

# Towards an Infrastructural Framework for Secure Electronic Publication

Jock Forrester and Barry Irwin, *CISSP*

J.Forrester@ru.ac.za, B.Irwin@ru.ac.za, Computer Science, Rhodes University

**Organisations are increasingly publishing electronic documents on their websites and 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 parties once the document has been published.**

**To prevent the accidental, or deliberate, publication of documents containing hidden information the organisation needs to have an Electronic Publication Policy, more importantly though, it needs to have the Technical Infrastructure in place to enforce the policy.**

**This paper outlines such a Technical Infrastructure.**

**Keywords: Hidden Data, Electronic Publication, Metadata**

## I. INTRODUCTION

**H**IDDEN data within documents is becoming more common place as the organizations end users are being more and more competent with the applications used to author documents. The users are also taking advantage of more advanced features that rely on the use of metadata.

There have been several publicised occurrences of hidden data being discovered in electronic documents that are freely available which has tarnished the reputations of well known organisations and governments [1].

Organisations need to be aware of the danger that hidden data can pose and implement policies and the infrastructure to support the policies to prevent the publication of documents containing hidden data.

The proposed framework outlines the methods and processes to be used to clean the document being published, convert the document to a safer format or to block the publication.

## II. HIDDEN DATA

The study that Byers [2] conducted on approximately one hundred thousand Microsoft Word Documents retrieved from the internet revealed that nearly 50% of the documents

This research is funded by the Centre of Excellence in Distributed Multimedia in the Rhodes University Department of Computer Science.

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.

### A. Types of Hidden data

Types of data that can be recovered from a document include the following [1]:

- Multiple Versions of the document can be embedded within the file.
- The application can also store comments and changes made by the authors and reviewers.
- The Authors details are stored in the document. These details include Usernames, File locations, File Servers and Manager's details.

The framework proposed below will need to be able to identify all types of hidden data.

### B. Real world example

The analysis of hidden data in a report of Weapons of Mass Destruction in Iraq produced by the UK government revealed names of the four employees who worked on 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 to minimise such leakage [3]. 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 [4]

### C. Policy Requirement

As can be seen from the real world example, it is important that the organisation controls the publication of documents that will be exposed to the internet, corporate partners or competitors.

## III. PUBLICATION POINT TYPES

There are several exit points for documents generated in the organisation to the outside world. The electronic publication point types can be categorised as follows:

- Electronic publication: the publication of data via email, posting on websites, via Instant Messaging applications, shared via blogs and Peer to Peer Applications.
- Manual publication: Copying the documents to CDROM, Diskette or USB Flash Stick for transportation to a presentation at the client's venue. Such documents are usually then copied to the client's computer.

The proposed framework will need to cater for both types of publication points.

## IV. THE PROPOSED FRAMEWORK

Figure 1 highlights the proposed framework for

preventing the publication of electronic documents containing hidden data.

The Framework must be able to intercept and control the two publication points mentioned above as well as decide what action to take on the document.

#### A. Interception

Intercepting the publication of documents via the electronic method can take place in one of two techniques; either the Publication Policy Server (PPS) is placed between the organisation and the firewall, or the main exit point to the extra-organisational network, as in Figure 1.

Alternatively, server agents can be deployed on specific application servers, such as email [5], or web servers to intercept documents with hidden data.

The PPS will also need to be able to intercept documents being written to CDRom, USB flash stick (or other removable media) and to smart devices such as Pocket PC's and Smart phones. This can be accomplished by installing a desktop or mobile agent on every computer within the organisation.

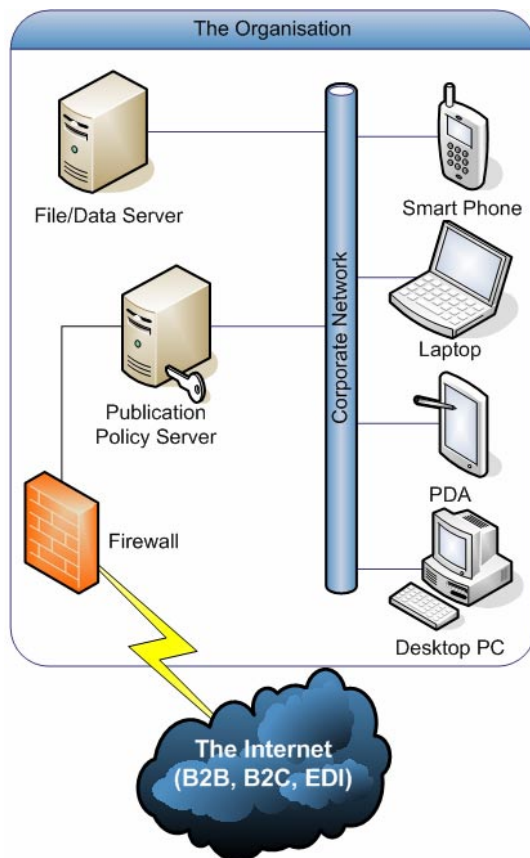


Figure 1: Proposed Framework

#### B. Action

The PPS will be capable of performing one of three actions on intercepted documents.

Documents containing hidden information will be outright blocked and permission will be denied to publish the document until the author has sanitised it. The PPS will clean the document on the author's behalf or lastly the PPS will convert the offending document to a format that does not contain hidden data, for example: PDF.

#### C. Caveats

##### 1) Agents

The PPS solution will have to be constantly updated with new server agents for new application types, document formats and for multiple operating systems. Likewise desktop agents will need to be developed for different desktop operating systems.

##### 2) PPS Server

The PPS Server will need to be constantly updated with new document format definition. It should also be considered a core service as it may potentially handle high volumes of data such as email.

The process of intercepting and cleaning documents should be as transparent as possible to the end user.

##### 3) Reviewing Channels

The PPS Server Policy logic needs to be able to distinguish between documents being published and documents being sent for review and collaborative authoring through the two publications points. The PPS server should not, for example, remove the reviewing comments from a partnership negotiation document.

#### V. CONCLUSION

The danger of hidden data should be of concern to all organisations. The advanced features of current document editors allow users to be more creative and productive, however, there may potentially be sensitive information stored in the documents.

It is important for organisations to have an Electronic Document Publication Policy which controls how and in what format documents are published.

Without a technical framework enforcing the policy, adherence to it will be minimal. The proposed framework attempts to cover all publication points within the organisation, provide centralised management and be adaptive to new document formats and publication mediums.

#### REFERENCES

- [1] J. Forrester and B. Irwin, "An Investigation into Unintentional Information Leakage through Electronic Publication," in *Information Security South Africa*, 2005.
- [2] S. Byers, "Information Leakage Caused by Hidden Data in Published Documents," *IEEE Security & Privacy*, March 2003, vol.2, issue 2.
- [3] M. Ward, "The Hidden Dangers of Documents," *BBC News*, August 2003, <http://news.bbc.co.uk/2/hi/technology/3154479.stm>.
- [4] R. E. Flinn, "Forensic News," *Journal of Forensic Accounting*, vol 5, pp249-254, 2004.
- [5] Workshare Protect, Product overview, <http://www.workshare.com/products/wsprotect/default.aspx>.

**Jock Forrester** is currently employed at Rhodes University as the Senior ICT Specialist for the Departments of Computer Science and Information Systems and is part time towards a MSc in Computer Science specialising in the field of Digital Forensics.