The Race to Create a Digital Library: Google Books vs. the Open Content Alliance
Klara Maidenberg, FIS2309, Design of Electronic Text

Abstract
In recent years, a number of organizations have begun the task of digitizing great numbers of books and making them accessible via the Internet. The two largest and best publicized initiatives are the Google Books and Open Content Alliance (OCA) projects. While they both aim to make large numbers of books accessible to users online, the two initiatives have several important differences, such as the motives of the parties involved, the transparency of the digitization process, and the way in which copyright issues are handled. This paper provides readers with an overview of the history and current issues surrounding these two consortia projects, and proposes some potential implications that they might have on libraries and their users.

Keywords: Electronic books, Google Books, Open Content Alliance, Digitization, Open Access

Introduction
The short history of electronic books can be traced back to 1971, when Project Gutenberg (www.gutenberg.org) volunteers started digitizing books before there was even a widely available Internet on which to distribute them. The first commercial packages of electronic books appeared around the same time as the first CD-ROMs, because it was now possible to scan full text into a computer, and convert it to digital files. The Library of the Future was one of the first to distribute these products. The first edition appeared in 1991, a single CD-ROM with about 300 public-domain literary works displayed in ASCII text, selling for $695.00. Later editions increased the number of public-domain works included, and reduced the price. Today, CD-ROM compilations inspired by the Library of the Future and containing more than three thousand public domain items can be purchased on eBay for less than three dollars. In the early 90’s, eBooks moved further towards wider acceptance when the Library Journal published a cover story about them in September 1992, and Adobe Acrobat first appeared in November 1992 (Mullin, 2002). Since that time, the idea of reading a book in electronic form has gained mainstream acceptance and many libraries have added electronic books to their collections.

More recently, as libraries learned that the equipment and staffing needed to perform large scale digitization are expensive, and as technology increased the capacity of servers to store massive digital files, a number of collaborative efforts have formed to collectively share the expenses and reap the benefits that mass digitization offers. Two such efforts, the Google Books project and the Open Content Alliance (OCA) have garnered the widest public and expert support. This paper provides readers with an overview of the history and current debate concerning these two initiatives, as well as an analysis of how these projects impact on libraries.

Google Books
Google Print, one of many services Google has introduced as extensions of its popular search engine, was announced in December 2004. The program, collectively called Google Books, has two parts: Google Print Publisher and Google Print Library (Dye, 2006).
In Google Print Publisher, publishers can sign up for free to submit their books for inclusion in Google's search index and receive half the revenue from contextual ads that Google pairs with search results. As books in Google Print Publisher are searched, a bibliographic record appears, and users can view the page on which the search term is located, plus up to two pages on either side of the keyword. Also displayed with search results are links to Web sites selling the book, including the publisher itself, along with book vendors like Amazon.com. Although Google scans and stores the full text of each book on its servers, a few pages are purposely excluded, and users cannot print or copy images. Print Publisher has received largely positive reactions from publishers, authors, and users. Penn State Press, a nonprofit, scholarly publisher, agreed to put a significant portion of its catalog into Print Publisher during test stages of the program, and Tony Sanfilippo, marketing and sales director at Penn State Press, said that he would recommend it to his nonprofit and commercial peers (Dye, 2006).

In Google Print Publisher, publishers have a proactive say as to which of their books are scanned, but with Google Print Library, Google delves into the stacks of major libraries at the University of Michigan, Harvard University, Stanford, Oxford, and the New York Public Library and scans the collections regardless of copyright status. Google provides the labor and financial backing in exchange for access to the books, and it creates two digital copies, one going into Google Print library and the other going to the university that supplied the item. Google has committed an estimated $200 million to scan and index 15 million books by 2015.

Print Library has been met with widespread criticism. Publishers have complained that the project robs them of revenue in the digital book area of their business. Since libraries can now get digitized copies of print books from Google, they will no longer buy those same ebooks directly from publishers (Dye, 2006). In response to this criticism, Google has described the program as being an “electronic card-catalog” that helps users locate information. Google claims that the program benefits copyright holders and publishers by making books more discoverable, and therefore, more likely to be purchased (Balas, 2007).

So far, many prominent libraries have accepted Google’s offer, such as the New York Public Library and libraries at the University of Michigan, Harvard, Stanford and Oxford (Hafner, 2007). However, the resistance from some libraries, like the Boston Public Library and the Smithsonian Institution, suggests that many academic and non-profit institutions are intent on pursuing a vision of the Web as a global repository of knowledge that is free of business interests or restrictions. Even though Google’s program could make millions of books available to hundreds of millions of Internet users for the first time, some libraries and researchers worry that if any one company comes to dominate the digital conversion of these works, it could exploit that dominance for commercial gain. This concern is heightened by the fact that Google requires participating libraries to install a technology that would block commercial search services unaffiliated with Google from indexing the books (Hafner, 2007).

Another criticism of the Print Library initiative reflects its liberal interpretation and application of copyright law. In 2005, less than a year after Google announced its intention to scan library books, both the Authors Guild and the American Association of Publishers (AAP) filed separate lawsuits challenging that Google was violating the Copyright Act by reproducing copyrighted material for commercial gain (Balas, 2007, Tennant, 2005). While attempting to negotiate a compromise with members of the AAP, Google voluntarily agreed to stop scanning copyrighted material. The AAP proposed a solution based on the unique ISBN number that has been assigned to every book published since 1967. Using ISBN numbers, the AAP argued, Google could determine which works are under copyright and contact the publisher and author to
obtain permission before scanning. When Google rejected the ISBN proposal, talks broke down and Google resumed scanning the books in question (Dye, 2006). However, in response to these pressures, Google has begun to allow publishers and copyright holders to opt out of participation (Dye, 2006; Albanese, 2005).

**Open Content Alliance**

The previously mentioned lawsuits, which challenged the legality of Google's book digitization plan, allowed enough of a delay in Google's scanning process to provide engine rival Yahoo! an opportunity to roll out its own book digitization project (Dye, 2006; Quint, October 3, 2005). In 2005, Yahoo! announced a collaborative initiative amongst a number of international cultural, technological, non-profit, and governmental organizations, who all began working on a new mass digitization project with the goal of establishing a flexible, open infrastructure for bringing large collections of digitized material into the open Web (Quint, October 3, 2005). This initiative, conceived by Brewster Kahle, founder of the Internet Archive, in collaboration with Yahoo!, was first announced in October of 2005. The project, called the Open Content Alliance (OCA), aims to permanently archive librarian selected digital content, via a new model of collaborative library collection building. Brewster Kahle explained the alliance and the contributions of its initial partners as follows: “The Internet Archive will host the material and sometimes help with digitization; Yahoo will index the content and is also funding the digitization of an initial corpus of an American literature collection that the University of California (UC) system is selecting. Adobe and HP are helping with the processing software, University of Toronto and O'Reilly are adding books, Prelinger Archives and the National Archives of the UK are adding movies, and other items” (Albanese, 2005, p. 14).

The OCA’s content collections cover a wide range of material, including digitized print and multimedia content that will range from fiction to children's books to engineering white papers. These collections, posted on the Internet Archive’s website, even include the contents of T-Space, the digital archives from the University of Toronto and other Canadian universities, which were built using MIT’s DSpace format (Quint, October 3, 2005).

The OCA’s governing group, which is comprised of representatives of the collaborating institutions, has laid out a number of guiding principles, which include the goal of encouraging the greatest possible degree of access to, and reuse of collections in the archive while respecting the rights of content owners and contributors. The OCA also commits to offering collection and item-level metadata of its hosted collections in a variety of formats such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and RSS (www.opencontentalliance.org). The OCA encourages members to create and share tools (including finding aids, catalogs, and indexes) that will enhance the usability of the materials in the archive. Finally, copies of the OCA collections will reside in multiple archives internationally to ensure their long-term preservation and accessibility to all (www.opencontentalliance.org).

The content in the OCA archive is made accessible through the OCA website and through the Internet Archive. Additionally, Yahoo! indexes all content stored by the OCA in its search engine, to make it findable and accessible to Internet users. Meanwhile, the policies on use and access of individual items vary based on the parameters set by the contributing institutions. For example, the collection of American literature contributed by the Internet Archive, the University of California, and Yahoo! carries no restrictions and may be downloaded and reused for any purpose. There is even some interest in the publishing world about the opportunities for exposure created by the OCA’s work. O'Reilly Media, a relatively new commercial publisher, has agreed to make certain content available to the OCA in the hopes of making it more discoverable by readers (Quint 2005).
The Research Libraries Group (RLG; http://www.rlg.org), a major library bibliographic utility, has also joined OCA, contributing its bibliographic metadata. RLG plans to supply bibliographic descriptions to OCA digitizing operations from the more than 48 million titles in its RLG Union Catalog (available for direct searching at http://www.redlightgreen.com).

Though much smaller in membership than OCLC, the other major library bibliographic utility, RLG's membership of more than 150 research libraries, archives, and museums have a breadth of subjects, languages, and content types in their collections that should assist OCA in handling archives of older, public domain material (Quint October 31 2005).

In contrast with Google Print's secretive policy toward its proprietary digitization equipment (Ashmore & Grogg, 2008; Peek, 2007), the Open Content Alliance has released extensive details on its Scribe system, as well as other options for participants and users. The OCA also learned from Google's mistakes in the sphere of copyright law (Albanese, 2007).

The second similarity between the two initiatives is their archival function. Through digitization, both projects contribute to creating durable digital copies of materials, which can be used as backups of brittle and deteriorating items, and can be invaluable in cases where fires, floods and other disasters damage or destroy the hard-copies of books (Albanese, 2006).

At OCA's inaugural event, Brewster Kahle stated that OCA would try to target the 80% of books published between 1923 and 1964 that are out of copyright (Tennant, 2005), then expand to include orphaned books, where the publisher and author cannot be found, then out-of-print works, and finally in-print material (Quint, October 31, 2005). For the distribution of copyrighted content, the OCA uses Creative Commons licenses (http://www.creativecommons.org), which offer a number of licensing models that encourage personal use, reuse, and flexible access to digital content (Quint October 3 2005).

**Similarities and Differences**

So what features do both endeavours have in common? Firstly, both benefit from the financial backing and technological expertise of large corporations. While Google Books is funded and administered by the search engine giant, Microsoft had initially joined the Open Content Alliance with an estimated $5 million promise to digitize approximately 150,000 books (Quint, October 31, 2005). However, Microsoft subsequently withdrew its participation in the OCA initiative (Ashmore & Grogg, 2008).

While both groups essentially perform the task of digitizing and making available the full texts of books, a number of key differences between them have to be noted. These differences include their interpretations of copyright, the extent to which they are willing to share their methods and technology with the public, their motives for undertaking the task of digitization, and finally, the needs of the end-users that they seek to fulfill. For example, the needs of leisurely web-surfers looking for suggestions on books to purchase are very different from the needs of researchers and historians interested in carefully reading and examining the entire contents of historically significant books. While the functions of Google Books are sufficient for the needs of the former type of user, the latter user would need unobstructed access to the full text of a book, a use that is not always possible through Google Books.

With regards to copyright, as has already been mentioned, the OCA currently restricts its efforts to works in the public domain, such as those works whose copyright period has ended (Ashmore & Grogg, 2008). Google, in contrast, scans both copyrighted and public domain works. The question of whether Google's digitization project constitutes fair use and the way it applies
to digitized material is one of the issues in the debate over Google’s digitization initiatives (Balas, 2007).

Quint (October 31, 2005) notes that the OCA is much more open about its equipment and sharing the resulting files than Google Print. In fact, it was the observation that Google does not want the books to appear in any other search engine’s listing but their own that fuelled Brewster Kahle into creating the OCA (Balas, 2007). Johnson (2007) notes that libraries that wish to work with Google must agree to a set of protective terms, which include a commitment to making the digitized materials unavailable to other commercial search services. The Open Content Alliance, by contrast, is making the material available to any search service (Hafner, 2007).

Where underlying motives are concerned, Google, which is a commercial, profit-driven enterprise, obviously benefits from including book searches in its arsenal of services. So significant is the revenue from the search traffic added by this product, that Google in fact pays the libraries to gain access to the books (Hafner, 2007). The OCA, in contrast, has no monetary gain. It costs the OCA as much as $30 to scan each book, a cost shared by the group’s members and benefactors (Tennant, 2005), so there are obvious financial benefits to Google’s offer (Hafner 2007). Libraries that sign with the Open Content Alliance are obligated to pay the cost of scanning the books but do not need to pay the Internet Archive for storing and sharing them.

**Implications for Libraries**

Overall, the existence of initiatives like the OCA and Google Books benefits all libraries and their patrons, even when the library does not participate in the digitization effort. Because full texts of an increasingly large number of books are available for free online, through Google or the Internet Archive, patrons looking for the mention of a particular term no longer need to flip through pages of books or use a printed index to find it. Instead, they can use the built-in search features to find the occurrences of the term in context. This results in saved time, and decreased damage to rare and brittle books.

Additionally, those patrons who do not have access to a library or to specific rare or historically significant books, can now read these books, in whole or in part, for free over the Internet. This accessibility to valuable knowledge is a welcome breach in the cultural and digital divide. Society as a whole benefits from the addition of cultural treasures to the public domain, where more individuals can enjoy and learn from them and utilize them for research that will continue to enrich humanity.

Clearly, these two initiatives, and others like them, are transforming the way people access information. In this changing landscape of information access, librarians have to decide whether they will join the digitization movement, and to what extent. The awareness of the existence of these electronic copies of valuable books is a small step towards an increased incorporation of ebooks into the library’s repertoire, and may be all the changing some libraries will do for now. Many libraries may have already begun small-scale digitization projects, done in-house, to preserve brittle or particularly valuable books, or to widen access to important materials for their remote users. In such local projects, the products of digitization are often not widely shared beyond the library itself (Ashmore & Grogg, 2008). However, as many libraries have already realized, since the expense of acquiring and staffing the equipment required for digitization is relatively high, it beneficial to become a part of a consortium effort where resources and costs can be shared. If the decision to join a consortium has been made, the further question seems to be “which one?”

To help in making this decision, librarians need to become educated about the differences and similarities of various digitization projects, as was outlined in detail above. Libraries must iron out questions about how the copyright of their materials will be handled, what restrictions on use
and users will be in place, and finally, what kind of benefit (monetary or other) they hope to reap from making their books available online. In the words of University of California’s Daniel Greenstein: "There is a huge window of opportunity based on the search engines' competition with each other. However, the race to add content also brings with it a challenge. If the academic and library communities don't begin to define what our basic requirements are for a massively digitized openly accessible file, we could find ourselves just being taken advantage of" (Albanese, 2005, p. 15).

Conclusion

As Brewster Kahle observes, today’s readers and researchers expect that most things will be available electronically. This expectation is reinforced by the fact that e-journals and newspapers are now widely accessible in electronic form. It is Kahle’s opinion that librarians now have the responsibility to take books into the new technological age by facilitating their discovery and access online (Albanese, 2007). The work being done by Google Books and the OCA to digitize books and make them accessible online is thus an exciting sign of progress in the world of information retrieval. From the interest these projects have received in the mass and professional library media, as well as in live and electronic discussions, it appears that most people are embracing digitized collections as a welcome addition to our information retrieval options. However, it is important to clarify the distinctions between the two initiatives, and for professionals involved in the fields of publishing and librarianship to understand that the methods and goals of Google Books and OCA are not the same.

Google, a commercial enterprise, created its Books feature to gain more traffic to its search engine, and therefore, gain more revenue. The OCA, by contrast, is a non-profit operation focused on the humanistic goal of preserving and disseminating important resources as widely as possible. Additionally, the two consortia differ in their views and treatment of copyright and in their willingness to share the products of digitization with other online enterprises. In the words of Paul Duguid, adjunct professor at the School of Information at the University of California, Berkey: “There are two opposed pathways being mapped out. One is shaped by commercial concerns, the other by a commitment to openness, and which one will win is not clear” (Hafner, 2007).

Works Cited


Resources Consulted


