

# PERSONANONDATA

PUBLISHING INDUSTRY NEWS, TRENDS AND STRATEGIES IMPORTANT TO PUBLISHERS AND INFORMATION PROVIDERS.

## Google Print Project

Thomas Rudin, the associate general council for Microsoft, lambasted Google's approach to copyright protection characterizing it as 'cavalier' in comments delivered at the Association of American Publishers conference in New York. Those of us in publishing have a first-hand understanding of this opinion and other segments of media are rapidly coming to a realization that even obvious content ownership isn't enough to preclude Google from adopting and, more importantly, making money from content under copyright. Google is probably the only company that was willing to take the significant legal risks associated with the purchase of YouTube, for example.

Publishers have elected to sue Google to protect their content rights and those of their authors. At the same time, publishers have engaged with Google by participating in the Google Scholar program. Here publishers are equal partners and (I assume) negotiations for the acquisition of content by Google were conducted in good faith, and the results have been good to great for both parties (i.e., Springer, Cambridge University). It is also no bad thing that Google's content (digitization) programs have encouraged other similar content initiatives- particularly those of some of the larger trade and academic publishers.

The continued area of friction is the digitization project Google initiated to scan every book in any library willing to participate. This is where publishers got upset. They were not consulted or asked permission; they cannot approve the quality of the scanning; they will not participate in any revenue generated; they cannot take for granted that the availability of the scanned book will not undercut any potential revenues they may generate on their own. The books in question are the majority of those published after 1923, which are still under copyright protection.

Having said that, let's get one thing straight: Having all books which exist in library stacks (or deep storage) available in electronic form so that they can be found, indexed, searched, or even reassembled and generally resourced in an easy way is a good thing - an important step forward and opportunity for libraries and library patrons. Ideally, it would lead to one platform (network) providing equal access to high-quality, indexed e-book content which any library patron would be able to access via their local library. Sadly, while the vision is still viable, the execution represented by the Google library program is not going to get us there.

Setting aside the copyright issue, the Google library program has been underway for approximately 24mths and results and feedback are starting to show that, in reality, the program is not living up to its promise. According to this post from Tim O'Reilly, the scans are not high quality and, importantly, are not sufficient to support academic research. Assuming this is universally true (?), the program represents a fantastic opportunity lost for patrons, libraries and Google. BowerBird via O'Reilly states:

"umichigan is putting up the o.c.r. from its google scans, for the public-domain books anyway, so the other search engines will be able to scrape that text with ease. what you will find, though, if you look at it (for even as little as a minute or two) is that the quality is so inferior it's almost worthless"

Could Google suffer more embarrassment as disillusion grows over the program? Perhaps; but I doubt it will force them to rethink their methodology. It would be humbling for Google to 'return to the table' with publishers and libraries to work with them to rethink the project with the intention of resolving the copyright issues, and devising a better way to process and tag the content. To suggest that Google become less a content repository and more a navigator or 'switchboard' (which is how O'Reilly phrases it) is beyond expectation; however, were they to change course in this way, they would immediately reap benefits with all segments of the publishing and library communities. O'Reilly - a strong supporter of the Google program - believes the search engines (Google, Yahoo, others) will 'lose' if they continue to create content repositories that are not 'open'.

Ironically, the lawsuit brought by the AAP could actually have a beneficial impact on the process of digitization. As some have noted, we may have underestimated the difficulty in finding relevant materials and resources once there is more content to search (this is assuming full text is available for search). Initiatives are underway, particularly by Library of Congress, to address the bibliographic (metadata) requirements of a world with lots more content. It is possible the results of some of these bibliographic activities will result in a better approach to digitization of the more recent content (post 1923). Regrettably, some believe that since there may be only one opportunity to scan the materials in libraries, we may have lost the only opportunity to make these (older) materials easily accessible to users.

***Tomorrow: Just what is the universe of titles in the post 1923 'bucket'? The supporters of the Google project speak about a universe of 30million books but deeper analysis suggests the number is wildly exaggerated.***

# PERSONANONDATA

PUBLISHING INDUSTRY NEWS, TRENDS AND STRATEGIES IMPORTANT TO PUBLISHERS AND INFORMATION PROVIDERS.

## Google Print: A Numbers Game

*The following post is written by Andrew Grabois who worked with me at Bowker and has (among other things) compiled bibliographic stats out of the Books In Print database for a number of years. His contact details are at the bottom of this article.*

On February 6th, Google announced that the Princeton University library system agreed to participate in their Book Search Library Project. According to the announcement, Princeton and Google will identify one million works in the public domain for digitization. This follows the January 19th announcement that the University of Texas libraries, the fifth largest library in the U.S., also climbed on board the Library Project. Very quietly, the number of major research libraries participating in the project has more than doubled to twelve in the last two years. The seven new libraries will add millions of printed items to the tens of millions already held by the original five, and more fuel to the legal fire surrounding Google's plan to scan library holdings and make the full texts searchable on the web.

The public discussion has been mostly one-sided, with Google supporters trying to hold the high moral ground. Their basic argument goes something like this: The universe of published works in the U.S. consists of some 32 million books. They argue that while 80 percent of these books were published after 1923, and are, therefore, potentially protected by copyright, only 3 million of them are still in-print and available for sale. As a result, mountains of books have been unnecessarily consigned to obscurity.

No one has yet challenged the basic assumptions supporting this argument. Perhaps they've been scared off by Google's reputation for creating clever algorithms that "organize the world's information". This one, though, doesn't stand up to serious scrutiny.

The figures used by supporters of the Library Project come from a [2005 study](#) undertaken by the Online Computer Library Center (OCLC), the largest consortium of libraries in the U.S. According to the OCLC study, its 20,000 member libraries hold 31,923,000 print books; the original five research libraries participating in the Google library scanning project hold over 18 million.

OCLC did not actually count physical books. They searched their massive database of one billion library holdings and isolated 55 million catalog records describing "language-based monographs". This was further refined (eliminating duplicates) to 32 million "unique manifestations", not including government publications, theses and dissertations. The reality of library classification, however, is such that "monographs" often include things like pamphlets, unbound documents, reports, manuals, and ephemera that we don't usually think of as commercially published books.

The notion that 32 million U.S. published books languish on library shelves is absurd. Just do the math. That works out to more than 80,000 new books published every year since the first English settlement in Jamestown in 1607. Historical book production figures clearly show that the 80,000-threshold was not crossed until the 1980's, after hovering around 10,000 for fifty years between 1910 to 1958. The OCLC study showed, moreover, that member libraries added a staggering 17 million items (half of all print collections) since 1980. That averages out to 680,000 new print items acquired every year for 25 years, or more than the combined national outputs of the U.S., U.K., China, and Japan in 2004.

Not only will Google have to sift through printed collections to identify books, and then determine if they are in the public domain, but they will also have to separate out those published in the U.S. (assuming that their priority is scanning U.S.-based English-language books) from the sea of books published elsewhere. The OCLC study clearly showed that most printed materials held by U.S. libraries were not published in the U.S. The study counted more than 400 languages system-wide, and more than 3 million print materials published in French and German alone in the original Google Five. English-language print materials accounted for only 52% of holdings system-wide, and 49% in the Google Five. Since more than a few works were probably published in the United Kingdom, the total number of English-language books published in the U.S. will constitute less than half of all print collections, both system-wide and in Google libraries.

So how many U.S.-published books are there in our libraries? Annual book production figures show that some 4 million books have been published in the 125 years since figures were regularly compiled in 1880. If, very conservatively, we add an additional 1.5 million books to cover the pre-1880 years, and another 1.5 million to cover books published after 1880 that might have been missed, we get a much more realistic total of 7 million.

Using the lower baseline for published books tells a very different story than the dark one (that the universe of books consists of works that are out-of-print, in the public domain, or "orphaned" in copyright limbo) told by Google and their supporters. With some 3 million U.S. books in print, the inconvenient truth here is that 40% of all books ever published in the U.S. could still be protected by copyright. That would appear to jive with the OCLC finding that 75% of print items held by U.S. libraries were published after 1945, and 50% after 1974.

If we're going to have a debate that may end up rewriting copyright law, let's have one based on facts, not wishful thinking.

*Andrew Grabois is a consultant to the publishing industry. He has compiled U.S. book production statistics since 1999. He can be reached at the following email address: [agrabois@yahoo.com](mailto:agrabois@yahoo.com)*

Information Media Partners: [michael.cairns@infomediapartners.com](mailto:michael.cairns@infomediapartners.com)  
[www.infomediapartners.com](http://www.infomediapartners.com)

Phone: 908 938 4889  
[personanondata.blogspot.com](http://personanondata.blogspot.com)