Google Book Search and The University of Michigan

Anne Karle-Zenith, Special Projects Librarian, University of Michigan University Library

818 Harlan Hatcher Graduate Library
Ann Arbor, MI 48103-1205
Email: annekz@umich.edu

Abstract/Background

In December 2004, Google and the University of Michigan announced they were partnering in an ambitious project to digitize the bound print volumes of the University Library collections. The digitization project, known locally as the Michigan Digitization Project, will provide scholars and the general public with an unprecedented ability to search for and locate books from the University's vast collection. This initiative has the potential to revolutionize the way the world's knowledge is transmitted and to democratize access to information. In addition, university libraries are uniquely tasked by the public to be repositories of human knowledge and information. The digital archive resulting from this project will significantly advance the University of Michigan's ability to meet that responsibility. This paper is an update on the status of the partnership between Google and the University of Michigan, including how and why the partnership came about, what has taken place since the project was announced, and future developments.

History of Digital Initiatives at the University of Michigan

The University of Michigan (UM) Library has been involved in digital initiatives since the late 1980s. At that time, the Library was already experimenting with how to provide access to primary information in electronic formats. In 1990 we introduced UMLibText, a command line-based service that provided campus-wide access to Middle- and Early Modern-English works via a textual analysis server. The works were marked up using SGML, based on the then-emerging Text Encoding Initiative guidelines. By 1994, we had implemented a gateway from a CGI-compliant web server to PAT, Open Text's text search engine and began putting full text resources online. These efforts eventually evolved into the Humanities Text Initiative (HTI). With a wider variety of collections and a broader user base than UMLibText, the HTI was designed as an umbrella organization for the creation and maintenance of online text, and as a mechanism for furthering the Library's capabilities in the area of electronic text. HTI supported the delivery of externally created SGML collections, and collaborated with publishers and other academic operations to design and build local access mechanisms for their titles. By 1996, we had already managed to put almost 2 million pages of encoded text online.

Around this same time there were several major digitization projects going on at the Library in addition to HTI. PEAK (Pricing Electronic Access to Knowledge) was a project to set up an experimental service for electronic journal delivery and explore new possibilities for pricing of subscriptions. The PEAK information service provided access to over 1100 journals published by Elsevier Science, including search and retrieval functions. The Making of America Project was a partnership between UM and Cornell University to digitize and provide access to primary source materials documenting American social history from the antebellum period through the reconstruction. In addition, JSTOR was originally based at the University of Michigan before it split off as a non-profit.

These digital initiatives were primarily being supported through grants, and were operating simultaneously but independently of one another. It was at that point, in 1996, that the UM Digital Library Production Service (DLPS) was formed. DLPS was intended to provide infrastructure for campus digital library collections, including creating access and providing digitization services. Originally jointly funded by the Library, the University Information Technology Unit, the Media Union and the School of Information, DLPS gathered existing units and projects under one umbrella organization focused on production (as opposed to projects), as well as long term access and preservation of our digital collections.

DLPS developed several significant products and services for library digital collections. Some examples include:

- DLXS middleware – a set of open source tools for mounting collections of digital library content
- The XPAT search engine
- OAISTER, a collection of metadata describing digital objects, which is

“harvested” from libraries and other institutions. OAISTER currently comprises almost 10 million records from over 700 institutions around the world

DLPS also provides hosting services for other academic institutions' and non-profits' digital collections.

Today the UM Library digital collections include text, images, numeric spatial data, finding aids, and bibliographic and reference collections. Even prior to the Michigan Digitization Project, we had 141 text collections online comprising 25 million page images, and 89 image collections with approximately 200,000 images. Clearly, the UM Library has been at the forefront of library digital initiatives, and has understood the value of investing in such initiatives, for a long time.

Google-University of Michigan Partnership

In December 2004, Google and the University of Michigan announced that they were partnering to digitize the collections of the UM University Library System, which includes 19 libraries with holdings of over 7 million volumes. The materials would come from UM's bound print collections (i.e., monographs and serials), the exceptions being Special Collections materials and very large items such as folios. The digitized volumes would be searchable via Google's search engine, and the full text of the out-of-copyright materials would be available for viewing and downloading. In addition we would receive our own digital copy of each volume, which we were permitted to integrate into services available via the Library website. Google agreed to cover most costs related to digitization of materials. (University of Michigan is a publicly funded university, therefore our contract with Google is publicly available and can be accessed on the Library website.) In order to distinguish between our locally digitized materials and the materials being digitized by Google, we decided to refer to the Google project at UM as the “Michigan Digitization Project” or “MDP.”

The Library was excited about engaging in the partnership with Google for several reasons. This is the UM Library mission statement:

> *To support, enhance, and collaborate in the instructional, research, and service activities of faculty, students, staff and contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge.*

When Google proposed this partnership, we realized that the project fit perfectly with the Library’s mission. The instructional and research activities of faculty, students and staff will be enhanced through this project by providing new ways for users to access library materials (e.g., through Google). We planned to create our own interface to the electronic works, which would be specifically designed to support the needs of scholars. We will be able to use the project to reconceive how library services should be provided based on the increased availability of electronic content in our collections. In addition, the Library will be contributing to the common good by making the works that aren’t restricted under copyright freely accessible to the world, and by creating a digital archive of library materials that can be curated and preserved for the future.

Finally, one of the primary reasons we wanted to engage in this project is because it was clear it could never do this on our own. Mass digitization projects on this scale are only possible through partnerships such as this one. Although the UM Library has been involved in digitization initiatives for many years, at our current production rates, it would take us more than a thousand years to digitize what Google will be able to accomplish in less than ten.

Progress and Current Status

As of this writing, it has been approximately two and a half years since Google and UM announced the partnership. Much progress has been made since that time.

Digitization of materials started at about the time of the project’s announcement. Work began with collections in the Library’s off-site storage facility (Buhr) primarily because it is a closed stacks environment, so processes and procedures could be worked out without disruption to our patrons. We worked closely with Google to develop workflows on how to get materials from the shelves to the scanning location and back, and to develop systems for tracking the status of the volumes as they moved through the process.

As capabilities improved, capacity increased. Refinement of processes and procedures brought about decreases in turnaround time. The flow of materials in and out has now stabilized, and production levels will remain at the current rate for the foreseeable future. While we do not discuss the total number of volumes scanned, we can say that at this point tens of thousands of volumes go in and out on a weekly basis. Books are typically off the shelf for only 5-8 days.

The Library also had to set up mechanisms to receive the digitized volumes from Google, as well as a storage infrastructure to support the archiving of the digital copies. This “flow” of digitized material from Google to us also continues to increase, to the point where we are getting content from Google on almost a daily basis. The images we receive are generated through a derivation process. Google creates a high quality master, and then extracts from that master to create copies that adhere to library preservation standards. The Library currently receives 600 dpi bitonal TIFFS for text pages, and 300 dpi JPEGs for pages with illustrations (with plans to switch to JPEG2000 in 2007), as well as OCR (Optical Character Recognition) text files. Automatic quality assurance checks are performed on the files as they are received, and we have also put a sampling process in place, which allows staff to do “manual” quality review of the images in order to provide feedback to Google on a continual basis.

In September of 2006 we released our local implementation of the project at UM - we call it “MBooks.” We built this system to provide access to our copies of the materials digitized by Google. Access is through MIRLYN, the UM Library OPAC. Currently MBooks only includes materials that are part of the MDP, but eventually we plan to move most of the Library’s existing digital text collections into the MDP repository and make them accessible through the MBooks interface.

We also created a rights database that stores rights information for each digitized volume that is part of MBooks. The rights information determines who will have access to which works. This was set up using an algorithm that takes several factors into account:

- Copyright status, and/or explicit access controls associated with a volume
- The volume’s source (meaning digitized by Google or digitized by UM)
- The identity of the user (if known)

For works not protected by copyright, such as works that have fallen into the public domain, or US Government documents, users are able to view the full text of the work by clicking on a link in the bibliographic record. We use The Handle System to provide links, instead of URLs. The Handle serves as a permanent pointer to the resource, so even if we move the volumes around on our servers, the Handle will always point to the then-current location. This is important for scholars who include citations to these works in their publications, as it will ensure that the citation will always be correct.

Some features of the MBooks interface include the ability to navigate through the book (page by page, or jump around), to change the size of images, and to rotate the images. It includes options to view the page images as high-resolution image files, as PDFs, or as text files. The text files are generated from the OCR, and allow users to cut and paste text, and also can be used with screen readers (for visually impaired users). In addition, users can search for a term or phrase within the book and the results will return all the pages where that term or phrase appears, with links to those pages.

For works not in the public domain, users cannot view the full text. We should emphasize that **no one** can see these in-copyright works – they are not accessible on the UM campus, they are not accessible in the Library – only a very small number of Library staff even have access to them.

The link in the OPAC record is still there, but it leads to a page stating that the work is restricted under copyright. Users can still search within the in-copyright books, and the results will again display how frequently and on which pages the search term appears, but there are no links to the page images. Even without the ability to view the text, knowing how frequently a search term or phrase appears in a work can help users judge if a resource is relevant enough to go to the library and pick up the book, or order it through interlibrary loan.

Future Plans

One of the Library's top priorities at this point in the project is acquiring the storage space to store the digitized volumes. The current version of MBooks is a prototype that relies on less expensive and less reliable storage. We are estimating that we will ultimately need to store about 380 TB of material. The Library has recently undertaken an RFP process to secure more reliable and scalable storage. We plan to select a vendor and bring the larger storage system online in mid-2007.

We are starting to plan a timetable and strategize about how to move the operation from the Buhr remote storage facility into other libraries within the University Library system. As the other libraries are primarily open stacks, there are many questions about how to facilitate the operation, and how to schedule the work in order to avoid taking materials off the shelves at crucial times for patrons.

We have plans to further develop the MBooks interface. Preliminary discussions with users and with public service Library staff helped us to identify the need for mechanisms to support “collection building” within MBooks. We will devote significant resources to this development in the coming year. With such a mechanism in place, it should be possible for several types of collection-oriented actions, including, for example, the ability of faculty to create collections for courses.

Working with guidance from the Office of the General Counsel and the Office of Services for Students with Disabilities, and in conversations with the National Federation of the Blind and with legal scholars, the Library has also drafted access mechanisms for “reading” in-copyright works by approved students with disabilities. We hope to fully define and begin deploying those mechanisms before the Fall 2007 semester.

Finally, we hope to take advantage of US copyright law’s exceptions for libraries and archives in order to provide access to more of the in-copyright materials. Again, working with guidance from UM’s Office of the General Counsel, we will provide access to many brittle and damaged in-copyright volumes as permitted under the Section 108 provisions of the copyright law. Although the law requires us to restrict full-text access to these materials to users in University Library facilities, this approach will constitute a significant enhancement to access. We have also reallocated some staffing to support research into the rights status of materials that appear to be in copyright but may have fallen into the public domain. In the future we hope to be able to take advantage of legislation regarding orphan works (not yet through Congress), and we also plan to pursue permissions from publishers to provide access to their materials.

Conclusion

With its long history of involvement in major library digital initiatives, the University of Michigan Library was a natural choice to partner with Google to digitize its collections. Two and a half years into the project, the partnership is already yielding results. The UM Library and Google have already made the texts of many thousands of volumes available for searching, by anyone with access to the Internet, through MBooks and Google Book Search. In addition, users can view the full text of the works that are not protected under copyright. University of Michigan President Mary Sue Coleman summed up the University's view of the project in a speech she gave at last year's "Scholarship and Libraries in Transition" symposium, held in Ann Arbor in March 2006:

> *At its essence, digitizing books and widening their exposure is about the public good. I believe it transcends debates about snippets, and copyright, and who owns what when, and rises to the very ideal of a university – particularly a great public university like Michigan. Our work is about the social good of promoting and sharing knowledge. As a university we have no other choice but to make this happen.*

Useful Links:

Michigan Digitization Project – http://www.lib.umich.edu/mdp

Mirlyn (UM online catalog) – http://mirlyn.lib.umich.edu

University of Michigan Digital Library Collections –

http://www.hti.umich.edu/cgi/c/collsize/collsize

Humanities Text Initiative - http://www.hti.umich.edu/

PEAK - http://www.lib.umich.edu/retired/peak/

Making of America - http://www.hti.umich.edu/m/moagrp/

DLXS - http://www.dlxs.org/

XPAT - http://www.dlxs.org/products/xpat.html

OAISTER - http://oaister.umdl.umich.edu/o/oaister/

The Handle System - http://www.handle.net/