Frequently Asked Questions
About the Million Book Project

- Once you've scanned a title, how soon will you return TIFFs to the publisher?
- Who will determine the pricing of value-added components of the MBP?

What is the current status of the Million Book Project?

Use Internet Explorer to access the Million Book Project/Universal Library sites:

1. Million Book Project [*The Universal Library, China site*], available http://www.ulib.org.cn
2. Million Book Project [*Digital Library of India*], available http://dli.iiit.ac.in
3. Million Book Project [*The Universal Library*, *U.S. site*], available http://www.ulib.org

As of June 2004 -

- Project partner OCLC has provided guest IDs to Indian partners to enable them to capture metadata from the OCLC online catalog and has agreed to help identify source libraries for acquiring selected materials. (Partners in China are members of OCLC so have no need for guest IDs.) OCLC is developing a digital registry that will eventually be used in the Project to help prevent duplicate digitization of books. OCLC might also provide a permanent archive of the Million Book Collection and enhanced access to the books via links in WorldCat.
- The Internet Archive is a project partner, providing a permanent archive for the Million Book Collection, quality control tools, and assistance with acquiring books.
- Indian and Chinese partners have had extended visits at Carnegie Mellon to discuss project logistics and technologies. Carnegie Mellon participants and other project partners in the United States have visited India and China to train the scanning personnel and meet with project administrators to develop project plans. Carnegie Mellon University Libraries prepared and distributed a manual detailing the workflow and standards for metadata capture and digitization.
- Scanning equipment and related software have been purchased and delivered to India and China. Recently purchased equipment includes color scanners and microfilm scanners. Additional equipment will be purchased in the future.
- Over 50,000 books have been scanned to date. Fourteen scanning centers are currently running in India alone. The goal is to have 100 scanning stations digitizing books, each operating two shifts a day, producing 3000 scanned pages per day.
- Chinese partners are digitizing unique collections in Chinese libraries and books in their collection for which the Million Book Project has secured copyright permission.

Chinese partners are coming to the United States in February 2004 to select additional materials. Indian partners are digitizing unique collections and government textbooks that were published in eleven of the eighteen official languages in India. They also scanned a pilot shipment of thousands of books from the United States. Carnegie Mellon University Libraries is digitizing Carnegie Mellon technical reports and other materials selected for the Million Book Project that for various reasons cannot be shipped abroad.

- Kiosks and Internet cafes have been developed and installed in India to provide people with Internet access.
- Carnegie Mellon School of Computer Science has developed a system to support free-to-read access to the digitized books on the web. The system provides tools to add books to the collection and generate PURLs and usage reports. Future developments will enable the Million Book Collection to be indexed by popular Internet search engines like Google and harvested via the OAI protocol. Spring semester 2004 Carnegie Mellon University Libraries is working with students in a Human-Computer Interaction course to prepare specifications to enhance the design and functionality of the system.
- Project partners at the University of California, Merced, have funded the hiring of a full time employee for one year to initiate copyright permission requests. UC Merced will also be a mirror site for the Million Book Collection.
- Investigations are underway for one or more partners to provide print-on-demand service for the Million Book Collection. Meetings were held with ProQuest in November 2003. America OnLine (AOL) has expressed interest in the Million Book Collection. top

What Purpose Does the Million Book Project Serve? What Problems Does the Project Address?

Research reveals that students and faculty look online first when they need information because of the speed and convenience of online access. They prefer remote access to electronic resources rather than having to go to a physical library facility. Though faculty and graduate students often turn to a library web site or licensed electronic resources when they need information, undergraduate students tend to start with popular Internet search engines like Google because these search engines are more convenient and easier to use than library databases. Most students believe the information they find on the open Internet is good enough to use in their coursework. Unfortunately, only about 6% of the surface web content indexed by popular search engines is appropriate for student academic work. Faculty are concerned that the lack of quality resources on the surface web is having a negative impact on the quality of student learning.

Meanwhile, the increasing availability and use of online bibliographic databases, the increasing number of scholarly publications, and the increasing cost of library materials have created a situation wherein libraries are spending more money but acquiring fewer materials. Interlibrary loan is increasing, but the turn-around time is often inadequate for both the highly competitive research conducted by faculty and the shorter deadlines of students. Consequently, user satisfaction is decreasing. Research recently conducted by Carnegie Mellon University Libraries to improve our understanding of the graduate student experience exposed their frustration with the amount of time it takes to get the materials they need for teaching and coursework because the Libraries' electronic resources are not easy to use. To save time and aggravation, they often turn to an Internet search engine first. Among the concerns they expressed was the difficulty of acquiring old journals and out-of-print books. Collection size, the turn-around time required for interlibrary loan, and the cost of document delivery constrain their selection of research topics, the quality of their work, and their grade point average. Lack of free and speedy access to quality resources has a negative impact on the timeliness and success of academic work. Research indicates that most students and faculty perceive a significant gap between their need for speed and convenience and the service their library is providing.

Beyond the boundaries of these problems, tremendous disparity exists across the nation and around the world in the size and accessibility of library collections. Some single institutions, like Harvard and Yale, have more books in their libraries than some entire states have in all of their libraries combined. In our rapidly changing world, lifelong learning and access to books have become essential to employment, health, peace, and prosperity. Greater public access to information is consistent with the goals of education and deliberative democracy. The expectation is that greater access to information will enhance respect for diversity and pluralism, alter the ways in which people work and deliberate together, and better equip people to understand and challenge the world around them. The Million Book Project will digitize a large body of published literature and offer it free-to-read on the surface web - providing students, faculty, and lifelong learners with rapid, convenient access to quality resources. Equitable, world-wide access to the Collection will contribute to the democratization of knowledge and empowerment of a global citizenry. An important byproduct of the Collection will be the existence of a test bed that stimulates and supports much-needed research in information storage and management, search engines, imaging processing, and machine translation. top

For More Information

"How students search: Information seeking and electronic resource use" (EDNER [Formative Evaluation of the Distributed National Electronic Resource] Project, Issues Paper 8, 2002). Available: http://www.cerlim.ac.uk/edner/ip/ip08.rtf

S. Jones and M. Madden, "The Internet Goes to College: How Students Are Living in the Future with Today's Technology" (Pew Internet & American Life Project, September 15, 2002). Available: http://www.pewinternet.org/reports/toc.asp?Report=71

S. Lawrence and L. Giles, "Accessibility and Distribution of Information on the Web." *Nature* 400 (1999): 107-109. Summary available: http://wwwmetrics.com/

LibQual+TM Spring 2002 Survey Results, Association of Research Libraries (Texas A&M

University, 2002). LibQual+TM Spring 2003 Survey Results, Association of Research Libraries (Texas A&M University, 2003). See http://www.libqual.org

D. B. Marcum and G. George, "Who Uses What? Report on a National Survey of Information Users in Colleges and Universities," *D-Lib Magazine* 9, 10 (October 2003). Available: http://www.dlib.org/dlib/october03/george/10george.html

OCLC, "How Academic Librarians Can Influence Students' Web-Based Information Choices" (White Paper on the Information Habits of College Students, June 2002). Available: http://www5.oclc.org/downloads/community/informationhabits.pdf

top

What are the research issues in the Million Book Project?

- ***Information storage and management***--The MBP when completed will produce approximately 250 million pages or 500 billion characters of information. The storage requirements for the image files will be approximately 50 petabytes--an order of magnitude larger than any publicly available information base. Creating and managing such a vast information base poses many technological challenges and provides a fertile test bed for innovative research in many areas (described below). The MBP is a multi-agency, multi-national effort that will require the database to be globally distributed. For location independent access, this globally distributed database should appear to be a virtual central database from any place around the world. Mirroring the database in several countries will ensure security and availability. The network speeds at the various nodes would be different. Research in distributed caching and active networks would be needed to ensure that the look and feel of the database is the same from any location.

- ***Search engines***--The search engines of today work on the principle of keyword matching and perform searches in one language at a time. With a large corpus of multilingual data provided by the MBP, along with multilingual summarization and translation tools, a well-directed research effort would be needed to ensure concept- and content-based retrieval of knowledge from across multilingual data.

- ***Image processing***--The accuracy of Optical Character Recognition (OCR), even in some of the most developed languages, is hindered by the bad quality of the images. This is particularly true for older books and those that use ancient fonts for which the OCR is not tuned. Even the very best OCR accuracy of the order of 98% may not be acceptable in some cases. In order to obtain an improved accuracy close to 100%, advanced image processing research that will perform recognition beyond the character level will be needed. With the availability of large test data from the MBP and the exponentially increasing computing power of the microprocessors, well-directed image processing research would lead to near perfect optical recognizers.

- ***Optical Character Recognition (OCR) in non-Romanic languages***--The MBP will have considerable content in many Indian and Chinese languages. The development of OCR in many of the Indian Languages is far more complicated. For example, some of the problems in the development of OCR for Indian languages are:

- There are 1500 spoken Indian languages and 17 scripts.
- Unlike English, where the number of characters to be recognized is less than 100, Indian scripts have several hundred characters to be recognized.
- Non-uniformity in the spacing of the characters within a word because of the presence of Consonant Conjuncts (vowel + consonant) makes OCR more difficult. Also, the presence of Consonant Conjuncts results in improper line segmentation. Programs will have to do further processing to segment the lines.
- Consonants take modified shapes when attached with the vowels. Vowel modifiers can appear to the right, on the top or at the bottom of the base consonant. Such consonant-vowel combinations are called modified characters. In addition, two, three or four characters can combine to generate a new complex shapes called compound characters. These characters are very difficult for a machine to recognize.
- In scripts like Bangla and Devnagari, all the characters in a word are connected by a unique line called shirorekha (also called head line). In these scripts, character segmentation is especially difficult.
- In south Indian scripts, vowels occur only at the beginning of a word as against the vowels in Oriya, where they occur anywhere within a word. So, the language morphology for some groups of scripts is different from the others.
- There is no universally acceptable standard encoding scheme for Indian scripts. This necessitates a scheme where the output labels from the OCR system can be mapped to the labels used by the typesetter through a mapping table. Because of the non-availability of quality-segmented data, the recognition rate of the Indian language OCR cannot be pushed beyond 90% using character level recognition. To obtain a higher recognition rate, word level information in the form of dictionary has to be used. A word corpus is essential, but such a word corpus for most of the Indian scripts is not available today. The MBP will be able to provide this important missing piece. Once such data is available, it will be possible to use advances in image recognition to develop the OCR in Indian and other non-Romanic languages.

- ***Copyright laws and digital rights management***--In the new digital economy, providing democratic access to information while suitably and reasonably rewarding the innovator is possible. The largest repertoire of free software available on the web in many cases has been the outcome of state supported research. This free availability of software has in fact contributed to more developments and hence an exponential growth of knowledge. Even in literary and scholarly publications, authors have experienced increase sales of their work whenever they are made freely available on the web. This is in tune with observations in the new economy that the companies that make more and more of their software freely available on the web, have their market capitalization enhanced. The MBP, with its proposed plans to make a large knowledge base freely available, will provide useful statistics for testing many economic and sociological models.

- ***Language processing***--The MBP will produce an extensive and rich test bed for use in further textual language processing research. It is hoped that at least 10,000 books among the million will be available in more than one language, providing a key testing area for problems in example based machine translation. In the last stage of the project, books in multiple languages will be reviewed to ensure that this test bed feature is accomplished.

Many believe that knowledge is now doubling at the rate of every two to three years. Machine summarization, intelligent indexing, and information mining are tools that will be needed for individuals to keep up in their discipline work, in their businesses, and in their personal interests. This large digitization project will support research in these areas. This will be of greater significance for the Indian languages where new tools for summarization, grammar and spell checking, thesaurus and translation dictionaries need to be developed ab initio.

The data provided by the MBP with the right research inputs will facilitate the development of language- and location-independent intelligence amplifiers for furthering information creation. top

What content will be included in the Million Book Project? What scanning is currently underway? What about copyright permissions?

Our initial thinking was to take a staged approach to collection development on a discipline-by-discipline basis. However, discussion with project partners and potential partners in November 2001 at a collection planning meeting funded by NSF resulted in the decision to focus on providing free-to-read access to multiple collections. Copyrighted works will be digitized upon receipt of permission from the copyright holder to include the works in the Million Book Project.

Our partners in India and China are currently digitizing local materials. Our Chinese partners are digitizing unusual and unique rare collections in Chinese libraries. Our Indian partners are digitizing government textbooks published in eleven of the eighteen official languages in India.

- U.S. works that aren't copyrighted, like U. S. government documents.
- U.S. out of copyright works

As of August 2002, Carnegie Mellon University Libraries is pulling books published before 1923 and boxing them for shipment to India. We estimate digitizing 45,000 titles in the Carnegie Mellon collection, including biographies and science books.

We have identified approximately 2000 titles in Books for College Libraries that are out of copyright. The next step is to locate these books in Carnegie Mellon's library collection or the collections of our project partners and ship them abroad.

- U.S. copyrighted works

  We've received permission from the National Academy Press to digitize all of their titles published through 1994--approximately 3500 titles. As of August 2002, hundreds of these books have been pulled from Carnegie Mellon?s library collection and boxed for shipment to India in the immediate future. A future step is to locate the remaining titles in partner libraries. Those that are not already available in Chinese libraries will be located in the U.S. and shipped abroad for scanning.

  In June 2002 we began sending letters to U.S. publishers seeking permission to digitize all of their out-of-print or remaindered in-copyright works. Negotiations are underway with many academic presses. We are preparing to send letters to publishers seeking permission to digitize all of their copyrighted titles listed in Books in College Libraries, approximately 58,000 titles. When permission is granted, we will locate the books in partner libraries. Books that are not available in Chinese and Indian libraries will be located in U.S. partner libraries and shipped abroad for scanning.

  Carnegie Mellon has committed to digitizing the university's many technical report series--approximately 5500 titles. Permission to digitize the reports is being sought from the copyright holders. Scanning the Robotics Institute's technical reports began in August 2002. Permission is also being sought to scan technical reports published by IBM and Microsoft.

- Other

  We want to digitize the British Parliamentary Papers and are currently investigating whether these documents are copyrighted. We believe that many of these works are available in libraries in India.

top

Who are the key U.S. participants in the Million Book Project?

- Dr. Mark Kamlet, Provost, Carnegie Mellon--Kamlet led the delegation to China in 2002 and will lead the delegation to India in 2003.
- Dr. Raj Reddy, Simon University Professor, Institute for Software Research International (ISRI), Carnegie Mellon
- Dr. Gloriana St. Clair, Dean of University Libraries, Carnegie Mellon
- Dr. Ching-chih Chen, Professor, Simmons University
- Dr. Michael Shamos, Distinguished Career Professor and Principal System Scientist, School of Computer Science, Carnegie Mellon

- Dr. Jaime Carbonell, Director of the Language Technologies Institute and Professor, School of Computer Science, Carnegie Mellon
- Mr. Eric Burns, Director of Software Development for the Universal Library (parent project of the Million Book Project), School of Computer Science, Carnegie Mellon
- Ms. Gabrielle Michalek, Head, Archives and Digital Library Initiatives, Carnegie Mellon University Libraries
- Ms. Denise Troll Covey, Principal Librarian for Special Projects, Carnegie Mellon
- Ms. Erika Linke, Senior Librarian and Associate Dean of University Libraries (Collection and User Services), Carnegie Mellon
- Mr. Brewster Kahle, Internet Archive
- Phyllis Spies, Andrew Wang, and Lorraine Normore, OCLC
- Bruce Miller, University Librarian, University of California at Merced

  top

Who are the other partners in the Million Book Project?

- China--Representatives from these institutions met with key U.S. participants in the U.S. in 2001 and in China in 2002:
  - Chinese Ministry of Education
  - Chinese Academy of Science
  - Fudan University
  - Nanjing University
  - Peking University
  - Tsinghua University
  - Zhejiang University
- India--Representatives from these institutions met with key U.S. participants in the U.S. in 2002 and in India in 2003:
  - Indian Institute of Science, Bangalore
  - International Institute of Information Technology
  - Indian Institute of Information Technology
  - Anna University, Chennai
  - Mysore University, Mysore
  - University of Pune, Pune
  - Goa University, Goa
  - Tirumala Tirupati Devasthanams, Tirupathi
  - Shanmugha Arts, Science, Technology & Research Academy, Tanjore
  - Arulmigu Kalasalingam College of Engineering, Srivilliputhur
  - Maharashtra Industrial Development Corporation, Mumbai
- United States--The following institutions have agreed to participate in the early stages of the Million Book Project. Additional institutions have agreed to join the project when we have gathered data on the time and cost of sending materials abroad, having them safely returned, and delivering electronic copies of the

scanned materials to the home institution.

- Indiana University
- Pennsylvania State University
- Stanford University
- TriColleges (Swarthmore, Haverford, Bryn Mawr)
- University of California, Berkeley
- University of California, Merced
- University of Pittsburgh
- University of Washington

top

What university/scholarly presses are participating in the program?

The University of Texas Press, Brookings Institution, the American Meteorological Society, American Institute of Biologocal Sciences, and Rand McNally are among the presses that have given permission to digitize their out-of-print in-copyright books. National Academy Press has given us permission to digitize all of their books published prior to 1995. As of June 2004, we are in negotiation with many other presses, including Johns Hopkins, Duke, Penn State, and the Russell Sage Foundation. top

How is the Million Book Project supported?

To date, two grants totalling $3.6 million have been received from the National Science Foundation to purchase equipment.

The Chinese Ministry of Education, Chinese Academy of the Sciences, Indian Institute of Science, and Carnegie Mellon University Libraries and School of Computer Science are providing personnel and facilities, and participating in collaborative research. Carnegie Mellon University Libraries is training the scanning operators.

University of California, Merced, will be a mirror site for the Million Book Collection. They have also contributed funds and personnel for copyright permissions work.

Brewster Kahle (Internet Archive) is providing disk storage.

OCLC is providing project partners with metadata at no charge, will support a registry to track progress and avoid duplicate scanning, and might become a sustaining host of the final Million Book Project collection.

Additional grant proposals are planned to support seeking copyright permissions, further collection development, the management of project logistics, and shipping costs. top

Can users of the Million Book Collection print or download the books?

The delivery systems for the Million Book Collection might restrict Print and Save functionality to one-page at a time. netLibrary's experience indicates that this is sufficient deterrent to prevent users from printing or downloading entire books. This restricted functionality is required for copyrighted books in the Collection. (Note that copyrighted books are included in the Collection only with the permission of the copyright holder.)

To Print or Save a displayed page, move the mouse over the page image. A little toolbar will appear, with icons that enable users to Print, Save, and Email the page. Just click on the appropriate icon to Print, Save, or Email the page.

top

Will the Million Book Project preserve the fixed format of the initial publications?

Yes, the digitized works will preserve the fixed format of the initial publications. top

What metadata is being captured about the digitized works?

MARC records and administrative metadata are being captured following existing standards. Dublin Core is being used for materials that have not previously been catalogued or where MARC is inappropriate, for example, for photographs and three-dimensional cultural artifacts. top

Publishers might not give the MBP blanket permission to digitize and make available all of their out-of-print, in-copyright titles, but might entertain requests for permission to digitize specific titles. Is that possible?

The MBP approach is to request permission for a range of years, for example, everything published prior to 1990. A publisher could specify the cut-off year or, alternatively, specify the list of titles for which they grant non-exclusive permission to digitize in the MBP. top

What value-added services will the MBP develop and what formula will be used to calculate publisher royalties? When might participating publishers begin to see income from the project?

The MBP is not developing a for-profit system. All of the content will be available free-to-read on the Internet. Participating publishers will get copies of the digitized books and metadata, and can themselves provide or enable others to provide value-added services to access the digital books. Permission granted to the MBP is NON-exclusive.

Reading the case study of the National Academy Press's experience putting their books online free-to-read could facilitate understanding and appreciation of the benefits of this approach. See: Barbara Kline Pope, "How to Succeed in Online Markets: National Academy Press: A Case Study," *Journal of Electronic Publishing* 4, 4 (May 1999). Available: http://www.press.umich.edu/jep/04-04/pope.html

top

What kind of accuracy will the MBP achieve in scanning?

Carnegie Mellon has established a workflow (based on pilot, 100-book and 1000-book projects) that includes steps to insure capture of high resolution images and essential metadata, post-processing to correct skewing and crop dark borders surrounding the page images, and OCRing to create searchable ASCII text with 98% accuracy. top

Will the TIFFs meet the Print-On-Demand (POD) standards of Replica and Lightning Source?

The MBP follows the standards and best practices supported in "A Framework of Guidance for Building Good Digital Collections" Developed by the Institute for Museum and Library Services in 2001 and endorsed by the Digital Library Federation in 2002. See:

http://www.diglib.org/standards/imlsframe.htm
http://www.imls.gov/pubs/forumframework.htm

More specifically, our guidelines for data production (excerpted from the MBP NSF proposal and based on pilot projects) are:

- Bitonal images with a pixel depth of 1 bit-per-pixel scanned at a resolution of 600 dots per inch (DPI). Images will be stored as "Intel" TIFF (Tagged Image File Format) files with the header content specified. The compression algorithm used is ITU (Formerly CCITT) Group 4.
- TIFF version 5.0 is acceptable. Subject to testing, version 6.0 (or later) may also be acceptable.

- The initial-capture system includes dynamic thresholding or a similar feature to capture variability of darkness in the imprint and possibly darker (e.g., foxed) backgrounds from decay. Images should be as readable as the original pages.
- Typical expected data will be provided for most TIFF tags (normally, the data supplied by software default settings). A specification for the TIFF header will be produced to include scanner technical information, filename, and other data, but to be in no way a burden on the production service.
- Images will be written in sequential order, with corresponding 8.3 filenames, e.g., 00000001.tif as the first image in volume sequence and 00000341.tif as 341st image in volume sequence.
- Volumes provided to the MBP will be assigned unique identifiers that conform to 8.3 format. The images will be in directories named with the corresponding identifier (e.g., the volume identified as akf3435.001 will have a directory with the same name, and 00000001.tif through 0000000N.tif files within that directory).
- Images and directories (as specified above) will be written to gold CD-ROM according to agreed upon specifications and using ISO9660 format.
- Skew will be within a specified range of degrees allowed.

top

Once you've scanned a title, how soon will you return TIFFs to the publisher?

The timing depends on many factors, including how long it takes us to locate copies of the books for which permission was granted, how many books are involved, what's already in the queue of books waiting to be scanned, etc. top

Who will determine the pricing of value-added components of the MBP?

The publishers or vendors who develop the value-added components will determine the pricing for the services they provide. top

---

April 28, 2005 -- http://www.library.cmu.edu/Libraries/MBP_FAQ.html
Denise Troll, Associate Dean of University Libraries, troll@andrew.cmu.edu