Identifying and Selecting Content for the Million Book Project

Christina Birdie, Indian Institute of Astrophysics, Bangalore, INDIA
Erika Linke, Carnegie Mellon, Pittsburgh, PA USA

Abstract

This paper focuses on collection development and implementation issues and challenges of the Million Book Project (MBP), an international digital library project. The project's objective is to create a free-to-read, universally accessible million-book digital resource. The creation of this digital storehouse will also provide a test bed for research and investigation on data mining, automatic summarization, machine translation, and development and transformation of digital library tools. Initially the project had fifteen partners in China, India, and the United States. As the project has evolved, additional project participants have joined the effort. The collaboration of both old and new partners is essential to the creation a large-scale, multi-national digital library. Librarians hailing from two of the partner countries, India and the United States, will focus on collection development and content selection among partners physically separate but virtually connected.

The Million Book Project

In its most basic form, a digital library has two main thrusts – one is the content of the library and the other is the technology used for digital library tools that can reliably deliver or mine the content. Digital library projects have become part of the library landscape over the last ten years. Early projects rightly showcased unique or unusual offerings or projects that had a direct connection to campus constituencies and that were attractive to funding agencies, alumni and community.[1] They not only focused on content but also on proof-of-concept. Today not much has changed—institutions and organizations want to provide the best content online using state of the art technology.

The Million Book Project stems from the vision of Dr. Raj Reddy, former dean of the School of Computer Science at Carnegie Mellon. His vision of a universal library available to all stems from a deep rooted belief that access to knowledge is a bridge to economic development and improved social conditions. The application of computing research, technology and tools can be the vehicle that makes the universal library universally available. These two notions form the foundation and driving vision of the project. Based on this vision, a proposal was sent to the National Science Foundation (U.S.) to seek support for creating this digital library and the test bed. Grants totaling $3.5 million have been secured. The project calls for scanning workshops to be set up in China and India. Scanners, computers, servers and software are funded from the grant, and manpower for scanning, indexing and hosting are provided by China and India. Partners in the project include universities, institutes and government agencies.[2] Early meetings of the project partners laid the foundation and groundwork for international cooperation. The content for the scanning project comes from partner libraries in the

United States, China and India. Scanning standards were established and a procedure manual to use at the scanning centers was produced.[3]

The Million Book Project focuses on content on a scale that would allow computer scientists and technologists a large test bed to continue developing tools to mine, explore, and further explore and test how new technology could improve access to digital libraries on a large scale. Tools to be developed and refined include machine translation, automatic summarization and information retrieval. To provide that test bed, a target of one million scanned books was established and computer scientists then turned to the library profession to assist in amassing the collection.

Collection Rationale

Building the Million Book Project collection bears an historical relationship to the development of early libraries, in the sense that it is systematic when possible and relies on the generosity of donors, the availability of the marketplace and the need to build a coherent collection. Just as we build paper collections today with users in mind, the modern digital library must also listen to users' needs. At the same time, intellectual property and copyright laws may be an obstacle to reaching greater mass of digital content. Like our forebears, the contemporary digital library builds on the cultural heritage of society, the great writers and philosophers of the world, and scientific and technology works. In specific, the collection begins as a collection of collections as project partners make choices related to local and or national need, precedent or requirement.

The Million Book Project relies on a host of factors in developing content. In the interest of managing time and keeping costs down, this project asks the question: If an item has already been selected for the library collection or transferred to storage, why should additional time and money be spent to select content on a title by title basis. We assert that librarians and curators have already selected titles once, and sometimes two or three times. In a project of this scope with manpower available for scanning, the revisiting of collection decisions does not make economic sense.

If one can accept that selecting content on a title-by-title basis is redundant and expensive, how then do we choose what to tap first for digital library content?

Content Selection

Each partner in the Million Book Project determines the parameters of their contribution to the content, determines the scope of what is to be included and selects the materials to be scanned. There is no overarching theme about what should be selected. The only requirement is that materials to be scanned are allowable—that they are out of copyright or that permission has been secured for materials still in copyright. To date, scanning in China has focused on cultural materials.[4] Partner countries have initially highlighted

cultural materials and expressed interested in building depth in the sciences and engineering.

Collection Activities in India

Collection development and coordination activities are in the very early stages as logistical and technical issues have center stage. Librarians at member partners in India have yet to confer on collections and collaboration.

At partner institutions and organizations in India, the following collections were broadly identified during the project development phase and during recent high level meetings of institutional and government officials of the three participating countries.

- Indian Institute of Science: 40,000 books no longer in copyright
- International Institute for Information Technology (Hyderabad) and the Government of Andhra Pradesh: Telugu textbooks
- Indian Institute of Information Technology: Sanskrit literature and science & technology books in English and Indian languages from Bose Library, Allahabad
- Pune University: Maharastrian literature and books
- Goa University: Portuguese literature and books
- Tirupathi and Tirumala Devasthanam: Sanskrit and Telugu literature and vedic documents, including palm leaf books
- Anna University: Tamil literature and palm leaves containing ancient Ayurveda medical practices
- National Centre for Software Development and the Government of Maharastra: Textbooks in Marathi and science & technology books
- SASTRA (Shanmugha Arts, Science, Technology and Research Academy): Sanskrit and Tamil literature from Tanjore Library dating to the 4th century B.C.
- Avinashalingam College: Books and manuscripts from old libraries in the Tamil Nadu regions in Tamil, Telugu, English and Sanskrit

A major scanning center has been established in Hyderabad with complementary sites established elsewhere. Librarians in India have not yet had an opportunity to network or discuss the collection issues related to the project. It would be beneficial to project and collection development for this to occur.

Collection Activities in the United States

Carnegie Mellon, along with some member libraries of the Digital Library Federation (DLF)[5] have held discussions about the issues inherent in a digital library of this magnitude, about the challenges in selecting material of benefit to members and about the questions around sending materials outside the U.S. for scanning. The compelling issue under discussion focused on the types of materials to be selected. The most serious concerns that participants voiced centered on the distance that books would travel to the scanning centers established in India and China.

Out of Copyright in the U.S. Materials that are out of copyright in the United States are an obvious group of materials that can be tapped for content. Those materials published before 1923 are in the public domain and no longer protected by copyright and thus may be freely scanned. An early pilot project of the Million Book Project was to identify a group of out of print materials to test the logistics of sending materials off site for scanning.

*U.S. Government and State Documents***.** Another group of materials that are not copyrighted are U.S. government documents and state documents. The U.S. Government Printing Office has had a rich history of publishing materials for a broad range of U.S. government departments and agencies on a wide variety of topics.

Technical Reports and Theses. Scientific and technical literature is highly sought after and project participants have expressed interest in seeing technical reports and theses included in the digital collection. At Carnegie Mellon, authors are being contacted to secure the necessary permissions to digitize and include earlier technical reports and include them. The project will pay close attention to this need and seek ways to include the broadest range of scientific and technical literature.

Best Books Approach. This approach centered on drawing on selections from standard recommended book lists. For example, **Books for College Libraries**,[6] a best books reference work, was used to identify titles of interest to the academic community and would be a good beginning for establishing a core collection. Only a small percentage of books are in the public domain so including other titles identified in the work posed a challenge.

Participant's Choice. Some DLF members suggested that individual institutions offer to take responsibility for selecting content in particular subject areas. Coordination of this effort would be very intensive and the need was expressed for some method or means of identifying titles that had been digitized or were going to be digitized. Furthermore, participants had some reservation and concern about the logistics related to scanning at a remote site. The participant's choice option has been put on hold for the moment for two reasons: One is that there is an initiative underway that would create a digital registry, and secondly, there is a pilot project to address issues related to the transport of materials.

Donations. As the Million Book Project has expanded, a new source of content, albeit very small, has been the digital donations of individuals who are seeking a home for materials that they have scanned. These materials of course must be in the public domain or copyright-cleared.

In Copyright Materials. A recurring topic and challenge centered on the copyright question throughout the early discussion about what to send and how content might be identified. At Carnegie Mellon efforts were made to address the problem. An early pilot project on copyright permissions sought to shed some light on the question.[7] The pilot project showed that it was possible to secure permission to digitize works still in

copyright provided publisher stipulations were met. Although some publishers granted permissions freely, others want to limit use to the campus community or granted permission for only a few years. Others sought a payment for the permission to copy.

As the Posner Family Collection[8] was being scanned, the need to secure copyright permissions arose again with some urgency. Previously, the National Academy Press had given us blanket permission to digitize all of their earlier publications. Their more recent publications were already in a digital form. It slowly became clear that the most productive way to secure copyright permissions would not be to ask for permissions on a title-by-title basis, but would best be secured by asking for broader permission to digitize materials. Thus, a new initiative was begun—to ask for permission to digitize out-of-print, but in copyright books. This very new approach has been met with some success.[9]

One outcome of the discussion among DLF members was the stated need for a pilot project to ascertain and understand the issues, costs, and uncertainties inherent in shipping books off site for scanning and, in this project, sending them out of the United States for scanning. That pilot project is currently underway with approximately 3,000 books from the Carnegie Mellon collections having been sent to India for scanning.

An issue that continues to surface around digital projects is the question about whether a title has already been digitized. The growth of digital libraries would be well served by creating a way to identify when materials have been digitally reformatted or to register the intent to reformat an item. The Digital Library Federation devoted some effort to addressing this topic[10] and have now partnered with OCLC. It is OCLC's intent to create a Digital Registry that would enable libraries to have a site where details about a digital object can be recorded. Data elements would include information about the format and scanning standards used for the object. This information would then be available so that a library or organization could consider whether or not to scan a title. At the present time, the OCLC Registry is in prototype and is being tested.

Although the Million Book Project is still in its infancy, issues about preservation and sustainability have been considered. In the short term, mirrored sites will offer a current way to have multiple copies. Various national agencies have been contacted to gauge their interest in supporting the project for the longer term. Furthermore, the preservation challenge is another of the many test bed applications to be developed.

As distant partners in this project, it is our observation that there are some special challenges inherent in this work. The first challenge is the collaborative aspect of the project. Working with another profession or academic group, in this case computer scientists, casts some of the terms and parameters of the project in a different way. For some of our computing partners, this project is a test bed, a potential place for discovery and experimentation. For librarians, however, it is important to remember that this is a project in its very early stages, that the tools that will enhance its use are only being developed now and that, in this early state, the importance of 24/7 access is not always as well understood by our research partners. Securing content is not as simple as computer

scientists would have hoped. Libraries are stewards of their collections and much has been invested in the growth, development and stature of the collection. For some, this project is perceived as a radical departure from past collection development practices. Developing the collection and securing the content is not easy; with a project of this scope, it can be a challenge to find partners to take part in this effort. As this is a joint project, its character is shaped by the social and cultural traits of the partners. Melding the computer and library culture to build this universal library is and will be challenging. The success of the million book project will be a worldwide resource available to all.

---

[1] Smith, Abby. *Strategies for Building Digitized Collections.* Washington, D.C., Digital Library Federation and Council on Library and Information Resources. September 2001.
See also: <http://www.clir.org/pubs/reports/pub101/pub101.pdf>

[2] These include in 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; and in 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 and Research Academy, Tirumala Tirupati Devasthanam, Maharashtra Industrial Development Corporation and the University of Pune.

[3] *Million Book Universal Library Project: Manual for Metadata Capture, Digitization, Post Processing and OCR*. See http://www.library.cmu.edu/Libraries/MillionBookManual.pdf [Accessed April 10, 2003]

[4] *Beijing University*: Ancient rare books including Song and Yuan Dynasty rare books, family trees, paintings and inscription rubbings. Chinese periodicals before 1949 in politics, law, culture, education, finance, economics, students, women, academics, technology, religion, folk customs, and natural sciences.
Tsinghua University: Ancient engineering technology history and study in China
China's contribution to science and technology, including engraved bone texts
Fudan University: Full text of documents in the Chinese Culture Documents Database;
Full text of documents in the Chinese Classical Literature Database; Full text of documents in the Chinese Classical Art Vision Database
Nanjing University: Full text of documents in the Jinling University Technical Periodical Database; "Six-dynasties Culture" multimedia database in cooperation with the library of the Nanjing Normal School
Zhejiang University: Dunhuang documents, including hand written and carved ones;
China's Southeast countryside area with local area geography, commercial town materials, and history and technical articles; Tea culture materials, ancient documents, periodicals, texts; Silk work and silk materials, including ancient documents, technical articles, texts.

[5] The Digital Library Federation is a consortium of libraries and other agencies committed to deploy information technology in support of collections and services. Three efforts of focus include the development of 'best practices,' coordination of leading edge digital library development and initiation of collaborative projects. See http://www.diglib.org/ [Accessed April 10, 2003]

[6] *Books for College Libraries: a Core Collection of 50,000 Titles*. 3rd edition. Chicago, American Library Association, 1988. 6 volumes.

[7] George, Carole A. "Exploring the Feasibility of Seeking Copyright Permissions. Final Report." http://zeeb.library.cmu.edu/Libraries/FeasibilityStudyFinalReport.pdf [Accessed April 10,2003]

[8] The Posner Family Collection of 622 titles includes landmark titles of the history of western science, beautifully produced books on decorative arts and fine sets of literature.
http://posner.library.cmu.edu/Posner/ [Accessed April 10, 2003]

[9] Covey, Denise Troll. “Copyright Permission: Turning to Dust or Digital?” http://www.library.cmu.edu/Libraries/DustOrDigitalREV.ppt [Accessed April 16, 2003]

[10]“ Registry of Digital Reproductions of Paper-based Books and Serials Functional Requirements.” See http://www.diglib.org/collections/reg/regfunc.htm [Accessed April 10, 2003]