manual for performing formal redactions from official documents, NSA’s tips can be useful for informal document sharing.

To cleanse a document of metadata and other hidden forms of data, NSA advises users to first turn off the track changes feature in the document to be released, then remove the sensitive data. After all the sensitive data is expunged, open a new document, giving it a nonrevealing name, and paste the allowed information into that document (NSA calls this “residual document composition information”). After that, NSA recommends that users convert the Word document into a PDF file.

This entire approach, although fundamentally sound, can be cumbersome, Carvey noted. The manager that enacts these rules as standard office procedure could find employees spending a lot of time slogging through each step.

When NSA’s guidance gets too burdensome, it could be time to consider automated tools. Experts say electronic redaction, in particular, is a high-stakes, sensitive activity, prone to human error, and not something that should be left to human judgement. Companies such as ZyLAB North America LLC of McLean, Va., offer redaction software that covers the dumb gotchas and can automate common tasks. Microsoft plans to offer a robust redaction tool in its upcoming Office 12 feature.

But as agency managers come to understand the volume of documents under their control, each one a potential embarrassment, they’ll probably start considering an officewide document cleansing process. Workshare and other companies offer software that can check documents as they leave the internal network environment, stripping them of unnecessary metadata and sensitive information. And several e-mail security appliances can now be configured to block certain outgoing messages as easily as they block incoming spam.

Said Workshare’s Rutsky, “[Agencies] come to us when they realize the magnitude of the risk.”

© 1996-2006 Post-Newsweek Media, Inc. All Rights Reserved.

[http://www.gcn.com/25\\_4/tech-report/38253-1.html](http://www.gcn.com/25_4/tech-report/38253-1.html)