Personally Identifiable Information (PII)

And RUST Technology

in Open Document Format (ODF 1.0)

Introduction

> "The key concept for understanding the issues that lead to the inadvertent exposure is that information hidden or covered in a computer document can almost always be recovered. The way to avoid exposure is to ensure that sensitive information is not just visually hidden or made illegible, but is actually removed from the original document. Thus any sensitive information must be removed from the document through deletion."
>
> -- from [RWC]

This paper is about a simple trick every clerk typist should know and every author should understand. But make no mistake, there is no magic here and no unverifiable results. In fact, it is very possible to stumble onto a desired result by accident, but there is no harm in knowing what to look for, and with a nod to the virtues of ISO Standards, knowing when you are finished looking.

The Word Processor used is the OpenOffice.org Writer[OO]. This application will produce OpenDocumentFormat (ODF 1.0)[ODF], an XML format and an ISO Standard[ISO].

Bibliographic Marks

A Bibliographic Mark placed in a document represents several pieces of information (fields), normally a reference to a book, article or some other resource. It has an iconic footprint in the document, yet, individual fields may be edited from the Word Processor and the fields populated from a (data base) table. Once populated, the field data is incorporated into the document. The reference is not a Hyperlink to the data, although the fields themselves may be URL's.

In the terminology of Creative Commons[CC], the Bibliographic Mark can be thought of as:

OpenOffice.org Writer display : Bibliography entry <- The icon

Meta Data : <dcterms:bibliographicCitation>[fields]</dcterms:bibliographicCitation>

Personally Identifiable Information (PII)

Personally Identifiable Information is a meta data property in that it's veracity is not determined by the document, nonetheless its "class" (common type) is well defined and definitions can be re-linked if missing. While the prototypical example is a person's Social Security Number, many other types of information have the same characteristics.

This is just a somewhat simpler case of a bibliographic citation. In this case the [fields] are fixed and provide a citation to the classification and definitions of the PII, as well as the PII itself. But notice that if the PII "note" field is blank or *null*, then processing, and in particular copying (propagation), proceeds without exception.

Meta Data : <dcterms:bibliographicCitation>[PII fields]</dcterms:bibliographicCitation>

Meta Data : <dc:identifier>[PII (note field)]</dc:identifier>

Meta Data (XHTML): <rust method="rust" mark="*">[PII]</rust> See Also [TWH]

A generalized solution to the non-proliferation of PII is preferable for reasons stated below. Nonetheless, the name space[W3C] or tag space[MF] concepts are helpful, compatible, and an intrinsic part of the generalization.

The RUST Principle

RUST is an acronym which stands for "Redact Unless Static Text", and a principle governing the behavior of machines that generate human viewable text renderings although a picture of a page or a CRT display etc. must be included for practical reasons. Simply put, a mark should always be rendered (and copied) as a mark only. The thought of "lost data" horrifying as it is to Computer Science simply does not apply to meta data; that whose definition cannot be lost (as long as the name space, the data definition "space" exists). Applications processing XML normally follow the convention that element content is visible and may be styled and attribute content is not visible and not stylized. CSS transformations can bend this convention leading to well known redaction problems[RWC]. Specifically, if the style sheet creates another opaque layer element to "cover" text, a simple copy/paste operation will reveal the text glyphs underneath. Normally, the element content would be copied but a Hyperlink to context sensitive help (an implied attribute) is an acceptable substitute as would be a reference to a normative or informative specification. Although, bibliographic citation elements may define many more attributes (fields) than necessary to represent PII, storage of empty attributes not an issue as a further consequence of their lack of visibility and style. Search terms of a Controlled Vocabulary and subject headings can be stored in the "extra" attributes to facilitate analysis of multiple document archives.

We differentiate here between analysis of the statistics of a group of documents and data mining – the logical AND of two or more (partial) identifiers to lower the probability of misidentification; the logical OR of two or more (partial) identifiers to collate a class; NOT; XOR etc..

Data Mining and PII

Information about people is valuable to collect, with the implied premise that PII, as a (often partial) identifier, is a step closer to identification of needs, wants, proclivities etc.. The premise is problematic at best, regardless of the information quality. A thin rationale for a marketing campaign never stopped a spammer who was paid up front, and therefore the often heard argument that "Somebody must be Buying" is just not true.

Partial information in the hands of a mischievous collector is likewise problematic because the quality may be good enough to do great damage to both misidentified individuals or misclassified groups. Moreover, the cost of retroactively protecting a group increases with the size of the group. To assume that a subset is at higher risk than another subset is impractical, if not impossible.

Security and PII

Security should be technology driven and responsive. Responsive in the sense that an animal might show a "fight or flight response", not a cognitive decision, rather a single response which may be evidenced by different actions accomplished with available technology. The winged monkeys in "The Wizard of OZ" terrify us because we know that rattlesnakes are shy, and not because we know rattlesnakes bite. In the context of security for PII, encryption technology is useful, but it is "all fight and no flight", and there are other effective responses.

Still, there is always the danger with technology driven security issues that simple physics that was beneficial to begin with, can be reversed. Computer Technology, often in the interests of "user friendly technology", is very dangerous in this respect. A choice of fonts is not an acceptable solution to the unwanted disclosure of PII, for example. Only a combination of both technology and responsiveness can instill security, technology alone is insufficient because, in the main, if the same threat will approach and recede at exactly the same speed then only response matters. In the particular case of PII, propagation avoidance, the "flight" response, is possible, practical and a useful adjunct to the perimeter defenses common in most Information Systems.

Methods

Types of PII

(text)	Description	Examples
[aka]	Alias	Maiden Name, Pen Name, DBA, AKA etc.
[alpha]	Alphanumeric Identifier	SSN, Military Service ID, Postal Code etc.
[birth]	Circumstances of Birth	Age, Birthday, Place of Birth etc.
[bystander]	Affiliated Person Of	Spouse, In-Law, Friend, Acquaintance etc.
[car]	Asso. Personal Transportation	Car, Car License, Driver License etc.
[contact]	Contact Information	Telephone, IP Network, Geographic etc.
[death]	Circumstances of Death	Day, Place, Cause etc.
[dna]	Blood Relative Of	Blood Relative (see also:bystander)
[groups]	Group Affiliations Of	Group Membership, Professional Licenses etc.
[health]	Asso. Health Data	Genetic, Communicable, Chronic Disease etc.
[ice]	In Case of Emergency	Third-Party Contact Information
[location]	Asso. Location	Workplace, Network URL etc.
[marks]	Identifying Marks	Race, Gender, Tattoos, Scars etc.
[misc]	miscellaneous	None of the above
[money]	Asso. Financial Transaction	Bought, Sold, Payment Type etc.
[witness]	Asso. Public Event	Rode Public Transportation, Concert etc.

Table 1: ICON names were chosen per the author's preference. Suggestions welcome mailto:gannon_dick@yahoo.com?subject=Types of PII

Bibliographic Marks

It is suggested that the bibliographic data base be used to base load the entry(see end note[i]). A considerable amount of tag specific useful information can be stored and be available without re-typing. Each row of the data base is itself a bibliographic citation. Documents may contain multiple instances of the same citation, but it is up to the user to insure the integrity of the row identifiers (aka Short Name or class name).

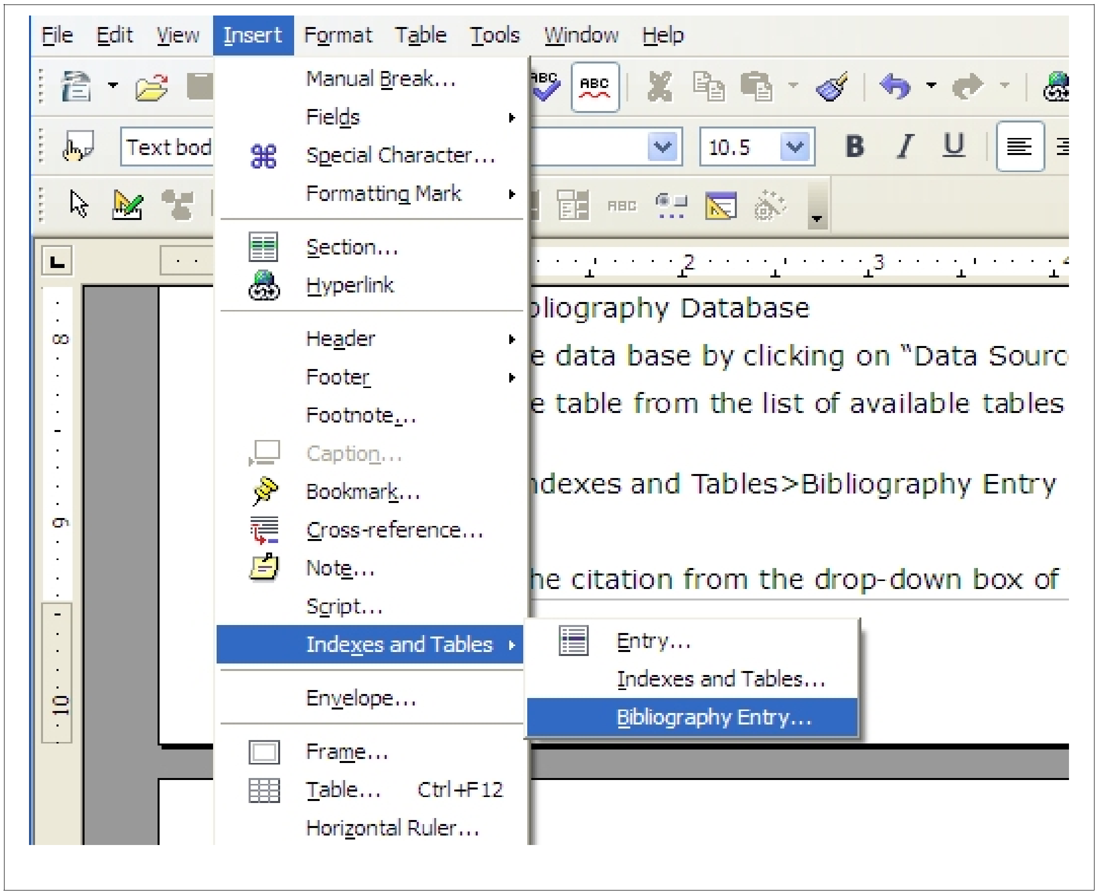
1. If the data base is used, select the data base and table containing the PII entries.

   Tools>Bibliography Database

   Select the data base by clicking on “Data Source” to the right near the top

   Select the table from the list of available tables in the drop-down box to the left near the top.

2. Insert>Indexes and Tables>Bibliography Entry



bliography Database

e data base by clicking on "Data Sourc

e table from the list of available tables

ndexes and Tables>Bibliography Entry

he citation from the drop-down box of

3. Choose the citation from the drop-down box of “Short Name” and push the Insert Button. Important: Close the dialog. This copies all auxiliary information from the data base to the document.

4. Right Click on the icon and select Bibliography Entry...

   Push Edit. The citation fields will be displayed. Type the data into the “Note” field, e.g.

   [alpha] -> '123-45-6789',

   [witness] -> 'AA Flt 9088 lv. DFW 0800 ar. LGA 1017'

   When you have finished editing, close the dialog and push Modify.

   You will receive a warning, choose “yes”. The wording of the warning is perhaps unfortunate. The purpose of the warning is to accept or reject your editing of only the entry selected, not of all entries with the same Short Name.

Source and Archives

The “source” file is the file you can launch (and edit) in your Word Processor. An “archive” file is made with an ID Transform (XSLT Output Filter). While it does contain all the bibliographic information including PII, pictures or other embedded objects are uuencoded, that is, for transport but not for easy redisplay. The advantage to using an archive file for further processing is that it cannot be easily reloaded into the Word Processor, so, in the absence of the “source” an outright forgery of the source is very difficult.

A further advantage of using an archive proxy is that a source file is constantly being updated and in the case of temporal meta data this should be of some concern. For example, an issue date is logically later than a modification date. An overly user friendly word processor might not see it that way, and automatically update the issue date leaving the modification date unchanged or setting it equal to the issue date.

An archive file, and any condensed version, will lessen the amount of encrypted storage necessary, and since XML files are text and a standard, the data will persist for the lifetime of the storage media.

The very existence of an archive file as different from the source file identifies the boundaries of a common quarantine. Archive files from many sources may be commingled in protected access network storage as long as the file names are unique or made (mangled) unique by the platform Operating System. Indexing source files can take considerable time, but adding an archive version to an index much less time. Standardized associated keywords reside with the bibliographic information.

Results

Only a "hands on" user of the Word Processor Application is able to view the PII. A facsimile, a print preview, a photograph of the screen or a projection of the screen will show only the icon,

Bibliography entry.

Furthermore, a HTML file, a Rich Text Format (RTF) file, a TXT file (with any character encoding) or an Adobe PDF file will only contain a text icon, a text rendering of the class name. A copy/paste operation from the Word Processor to a Spreadsheet will only copy the text icon. None of the PII values will propagate to the new format. **PII will be redacted from source documents by deletion.**

The only ways to proliferate bibliographic data are to have an electronic copy of the source or a printed or electronic copy of an archive. The archive is text which can be read and modified by any number of applications except the source Word Processor. The source is somewhat "tamper-proof". Therefore the archive file is useful for both verification and forensics. **Completely stripping PII from an archive file is a straightforward XSL Transform.**

Templates may include unpopulated bibliography entries in content.xml which can then be populated with XSLT or scripting. Once inserted, a bibliography entry belongs to the document. Master Documents may contain sections set to read-only or read-write. It is not possible to make an archive file directly from a Master Document because export filters are not available. The edit ability of an entry is indirectly controlled by the read status of the Master Document section containing it and resolved when the Master Document is converted to an ODT file (whence exported to an archive).

The RUST Principle, or circumstantial abstractions of it, is followed in all cases.

Documents and Document Templates prepared in this fashion are like forms, but have some distinct security and processing advantages.

- Unlike forms, documents have a full complement of captured but hidden meta data. Successful analysis of the data need not depend on initial form design requirements. Nonetheless, that the **data can be compacted**, as with a form, means lower requirements for encrypted storage.

- This method gives every computer, regardless of environment, some of the privacy/security features of an a Automated Teller Machine(ATM) or a pay phone. For example, any edits are done with a small screen dialog box which limits the distance from which data can be seen. **Privacy enhancements are cross-platform.** By contrast, web forms are always in edit mode, and of variable size.

- **Sensitive data, not necessarily PII, can be masked as well.** For example, a (source copy) serial number, a link to a network storage location, a link to an authorization document or a link to a requirements document may be included automatically in a template. For example, **it is possible to "pre-redact" a document** without a need for the long term storage of two distinct source documents. With RUST Technology, redacted data co-exists with public text. Decisions regarding redaction or exposure of the data reside with the last person to possess a source copy, even if that is not the original author. This feature comes into play, for example, with a Freedom Of Information Act (FOIA) fulfillment with documents from several governmental agencies each with its' own Virtual Private Network (VPN) or intranet. Collation, searching, sorting and indexing are standardized activities in aggregate, but with local control over specific redactions.

- The less an unauthorized intruder knows about the site map of an intranet the harder it is for them to move around the network undetected. It is likely that an automated HTTP client would not be able to discern the difference in a redacted or original source document, although an HTTP 404 (page not found error) is easy for an intruder to detect. Hackers do not take rejection very well, they just try harder and, crucial to the odds of detection, more often. Every separate network penetration and every time an intruder “goes away” thinking they have what they wanted to steal increases the probability they will be caught “going back in”.

Discussion

Desired Behavior of Word Processing Tools

Personally Identifiable Information (PII) is meta data. Whether it is *null* valued, hidden, in a file header or embedded in the common stream it maintains semantic properties which can be linked (or more often re-linked) to external classification schemes. Especially in reference to people, identifiers and partial identifiers are subject to abuses which long predate the electronic age.

Crimes which depend upon misidentification or a confusion of identity are ages old. The electronic age has made such crimes faster and and more pervasive, but it would be a mistake to think that technology has made the crimes somehow "better", less subject to detection. Ironically, the same can be said of Authorship, Journalism and a host of other nominally legitimate pursuits! The commission of Journalism remains almost undetectable in some cases, although the number of independent mastheads is at a low point <wink />.

It is not unreasonable to suggest under these circumstances that the tools and technology of both Identity Theft and document production are simple multipliers. If we could change the habits of individual authors we could reduce Identity Theft per capita. The unique nature of meta data is such that tools and technology can be optimized for desired behaviors over and above a generalized dictum that the technology be "user friendly". The RUST Principle, as defined above is an example of this optimization. The author recently saw an advertisement for a laptop whose screen had a restricted viewing angle for security reasons. As far as he knows, this security feature was present in every laptop made before the year 2000, demonstrating that some security features are simply a reclamation of physics that was beneficial to begin with.

It is important that an author understand that a meta data "teaser" tag, that is *null* valued, is not per se dishonest. External references and definition data of the name space are static and correct. If a tree falls in the forest and no one hears it does it ~~make a sound~~ have a Social Security Number? Well, sure it does, or at least it has the same one it did before it fell. Maybe we should let identity thieves worry about how to get that information. As Napoleon put it – if the enemy is doing something stupid, *a good general leaves him alone.* Rather than pretending that paper will someday grow hyper links, we take the failure of WYSIWYG as a starting point to protect meta data, and in particular PII, from propagation.

References

[RWC] - Redacting with Confidence: How to Safely Publish Sanitized Reports Converted From Word to PDF;Report # I333-015R-2005;12/13/2005;NSA (author's note:not classified)

[OO] – http://www.OpenOffice.org

[ODF] – http://www.OASIS.org

[ISO] – Information technology **—** Open Document Format for Office Applications (OpenDocument) v1.0**,** ISO/IEC 26300:2006 First Edition 2006-12-01

[CC] - Extending Creative Commons Metadata; http://creativecommons.org/technology/metadata/extend

[TWH] – Taking Work Home; GJ Dick; http://www.geocities.com/gannon_dick/TakingWorkHome.pdf

[W3C] – http://www.W3C.org

[MF] - http://microformats.org/wiki/rel-tag

[i] To “install” the database table to the stock OpenOffice 2.x application, copy the PII.dbf and PII.dbt files to the same location as the database preset biblio.dbf and biblio.dbt files. The PII table files are found in the examples file, a zip file.