Security Protocol for Ebook Production

Just something I've been working on, I would particularly be grateful for any suggestions y'all may have for viable methods of defeating the Natural Language Watermarking (NLW) techniques that are briefly mentioned towards the end of the text.

By: DIzzIE [antikopyright 2008]

Disclaimer

All content below is presented purely for informational purposes and is not meant to condone any activity which may be illegal in the reader's jurisdiction. The myth of Intellectual Property is serious business, m'kay? Please sign here in either yak blood and/or horse semen __________________ to indicate your comprehension of, and agreement to, the terms in this Disclaimer prior to reading any further.

Intro

I've written up a couple texts on making ebooks, but lately I've been noticing a great deal of lackluster security efforts in existent ebook production which could all too readily lead to the apprehension of the original scanners and suppliers. Here then are a few tips that should hopefully pave the way for a security protocol of sorts that one might follow when producing ebooks; basic steps towards protecting the source of the book so that the supplier has a reduced risk of getting caught. This text will address both security measures to take when scanning treeware tomes and when making already existent ebooks available to the public, though since the tips for both often overlap, they'll be intermingled with one another below.

Book Covers

Let's start with the first point of contact: the book covers. As a general rule of thumb, always use an already-available online cover image (found on Amazon or through a standard web image search) as opposed to scanning in the covers of the book in your possession. While seemingly innocuous, the book cover can reveal everything from its source of origin (i.e., a library stamp), to your identity (i.e., a fingerprint). If you're scanning a particularly rare or highly specialized book, then even the particular creases and folds present on the cover can lead to identification of the source and, in turn, the identification of the person who scanned it. Also keep in mind that the security precautions taken with book covers apply to inside pages as well (which is why you should always strive to OCR (and proofread) as much of the book as possible).

Stickers, Stamps, and other Insignia

This is particularly pertinent to library books, but also applies to bookstore-specific stickers and the like. All too often I have seen book scans complete with unobfuscated (let alone removed) library barcodes and stamps that reveal the name of the library. Investigators could then simply analyze the file's timestamps and other pertinent metadata and then proceed to subpoena the library's records for the book in question and see who had it checked out within the timeframe indicated by the timestamps found in the file. Furthermore, if several books start popping up online from the same source, then that would obviously present even stronger evidence that would allow the identification of the suspected scanner.

Also keep in mind that, for instance in the US, when libraries or bookstores receive National Security Letters requesting information on a patron or customer, the letters come with gag orders that legally prevent the library or bookstore in question from talking to anyone about the letter (as well as making a point that service shouldn't be interrupted to the personage who is under investigation, lest z becomes suspicious). In 2005 alone, over 9,000 NSLs were issued (though these weren't necessarily limited to libraries).

It is not enough, however, to simply blank out all identifiable stamps or stickers, as the remaining blank rectangles can likewise be used identify a copy of the tome in question by comparing all available library copies of a certain book (again, this is obviously of greater concern for those involved in the digitization of obscure material rather than mass market productions). Thus, when blanking out identifiable insignia in the text, be sure to overcompensate and blank-out larger section then necessary. Ideally, you should always strive to OCR the text so that no extraneous insignia remains at all. While it may seem improbable that anyone will go through all of these steps to identify a book scanner, our motto for the duration of this textfile will be: better paranoid as fuck than dead as a duck.

Fingers and other Bodily Identifiers

When scanning or taking photos of a book, be sure that your wrists, hands, and fingers are nowhere to be seen in the proximity of the photo frame or on the scanner bed. If this is not possible, then be sure to wear gloves along with long sleeves. Leaving your limbs exposed has the potential to reveal such vital elements as the color of your skin, your sex, your age, as well as any personally-identifiable characteristics like scars, rings, or tattoos (not to mention other more significant data if you for instance have a bracelet with your name on it). Along the same line, particularly if taking digital camera pictures of a book, don't leave anything silly lying around within the frame of the shot, like your school ID.

Nota Bene: Even if you think that you have cropped out all potentially incriminating bodily markers from your page scan, be sure that there are no embedded page thumbnails in the resulting file. For instance, if you were cropping out the edges of the book scan that had your fingers in ABBYY FineReader, and then proceeded to save the ebook as a PDF, go ahead and open the resultant PDF in Adobe Acrobat and click the Pages icon on the left-hand side (next to the Bookmarks icon). Be sure that the thumbnails shown there match the cropped main page, and are not thumbnails of the original uncropped scans (earlier versions of FineReader had this security flaw). If the thumbails do not match, select all the page thumbnails (click on the first one, then scroll down to the last one and hit Shift as you click on the last one), right-click, and select 'Remove Embedded Page Thumbnails.' This should generally make for good practice as it will also reduce the filesize, while at the same time strengthening your security level.

Serial Numbers

Now moving past the covers to the front matter of the book, be vigilant for any potentially incriminating serials lurking about. Relatively rare in treeware productions (though I have on occasion seen them in treeware books as well as periodicals), serial numbers are much more common in existent ebooks. The trick here is to separate the useful numbers that you should leave in place (ISBNs, print run indicators, Cataloging in Publication data, etc) from the potentially incriminating numbers which will make you susceptible to unique identification and apprehension. The rule of thumb here is if the number returns many hits when you plug it into a search engine (via a proxy, of course), then it is probably a standard cataloguing number. If, on the other hand, there are no hits for the number, and it doesn't follow a noticeable pattern that other cataloguing numbers do, then it is likely a serial number and thus poses a security risk and needs to be removed.

Let's take a look at an excerpt from the edition notice page (the page with all of the copyright garble) of a random book (with me adding the letters in brackets for ease of reference):

Code:

```
K1401 .L47 2001                [ A ]

346.04'8'0285-dc21    [ B ]    2001031968     [ C ]

ISBN: 0-375-72644-6    [ D ]
```

10 9 8 7 6 5 4 3 [E][A] The LCC (Library of Congress Classification) call number is the number used by various libraries to categorize and shelf books; they always start with one to three letters and end with numbers, which serve to classify the book into subjects and subjects-within-subjects (for example science-->physics-->etc..). The second set (technically referred to as the cutter number) serves to identify the author and the book title, and finally the third set is the year of publication. Simply being a call number, [A] thus poses no threat to us and can be left intact.

[B] The DDC (Dewey Decimal Classification) is another number widely used by libraries to sort books. It is typically easily spotted by its use of decimals and apostrophes (technically called prime or hash marks). The number poses no threat, though, as a totally irrelevant aside, it is noted for its inherent bias (for instance a way disproportionate amount of numbers designated for the religion class are dedicated to Christianity, with most other religions getting lumped into the 290 'Other' section). Meaning that if you want to be all PC and shit, go ahead and delete this numerical symptom of ethnocentrism from your ebook; that'll show 'em...

[C] The LCCN (Library of Congress Control (or Card) Number) is another cataloguing number assigned by the LoC, and is 10-8 digits long, wherein the first two or four digits are the year of publication, followed by a six digit serial number, with a hyphen at times separating the year from the serial. Once again, the LCCN poses no risk.

[D] The ISBN (International Standard Book Number) is a 10 or 13 digit number to commercially identify the book and poses no threat. It is also usually the only number that's actually labeled for us (as in the above example), making our work even easier.

[E] The last number in this count-down sequence is the printing run of this particular book (for instance if someone sells you a book which they claim is a first printing, you can flip to the bottom of the edition notice page of the book and double check. If the last number in the row isn't 1 then you're being swindled). The risk posed by this print run row is minimal, yet existent. For instance, if it is known that a certain print run was particularly rare (due to limited quantities or printing errors in the run), then the pool of possible locations for this particular treeware copy will be significantly narrowed. It is thus best to either modify the last digit in this row, or delete it entirely (though please note that significantly different print runs should also have different ISBNs, so be sure to make sure that your ISBN is a common and not a rare one by performing a proxied web search for it).

{ F } The aforementioned five numbers are some of the most common (Western) numbers and are, for the most part, fairly innocuous. You may, howver, also encounter other classification numbers that are also fairly harmless not listed here (for example the DOI (Document Object Identifier), which will be discussed below in the e-journal section), so remember our aforementioned rule of thumb: run a proxied web search for the number in question, and if in doubt, always err on the safe side and delete the number to cover your ass, just in case.

Metadata

The risks and dangers of metadata are significant enough to warrant a textfile of their own, which it would be in your best interests to read, as it is of the utmost importance to ebook suppliers. Suffice it to say that everything from the serial number potentially embedded in the Exif data of the photos of the book you just took (as was brought to the public's attention when the last Harry Potter ebook was made available), to your real name in the Word document where you proofread your recent scan, can all be used to readily identify you.

Journals and other Electronic Sources

Occasionally when you download academic journal articles or ebooks, the file (typically a PDF), will come accompanied with a cover page that identifies the name of the subscriber to the journal database (typically the academic institution at which the article was downloaded), along with a timestamp. Obviously, prior to distribution you will want to remove this cover page by either opening up the file in Adobe Acrobat Professional and deleting the page, or using the free GPL Ghostscript to convert the PDF to a new PDF, selecting which pages you wish to keep.

Nota Bene: Prior to modifying the PDF in any way, change your system clock to an earlier date/time so that the resultant file timestamps cannot be successfully cross-referenced to server download times by any investigators who are trying to track down the source of the leaked ebook or journal.

If the PDF file has any sort of security features which Ghostscript cannot bypass which prevent you from modifying the file, try removing the protection using Elcomsoft's Advanced PDF Password Recovery program.

If there is a serial number or watermark along the margins of each page, you can either manually remove them using Adobe Acrobat Professional's TouchUp Text/Object tools (though employing this manual approach will be a painstaking process if this is a large article or ebook), or you can remove them automatically by cropping the pages of the PDF. To do so, hit Ctrl-P in Acrobat to access the print menu, select Adobe PDF Printer (you'll need to have Acrobat Professional installed), and hit Properties. Next, select Adobe PDF Properties, and in the Adobe PDF Page Size drop-down box, select Custom and click Add. Modify the page width/height and hit OK to get back to the main printer menu.

In the Page Handling area, change the Page Scaling drop-down menu to None. The portions of the page that will not appear when you create the new PDF will now be greyed out. Make sure that the serial numbers/watermarks are in the greyed out region (but that none of the actual text is), and then hit OK to print the new clean PDF.

Note that many electronic journal articles (and to a lesser extent ebooks) now have a DOI (Digital Object Identifier), that looks frighteningly similar to a serial number, but is in actuality the same for all copies of the digital object in question, and as such poses no risk to us (though in theory the same object may have several DOIs if registered through different Registration Agencies, but this would be impractical to assign a unique DOI to each downloaded instance of an object). The DOI typically looks like 10.*****/*** and can be looked up by appending the DOI to the following URL: dx.doi.org/. For example, dx.doi.org/10.1000/9 if the DOI in question is 10.1000/9 (in case you need to check if the DOI is valid or not).

Obfuscation and Disinformation

While the previous tips have generally centered around ways to remove identifiable markings from your ebook productions, you may also encounter times at which it may be more beneficial to obfuscate your source by adding nuggets of disinformation to the ebook. For instance, you may consider adding in a different library stamp, or writing in 'your' name on the inside cover and leaving it in the scan production.

If you are an insider at a publishing house, or a book reviewer who receives electronic ARCs (Advance Reading Copies), it may be to your benefit to make the ebook look as if its source was actually a treeware copy that was subsequently scanned and OCRed. In other words, you'll need to perform a sort of reverse proofreading and add in 'uncorrections' to the text so as to replicate common OCR errors. These include changing a c's to e's (and vice versa), 1's to l's and I's, h's to b's, d's to cl's, and so on. You may also add more blatant clues like adding 'OCR v1.0' to the file name.

Natural Language Watermarking (the hardcore shit you really need to be worried about)

NLW, the potentially deadly future of ebook distribution, marks a text not by inserting any sort of background image or serial number, nor by altering the appearance of words via character thickness or the spacing between characters, words, and/or sentences (these practices are collectively known simply as text watermarking, with NL text watermarking being a particular subset), but instead by making intricate semantic and syntactic modifications within the text itself.

That is to say, while format-based text watermarking can be defeated with relative ease by converting the file to another filetype (for instance a PDF to a TXT, or a LIT to HTM), which thus collapses any word or line-shifting injected into the original ebook, no two transcripts of a book that employs NL watermarking will ever be the same, as sentence structures are themselves modified for each copy of the book (the modifications are highly sophisticated, going beyond mere synonym substitutions/intentional spelling or punctuation errors, here's a particular example of syntactic and semantic natural language watermarking). If the point isn't clear: this is some motherfucking sneaky-ass shit.

Now, while there has been a great deal of research done on text watermarking in general, and natural language watermarking in particular, there's not a lot of clear evidence that publishers are actually using this mode of watermarking. But then again, its usage isn't exactly something that publishers would be broadcasting to the world. Thus, it is uncertain whether certain ominous phrases coming out of publishing houses are mere scare tactics or hints at implementations of text-based watermarking techniques. For instance, here are two entries from the FAQs of a prominent web-development publisher, Pragmatic Bookshelf, and from an ebook distributor WOWIO, respectively:

Quote:

Are the PDF files restricted?

There is no copy protection or functionality restrictions in the PDF files. You may view or print them for personal use as you see fit.

You may not give your PDF version to other people. The PDF file you order is personalized with your name and other identifying information. [emphasis added]

You can buy multiple licenses of a PDF file for your team or organization, in which case the PDF will be stamped with the number of allowed licenses. We'll only send you one, so as to conserve everyone's bandwidth.

(Source)

Quote:

Does WOWIO use any kind of digital rights management (DRM)?

Since anyone can defeat the most "sophisticated" DRM with the print screen button, we believe that technology-based DRM is essentially a fraud. Our approach takes the market incentive out of misbehaving, rewards people for doing the right thing, and tries to stay out of the way of honest users. To help keep everyone honest, however, readers must authenticate their identity and agree to a licensing agreement when they set up their account. Then, each ebook is serialized with the reader's authenticated name and a unique serial number, as well as other less visible markers. [emphasis added] WOWIO will immediately terminate the account of anyone caught illegally distributing ebooks, and will prosecute serious offenders.

(Source)

What these two policies indicate is a shift away from overtly obtrusive DRM measures akin to time-bombed files bogged down with restrictive licensing that expire after a week or won't load on more than two computers, to a much more malign form of copy-protection that is likewise much more difficult (though not impossible) to isolate and remove.

How then, are we to deal with these new watermarking measures?

Here is a general proposed protocol to determine whether the ebook in question employs NLW and to further isolate all potentially modified sentences. The protocol is in dire need of refinement through active development and testing, but might act as a spark for future counter-watermarking techniques that help liberate all locked-down and commodified information:

1. Strip all overt copy protection schemes from the file that prohibit tampering with the file. (e.g. using Elcomsoft's Advanced PDF Password Recovery or Ghostscript for PDFs or Convert LIT for LIT files).

2. Convert the file to plaintext ASCII to collapse line/word/character shifting-based text watermarking measures.

3. Remove any remaining visible serial numbers, content purchaser names, and other visible identifiers that remained after the TXT conversions in the previous step.

4. Procure a second copy of a suspected watermarked ebook and repeat step 1-3 for it. The more copies you can obtain, the more accurate the analysis will be.

5. Compare MD5 hashes of the two textfiles. If a discrepancy exists, the ebooks either employ natural language watermarking, or you fucked up on the spacing somewhere when deleting visible identifiers in Step 3.

6. If MD5 hashes do not match, run a side-by-side analysis of the two textfiles. Look over all found discrepancies to see if they were errors in spacing/conversion, or if they constitute possible examples of NLW. (There are a number of free programs to aide you in the side-by-side analysis, merely do a web search for 'compare two textfiles').

7. If cases of NLW have clearly been isolated in the ebooks, the question of what to do next is debatable. If only two copies of the ebook are available, it would be difficult to determine which of any two differing phrases is the 'original' and which is the modified, watermarked version. Furthermore, if one manually modifies all differing sentences in both versions, and each version only has certain sentences that are watermarked, then the modification of all formerly-watermarked sentences would still preserve the uniqueness of the text, because it will still be the only text with differing sentences in place of the 'original' ones, that--while different from the original publishing-planted watermarks--will nonetheless be totally unique from any other copy of the text anyway.

The copyright holders would have to run their own side-by-side comparison to isolate the differences and then cross-reference them to find the matching watermark key, thereby again potentially isolating the specific copy and purchase of the ebook. Yet, even though it is uncertain as to how successful this sort of counter-counter-move on the part of the copyright holders would be, at this point, this general protocol itself can at best be used merely to identify whether or not an ebook employs NLW.

Wrapping Up

All of this shit is most certainly not meant to discourage anyone from scanning or distributing ebooks. On the contrary, it is meant to ensure the continuation thereof through taking the proper security precautions to avoid apprehension. The key take-away points are:

~Don't leave any personally identifiable markers on the book. This includes fingerprints, creases and folds, library stickers, serial numbers, and metadata.

~OCR the content whenever possible.

~If dealing with distributing existent ebooks (whether purchased or Advance Reading Copies), don't fucking trust proprietary formats like PDFs or LITs. Always convert to ASCII TXT to minimize the possibility of format-based watermarking...

~...but don't fucking trust text either! Always perform side-by-side analyses whenever possible to make sure no natural language watermarking is present in the text prior to distribution.

And remember, better paranoid as fuck than dead as a duck.

-

Comments? Get in touch: xcon0 @t yahoo \/d0t/\ c||o|m (or call +1 (610) 887-6072)

For more knowledge check out www.rorta.net and www.dizzy.ws

____

Under the pleasant norms of Parisian life, beneath the veneer of culture and civilisation, one of the bitterest and most sadistic underground wars of modern history was fought out.

Last edited by DIzzIE; 25th June 2008 at 08:00 PM.