MARYLAND STATE BAR ASSOCIATION, INC.

**April, 2004**

# Document Security

By John Anderson

Usually with Microsoft Word, what you see is what you get.

If you make a change to a document, then that is what you see when it is printed out. But in fact in many cases it is what you cannot see at first glance that proves more interesting. There is a function in many versions of Microsoft Office programs (which include Word, Excel and PowerPoint) that means that fragments of data (which Microsoft refers to as metadata) from other files you deleted or were working on at the same time could be hidden in any document you save.

If you work on a file with others trading edits and tracking your changes, Word gathers metadata about your work, and this information can be accessed by anyone who opens the document. Even if you work by yourself, it tracks your edits. With the right tools this hidden data can easily be extracted.

## What's at Risk

Computer researcher Simon Byers has conducted a survey of Word documents available on the Internet and found that many of them contain sensitive information. He gathered about 100,000 Word documents from sites on the Web; every single one of them contained hidden information.

In a research paper about the work, Byers wrote that about half of the documents gathered had up to 50 hidden words, a third up to 500 words hidden and 10 percent contained more than 500 words concealed within them.

The hidden text revealed the names of document authors and their relationship to each other as well as earlier versions of documents. Occasionally it revealed very personal information such as Social Security Numbers - the lifeblood of identity thieves. Also available was useful information about the internal network through which the document traveled, which could be useful to anyone looking for a route into a network.

The seriousness of this problem become apparent when Secretary of State Colin Powell, in front of the United Nations, cited a British government high-level intelligence dossier about Iraq as providing one of the reasons why the world should go to war against Iraq. Not long after the dossier was publicly released, a lecturer in politics at Cambridge University did a little bit of sleuthing and found that the dossier was little more than a cut-and-paste job (rather than containing high-level intelligence, as the government claimed) that had been primarily copied directly from three publicly available articles, one of which had been written by a postgraduate student in the U.S. In fact, the dossier even had the same typographical errors found in the student's original article.

The dossier was a Word .doc file. The lecturer was able to extract the history of the last 10 edits to the file, including the names of the people who had edited it.

## What to Do About It

There are several methods which will reduce or eliminate the chance of passing along unwanted data. The first is to not send Word files electronically. Send a hard copy either by mail or fax, or scan the document and save it in another format. These methods completely eliminate the hidden data.

If you want to stay with Word, there are still ways to reduce the amount of hidden data. One method is to copy and paste the information into a new, blank Word document before e-mailing it and then send the new document along. Doing this will eliminate the revisions, comments and earlier versions of the document.

Another approach is to save it in Rich Text Format (.rtf). Doing this preserves all the fonts and formatting while removing some of the hidden data. To do this, click on "File," "Save As" and in the "Save as type" box select "Rich Text Format."

Microsoft issued a download for Office 2003/XP to allow users to "permanently remove hidden data and collaboration data, such as change tracking and comments, from Microsoft Word, Microsoft Excel and Microsoft PowerPoint files."

To use the add-in, install it, open the file you want to modify and choose Remove "Hidden Data" from the "File" menu. You can also modify multiple files from the command line. You'll find the instructions in the 1033 subdirectory of the removal tool's folder. [For example: C:\Program Files\Microsoft Office\Remove Hidden Data Tool\1033]

Another method to ensure that no one gains access to revision history is to convert your document to a PDF document using the tools available on Adobe's website (www.adobe.com), which will create an un-editable document that is free of personal data and revision history.

## Even PDF May Not Be Safe

The Washington Post published a scanned-to-PDF version of a handwritten letter left at the scene of one of the recent sniper shootings, allegedly written by the killers and intended for the police. The Post published the downloadable version of the "Ashland Sniper letter" to illustrate its article explaining how police were able to decipher significant additional information from both revealed and unintended clues communicated by the letter's author. In the letter, the sniper demands a $10 million dollar ransom and explains how that money should be delivered – deposited in a specific, stolen credit card account. Certain personally identifying details are blacked out in the PDF file.

However, the creators of the scanned PDF (it's unclear whether it was produced at the newspaper or elsewhere) have themselves revealed more information than they intended.

Anyone using the full commercial version of Adobe Acrobat software (NOT the free Acrobat Reader) to display the PDF can very easily remove the blacked-out areas intended to hide certain details. The PDF is simply an image file to which an added layer of black has been added. By choosing Acrobat's TouchUp Object Tool, then selecting a particular section of the darkened area, one can easily drag the overlay away from the text it is meant to protect, clearly revealing details such as the name and account number of the person whose credit card the snipers were attempting to use to stash their ransom cash.

There is commercial software available (such as Redax from Appligent - www.appligent.com/products/plug_ins/redax/redax.html) that works with Adobe Acrobat for this exact situation. Many government agencies commonly use it to redact or extract certain bits of information from private documents so that they can be made publicly available. Redax actually removes the selected text (including text within a graphic) and replaces it with meaningless blocks; the redacted information is "permanently removed from the PDF stream," according to Appligent.