AN INVESTIGATION INTO UNINTENTIONAL INFORMATION LEAKAGE THROUGH ELECTRONIC PUBLICATION

Jock Forrester

Departments of
Computer Science and Information Systems
Hamilton Building, Rhodes University,
Grahamstown, 6139

Email: J.Forrester@ru.ac.za

Barry Irwin (CSSIP)

Department of Computer Science
Hamilton Building, Rhodes University,
Grahamstown, 6139

Email:B.Irwin@ru.ac.za

ABSTRACT

Organisations are publishing electronic documents on their websites, via email to clients and potentially un-trusted third parties. This trend can be attributed to the ease of use of desktop publishing / editing software as well as the increasingly connected environment that employees work in. Advanced document editors have features that enable the use of group editing, version control and multi-user authoring. Unfortunately these advanced features also have their disadvantages. Metadata used to enable the collaborative features can unintentionally expose confidential data to unauthorised users once the document has been published.

There have been numerous well publicised occurrences of hidden data being discovered in freely available electronic documents tarnishing the reputations of well known organisations, newspapers and governments.

Hidden data in documents is also of interest for digital forensic investigators. Depending on the type and quantity of metadata left in the examined documents and application that the documents were created with, the investigator can prove who viewed or edited the document and attribute a document to a particular machine and / or person.

This paper outlines the types of data that can be extracted from documents through the use of freely available tools. It has also outlined potential uses for this information from the point of view of a digital forensic investigator and for possible exploitation. Finally it recommends the types of organisational policies and procedures to be put in place to prevent unauthorised data from being inadvertently published to the outside the world.

KEY WORDS

Metadata, Hidden Data, Hidden Information, User awareness, Electronic Publication, Document Attribution

AN INVESTIGATION INTO UINTENTIONAL INFORMATION LEAKAGE THROUGH ELECTRONIC PUBLICATION

1 INTRODUCTION

Early document formats contained formatting structure what interspersed in the text of the file, similar to HTML, when WYSIWYG (What You See Is What You Get) editors were developed the formatting commands became embedded in the file. This vastly improves the editing process and the functionality offered in editing documents however it does mean that information may be hidden in the document [1].

Advanced document editing applications, such as Microsoft Word and Corel WordPerfect often store metadata in the document's format that improves the functionality offered to the end user whilst authoring documents. This hidden data can range from the Author's name to complete previous versions of the document.

Byers [4] notes that the hidden data is crucial to offering an advanced editing application, however, if some of the metadata used to provide the advanced functionality is sensitive, there must be methods in place to ensure that the meta data is not distributed with the document. These methods include software placed on the exit points of the organisation's network and organisational policies regarding electronic publication and distribution.

Hidden data in electronic documents has arisen due to software that the user does not fully understand and the propriety file formats that are complex in nature and the near common place exchange of documents via electronic means [5].

Electronic document preparation and distribution is a routine and inevitable component of current business trends [2]. Payne [3] estimates that more than 75% of all newly created documents are shared, or collaborated on electronically. Copies of files can be made swiftly and documents can be sent nearly instantly to recipients via email.

Users need to be trained and equipped with the tools to remove hidden data from their documents before publication. Ideally this should be an automatic process that happens when the document passes outside of the organisations digital borders.

2 WHAT INFORMATION CAN BE FOUND?

Byers [4] discovered that out of one hundred thousand Microsoft Word documents retrieved from the internet nearly 50% of the documents had 10 to 50 hidden words, one third of the documents revealed between 50 and 500 words and 10% had more than 500 words embedded in the document.

The information that can be stored in Microsoft Office documents will not always obvious or visible to the user [2]. Common causes for hidden data being included in documents are briefly discussed below.

- Versions:
  Word can store multiple versions of the document as created through the editing process. Where there are multiple versions of the document, word displays a small icon on the bottom task bar, which can be easily overlooked.
- Track Changes:
  Word can track the changes made to a document by the various authors that edit the document. An editor, or author, can also insert comments into the document. Sometimes this extra information would have been hidden from view, if the publisher doesn't check

for the existence of the tracking information and remove it, it will be visible to anyone who has the document and unhides the tracking data.

- Metadata [6]:
  A large amount of metadata [6] can be stored about the author, the document and the organisation. The following is just some of the data that is stored:
  - Author's full name, Manager's name, Company name, document keywords, template used, computer username, previous authors and printer details.
  - The following document statistics are kept as well: Date printed, created, modified, accessed, last saved by, revision numbers, revision comments.
  - Full path to the documents location, either on the local machine, or a network server.
- Fast Saves:
  When Fast saves are enabled, any changes to the document are appended to the bottom of the file regularly. Although not visible in the application, the changes are visible when the document is converted to plain text or when viewed with a hex editor.
- Password Hashes:
  When MS Office Documents are password protected, the password is stored inside the document and as such can be extracted and subjected to brute force attacks and/or replacement.
- Microsoft Office 97 GUIDs
  All Office 97 documents had a GUID (Globally Unique Identifier) embedded into the file. The GUID was originally intended to aid in the merging of two documents. The GUID usually ended in the computer's MAC address, so the document could be traced back to and attributed to a specific computer [7].
- Outlook 2002 / XP:
  Any Office file that is sent as an attachment to an Outlook 2002 / XP email message contains a 10 digit number that can easily be traced to the machine on which the message originated. Outlook 2002 / XP does not respect the system privacy settings of the document, it adds an email address, and the 10 digit number [8].
- Microsoft PowerPoint:
  Speaker notes in Microsoft PowerPoint are linked to slides and usually contain the speaker's private thoughts regarding the particular slide.
- Hidden portions of the document:
  In Excel, one can hide entire sheets, or columns or rows or even a single cell. In Word text can be hidden by either marking it as hidden, or simply making the text the same colour as the background.
- Routing Slip information:
  When a document is emailed using the editors "Email document" function, the sender and receiver's email address will be stored in the document as well as all the mail headers.
- ODBC Settings:
  ODBC Settings can be embedded into a document, database locations, usernames and passwords can all be hidden.

2.1 Real world examples

There have been numerous well documented real world examples oh hidden data being discovered in documents that have been made available to the general public, business partners and competitors.

2.1.1 PDF Redaction failure

The Washington Post released a PDF document that contained a scan of the letter sent to the police by the Washington Sniper. The police attempted to cover, though the process of redaction, the sniper's address and telephone number and the account number he wanted the money transferred to. The police drew a black box over the censored details; however the data still existed in the PDF [9]. See Figures 2 and 3.

2.1.2 Tracking changes not removed

SCO filed a lawsuit against DaimlerChrysler, following an examination of the metadata in the Microsoft Word Documents associated with the case the metadata revealed that the complaint had originally been prepared against the Bank of America [12].

2.1.3 Metadata not scrubbed

Analysis of hidden data in a report of Weapons of Mass Destruction in Iraq produced by the UK government revealed names of the 4 employees who worked on the document, the location of the document and the location of the auto saves of the document. The UK Government has since moved to using PDF documents for electronic document publication [13].

The hidden data found in the document also helped to prove that a majority of the content of the report was plagiarised from an US Researcher on Iraq [14].

2.1.4 Text not really deleted

A WordPerfect document of the Kenneth Starr report detailing President Clinton's involvement with White House intern Monica Lewinsky that was published in HTML on the Internet contained more footnotes than the printed copies delivered to Congress. This occurred due an error in the document conversion process, however it does highlight the fact that just because one cannot see it on the screen, does not mean that it is deleted [4], [15].

Another example of text not really being deleted is that of a student who wrote a cover letter attached to a CV which was emailed to a prospective employer. This covering letter contained the names and addresses of other employers that the student had contacted previously [16].

2.2 Hypothetical examples

The real world examples highlight the danger that hidden data has already presented to organisations already, below are a few hypothetical examples that could occur.

2.2.1 Versions not deleted

An attorney emails a client a proposed settlement to review, the client makes comments regarding financial upper limits and comments on suitable terms and conditions regarding the settlement. The client's attorney then saves the comments made by their client in a version of the file, edits the new version slightly, and then emails it to the opposing counsel. The opposing counsel, noticing the version icon on the tool bar, views the client's comments is now in an advantaged position. [10]

2.2.2 Quote Tampering

It is possible to password protect a file in order to prevent changes being made to it, so when a sales rep emails a Microsoft Word document containing a quote to a potential client they need to email the read only password as well. To read the document, the client opens the document and enters the password when prompted for it.

Depending on the type of password protection, the client can remove the password from the document by creating a new word document, password protecting it with the same password and then opening the file in a Hex Editor and knowing where to look, the client can find the password hash. Knowing the password hash, the client opens the quote in the Hex Editor and searches for the password hash, upon finding it the client deletes it. Effectively deleting the password hash removes the password protection.

The client can now edit the quote to a new figure, password protect it and email acceptance back for the modified amount.

3 HOW TO EXTRACT HIDDEN INFORMATION

There are many ways to view the hidden data in a published document. There are commercial tools and open source tools available to examine documents.

3.1 Viewing Metadata

Viewing metadata is relatively easy, there are tools, such as catdoc [18] and antiword [19], available for free download from the internet that will extract the metadata from a document for you, or one can convert the document to plain text and look for the metadata.

Some of the methods discussed by Zall [10] for viewing metadata are:

- MS words basic interface – "show mark-up" to see the tracking changes and file – versions to see different versions of the document.
- Computer Mouse – hovering over embedded data or comments or editing changes will reveal hidden data as shown in figure 1.
- A Metadata viewer like WorkTrace [11] or Metadata Miner Catalogue PRO [12] will extract the metadata hidden in the document and present to the user of the software.
- Converting the documents to plain text will reveal all the text in the document. Microsoft Word's "recover text from file" open option provides a powerful tool to find data that has been hidden in files.

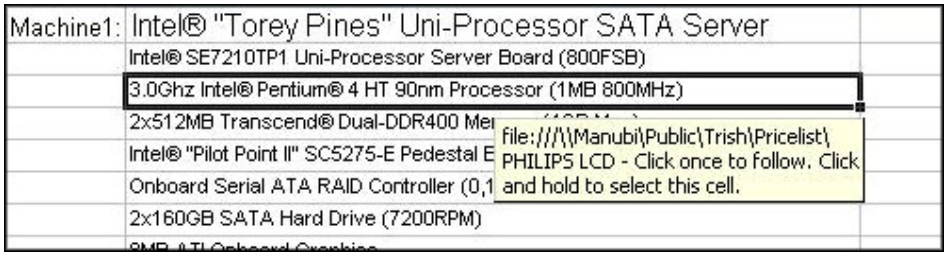

Figure 1: Hidden Data revealed by hovering the mouse over an Excel cell.

3.2 Circumventing Redaction in PDF documents

Common methods of redaction all try to prevent the user from seeing the data, which is not necessarily the same as removing the data from the file. Some of these common methods include:

- Placing black boxes over the text
- Changing the text colour to white, if the background colour is white
- Changing the background colour to black, if the text colour is black.

Circumventing the above methods of redaction can be as simple as copying the text below the black box and pasting it into another text editor, or opening the published file in Adobe Acrobat and changing the text colour and/or background colour [5].

In the case of the Washington Post, the document made available was a PDF of scans made of the original document. The Washington post drew black boxes over the sensitive information, Figure 2. The data did not exist in text form, but it did still exist in the PDF, it was just covered. By converting the PDF to postscript and then running a Perl script to remove the covering box, the information was visible, Figure 3 [5].

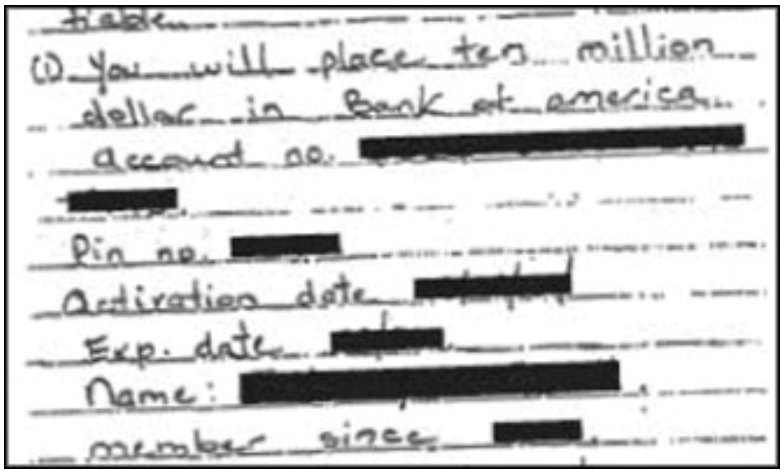

Figure 2: Redacted Washington Sniper letter [5]

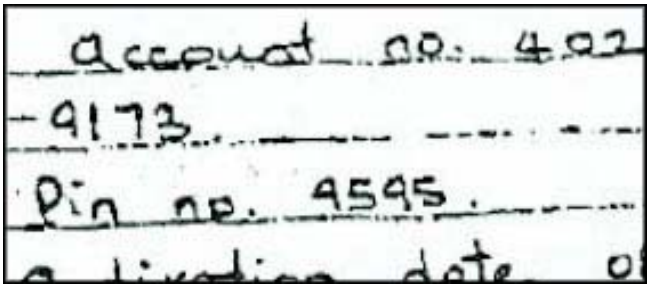

Figure 3: Un-Redacted Washington Sniper letter [5]

4 WHAT CAN THE HIDDEN INFORMATION BE USED FOR?

Hidden information has many uses. These can include identity theft, digital forensic attribution and even hacking.

4.1 Illicit of hidden information

4.1.1 Corporate espionage [9]

Hidden information in a document can also be to stealthily extract corporate information from an organisation. A user could open a sensitive document that they were working on, save the document as a different name. Then within the new document, save a version of the sensitive data and then delete it and start a new report / letter in its place.

At first glance the document that the employee is emailing looks like a family letter or an insurance claim, however the sensitive data is saved under a different logical version in the document.

4.1.2 Identity theft [9]

Byers [4] remarks that personal information is usually freely available in documents, particularly résumés. Concerned about privacy and identity theft, applicants remove their ID number, social security number from the document, however the information is still available.

4.1.3 Email addresses can be sold to spammers

A similar method used by Byers [4] to search for various document types on the Internet could also be used by spammers to locate email addresses.

4.1.4 Discovery and Reconnaissance

The initial phase of every breach of network security is discovery. An attacker needs to know information about the target's network. Hidden text in documents published on the target's website can provide the following:

- Network user names for brute force password attacks.
- Names of servers and information on the role of the server within the organisation i.e.: web server, document server, print server.
- Finger printing operating systems and application versions.

4.2 Forensic use of hidden information

4.2.1 Attribution

If an email with an attachment was sent from a machine with Outlook 2002 installed, the attachment will have a 10 digit number embedded in the document and on the originating machine there will be a file called "adhoc.rcd" in the following directory *c:\documents and settings\user\application data\microsoft\office*. In this file, next to the corresponding number will be the full path and file name to the sent document.

Similarly, documents created in Microsoft Office 97 will contain a unique GUID that will link the document to a particular computer.

4.2.2 Timeline

The file creation, modification and last accessed date and time information can be useful in establishment of a timeline and/or a supporting timeline for an incident [11].

4.2.3 Plagiarism Detection

Byers [4] notes the that methods that he developed to identify hidden data in various advanced document formats could be used to detect whether or not a document has been cloned or plagiarised from another document. At first keywords could be identified and compared between the two documents and then the hidden text can be used to gather stronger evidence.

4.2.4 Supporting Evidence

Hidden information sourced from documents can contain evidence that can assist a Digital Investigator in building of an effective and complete case.

5 HOW TO MINIMISE THE RISK OF EXPOSURE?

It is possible to minimise the amount of hidden data in documents that are published electronically. Issues arise when mistakes made when the document is reviewed or when the document is published / distributed [2]. The solutions below address the mistakes mentioned above, through training, tools and policies.

5.1 Alternative Publication Formats

An alternative to publishing documents on the web in a format that can contain hidden data is to export the final documents to a format that cannot contain hidden data. Documents can be exported to one of the following formats:

- Rich Text Format (RTF): This will retain most of the formatting of a Microsoft Word or Corel WordPerfect document but contains very little metadata.

- The document can be exported to plain text, however all of the formatting will be lost, which defeats the purpose of using an advanced editing application to author the document.
- The document can also be converted to HTML, however in many instances the HTML will need to sanitized further.
- The document can be converted into a PDF. Caution should be used, as PDF's can contain metadata which can disclose sensitive information.
- To ensure that no metadata is imported into a PDF, the Word document can be printed, then scanned and then converted to a PDF document [10].

Another alternative is to not use editors that have advanced editing functions and the capability to embed hidden information or to not use them to author documents that will be electronically published [4]. This is not ideal, as the advanced features are extremely useful and boost end user productivity.

5.2 Tools and Software Configuration

Microsoft have an add-in tool for Microsoft Office XP and Office 2003 [20] that can be used to remove personal and hidden data that might not be visible to the author when viewing the document in an Office application. Upon the tool completing its scan and removal of the hidden data, a log file is generated which contains what type of information was removed and any errors, if any [17].

Workshare have two products aimed at protecting organisations from the danger of hidden data TRACE is a freeware application that monitors documents that a user is working on and alerts them to the level of hidden data in the documents [11]. Workshare PROTECT integrates with popular email servers and strips hidden data out of documents sent as attachments via email. As an extra layer of security, the software converts the document to PDF [11].

Email however, is only one of many exit points that electronic data maybe published through. Future applications will need to be able to intercept the posting of documents on the organisations Internet web site, Intranet We Site as well intercepting documents being sent through Instant Messaging protocols.

Additionally the pre-configuration of end users' document editing applications may also reduce the amount of hidden data generated, for example: disabling the auto save function in Microsoft Word and the document tracking functionality in Microsoft Outlook XP/2002.

5.3 User Education

The first line of defence in the prevention of unintentional leakage of hidden data is user education. Payne [3] emphasises the importance of user awareness in preventing the publication of documents with hidden data.

In many users' minds, the content of the file of the document that they are working on is the virtual piece of paper that they see when they are editing the document [16]. Pre-configuring the user's computer not to use any of the auto save functions and to remove personal information on saving will reduce some of the hidden data that gets embedded in the files. The fault does not lie with the ability to track changes, or saving multiple versions of the document in the same file, rather the fault lies with user's lack of understanding of the functionality offered to them and the hidden implications.

In-house training on document publication or document editing should be compulsory training when a new employee joins the firm. Users need to be trained to be in a "Don't send, Publish" mindset. The organisation's computing users must follow appropriate procedures when sending or sharing an electronic document with parties outside of the organisation. As opposed to sending a document, publishing means a document following the electronic cleanup procedures,

such as removing versions, fast saves, comments and clearing all the tracked changes to the document.

5.3.1 Organisational Policy

Countering the problem of hidden data in documents that are published to the outside world can only be addressed through the application of comprehensive set of policies throughout the organisation. These policies need to be linked to the documents preparation and distribution cycle, however for these policies to be enforced, the document inspection should be automatic [2].

The determination about what to clean from documents that are intended for publication should be made at the firm wide level and not left to an individual employee's discretion.

Organisations should create a policy that compels employees to use a metadata removal tool before they publish any document, however enforcing such a policy is difficult without a mechanism to enforce it technically.

6 CONCLUSION

Hidden information can be costly. Unintentional data leakage can provide competitors with an advantage, hackers with information regarding an organisation's internal network and unintentionally tarnish the organisation's reputation. However, there are tools freely available to extract the hidden information from documents that can be used by the organisation to potentially identify hidden information in documents.

There is a distinct need for document extrusion detection tools to detect documents that contain hidden data in order to prevent the document from being published. These tools will need to monitor outgoing communications from the organisation's network. Such a solution should also be able to distinguish between documents that are internal, documents that are being shared with collaborators outside of the organisation and documents that are to be published to the outside world.

Ultimately it is user education that will prevent hidden data from being generated and distributed. Users need to know and understand the potential cost of using collaboration and versioning features of advanced editor and they must be provided with the tools to clean their documents of the hidden information, or at the very least warn of its presence in the document that they are working on.

7 REFERENCES

[1] B. Schneier (schneier@counterpane.com), "Hidden Text in Computer Documents," Crypto-Gram Newsletter, August 15, 2003, http://www.schneier.com/crypto-gram-0308.html#8.

[2] S. Wiseman, "Documents with Hidden Surprises", Autumn 2002, http://www.qinetiq.com/home/security/securing_your_business/information_and_network_security/white_paper_index.html.

[3] D. Payne, "Control metadata in your legal documents", http://office.microsoft.com/en-us/assistance/HA011400341033.aspx.

[4] S. Byers, "Information Leakage Caused by Hidden Data in Published Documents," *IEEE Security & Privacy*, March 2003, vol.2, issue 2.

[5] S. Murdoch and M. Dornseif, "Hidden Data in Internet Published Documents", *21st Chaos Communication Congress*, 2004.

[6] Microsoft, "How to minimize metadata in Word 2003", Article ID: 825576, July 2004, http://support.microsoft.com/kb/825576.

[7] R. M. Smith, "Fingerprinting of Office 97 files", 2002, http://www.computerbytesman.com/privacy/office97.htm.

[8] W. Leonard, “Woody's Office Watch: Outlook 2002's Privacy Busting "Feature",” vol. 7, issue 53, November 2002, http://www.woodyswatch.com/office/archtemplate.asp?v7-n53.
[9] W. Knight, “Online Document Search Reveals Secrets”, NewScientist.com, August 2003, http://www.newscientist.com/atricle.ns?id=dn4057&print=true.
[10] B.D. Zall, “Metadata: Hidden Information in Microsoft Word Documents and its Ethical Implications,” *The Colorado Lawyer*, vol. 33, issue. 10, pp 53-59, October 2004.
[11] P. Stephenson, “Using evidence effectively,” *Computer and Fraud Security*, vol. 2003, issue 3, pp 17-19, March 2003.
[12] L. Rowell, “Avoiding Snares and Gotchas in Word 2003,” *SAMS*, January 2005, http://www.samspublishing.com/articles/printerfriendly.asp?p=364262.
[13] M. Ward, “The Hidden Dangers of Documents,” BBC News, August 2003, http://news.bbc.co.uk/2/hi/technology/3154479.stm.
[14] R. E. Flinn, “Forensic News,” *Journal of Forensic Accounting*, vol 5, pp249-254, 2004.
[15] M. Eckenwiler, “Rasputin-like footnotes in Starr report,” *The Risks Digest*, vol. 19, issue 97, September 1998, http://catless.ncl.ac.uk/risks/19.97.html#subj3.
[16] S. W. Smith, “Probing End-User IT Security Practices — Through Homework,” *Educase Quarterly*, vol. 2004, issue 4.
[17] Microsoft, “The Remove Hidden Data tool for Office 2003 and Office XP,” March 2005, http://support.microsoft.com/default.aspx?scid=kb;en-us;834427
[18] Catdoc, 02 May 2005, http://www.45.free.net/~vitus/ice/catdoc/.
[19] Antiword, 11 December 2004, http://www.winfield.demon.nl/.
[20] Microsoft, “Office 2003/XP Add-in: Remove Hidden Data”, 14 July 2004, http://www.microsoft.com/downloads/details.aspx?FamilyId=144E54ED-D43E-42CA-BC7B-5446D34E5360&displaylang=en.