Polymorphic Information Tracers

By John C. Checco, CISSP

Date	Author	Description
2007-Oct-20	J. Checco	Initial Draft
2007-Nov-11	J. Checco	“Traceability, Durability and Integrity”
		“Data Separation”
		“Current Research”

Table of Contents

Information Protection

A marketing company's key to success is to have robust information about customers -- hard needs, soft needs, buying habits, et al. A research laboratory's capability to solve complex medical and scientific problems focus over the availability, accessibility and processing of huge amounts of data. Manufacturing and technology companies take extraordinary measures to protect the intellectual property that gives them distinct advantages in today's competing markets. Yet, it has only been in the last decade or so that IT funding has been increasingly focused on the importance of data retention, mining, warehousing, intelligence and protection. Have the corporate executive levels finally realized the intangible value in protecting information? Or, perhaps, was it out of regulatory changes for information governance?

Data Separation

-TBD-

Information Management Systems

In 2003, bioChec responded to a U.S. government RFP[1] with a proposal which included, among other technologies, "ubiquitous steganography techniques that can even be deciphered from document hardcopy." The underlying concept of this proposal was through the use of a document management system (DMS) augmented with biometric authentication and data watermarking, documents could be more controlled, more robust with meta-information and more traceable. While most DMS solutions rely solely on their protection methodology, the tenet of information security is not IF a system will be breached, but WHEN.

With this basic concept in mind, bioChec's proposed document management solution has the following qualities:

Document Access

These qualities protect the transport of documents from the safe zone (repository) to the volatile zone (users).

1. Authorization: Allow only access to document through role-based access control (RBAC). Role-based access controls are inherent in almost every enterprise single sign-on infrastructure. In the case where a client may not have such an environment, adding roles to the document management system is not difficult.

2. Authentication: Ensure the identity of the document requestor through forensically sound means – i.e. biometrics. Keystroke biometrics was chosen because of its ease of

[1] BAA03-02-FH

integration, software-only deployment and its seamless injection into the existing user experience.

Note that there is little or no differentiation in the document management market today with solutions on document access.

Document Identification

These qualities protect the document integrity and provide additional out-of-band features, such as document traceability.

1. Traceability: The ability to identify meta-information about the document such as:
    a. Who, where and when the document was last edited and/or published.
    b. Who, where and when the document was retrieved and/or printed.

2. Durability: Protection of the document information even through hardcopy. Many systems prevent the document from being printed or converted. Others deploy page-based watermarking.

3. Integrity: The document contents are verified to be in-sync with the system of record. With electronic documents, this is achieved through hashing codes – codes derived from arbitrary streams of data and considered to be non-repudiated. However, most solutions simply mark printed documents as "uncontrolled" – thereby avoiding the data integrity issue altogether.[2]

Data Watermarking

Most document management solutions in the market today utilize custom secure data formats that require a customized client application to read or manipulate the document. As of this writing, there have been no other commercially available innovations in the field of document security. bioChec's Data Watermarking solution extends traditional identification with the following requirements:

1. Embed identifying information into the document using steganography. This component requires several processes to be in place:
    c. Documents retrieved for distribution must be in non-editable format.
    d. Documents retrieved for editing must have its traceability information protected.

2. Identifying information could be garnered from the document even from hardcopy and/or scanning. This requires generating multiple distinct instances of the same document using polymorphism.

[2] Marking printed documents as "uncontrolled" is a necessary policy according to ISO9001 guidelines. It is unarguably a recommended policy. The distinction here is that applying this policy is NOT a solution to the data integrity issue.

Before one can readily discuss in-depth the solutions to the protection of information in documents, the user must be familiar with the basic concepts that the solutions are built upon.

Polymorphism

The use of polymorphic code engines has been used for both good intentions (securing code through obfuscation) as well as malicious intentions. For malware, polymorphism is used to avoid typical anti-virus signature detection. For commercial applications, it is used to prevent reverse-engineering of intellectual property. Although polymorphic viruses and worms have been in existence since the early 1990's, recent analysis of the Storm Worm has renewed interest in the advances of this technology.

Polymorphism is traditionally a zero-sum game. The purpose of a polymorphic code engine is to alter functional details without altering the logic of an application. Better polymorphic code engines are themselves polymorphic, to prevent polymorphism detection. Today, polymorphism extends to altering the functional logic as well -- as long as the end goal is accomplished -- a true exhibit of business intelligence.

Polymorphism is not only about software. The medical industry is increasingly warning about more and more strains of viruses that are drug-resistant. "Natural selection" is Darwin's term for describing nature's grand polymorphic mechanism.

Steganography

The art of "hiding in plain site" through the use overlaid media; steganography is said to have been used since the 1500's with Nicolas Poussin[3] to modern times for both profit (subliminal advertising) as well as terrorism[4]. In all known cases of steganography, still images as well as movies have been used to hide either other images or digitally encoded messages. Traditionally, images are used because our visual acuity with color is far narrower than the range of all visible light. And with the ascent of digital image compression formats, data is hidden in functionally non-visible areas (such as the high-order bits or alpha masks) of images. The latter method of steganography only works when an image is transported digitally in its original format – any hardcopy or re-compressing of the original image will lose the data it is trying to conceal.

---

[3] References interested in reading more about Nicolas Poussin:

- http://www.rennes-le-chateau.co.uk/html/mystery3.htm
- http://tracyrtwyman.com/blog/?page_id=54
- http://www.philipcoppens.com/arcadia.html

[4] Reference: http://www.wired.com/news/politics/0,1283,41658,00.html

Traceability, Durability and Integrity

The qualities of the Data Watermarking system are sensible; but the implementations to achieve these qualities raise serious business and technological issues – especially as several implementations are combined.

The main component, though, is to add traceability to the document. There are several ways to attach identifying information to a document. Each traceability method affects the durability and integrity of the document as well.

1. Format-defined attributes. If a document is stored in digital format, such as PDF, it is quite easy to add traceability into the document via application-defined attributes. Although this is quite easy to accomplish, it creates a maintenance issues if several formats are supported, or formats change over time. Also, document attributes are not preserved when a document is morphed out of the original format (i.e. printed, converted or scanned).

2. Page-based watermarking. From attaching printed barcodes to specific elements of a document[5] or using Xerox's patented DataGlyphs[6] imaging technology, this works well to preserve information when a document is either in electronic form or printed form. But it is quite easy to block out or corrupt the identifying information in the document – thus, it renders the document integrity as compromised and provides not traceability back to its origins.

3. Stylistic watermarking. This method relies on the use of components in the electronic format that reflect and algorithmic pattern of slight style changes known only to the originating system. This allows a printed document to be traced while minimizing the capability to overtly block the identifying information. This method was actually quite prevalent before the use of electronic word processing. Inherent in traditional mechanical typewriters was the variability in typesets – i.e. each typewriter had a specific set of letters that would be aligned slightly above or below the baseline. This along with other mechanical defects, such as interline spacing and intra-line spacing, one could deduce the signature of the typewriter that created the document.[7] In this digital world, such defects would need to be induced into the electronic form of the document.

---

[5] Refer to the author's work with Verizon Business' CRIT process managed by Icnivad.com

[6] Reference: http://www.parc.com/research/projects/dataglyphs/

[7] References:

- **"Landmarks in Typewriting Identification**," David A. Crown, *The Journal of Criminal Law, Criminology, and Police Science*, Vol. 58, No. 1 (Mar., 1967), pp. 105-111.
- http://books.google.com/books?id=CCh6hZUJB7oC&pg=PA116&lpg=PA116&dq=%22typewriter+identification%22&source=web&ots=S5WZjhZDEb&sig=B6-ECkIkk10iIjAPA0F3Atwhq1o
- http://www.dfs.virginia.gov/services/questionedDocuments/manuals/training/14%20-%2010.CourseH.Typewriters.pdf

4. Content watermarking. In mathematics, there are infinite iterations of algorithms that equate to an identity value (0 in addition/subtraction, 1 in multiplication/division, et al). If one wanted to add traceability to a particular mathematical formula considered to be intellectual property, it could theoretically be combined with any number of the identity algorithms to provide polymorphic instances of the same formula.

   The same concept can be applied to documents. Embedding hidden information into documents which contains unstructured data is immeasurably more difficult. Historically though, several types of content steganography can be found:

   - Messages embedded into structured verse where the content is imperative and the positioning of words and letters is paramount to extrapolating the message. Early cryptography essentially used this method.

   - Messages embedded into free verse where the content need not maintain its integrity. Most spam messages today are like this, they contain enough arbitrary information to hide its true message from the automated spam blocking heuristic engines. Of course, to the human brain, the "hidden" message is far from hidden.

   - Messages embedded into semi-structured verse where the style of writing can be extrapolated into a message. This includes such things as selection of phrases and/or idioms, grammatical choices, or sudden changes in literary styles.

   If each polymorphic instance was tracked by a database or in itself a pattern that can be decrypted into its identifying information, then one has achieved traceability without sacrificing durability or integrity.

   If a document contains content watermarking, it is feasible to ascertain identifying information about the document from hardcopy and/or scanned copies because the content is self-identifying.

Implementation Issues

The largest issue with data watermarking is integrity. Any document protected via content watermarking must preserve the contents of the original document. If the act of protecting the information alters the meaning of the information, the result is counter-productive.

Today's forensic graphologists can use methods to determine if a document or other written evidence has been tampered with through the culmination of handwriting analysis, typewriter analysis, and content heuristics.[8] But can the same heuristic patterns that identify fraudulent content manipulation also be used for steganography?

[8] Reference: http://www.handwritingexpert.com/

Content watermarking relies heavily on intelligent processing of document contents. Factors that affect any solution's ability to provide watermarking are:

- Inability of the heuristic engines to comprehend or ascertain true information meaning.
- Lack of cohesiveness in the content itself.
- Sparseness of contents.
- Variability in content format, media or style.

Testing and metrics play a large role in the acceptance of any new paradigm in information security. Not only must testing accurately reflect the problem being solved, but any algorithms must be open to public scrutiny.[9] "Security by obscurity" is not accepted by the information security community and should not be accepted by consumers as well.

Current Research

-TBD-

[9] Opening security algorithms to the public is a necessary evil. To protect intellectual properties, ensure all implementations of the algorithms are patented prior to publishing.