

Online document search reveals secrets

15 August 2003

NewScientist.com news service

Will Knight

Many documents published online may unintentionally reveal sensitive corporate or personal information, according to a US computer researcher.

Simon Byers, at AT&T's research laboratory in the US, was able to unearth hidden information from many thousands of Microsoft Word documents posted online using a few freely available software tools and some basic programming techniques.

Sophisticated editing programs will often store information in a document file that the end user will not see. Storing recently deleted text can, for example, make editing a more efficient process. But Byers says it could also expose unaware users to significant risks.

In his report, Byers suggests that a crook could analyse electronic documents to gather information that could help them carry out corporate espionage or steal someone else's identity to commit fraud.

"It is feasible that an individual may include their social security number on copies of a resume sent to prospective employers, but delete it from the version put online to guard against identify theft," Byers writes.

Random words

Using an ordinary online search engine and a random selection of keywords, Byers was able to find more than 100,000 Word documents including business documents and individual resumes. He chose to examine Word files because they are so common and stresses that other document formats can contain similar hidden information.

For example, in 2002 the *Washington Post* published a version of a letter sent by the Washington sniper in Adobe PDF format. Names and telephone numbers were visibly blacked out, but still found embedded in the file. However, Byers's new research reveals how widespread such problems could be.

After downloading the Word files, Byers used the free software tools "antiword" and "catdoc" to convert them to plain text. He then wrote a simple script to locate text that was not displayed in the original Word format. Byers discovered a wealth of deleted text and potentially sensitive information including people's names, email headers, network paths and text from related documents.

Bruce Schneier, of US security consultants Counterpane, discusses the research in the latest edition of his computer newsletter *Crypto-Gram*, published on Friday. He says it raises an important risk with using some document formats. "The worst is erased text," Schneier told **New Scientist**. "This has bitten people surprisingly often."

Blacked out

Neil Laver, UK group marketing manager for Microsoft Office products, says the software company is working to develop better ways for customers to ensure sensitive information is not inadvertently left in files.

He says hidden information can "incredibly useful" in improving the functionality of the software. "But if some of that data is sensitive, there have to be ways of ensuring that it isn't distributed where it shouldn't be," he says.

The next edition of Office 2003 will include tools that will allow users to remove personal information from a document. It will also include new "information rights management" that will let an author specify who can read or forward a document.

Other software programs can already be used to strip concealed text from documents. But Schneier says for the time being it may be best to convert documents to plain ASCII text before publishing online. "I don't know of any programs that effectively clean out the extra text," he says.

Byers' paper has been submitted for publication in the IEEE journal *Security and Privacy*.