Mass Digitization Projects: Celebration and Challenges

St. Clair, Gloriana

Dean of Carnegie Mellon University Libraries, Pittsburgh, Pennsylvania, 15213, United States

***Abstract* — This paper reviews mass digitization projects including the Million Book Project, Google Print/Book Search and the Open Content Alliance, noting key differences and common concerns (technology, metadata, legal issues). Challenges yet facing these projects may be greater than those already overcome. The author relates future digital library challenges to a human learning construct advanced by Nobel Laureate Herbert A. Simon (1916-2001).**

***Index Terms* — Educational technology, Intelligent systems, Learning systems, Libraries.**

I. Introduction

In my self-appointed roles as chronicler, director, and prophet of the Million Book Project, I want to talk about the history of the project, current concerns and challenges, and future issues, specifically the issue of human attention in the time of information glut. My thesis is that mass digitization projects are creating a revolution in information retrieval and that focusing human attention must be the new research agenda.

A. The Past: History of the Million Book Project and its Heirs

Million Book Project

Each of the early inventors of the computer imagined a digital full text library as an application. Certainly, such a library fulfills Vanevar Bush's vision of the Memex and figures peripherally in several science fiction stories, including Borge's dystopian "The Library of Babel." Equally certainly, several major digitization projects will contribute to an interactive whole.

As a project of the Universal Library group, headed by Dr. Raj Reddy, with the assistance of Dr. Jaime Carbonell, Dr. Michael Shamos and Dr. Gloriana St. Clair, the Million Book Project received its first funding for preliminary planning in 2000 from the National Science Foundation. The Million Book Project aims to make knowledge in many formats, languages and levels available worldwide free to read. To date, we have available on servers in India, China, Egypt and the U.S. over 800,000 volumes.[1] We are just beginning a project to scan a collection of about 100,000 volumes in Qatar.

Our collections are either out of copyright or scanned with permission; and are multilingual, including several Indian languages, Chinese, several European languages, Persian, Farsi and Arabic. In funding this work, the National Science Foundation intended to support initial research and activity that could lead to the revolution currently underway.

Our scientific purpose is to create a test bed for computer science research in such areas as:

- Machine translation
- Massive distributed databases
- Storage formats
- Use of digital libraries
- Distribution and sustainability
- Security
- Search engines
- Image processing
- Optical Character Recognition (OCR)
- Language processing
- Copyright laws

Progress continues in each of these areas at different speeds and with different participants.

Because of the emphasis on research, the National Science Foundation provided funding for equipment and travel. The governments of India and China have provided funding for work in their countries. The Qatar Foundation is funding the collection of Arabic heritage materials. Bibliotheca Alexandrina[2] has funded both digitization and research with some assistance for equipment from the Internet Archive.[3]

Google Print/Google Book Search

Google Print gained national attention in December 2004 through major press releases from its six central partners—Google, Inc., University of Michigan, Stanford University, Harvard University, Oxford University, and the New York Public Library. In 2006, the University of California system joined the Google group. Like the Million Book Project, Google endorses making all information available to the public. Google's practice of

financing this project through advertising continues to be the subject of major reservations among librarians, authors, and publishers.

The Google Print project has subsequently been renamed Google Book Search. August 2006 press releases note that uncopyrightable and out of copyright works are being linked to the University of Michigan's catalog and that about 200,000 volumes have now been scanned. For copyrighted materials, only 'snippets' (three lines of text) will be displayed, with a list of the number of times the search term appears in the book and the pages on which it appears. For more, readers will be directed to the books in traditional paper format.

Google™ has thoroughly captured the market among search engines with thirteen million search hours. The extent to which its ease of use shapes expectations for information retrieval will be discussed later in this article.

Open Content Alliance

Brewster Kahle, founder of the Internet Archive, founded the OCA with the purpose of bringing out of copyright material to the web. This alliance of universities and non-profits will use special equipment designed by Kahle and will work on site in the libraries. The Million Book Project is a member of the OCA.

Other digital library initiatives abound. Korea has an extensive digital library site with over a half a million volumes. Japan has been active in digital library initiatives since the field began. Several European countries have national digital library efforts, and Google Book Search announced an effort there at the 2005 Frankfurt Book Fair.

B. The Present: Current Concerns and Challenges

In several areas, ongoing digital library projects face significant challenges. Technology, metadata, and legal issues are three.

Technology

Technology seems to be keeping pace with the challenges of mass digitization projects. Open Content Alliance is using a new proprietary scanner developed specifically for its work. Only those working with it are aware of its details. Similarly, Google Book Search uses its own equipment and has not been forthcoming about its specifications. Scanner prices have fallen over the years, color scanners have become more prevalent, and standards are changing to reflect the use of grayscale and color.

Bandwidth for transmission of files continues to be a challenge, especially for those of us working internationally. The image files—OTIF and PTIF—are large and difficult to transmit without robust internet nodes. Dr. Raj Reddy is currently experimenting in India with the idea of cleaning up the OCR files so that corrected OCR can be transmitted instead of image files. That alternative would reduce the size of files to be transmitted by about 90 percent compared to image files. However, producing corrected OCR would be about ten times more expensive than producing image files.

It may be possible to balance bandwidth and cost issues by a new application of "the 80/20 rule." In libraries, we have long understood that roughly 20 percent of a collection accounts for 80 percent of circulation activity.[4] Further, we have been able to predict that the more recent materials would be a part of the active 20 percent.

Because of copyright restrictions, mass digitization efforts such as the Million Book Project have focused on older works or copyright-free materials. I think it will be difficult to guess what the most active part of these collections will be. Perhaps we need to develop a strategy that if a book is used a certain number of times, we will then think about correcting its OCR and beginning to serve it by HTML instead of TIF.

We should also study users' tolerance and comprehension of uncorrected OCR. Research questions include whether humans look past flaws and recognize words even though some letters did not OCR correctly, and whether search engines and other machines be able to read uncorrected OCR and infer what the correct words might be.

Metadata

Metadata also provides significant challenges for mass digitization projects. In the Million Book Project, we have wanted to use existing MARC records when possible. That has been somewhat effective for English-language collections, and we have transmitted metadata along with books sent from our collections at Carnegie Mellon. OCLC generously provided access to WorldCat for the project, but searching WorldCat from remote locations in India and China has proven difficult.

Sketchy metadata has been created for books in many different languages. Although training is provided on how to create accurate metadata, some operators have either been inattentive or not had sufficient education to do a good job with it. A particular problem is that some books have been mislabeled as to language; for example, a Farsi text was thought to be in Arabic. Subject descriptors also are difficult for scanning operators to assign. We have asked to have librarians in our partner countries create the metadata for their collections, but that does not appear to have happened.

The Qatar Heritage collection, a new addition to the Million Book Project, provides an interesting case in point. There is only a FileMaker Pro database for the 100,000 volumes in the collection. Yet the costs of doing traditional cataloging for it would be the biggest item in the project budget. Further, part of the high cost of traditional cataloging would result in call numbers and Library of Congress subject headings. While the latter would be useful as part of a metadata set, the western bias inherent in the headings might not make that approach optimal. Depending on where the Qatar Foundation eventually deposits this collection, and how they might choose to arrange it, classification numbers would seem an unneeded luxury, especially since the collection contains rare and non-rare books, manuscripts, newspapers, and other formats.

According to its web site, http://books.google.com, Google Book Search takes quite a minimalist approach to metadata. Metadata is illustrated as ISBN, title, author, and rights information. Those are the only elements that contributors are asked to supply for their contributions. However, I expect that library partners in the project will use their existing catalog records to provide a rich source of metadata for this important resource.

The Open Content Alliance, billing itself as the next step in Open Source and Open Network, will offer item-level metadata for a variety of formats. Since much of its initial content (such as the film archive) is non-book, OCA should experiment with the standards for metadata to support non-book formats.

Legal Issues

In the past five years, I have given more than 20 talks about digital book scanning projects. A persistent theme has been the barrier (or worry, or constraint) of copyright law in the United States and in the world. In the U.S., the copyright period was lengthened throughout the last century by various legislation. At the beginning of the 20th century, it was 28 years, with a possible 28-year renewal if the copyright holder applied for it. Now in much of the world, copyright lasts for the life of the author plus seventy years. In the U.S., all materials published before 1923 are out of copyright, and about 90 percent of the books published between 1923 and 1950s were not renewed and re out of copyright also.

U.S. book copyright renewal records are available on the web, and Michael Lesk (now dean of the library school at Rutgers University) has created a program to search them electronically. A library with an electronic list of books between those dates can have them machine matched against the list of books renewed. Books can also be checked title by title to see if they were renewed. This work is expensive.

Online copyright renewal records are not exhaustive. To be absolutely sure that a title is out of copyright, a search must be performed in the paper records of the Copyright Office in Washington, DC—at a cost of about $100 per hour. Fortunately, the process of checking the online copyright renewal records is considered a good faith effort to observe copyright.

In the course of the Million Book Project, Denise Troll Covey had the innovative idea that we should approach copyright on a publisher by publisher basis rather than title by title.[5] Our collection guide was a well-known U.S. library resource called *Books for College Libraries*, which lists the best books, recommended for every college library. We wrote letters and/or emails to the academic and scholarly association publishers included in that resource, asking permission to scan out of print titles. Over 20 percent of those publishers gave permission for multiple titles. The success of this strategy decreased the cost of seeking copyright permissions dramatically.

As soon as the Google Print project was announced, the American Association of Publishers brought them into court for copyright issues. Google Print had suggested that they would show only snippets (three-line excerpts) from those books that were still in copyright, but the publishers objected—and continue to object. Currently, the University of Michigan, which had planned to make its collection available digitally to its own constituency, will now only show snippets with a list of pages on which the search term appears. To read materials, Michigan students and faculty will have to view paper books rather than digital ones.

At the first ICUDL in China last year, Dr. Michael Shamos talked about using machine summarization to create a version of content that would not be constrained by copyright.[6] While there are several types of works that would not lend themselves to such an approach, an excellent machine summary might satisfy need in the case of many scientific works. At Carnegie Mellon, we continue to develop machine summarization and are eager to experiment with this approach.

In a recent issue of the *New York Times Magazine*, Kevin Kelly published a ‘manifesto’ entitled “Scan this Book.”[7] The cover advertising blurb said “What will happen to books? Reader, take heart! (Publisher, be very, very afraid.) Internet search engines will set them free.” Kelly is identified as the senior maverick at *Wired* magazine and is the author of several books. In the article, he weighs his own personal interests as an author and as a consumer of information. He believes that “the reign of the copy is no match for the bias of technology.” He thinks that the protocols of the screen will win over the conventions of the book, and concludes, “On this screen, now visible to one billion people on earth, the technology of search will transform isolated books into the universal

library of all human knowledge."[8] That sentiment summarizes the aspirations of the Million Book Project, Google Book Search, and the Open Content Alliance.

C. The Future: Human Attention

Even while we are in the midst of creating this 'universal library of all human knowledge,' we must think about how exactly this repository can contribute to learning, the first step in the creation of new knowledge, and with new knowledge, a better future. The late Nobel Laureate Herbert Simon, a Carnegie Mellon professor and colleague, spent fifty years as a university professor and forty years studying human learning. I will use his construct from "Cooperation between Educational Technology and Learning Theory to Advance Higher Education" to consider this problem in the categories of design principles, using information technology, identifying and organizing information for learning, and presenting knowledge.[9]

Design Principles

"Learning takes place in the head of the student, and depends entirely on the activities of the student," Simon advises.[10] Learning is student reaction to the experiences placed before them. A good analysis of the learning task will help educators to provide better experiences. The philosophy behind every large digitization project has been to free content from its locked, geographically-confined precincts and make it equally available to inquisitive students in all parts of the world.

Using Information Technology

Simon reminds us that while technology has magnified the information media—newspapers, magazines, books, television, telephone, fax, Internet, email—none of us has gained a single minute to add to the time we have to absorb information. As we focus on the head of the student, we must acknowledge that that head has only a few hours a day to attend to the experiences of learning. Attention is and will remain the scarce factor in education.

Rory Litwin, an outspoken critic of the Google Print project, speaks passionately about his fears of disintermediation. He worries that machines will be reading these digital books rather than humans.[11] Conversely, Simon thinks that we must ask expert systems to share the load of selecting intelligently the small fraction of information to which each of us should pay attention. Simon believes this is one of the most important tasks that researchers in artificial intelligence have before them.[12] Certainly, machine summarization, which Shamos proposes would provide the added benefit of pulling information out from under copyright restriction, should be a high priority.

Simon also believes that we must sample knowledge, not cover it. Courses and curricula must be designed so that students can partake of pieces of knowledge—because mastery of a whole discipline is no longer a viable educational objective. Now, education should focus on problem solving, in-depth understanding, and independent learning. Students must be prepared to learn just-in-time and to continue learning as fields develop. These basic skills will serve students as new bodies of knowledge emerge.[13] By contrast, Litwin worries about the decontextualization inherent in the snippet approach to content. But if the snippets can draw students to their areas of interest, then snippets will allow student attention to be focused where it can best be used for learning.

Identifying and Organizing Information for Learning

All existing knowledge about how humans absorb information, the rates at which they can absorb it, and the formats that make the most sense must be brought to bear on the body of information being created. Two important tools in this domain are *selected search* and *pattern recognition*. In each of these areas, research continues.

While current students prefer the convenience of the Google single search box, newer search engines are focusing on portal approaches that will allow students to see only the results that pertain to their domain of interest. Simon argues that recognition of patterns is what allows experts to use their knowledge to have intuition. His long experience with computers playing chess underlined the difference between the pattern a novice sees on a chessboard and the one an expert sees.[14] I see discipline-specific gateways and portals as being a significant area of innovation in the short term future.

Simon argues that the skills of the expert are often described in terms of IF-THEN. It is the expert who knows how to move from one step to another. He notes that textbooks often spend a great deal of effort in describing how to do certain steps, but do not focus on the conditions under which such steps would be employed.[15]

For the field of digital libraries, the challenge is to create the reference librarian as the intermediary who can direct students into the correct query techniques and into the discipline-specific gateways and portals that will make sense for them. This replication of human ability to hear and understand the reference question, and then translate the question into a set of resources to be explored will take a long time to develop. Meanwhile, the computer technique inherent in Amazon's comment to users 'other people who bought *x* book also bought *y* book' may be a good starting place. Simon lauds worked out problems as a technique that has helped many students to move from the steps themselves to a sequence of steps.[16]

Presenting Knowledge

The challenge to libraries is to develop web pages that can serve a similar pedagogical function. The lecture and all its analogues in PowerPoint presentations and televised alternatives cannot be seen necessarily to improve learning because they require the student to be passive. Simon argues that mental activity by students in doing something other than daydreaming is needed in order for learning to occur. For some librarians, presenting books on the shelves or in their digital counterparts was the whole task. But, for others, the IF-THEN sequence interaction with the student is at the heart of the librarian's job. Academic libraries had not only the responsibility for collecting and preserving the output of the disciplines but also the responsibility for acculturating students into those disciplines by helping them understand how new knowledge is produced, accepted, and archived.

Simon reminds us that few people can effectively do more than one thing at a time. Attention is serial.[17] Systems design should, therefore, focus on presenting one thing well. Although the thoughtfully designed picture or diagram may be worth many words, Simon recommends not presenting graphical representations but instead challenging a student to create her own visualizations, in her mind's eye.[18] The challenge for the digital library world is the representation of the relationships in the knowledge stored. The parabolic tree showing the major branches of knowledge with twigs for the related disciplines springing from them seems to offer a more organic approach than any variation on the idea of a file cabinet. Its downside is the amount of time it takes to reach the appropriate level of granularity. But even simple techniques, such as the familiar concepts of files and folders, are more helpful than the undifferentiated lists now created by many search engines.

Teaching The Teachers

Recently Carnegie Mellon librarians looked at a digital learning system to help envision how print, visual, spacial and auditory aspects could all contribute to a student's information search experience.[19] Using computer game open source software as well as extensive expertise from the digital gaming arena , Virtual Ancient Egypt creators had developed an attractive, extensible teaching tool that stimulates active learning. I believe that libraries must look to interactive technologies such as these to meet one of our biggest ongoing challenges—engaging students in the critical information seeking behaviors necessary to support learning.

Herbert Simon assumes that reading underlies all learning. Yet, in the Million Book Project, we have often talked about working with non-readers. In particular, in working with the Food and Agriculture Organization, we conceived an agricultural support data resource that would use a village knowledge officer to allow the unlettered to make queries and receive understandable answers. If we focus our attention on learners, we must recognize that there is a time *before* literacy for everyone.

II. CONCLUSIONS

Alternatives, such as gaming pods and elaborate immersive constructs, may help librarians turn passive students into active learners. In a sense, we came to this second ICUDL to celebrate, hoping that our task of scanning a million books would have been accomplished. Yet, as we consider our history, review our current challenges, and contemplate our future, I see that the challenges that lie before us are greater than the ones we have already faced. Humans must be our continuing concern. In every way that we can, we must help humans to deal with the oppression and elation of an overload of information. While the potential learner here in Alexandria has now been spared the physical trip to a building in the U.S. or India or China, that learner must still make the journey through an ocean of content to find that island of information specifically suited to her interest. We partners in this digital library group have many, many responsibilities to make that journey a successful one.

---

1 Use Internet Explorer to access the Million Book Project/Universal Library sites: Million Book Project [The Universal Library, China site], http://www.ulib.org.cn; Million Book Project [Digital Library of India], http://dli.iiit.ac.in; Million Book Project [The Universal Library, U.S. site], http://www.ulib.org; Million Book Project [Internet Archive site], http://www.archive.org/details/millionbooks.

2 Bibliotheca Alexandrina, http://www.bibalex.org/English/index.aspx.

3 Internet Archive, http://www.archive.org/index.php.

4 Richard L. Trueswell, "Some Behavioral Patterns of Library Users: The 80/20 Rule," *Wilson Library Bulletin* 43, 5 (1969) 458-461.

5 Denise Troll Covey, "Copyright and the Universal Digital Library," in *Proceedings of the 1st International Conference on the Universal Digital Library* (Hangzhou, China: Zhejiang University, November 2005), 9-26; "Rights, Registries, and Remedies: An Analysis of Responses to the Copyright Office Notice of Inquiry Regarding Orphan Works," in *Free Culture and the Digital Library: Symposium Proceedings 2005* (Atlanta, GA: Emory University): 106-140; *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, 2005).

6 Michael I. Shamos, "Machines as Readers: A Solution to the Copyright Problem," in *Proceedings of the 1st International Conference on the Universal Digital Library* (Hangzhou, China: Zhejiang University, November 2005), 1179-1187.

7 Kevin Kelly, "Scan this Book," *New York Times Magazine* (May 14, 2006): 43-49, 64, 71.

8 Ibid., 71.

9 Herbert A. Simon, "Cooperation between Educational Technology and Learning Theory to Advance Higher Education," chapter 3 in Paul S. Goodman, editor, *Technology Enhanced Learning: Opportunities for Change* (Mahwah NJ: Lawrence Erlbaum Associates, Publishers, 2002), 61-74.

10 Ibid., 62.

11 Rory Litwin, "On Google's Monetization of Libraries," *Library Juice* 7, 26 (December 17, 2004). Available: http://www.libr.org/Juice/issues/vol7/LJ_7.26.html#3.

12 Simon, 64.

13 Ibid., 65.

14 Ibid., 66-67.

15 Ibid., 67.

16 Ibid.,68.

17 Ibid., 69.

18 Ibid., 71.

19 Lowry Burgess, Lynn Holden and Jeffrey Jacobson, “Muses in the Library II: The Reality of the Virtual,” a lecture for the Digital Library Colloquium (September 14, 2006). The presenters discussed Dr. Holden’s Virtual Ancient Egypt (VAE) digital learning system, demonstrating the Virtual Egyptian Temple in a large-projection theater. Lecture video: mms://wms.andrew.cmu.edu/001/library_9-14-06.wmv. The Digital Library Colloquium lecture series is sponsored by Carnegie Library of Pittsburgh; School of Computer Science and University Libraries, Carnegie Mellon; and School of Information Sciences and University Library System, University of Pittsburgh (Pittsburgh, Pennsylvania).