Creating a free to read international digital library - five years later

Erika Linke
Carnegie Mellon University

Creating a Free-to-Read International Digital Library – Five Years Later

Erika C. Linke
Carnegie Mellon University, U.S.A.
el08@andrew.cmu.edu

Abstract:

The Million Book Project was begun in 2003 with the vision of creating a free-to-read, universally accessible, million-book digital resource by 2007 to provide a test bed for developing digital library tools and software. By spring 2007 1.4 million titles were digitized.

The project stemmed from a collaboration of the university library and the school of computer science at Carnegie Mellon University. From the project's inception partners included libraries, institutions of higher education and research organizations in China, India and the U.S., essentially computer scientists and librarians. Additional partners have been added selectively as the project became more visible in the international community.

This paper will discuss the challenges faced over the last five years through the partnership of organizations, some linked to national digital strategies and some not; highlight lessons learned from this project; and reflect on the future of this and other large scale digitization efforts.

Keywords: Million Book Project; International digitization projects

The Million Book Project began in 2003 with the goal of building a 1,000,000 volume digital library. The project achieved its millionth volume in 2006 and by September 2007, 1.6 million volumes had been digitized. Initially this international project included partners from China, India and the United States.[1] Midway in the project period the Bibliotheca Alexandrina in Alexandria, Egypt [2] joined the project as a partner bringing its distinctive collection, technical expertise and unique digital library challenges. The project has been underwritten by the U.S. National Science Foundation, the Chinese Academy of Science, the Chinese Ministry of Education, the Indian Institute of Science and the Indian Ministry of Communications and Information Technology. Throughout the project period additional government agencies and funders helped support the project.

The goal of the project was to build an international digital library that included out of copyright content in a broad array of languages. The primary purpose in setting a target of one million volumes was to create a significantly sized test bed to further research in machine translation, automatic summarization and other technical and strategic innovations. The international and diverse partnership insured that collection content included text in multiple languages, differing scripts and a wide range of subjects. Stephen Griffin acknowledged the project as one of few large scale models of international collaboration—across continents, language and type of material.[3]

During the course of project development and implementation another important objective emerged. The opportunity to showcase national content and pride in having important heritage collections have a presence on the web became a significant motivation. To showcase national heritage material both to citizens and the world influenced the range of collection materials made available for scanning. Initial plans that emphasized the desire to have the collection focus on scientific literature gave way to the hard reality of intellectual property and copyright. Project content included scientific literature when possible. Heritage collections, that is library collections that showcase literature, history and philosophy, formed another facet of the collection.

All the participants have been proud of reaching the collection milestone of 1,000,000 and exceeding the goal by 600,000 titles. The project partners envisioned, coordinated, championed and harmonized the dispersed efforts of the individual segments of the overall

project. Beneath the success story are the realities of the project and lessons gained from working together.

Challenges and Realities

Throughout the project, there has been unbridled optimism about the feasibility of reaching content goals and merging the content into one massive data file mirrored at partner sites. The commitment and enthusiasm remain, yet the reality is somewhat different in actual detail.

In the area of collection content, the numerical goal of one million volumes was achieved and surpassed. In that regard the content met project expectations. Much of the initial scanning content came from the libraries at Carnegie Mellon University. By using the report function of the online catalog, books published prior to 1923 were identified as candidates for scanning. Under U.S. copyright law, books published before 1923 are in the public domain. The plan was to pull the titles from the collection and ship them by container to designated scanning centers in India and China. Under the old copyright regime in the United States, books published between 1923 and 1963 held copyright for 28 years. For the copyright to be renewed for another 28 years, the copyright holder had to file renewal forms to secure the second 28 year period. These renewal forms were published in multiple volumes by the Copyright Office of the Library of Congress. The project digitized these volumes and with the assistance of Dr. Michael Lesk made these available for search through both manual search and batch process. This development allowed the library to create files of books held by the library and published between 1923 and 1963. These files could then be batch checked against the copyright renewal file that had been created. Those books whose copyright had not been renewed were made available for scanning.

Intellectual property remains a central issue or obstacle in identifying and securing permission to scan more current works. Interestingly China passed a law that has had a profound effect on materials available digitally at Chinese universities. The law permits the scanning of new materials published in China; display is only available at Chinese universities. These scanned materials are available bibliographically in the project but full text is only available to a partial segment, i.e. students and faculty at Chinese universities.

One of the stated goals by partner organizations was a desire to include current scientific and technical literature as well as other currently published content. At the onset of the project some partners did not fully comprehend the import of copyright. There was an unstated expectation that the project could provide a way to access online materials licensed by partner institutions, but this proved not to be the case. The project in general could not include the scanning of current, in copyright material or the inclusion of material licensed for local use. In the area of agricultural literature however, the project was able to borrow and scan agricultural documents from the U.S. National Agriculture Library and from selected agriculture universities in the United States. These materials were government documents that were in the public domain and were relatively current.

Carnegie Mellon engaged in a copyright project to ascertain the willingness of publishers to allow the scanning and OCR of in-copyright but out of print material. Denise Troll Covey of Carnegie Mellon studied the problem.[4] Through those efforts a range of publishers provided permission for scanning and posting the OCR'd files.

The challenges in drawing on these permissions were numerous. In some instances publisher permission was spelled out in general terms, leaving it to the library to identify the titles that fell into the stated guidelines. Whether the title was identified by guideline or spelled out precisely by the publisher, not all titles identified were owned at Carnegie Mellon. In fact few of the titles were actually held by the Carnegie Mellon library. The idea then was to secure or borrow copies from libraries elsewhere in the United States. The books would be gone for an extended period of time. These different paths to identifying and acquiring content created potential to send large shipments of books to partners for scanning. Therein was a major challenge to the project.

Imagine if you will a container filled with books en route to China and India destined for scanning centers. A container might have as many as 30,000 books. Clearly the time to scan, sort and process and then repack and reship would be time consuming. Before shipping a

second container of books and then, additionally, books from other collections, the Carnegie Mellon library decided to use the first shipment as trial—to understand better the logistical issues, the length of time that books would be unavailable, the condition of the returned books, and most importantly wanted to see the quality of both metadata and the scanned files.

It took well over a year for the books to be transported and returned. Traveling by container is hard on books; those books that were in any state of disrepair were further damaged by the shipping and handling. More problematic was the inability to gain access to the metadata and scanned files. One of the seeming surprises in the project was the difficulty of moving the scanned files from one site to another. In the optimistic vein, assumptions had been made about band-width for moving the files through the internet. This was not possible, partly because of a difference in world-wide band-width that had not been fully appreciated. When this method did not work as anticipated, partners resorted to burning CDs and shipping them. This too did not work as anticipated and files and metadata were not mounted in a system that could be readily accessed. Finally the solution appeared to be the transport of files simply on hard disk and moved by partners traveling from one site to another in the course of project business. Even now the project continues to pull together the files on disk and is attempting to mount them in a central location with fully mirrored sites.

Inability to assess the quality of scans and metadata quality was an impediment to collecting and sending more materials from the U.S. for scanning at partner sites. Without having access to work underway, it was difficult to speak confidently to other institutions about scan and metadata quality. The wear and tear on volumes sent was another factor that concerned potential lenders.

Though files were mounted on local sites there was no single site having all the content. It was from the partner sites that concerns arose about variations in metadata quality. Early plans took advantage of a partnership with OCLC who provided for the project a limited number of IDs and passwords that would permit access to OCLC's bibliographic records as the basis for metadata. Again the project did not fully comprehend the technical capacity and stability that would be essential to making effective use of OCLC to populate a Dublin Core record. As a consequence most metadata for the project came not from OCLC as envisioned but from keyed-in data.

Keyed-in data was only as good as the capabilities of scanning and quality control operators. It was later acknowledged that the educational level of the operators was not as high as it should have been. Knowledge of English was essential but was not as robust as hoped. Even within the range of languages of India, there were issues related to subject assignment and language identification. For example, some texts in Urdu were identified as Arabic; some languages of India are similar so that an individual, not a native speaker to both languages, might confuse them. Some texts in western languages were misidentified. Metadata for subjects were sometimes off target. Initially issues around metadata accuracy were not completely appreciated. It was late in the project that concerns finally surfaced around this issue and additional manpower was deployed to correct errors both in batch and by hand. This work is ongoing.

Annually a meeting of partners sets the stage for sharing of progress and problems—ICUDL (International Conference on the Universal Digital Library). It is at these meetings that disciplinary challenges and cultural issues within the project come to light. Overall there is agreement on the broad mission and goals of the project. At the disciplinary level, library science and computer science perspectives and approaches to problems lead to some tensions. Over the course of the project, each has come to better appreciate and to recognize the value of both perspectives to the benefit of the project. Disciplinary bonds are quite strong and commitment to the project is high. Each country has a different way of expressing agreement and disagreement. Some are more contentious, others are quite respectful of status and order. Even the way the project is viewed can be quite telling. The linkage of the Chinese funding, government connections and scanning targets are indicative of a construction project with expected attainment of five year goals. For others the project remains more of research project; targeted goals stem from research needs rather than governmental expectation and requirements.

The lack of consistent identification of the scanned files remains to be addressed. Without a uniform resource identifier the ability of users to locate, reuse and direct others to a resource is compromised. This weakness relates to the dispersion of the collection currently and

the different structures for the files. Thus far the metadata does not include this type of identifier. The display formats among the sites differ. Some sites rely on specific viewers for the text. Some partners are considering a plan to create PDF version of the files for users and relying on the OCR'd text for retrieval. What this points to is the freedom in the project that allows for organic development of the content. This freedom allows for changes to occur as technology catches up with expectation.

Computer scientists explore, develop and research. Librarians want the collection to be of benefit and use now and in the future. These objectives sometimes lead to heated discussion about the direction and timing of the project. The aims of partners at the discipline level sometimes create tensions and disagreements about critical issues. There is now agreement that one single interface and a single hosted server, with redundant servers, would be optimal. Redundancy is essential to maintaining daily availability. Though the project is collaborative, file ownership issues present one obstacle in implementation.

Future

One of the exciting features of the Million Book Project is pride among partners in the creation of a distinctive collection. It is highly diverse both in content and in language. Over twenty languages are evident in the content.[5] The documents are in Roman, Cyrillic, Arabic and Devanagari scripts and Chinese and Japanese characters. Traditionally in library catalogs, the search screen and subject terms are typically in the language of the home country. At the moment English is arguably the dominant research language. In this project there is a metadata bias for English and Chinese. What form should search and retrieval take when the content languages are so different and partner countries are from the west, East Asia, South Asia and the Middle East?

Content continues to grow and become more diverse. Though one might imagine the project coming to a conclusion, partners continue to be enthusiastic about the project. The Million Book Project has not been declared a completed project. Raj Reddy, Mozah Bint Nasser University Professor of Computer Science and Robotics in the School of Computer Science at Carnegie Mellon, declared that the future direction of the project will focus on all authored works of mankind, not just books. Indian partners have secured permission to digitize materials that were unavailable earlier. For example, the publisher of *Chandamama*[6] will allow the project to digitize the content of this children's magazine. Its significance is that the magazine is published in sixteen languages and provides a natural test bed for machine translation. As the project has gained recognition, new opportunities to add content continue to surface. Even as the initial content goal has been achieved, opportunities to add additional content are being pursued. In India the scanning of several newspapers is under discussion. The State Senate of Andhra Pradesh has expressed interest in having their proceedings digitized and included in the project. In November, 2007 Professor Jihai Zhao of Zhejiang University noted that among the next steps for the project are the second and third million books.[7]

Research opportunities continue to be developed in parallel to the project. Each year at the ICUDL meeting, participants update attendees on new directions at the regional level and on research stemming from the project or research related to project aims. At the most recent meeting[8], papers and panels covered a wide range of topics from quality assurance to intellectual property to technical and human factor issues. Beyond those papers research interests continue to focus on copyright law, distribution and sustainability, image processing, language processing, machine translation, massive distributed systems, optical character recognition, search engines, security, storage formats, and use of digital libraries.

Seven years ago many disbelieved that large scale digital projects were feasible or desirable. Yet along came projects such as this one followed by Google's massive set of digitization plans and projects. Large scale digitization projects, and initiatives by national libraries and major research institutions as well as the projects underway at Google and Microsoft are still infant projects. The task of digitization is a stepping stone to an unwritten future. Librarians, scholars, researchers and educators participate in projects but the real impact of this can only be speculated. What sort of future might we imagine is before us? How might our roles evolve and what steps might we take to secure a place at the table.

Library organizations continue to address the perception that everything needed for research is a click away. Digital projects like this one tend to reinforce the notion that everything is online and easy to retrieve. There is great promise in what online and digital resources offer to users. The lack of true integration of sources is a barrier to effective access to the realm of materials available. Current commercial tools that assist in resource discovery have limitations. There is a need to develop better retrieval tools to aid and improve virtual resource integration Librarians can take up opportunities to collaborate in developing improved and new tools to manage and retrieve information from the massive amount of content now online. Future roles for librarians and information professionals will move from traditional reference points to the creation of recast outreach efforts. Much has been written in the library literature about marketing library services. Those efforts are important but it is the product—the service that is offered that would benefit from greater scrutiny. Value-added services will result in greater impact. For example, Purdue's efforts in supporting the collaboration of librarians and researchers document one way to provide value-added service.[9] Discovery, search and retrieval seem easy and seamless. Library patrons are able to retrieve resources that are good enough or sufficient to complete a class assignment. The "good enough" model has limits when accuracy and precision are essential. It is in this realm that librarians and information professionals can create services that build on their knowledge and skills and find methods and models for delivery this new level of service.

Librarians can drive this new model through strategic partnerships and seek opportunities to partner in or influence the creation of digital library tools. Libraries have always been about collections and services. Librarians are in a strong position to aid scholars and researchers in finding what's most relevant, reliable and accurate. Promotion of that vision must be balanced with value-added services that move online library resources and discovery from good enough to excellent.

Notes

[1] China: Beijing University, Chinese Academy of Science, Fudan University, Ministry of Education of China, Nanjing University, State Planning Commission of China, Tsinghua University and Zhejiang University.
India: Arulmigu Kalasalingam College of Engineering, Goa University, Indian Institute of Information Technology – Allahabad, Indian Institute of Science, International Institute of Information Technology – Hyderabad, Shanmugha Arts, Science, Technology & Research Academy, Tirumala Tirupati Devasthanams, Maharastra Industrial Development Corporation and University of Pune.
U.S.A.: Carnegie Mellon University. Libraries providing collection assistance include the Carnegie Library of Pittsburgh, the National Agriculture Library, Cornell University.
Universal Library Websites: China: http://www.ulib.org.cn and http://www.cadal.zju.edu.cn/IndexEng.action; Egypt: http://dar.bibalex.org/; India (three sites): http://dli.iiit.ac.in/; http://www.new.dli.ernet.in/ and http://www.dli.cdacnoida.in/; United States: http://www.ulib.org/ and http://tera-3.ul.cs.cmu.edu/oldudl/

[2] http://www.bibalex.org/

[3] Opening remarks of Stephen Griffin, National Science Foundation, at the 2007 ICUDL conference.

[4] Troll Covey, Denise. 2005. Acquiring Copyright Permission to Digitize and Provide Open Access to Books. CLIR report 134. Washington DC: Council on Library and Information Resources and Digital Library Federation. http://www.clir.org/pubs/abstract/pub134abst.html

[5] Among them are Arabic, Bengali, Chinese, English, French, German, Greek, Hindi, Italian, Japanese, Kannada, Malayalam, Marathi, Norwegian, Persian, Russian, Sanskrit, Spanish, Tamil, Telugu, and Urdu.

[6] Chandamama website: http://www.chandamama.com/

[7] Zhao, Jihai. 2007. Annual Progress in the Million Book Digital Library Project in China. Presentation at the 2007 International Conference on Universal Digital Library

[8] http://tera-3.ul.cs.cmu.edu/ULIBConference.htm

[9] Mullins, James L. 2006. Associate Dean for Research: The Libraries Commitment to Interdisciplinary/Collaborative Sponsored Research in the Libraries and Throughout the University. IATUL Proceedings.
http://www.iatul.org/doclibrary/public/Conf_Proceedings/2006/Mullinspaper.pdf