October 2007
Rollin', Rollin': Google Book Search

Smart Libraries Newsletter

Smart Libraries Newsletter delivers hard data and innovative insights about the world of library technology, every month.

TO SUBSCRIBE

To reserve your subscription, contact the Customer Service Center at **800-545-2433, press 5 for assistance**, or visit **www.techsource.ala.org**.

The 2007 subscription price is just $85 US.

Rollin', Rollin': Google Book Search

Google Book Search (GBS), the growing online resource born in part from a massive book scanning project involving dozens of partner libraries around the world, just keeps rolling along, like Old Man River. You can love GBS, hate it, or be ambivalent toward it, but it may be the most significant information-related project since the creation of the Internet. Google Book Search already contains a wide variety of things, including the full scanned texts, snippets and limited previews of books, metadata, mashups, and user-contributed tags, reviews, and comments. Late summer was an unusually active period for the GBS project, with an unexpected flood of announcements, assessments, and new products.

Cornell Joins the Fold

In early August Cornell University announced that it had become the 27th library to join the Google project. In a portentous, if not downright pretentious, quote in the press release, attributed to Cornell President David Skorton, the project is placed in a context whereby Cornell's quest to become the world's land-grant university will be realized, resulting in a build-up of human capacity around the globe.

That's a lot of hullabaloo over the scanning of only 500,000 volumes. Google seems to have become a little particular about which volumes it wants to add to its growing digitized collection. Over the next six years approximately a half million books, both public domain and copyright protected works, from Cornell's Mann Library will be scanned. Overall, Cornell's twenty libraries hold close to eight million volumes, so this is little more than a drop in the bucket. Mann's collections are strong in the life sciences, agriculture, nutrition and food science, environmental sciences, human ecology, business, and applied social sciences.

Cornell already is involved in other digitization projects, partnering with Microsoft, Amazon.com, the U.S. Agriculture Information Network, and others. Cornell's registry of digital collections already contains over forty entries.

Trying to Do Good

Critical appraisals of the usability and usefulness of the massive GBS project—if it had been undertaken with government support, it would have been labeled a public works project—continue to roil and froth, too. For instance, in its August 2007 issue *First Monday* published a thoughtful piece by Paul Duguid. He argues that most assurances of quality with regard to web-based resources redound either to innovation or inheritance—relying on institutional authority and quality assurance techniques that antedate the emergence of the Web. In essence, many web-based resources basically buy and convert existing authority. Duguid suggests that

Receive *Smart Libraries* **via e-mail**

Subscribers that would like an e-mailed version of the newsletter each month should forward one e-mail address and all of the mailing label information printed on page 8 of the newsletter to jfoley@ala.org. Type "e-mail my Smart Libraries" into the subject line. In addition to your monthly printed newsletter, you will receive an electronic copy via e-mail (to one address per paid subscription) at no extra charge each month.

quality assurance in Google Book Search comes primarily via inheritance, because Google is drawing on the reputations of the research libraries involved, as well as the authors, editors, and publishers of the individual books being scanned. The basic question Duguid asks about the fate of inherited quality during digitization and reformatting is this, "Is quality necessarily inherited when old institutions provide established content in new digitized forms, or may the process of migration too easily leave behind significant aspects of the quality it was presumed to be carrying along?"

Because the Google Book Search project is vast, yet of undisclosed size at any given moment, Duguid suggests that "…we seem to have little option but to take its overall quality on faith (and on the reputation of the organizations involved)…" or attempt to get a sense of the overall quality through random sampling and examination.

Using Google Book Search proved to be very trying to Duguid. He uses his search for the delightful novel, *Tristram Shandy*, as his case study. What he found were some seriously mismanaged page scans, where, for instance, the left part of the text block on the page was completely missing from the scan, leaving the poor reader to infer the missing words, which probably causes Sterne to chuckle from his grave.

Duguid also discovered that the Google Book Search algorithm does a poor job, or perhaps no job at all, of distinguishing the various volumes in a multi-volume work. Duguid also

laments, almost to the point of lacrimation, that the production practices and the algorithm are causing inferior editions to float to the top of the digital heap of search results. "Google may or may not be sucking the air out of other digitization projects, but like Project Gutenberg before, it is certainly sucking better-forgotten versions of classic texts from justified oblivion and presenting them as the first choice to readers." With a project this size, Duguid could cry us a river.

Duguid concludes that it may be Google's digital technicians, not librarians, who are the great romantics of books as pristine, indistinguishable storehouses of wisdom.

Google Book Search in itself is a project of mind-boggling size. Contemplating ways to clean up the seriously flawed individual page scans that are being made throughout the project could lead to insanity. If there ever was a project aptly suited to the adage to grab the brass ring as it goes by, it is a mass digitization project. By the time a seriously flawed page scan is discovered, probably most often by users, it would be prohibitively expensive and time-consuming to find, retrieve, and rescan the troublesome pages.

Perhaps Google should open up the quality assurance aspect of Google Book Search and let individual users find and re-scan problem pages. Such an effort would open another can of digitization worms, but it may be the only hope of cleaning up the egregious scanning errors that are inevitable.

Kahle Opines

The August 15th issue of *Library Journal* contained an interview with Brewster Kahle by Andrew Albanese. Kahle is the founder of the Internet Archive and one of the leaders of the Open Content Alliance (OCA), another massive scanning project. For Kahle, the opportunity to "…merge the texts and traditions of our print past and our web future…represents a truly historic moment in our culture." OCA can be understood as the most viable alternative to GBS. OCA relies primarily on foundation support and in-kind contributions from the partner organizations, while GBS seems to rely primarily on revenues from online advertising, Google's bread-and-butter business. OCA is focusing on books in the public domain, while GBS also is including copyright-protected content, some scanned from books held by the partner research libraries, others provided directly by publishers. All OCA content is freely and fully available to everyone, while Google has to restrict access to some content in GBS, resulting in the infamous snippets, rising book sales at online and bricks-and-mortar bookstores, and, one can only hope, increased use of obscure books held in libraries.

Both Albanese and Kahle are wary of Google Book Search, which admittedly is scanning at a faster pace than is OCA or any other contender at the moment. Albanese observes, "But with its gold rush, "scan first, ask questions later" approach, Google's library program, despite myriad potential benefits, has also wrought confusion, lawsuits from publishers and authors, and serious concerns about how our shared, public domain heritage could be parceled away by commercial gatekeepers in the coming digital generations."

Kahle and others envision the Open Content Alliance as an open, inclusive initiative involving universities, public

> ## Duguid concludes that it may be Google's digital technicians, not librarians, who are the great romantics of books as pristine, undistinguishable storehouses of wisdom.

libraries, foundations, and commercial companies, all working together to create something out of scanned books in the public domain that will be a digital public good for centuries to come. His major concerns about Google's agreements with libraries are, first, the prospect of perpetual restrictions on public domain books, and, second, the fact that these negotiations are being conducted in secret, followed by triumphant press releases.

Kahle admits that he is a bit surprised that libraries continue to align themselves with the Google project now that the OCA has emerged as a viable, more open and socially friendly alternative. "The public domain is small enough as it stands… let's not clobber it again as it goes digital."

Kahle also makes some observations about the struggling ebook industry, which erred in putting the DRM-managed interfaces, devices, and content before the open content horse, contrary to how the Web achieved a critical mass of content. "DRM … didn't work for the software industry, it's not working for music, and it won't work for books."

Kahle is concerned that the library system in the U.S. is becoming too cozy with corporations and corporate values. He says, "I see the library system in this country as a $12 billion industry dedicated to preservation and access of materials that are not mediated through a corporate experience….This is one of the biggest issues facing libraries in the future: what services will they perform, and what services will be performed by companies or by nonprofits acting like companies."

My Library is Your Library

In early September Google announced that a "My Library" set of features had been added to Google Books. Whenever you conduct a search in Google Books and retrieve some books, you can select to add one or more of the books to your library. NB: you need to have created a Google account in order to set up and begin populating a Google Books My Library. Once you have built up a library, you can limit your search to only the books you have selected to include in your library. You can write book reviews, add tags (called "labels") in My Library, give the book a rating (the usual one to five stars) with one click, import books automatically by inputting a list of ISBNs, export your library information, and even set up an RSS feed. And apparently your library of books and book information selected from Google Book Search is available for all the world to see—friends, family, employers, identity thieves, et alia.

The My Library project harkens back to LibraryThing. Although LibraryThing is a richer experience with books and bookish people, in the long haul it may be unwise to dismiss Google's first effort in this field.

Google Book's My Library tool set makes me wonder if I have been focusing too much on the mass digitization aspects of this project. Yes, the scanning process has been unnecessarily secretive, the negotiations have been private, and the output is occasionally of woefully poor quality, but eventually the scanning will be largely complete. Perhaps obsessing over the mass digitization in order to understand the deeper meaning and future implications of this project is akin to trying to understand the meaning and importance of oil to 20th and 21st century developed countries by focusing on the refining process. Or, to return like a salmon to the river metaphor, the processes by which all the concrete that resulted in the Grand Coulee Dam was created, poured, and cured are fascinating, but, once the dam is complete, it becomes an enormous, undeniable fact of the landscape and the ecosystem. Interest in the process that created it wanes in the public imagination, and our collective attention turns to the more enduring question: What have we wrought? —Tom Peters

More Info. @:

Google Book Search:
 http://books.google.com/
Google Book Search My Library:
 http://books.google.com/googlebooks/mylibrary/
Cornell Press Release:
 http://www.library.cornell.edu/communications/Google/
Registry of Cornell's Digital Collections:
 http://rdc.library.cornell.edu/search/index.php?mode=browse&type=Collection
Paul Duguid's article in First Monday:
 http://www.firstmonday.org/issues/issue12_8/duguid/index.html
Andrew Albanese's Interview of Brewster Kahle in Library Journal:
 http://www.libraryjournal.com/article/CA6466634.html
Open Content Alliance:
 http://www.opencontentalliance.org/

Partial Management Buyout at Mandarin

Although the larger companies tend to monopolize the headlines, there continues to be a number of smaller companies that also participate in the library automation industry. These smaller companies often are more specialized, offering lower cost products to smaller-sized libraries. Broadly speaking, these companies sustain high levels of satisfaction in their client libraries.

Mandarin Library Automation, standing in the ranks of these smaller companies, recently took a step forward in the company's business evolution. In a transaction completed on July 1, 2007, management gained ownership of fifty percent of the company. Three of the company's executives, including Leo Lazo, President, Christine DiVito, Comptroller, and Art Graham, Vice President of Technical Services, acquired half of the company's stock. In this move, the executives responsible for managing the company gain a substantial stake in its ownership as well.

Although this shift represents only an incremental step in the evolution of Mandarin Library Automation, it presents *Smart Libraries Newsletter* with the opportunity to spotlight one of the smaller companies in the industry.

Mandarin Library Automation caters mostly to the K-12 school library market, though small academic and special libraries use its products as well. Mandarin Library Automation indicates that its products are used in libraries in 50 countries worldwide, but most of its clients are in Canada and the U.S.

The company's flagship product, Mandarin Oasis, was introduced in 2005, as a fully Web-based automation system. Mandarin Oasis fits well within the trend in the K-12 school library arena of moving away from PC-based systems installed in each library toward centrally managed district-wide Web-based systems.

M3, the company's primary product since 1999, continues to be widely implemented. Shortly after the launch of Oasis, the company began offering the M3 automation system as a free download. Any library can install and use the basic M3 product without charge and without support, but can optionally purchase add-on modules and support. The free Mandarin M3 download includes a 32-Bit client/server architecture, an online public access catalog, circulation and cataloging modules, configuration and group editors, and a data import/export utility.

Prior to this latest transition, Mandarin Library Automation was wholly owned by EGEG Holdings, which in turn is owned by Elliot Goldstein and Eleanor Goldstein of Boca

Raton, Florida. At its peak, the EGEG Holdings portfolio included SIRS Publishing, SCP Commercial Printing, and Mandarin Library Automation.

Elliot and Eleanor Goldstein started an educational publishing company operating under the name Social Issues Resources Series, Inc. The company's core products included SIRS Researcher, SIRS Government Reporter, and SIRS Discoverer, subscription products originally delivered on CD-ROM and later through the Web. SIRS products primarily targeted K-12 Schools.

SIRS involvement in the library automation arena began in 1994 when the company entered into a joint marketing agreement with Melchior Management Systems, a company based in Montréal, Canada, that created the Mandarin Library Automation system.

Melchior was founded to develop library software. In 1986, the company introduced the Mandarin library automation system, one of the early PC-based systems that included integrated cataloging and circulation modules and that used MARC records. Mandarin was a DOS-based system that gained a strong reputation among the early PC-based systems.

After about a year of joint marketing, EGEG Holdings, through SIRS, acquired the Mandarin software from Melchior Management Systems on July 1, 1995. Harry Chan, formerly President of Melchior, came on board as a Vice President for SIRS and continued oversight of the Mandarin software through a new division named SIRS Canada. The acquisition of Melchior by SIRS was positioned at the time as a natural transition for the company, given the success of the joint marketing agreement. This move greatly expanded the opportunities to market Mandarin in the U.S through SIRS' existing marketing infrastructure.

As interest in DOS-based systems waned, the company began development of a next-generation system, called M3, launched in 1999.

Following the acquisition of Mandarin by SIRS, the company operated as a merged business, operating under the name SIRS Mandarin. Once acquired, the library automation division reported to Leo Lazo, the Senior Vice President for SIRS. At this point, much of the development of the Mandarin software was transferred to Boca Raton, Florida.

In August 2001 EGEG reorganized its businesses, separating the publishing and library automation operations, forming two business units called Sirs Publishing Inc. and Mandarin

Library Automation, with Leo Lazo as President. While the two businesses continued to share some administrative and support resources, they began operating independently. This move paved the way for upcoming divergent business developments.

In November 2001 the company closed the offices in Montreal, the headquarters of Melchior prior to its acquisition. At this point Harry Chan exited the company and Leo Lazo stepped in as President of the newly-formed Mandarin Library Automation while continuing his role as Senior VP at SIRS.

Chan continues to be involved in the library automation arena as owner of Media Flex, Inc., which offers a number of products (furniture, supplies, labels, security strips, etc.) and services for libraries, including hosting and support for the open source OPALS-NA library automation system for schools, small businesses, and churches.

In a transaction that closed on July 21, 2003, EGEG sold SIRS Publishing, Inc. to ProQuest Information and Learning. The SIRS databases were a natural complement to the other information products offered by ProQuest and strengthened their presence in the K-12 school arena.

Following the sale of SIRS to ProQuest, Mandarin Library Automation continued under the ownership of EGEG Holdings. This latest move to sell half of its interest in Mandarin Library

Automation to this management group represents another step in the scaling down in its investments. Yet, the influence of EGEG Holdings remains strong since it still owns half of the company. Elliot Goldstein will continue to serve as the Chairman of the Board of Directors of Mandarin Library Automation, Inc.

In the current business cycle in the library automation industry we see increasing investments by private equity and consolidations through mergers and acquisition. This management buyout bucks that trend. A transition toward ownership by the executives who run the company seems to be a move that will allow Mandarin Library Automation to continue to prosper in its niche of the industry rather than become the latest victim on the mergers and acquisitions front.

—Marshall Breeding

More Info. @:

Mandarin Library Automation:
http://www.mlasolutions.com/
Media Flex, Inc.:
http://www.mediaflex.net/
OPALS-NA Library Automation System:
http://www.opals-na.org/

Hats Off to Fedora Commons

In mid-August the Fedora Commons announced that it received a four-year $4.9 million grant from the Gordon and Betty Moore Foundation to develop organizational frameworks and technological tools to change how knowledge workers and cultural institutions create, share, and preserve their digital intellectual creations.

Fedora Commons is a recently created non-profit organization, the child of the Fedora Project, a collaborative open-source software project involving Cornell University and the University of Virginia that has received grant support from the Andrew W. Mellon Foundation, and the grandchild of the Flexible Extensible Digital Object Repository Architecture (Fedora) developed at Cornell Computing and Information Science.

According to the press release, the open-source system will support a networked model of intellectual activity—creating new knowledge and building on, annotating, and refining the ideas of colleagues. Fedora Commons "…will focus on the integrity and longevity of the intellectual products that underlie this new form of knowledge work."

The Moore Foundation funds also will be used to expand the community of Fedora partner organizations who help develop the software, implement applications, and provide outreach to other organizations. Established in September 2000, the Foundation has three programmatic areas of support: environmental conservation, cutting edge, interdisciplinary, results-driven science, and the improvement of the quality of life in the San Francisco Bay Area. Already in 2007 the Foundation has awarded nearly $60 million in grants.

Fedora Commons is taking a soup-to-nuts approach to the information life-cycle, storing raw data, analyzing data in novel ways, expediting the peer review and publication processes, reusing published information to generate new knowledge, and storing it all in repositories that are sustainable and extensible.

—Tom Peters

More Info. @:

Fedora Commons:
http://www.fedora-commons.org
Press Release:
http://www.fedora-commons.org/about/news.
php#moore-grant
Gordon and Betty Moore Foundation:
http://www.moore.org

Google and Microsoft Want to Be Your Health Information Manager

Health information, like the weather and the source and substance of your next meal, is a topic that interests everyone. Increasingly, when people feel the need for more health-related information—before, during, after, or (egads) in lieu of a consultation with a healthcare professional—they turn to the Web.

Information seeking, health-related tests, diagnoses, and treatments create a personal health information file or database. Most of this information is still contained in printed files, and most of the information is controlled by hospitals, clinics, and health insurance providers, not by individual patients. If you think the quest for the paperless office has involved more than its fair share of tilting at windmills, imagine the quest of creating the paperless clinic, let alone a nationwide primarily paperless health information and record-keeping system.

Both Microsoft and Google want to change the basic system by which health-related information is created, stored, shared, and controlled. An article in the August 14, 2007 edition of the *New York Times* outlined their basic plans. Microsoft plans to unveil (unleash?) it consumer-centric health information system later in 2007. Google's initiative, according to the

NY Times report, has experienced some delays and probably will not launch (even as a perpetual beta phase) until sometime in 2008.

It will be interesting to see how consumers of health services respond to the idea of having large, generalist corporations like Microsoft and Google handle their personal health information and records. The current situation regarding personal health records is a wide variety of relatively small records-holding clinics and other organizations. Mid-sized companies specializing in health-related information, such as WebMD, may try to shove Google and Microsoft out of a sector of the information economy that they feel is theirs.

Although most libraries do not have a direct interest in this coming health records culture clash, if Google and/or Microsoft are successful in gaining significant market share, and redefining how most people think about and use their personal health information, the pattern whereby one or two large corporations essentially take over a segment of the information economy that has heretofore been largely decentralized, paper-based, and atomized may have libraries watching their backs.

—Tom Peters

Structure Beyond Shelter

Most people would agree with the adage that food, clothing, and shelter are the three necessary and sufficient external resources needed to keep body and soul together. Food implies liquids, and clothing in its elemental state is a form of shelter, so it may be possible to boil the absolute necessities of life down to two things: food and shelter. Everything else is not superfluous, but yet not absolutely essential. Although Thomas Jefferson once wrote in a letter to John Adams "I cannot live without books…." he probably never tested his hypothesis. He lived over eleven years after writing that letter, and he apparently never abstained from books for any length of time.

We humans not only construct shelters for ourselves, but also for our belongings and things we value. The "will to shelter" probably soon branched out to encompass livestock and furniture. Now we have three-car garages, safety deposit boxes, and all manner of specialized shelters.

Libraries serve a sheltering function, too. Real life libraries protect books and other documents—not to mention the librarians and users—from the elements. When the elements intrude on a library, through flood, wind, and fire, it's news. Library structures need to provide shelter, and they need to bear the weight of their contents. These two prerequisites have had a profound impact on library architecture.

As librarianship moves into virtual worlds, the purpose, role, and meaning of library as structure become problematic. In the short history of library architecture in Second Life, a three-dimensional virtual reality world, there already have been some interesting evolutionary developments.

Librarians who are active in Second Life are wondering why we even need library buildings, other than to give avatar-patrons a comforting sense of similarity with real-world libraries. Making avatar-patrons feel comfortable sounds like an unqualified good thing to do, but by trying to do so we may actually be retarding both their and our understanding of how this new information space should be structured based on its natural affordances.

Already there have been long discussions why doors and impressive staircases are needed in an environment that contains no weather, no ultraviolet radiation, no destructive fire, and no warping water, and where avatars can easily fly to the upper floors of the building, if the architect thought of that as a design possibility.

Quick, draw a detailed diagram of how the hard disk on your computer stores your files. Obviously, we need to distinguish between information storage structures and information presentation structures. In the amusing story (p. 99) in David Weinberger's recent book, *Everything Is Miscellaneous*, he relates the response of Brion Vibber, the chief technical officer of Wikipedia, to Weinberger's question about where the text information for the Wikipedia article on elephants actually is stored:

<brion> god only knows.
<brion> On the disk somewheres

For virtual world users of information resources, and even for designers of virtual world information presentation structures, they do not need to know exactly how and where the physical information (bits and bytes) is stored, but they need to think long and hard about how to organize and present information for optimal usability and usefulness. In other words, we need to think about how to structure beyond shelter.

These are just opening discussions and experimentations in what promises to be a long, professional exploration of how to organize and present information in a structured three-dimensional environment when the "will to shelter" no longer is a driving consideration in that conversation. Sure, the structure of essentially two-dimensional web pages may provide some guidance and cautionary tales in this coming quest, but perhaps the most interesting and useful information structures for virtual worlds will emerge in response to the basic question: If we could structure a three-dimensional information environment without worrying about the need to shelter and protect substantial information objects, how would we design that information interface? —Tom Peters