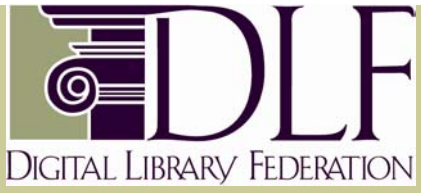

Carnegie Mellon University

Report to the Digital Library Federation
October 2004
DLF Newsletters, Volume 5:1 (2004)
Volume 5, Number 1. Fall 2004

http://www.diglib.org/pubs/news05_01/

TABLE OF CONTENTS

I. Collections, services, and systems

A. Collections

Charette, **http://www.library.cmu.edu/Research/ArchArch/Charette/index.html**

Size: 16,098 pages
Scope: A Pittsburgh architecture journal
System: DIVA, WolfPack
Migration: With other DIVA resources
Intellectual property: Permission
Funding creation: Library resources
Sustainability: Library resources
Use: Recently released; 97,161 pages served so far in 2004; active; metadata harvesting and copying into Million Book Project would increase use
Challenges: Coated paper, graphics, multi-volume serial

Diplodocus and Douglass Archives, http://diva.library.cmu.edu/CMNH/

Size: 2,610 pages
Scope: Carnegie Museum of Natural History archival material
System: DIVA
Migration: With other DIVA resources
Intellectual property: Permission
Funding creation: IMLS funding
Sustainability: Library resources
Use: 6,589 pages served since 2001
Challenges: Sustaining collaboration with another institution; need critical mass of materials to make collection more useful

Heinz, Newell and Simon Collections: http://diva.library.cmu.edu/HELIOS/, http://diva.library.cmu.edu/Newell/, http://diva.library.cmu.edu/Simon/

Size: 1,114,105 archival pages and growing
Scope: Papers of distinguished Carnegie Mellon affiliates
System: Original HELIOS (Heinz-funded software that became commercial) replaced with DIVA (commercial software configured locally); some functionalities lacking in new software
Migration: Already moved once
Intellectual property: Permission
Funding creation: Heinz endowments, family gifts, IMLS grant
Sustainability: Two family endowments
Use: 90,735 pages served since 2001; metadata harvesting would increase use
Challenges: Beginning project, different materials

Million Book Project: http://www.dli.ernet.in/ or http://www.dli.ernet.in/aui/ (India), http://www.archive.org/texts/collection.php?collection=millionbooks (Internet Archive), http://www.ulib.org/html/index.html (Carnegie Mellon), http://www.ulib.org.cn (China)

Size: 124,000 volumes (summer 2004)
Scope: Aall public domain materials; India, China, and other major international partners
System: Mini Universal Library (locally-customized commercial software)
Migration: Mirrored worldwide and migrated using post-doctorate fellow Dr. John Ockerbloom's typed object model techniques for translating data into new formats
Intellectual property: Public domain and permission
Funding creation: NSF, India and China governments
Sustainability: Negotiating with OCLC
Use: Active
Challenges: Feeding international scanning centers with content, copyright clearance, and others.

Posner Project/Posner Memorial Collection, http://posner.library.cmu.edu/Posner/

Size: 313,413 color book pages and archival correspondence files
Scope: Books and documents collected by Henry Posner Sr.
System: DIVA, WolfPack (locally-customized commercial software to perform data conversion tasks and post-processing quality control)
Migration: With other DIVA resources
Intellectual property: Public domain and permission
Funding creation: Posner family gift
Sustainability: Posner Fine Arts Foundation if required
Use: 59,827 pages viewed and 10,380 covers viewed Jan-June 2004; metadata harvesting and copying into Million Book Project would increase use
Challenges: Rare, fragile color materials

Swiss Poster Collection, http://swissposters.library.cmu.edu/Swiss/

Size: 584 images to date
Scope: Rare book room collection
System: FileMaker database access via the web
Migration: With other FileMaker databases
Intellectual property: Fair use low image dpi
Funding creation: Donor gift
Sustainability: Library resources
Use: 63,666 images viewed since 2002
Challenges: Library absorbs future costs to photograph collection additions each year (20-30 posters); posters are large and difficult to photograph; cataloging is challenging and time-consuming.

Technical Reports: Public access through library catalog, http://cameo.library.cmu.edu

Size: 82,103 pages
Scope: Technical reports produced at Carnegie Mellon
System: WolfPack; accessed through Cameo (library catalog)
Migration: With other digital collections
Intellectual property: Public domain and permission
Funding creation: Library resources
Sustainability: Library resources
Use: Recently released
Challenges: Identifying materials in public domain; no OCR

B. Services

Automated Resource Finder (ARF),
http://www.library.cmu.edu/Research/arf/index.html

Automated Resource Finder, a locally developed tool, assists the user in finding relevant materials in a variety of formats and at a variety of levels beginning with a basic level. ARF is customized so that for specific resources such as the library catalog (Cameo) or the Britannica Online, searches are already pasted into the resource search window. This tool is a starting point for a user. Links to "Ask a Librarian" are part of the implementation. Consultants, using data from user testing, recently revised the design of the ARF interface. Carnegie Mellon usage statistics indicate that the majority of online catalog and database use occurs by remote access. Without librarians to guide them, remote users often use inappropriate materials (indexed by popular search engines) or become confused and overwhelmed by the many e-resources provided by the library. ARF helps address these problems. ARF pages were hit almost 3000 times per month between January and September 2004.

Electronic Reserves: Restricted access (current students/faculty) in library catalog,
http://cameo.library.cmu.edu

Size: 89,281 pages
Scope: Changing set of materials for use in current courses
System: Locally-developed programs; accessed through Cameo (library catalog)
Migration: Created for one-semester use only
Intellectual property: Fair use and permission
Funding creation: Library resources
Sustainability: Library resources
Use: 151,506 pages served in 2003 (compared to 6,857 traditional reserves uses)
Challenges: New emphasis on multimedia

Onsite Authentication

Until recently library workstations were available to walk-in users with no authentication required. Beginning in spring 2005, all users of public library machines will be required to authenticate. A 24-hour ID and password will be available at the circulation desk. For specific research needs, a one week pass will be available. [Details for onsite authentication are not yet finalized.]

Remote Access

The library, along with the university, has grappled with the problem of remote user access. The library provided access to remote users using a proxy server. The use of this technology to access library resources was satisfactory part of the time but left some users dissatisfied and unable to access needed resources. We now use the IP extension service, also known as Virtual Private Network (VPN), to connect remote users to library

resources. The library also participated in beta testing of the Shibboleth software, specifically with JSTOR and FirstSearch.

C. Systems

Digital Information Versatile Archive (DIVA)

DIVA provides web access to many of Carnegie Mellon's digital collections. It uses commercial database software and open standards to provide a common framework for digital projects. For example, DIVA's full-text searching is provided by Oracle, interface customization is done using Java Server Pages, and metadata harvesting is provided through an OAI interface. DIVA allows students and researchers to search, browse, view and print digital images of books, journals, technical reports, and archival documents. With specifications developed by Carnegie Mellon librarians and archivists, DIVA provides conventional access to library and archival materials, and adds new functions for searching and retrieving documents, supporting multimedia, and customizing the structure and presentation of collections.

Publishers Database

This locally-created interface supports Copyright Permission Research. See Projects update below.

DIVA Scan

This locally-created interface enables scanning operators to digitize materials quickly into TIFF format, create or add metadata, and make initial decisions about the need for professional processing by archivists.

WolfPack Technology

This suite of locally-customized commercial software performs large data conversion tasks in a distributed manner. The WolfPack framework allows the best off-the-shelf commercial conversion programs to be used in an automated system. For scaleability, it runs the conversions in parallel on a large number of machines. WolfPack performs image cropping, de-skewing, de-speckling and OCR, and creates JPEG and Acrobat files from scanned images. Built-in quality control mechanisms ensure the integrity of the digital collection.

DOI Server

This locally-built system supports metadata harvesting. We use the Digital Object Identifier (DOI) standard to create persistent URLs for objects in our digital library. A server was assigned and specialized software was written to maintain a database of DOIs for our digital projects. This system allows the physical location of an item to vary throughout its lifetime while its identifier remains unchanged.

OAI Layer

To make our digital library useful to the widest audience, we support standards that allow remote tools to harvest metadata about our collections. Using Open Archive Initiative (OAI) metadata harvesting protocol in our digital library software allows content to be indexed remotely. OAI is scheduled for fall 2004 implementation.

II. Projects and programs

A. Projects

New Project Announcements

Jewish Serials Project

This project will digitize and make available three serials (periodicals) emanating from the Jewish community in Pittsburgh, Pennsylvania, 1895 to present. This digital collection will be valuable to historians, genealogists and religious scholars.

Shull Papers

Early in 2003, the University Archives received the papers of Clifford G. Shull, who shared the Nobel Prize for Physics in 1994. The American Institute of Physics has provided an $8,000 grant to process the papers and create a finding aid. The Shull family is providing additional funding to digitize and make available the collection (41 linear feet of materials representing Shull's work as a student, researcher and faculty member at MIT). Shull did his undergraduate work at Carnegie Tech. This collection will be an important resource for physicists and those doing research in the history of science.

Strategic Partnership with FAO for the Million Book Project

On October 14-16, 2004, Dr. Anton Mangstl, Director of Library and Documentation Systems Division of the United Nation' s Food and Agriculture Organization (FAO), and John Reid, Chief of Technical Services, met with the Million Book Project to discuss a strategic partnership. On October 16, Dr. Jacques Drouf, Director General of FAO, and Dr. Charles H. Riemenschneider, Director of FAO' s North American Liaison Office, joined the ongoing meetings. FAO produces documents in five languages-English, French, Spanish, Arabic, and Chinese. FAO and Carnegie Mellon want to develop a strategic partnership around several issues-content for the Million Book Project, enhanced provision of online question answering services, and placement of PCtvts in three different countries. Action items from these meetings include:

- FAO will begin to ship duplicate copies of about 20,000 pages of content for scanning in India. After that is completed, materials that must be returned will be shipped for scanning.
- Servers will be shipped to Rome for the scanning of materials too rare to ship. Eric Burns and Ed Walters will stop in Rome on their way to India to configure servers and to consult with FAO technical staff.
- A work plan for the next three years will be developed with the assumption that FAO and Carnegie Mellon will cover their own costs in achieving its goals.
- A workshop will be convened in Rome on February 9-12, 2005, to pull together individuals who can discuss how to bring FAO information into a service analogous to Google Answers or Ask Jeeves. A group of experts in different kinds of online reference services and in the technology that supports them will be convened.
- The Director General will provide a list of selected countries and villages for an experiment with PCtvt.

Synthesis by Morgan and Claypool

Carnegie Mellon will participate in Morgan & Claypool's Developmental Partner Program (DPP) which is designed to provide feedback from a selected group of 30 libraries worldwide on the next generation of digital products. Synthesis is an innovative product comprised of "Lectures" which are short documents (75-100 pages) born digital, that provide a current summary of a research topic in electrical engineering and computer science written by leading researchers in their field. Founded by Michael Morgan (who previously founded Morgan Kaufman) and Joel Claypool (who created the CRC Press Engineering Handbook Series), the company is not burdened with legacy systems and is free to develop new user oriented products and business models.

Update on Existing Projects

Copyright Permission Research

The Libraries began conducting research on acquiring non-exclusive copyright permission to digitize and provide open access to books in 1999. An initial study was conducted to determine the likelihood of publishers granting permission to digitize and provide free-to-read web access to their copyrighted books. A second study sought permission to digitize and provide open access to the copyrighted books and archival materials in the Posner Memorial Collection. A third study seeks permission to digitize and include out-of-print, in-copyright books in the Million Book Collection. Lessons learned from each study are applied in the next study in an effort to decrease the transaction cost and increase the success rate. For example, copyright permission efforts in the Million Book Project are focusing on books published by scholarly associations and university presses because the feasibility study and the Posner project revealed that they are more likely to respond and grant permission than commercial publishers.

University Libraries' research has begun to establish best practices for copyright permissions:

- Determining whether a book published 1923-1963 is still in copyright. Using a re-keyed version of the copyright renewal records prepared by Distributed Proofreaders, Million Book Project partner Michael Lesk has developed a way to submit queries from a batch of library catalog records and determine which books are in still in copyright.
- Identifying and locating copyright holders. Publishers often merge or go out of business or copyright reverts to the author or the author's estate. The U.S. Office of Copyright retains no records of copyright ownership. Often multiple letters are sent to different addresses seeking permission for the same title. Sometimes the copyright owner cannot be located. Carnegie Mellon legal counsel advised us to assume "permission denied" if the copyright holder cannot be found.
- Reducing the cost of acquiring copyright permission. Copyright permission work in the Posner study, not counting administrator time or legal consultation fees, cost approximately $78 per volume for permission granted. Consequently an entirely different approach is being used in the Million Book Project (MBP). Rather than requesting permission for designated titles, the MBP is requesting permission to digitize and provide open access to all of the publisher's out-of-print, in-copyright books. Using this approach, we have received permission to include roughly 47,000 copyrighted books in the Million Book Project at an estimated cost of $0.42 per book.
- Funding the copyright permission work. We want to include 500,000 copyrighted titles in the Million Book Project. A grant proposal that would have provided financial support failed.
- Increasing the response and success rates. Our research indicates that engaging publishers in a dialog and offering to give them copies of the digitized books increase the response and success rates. Seeking permission for older, out-of-print materials and providing a web site where publishers can see the quality of the scanned books are also helpful strategies.

Challenges: Partnering with others to conduct further research that would ultimately reduce the cost and increase the success of acquiring copyright permission to digitize and provide open access to books.

Digital Image Database

At the beginning of the fall 2004 semester, the University Libraries introduced to the campus community its Digital Image Database. The database consists of images requested by Carnegie Mellon University faculty and students as well as images purchased through commercial vendors. Images are made accessible via the web for study and use in Carnegie Mellon University classroom teaching and presentation. With the introduction of the Digital Image Database, we hope to improve authorized access to our image collection as well as eliminate contention typically associated with individual 35 mm slides.

Each image is scanned as the highest possible resolution TIFF from which various sized JPGs are derived (constraining tool will automatically adjust height and width). The sizes available are Thumbnail (128 x 99 dpi), Working (640 x 480 dpi), and Presentation (1024 x 768 dpi). Once the JPGs are generated the archival TIFF is moved to magnetic storage. Image resolutions were based on recommendations/standards from the Library of Congress (American Memory Collection) website, ARLIS conference presentations, various listservs and professional communiqués.

The intellectual content provided to catalog the images will be generated by library staff or, in the case of purchased collections, by the vendor. The Getty Research Institute's Vocabulary Databases, made available via the web to support limited research and cataloging efforts, may be consulted to standardize creator names, confirm dates, etc. The Image Database will be available to Carnegie Mellon users only through Internet Protocol authentication. Copyright and/or source information will be provided within the image's record.

Million Book Project

The project's vision is to bring one million books free to read to the internet. NSF has provided $3.6 million in funding for equipment and travel with China providing $4.8 million for research and scanning and India $25 million for a broad range of language initiatives including scanning.. The collection will be a collection of collections of out of copyright, public domain, and permission granted materials. Several university presses and scholarly associations have given permission to have their lists scanned, or are negotiating their permissions. Indigenous materials in India and China are being scanned and U.S. books are being shipped to and from international scanning centers. Research objectives for the project include security, copyright, digital rights management, optical character recognition accuracy, OCR of non-Roman languages and scripts, automatic metadata creation, summarization, intelligent indexing, machine translation, storage formats, and search engines. **Library challenges:** Identifying collections to be sent to international scanning centers, ensuring metadata standards, locating books with permissions given, and finding a sustainer.

B. Programs

None at this time.

III. Specific Digital Library Challenges

In 2004, Carnegie Mellon University Libraries completed the work outlined in our *Digital Library Plan, 2001-2007* (available below). Our current challenge is to create a new Digital Library Plan for the university.

IV. Digital library publications, policies, working papers, and other documents

- **Digital Library Plan, 2001-2007, updated for University Libraries Advisory Board Visit (October 2004). Available: http://www.library.cmu.edu/Libraries/DigiPlan.pdf**
  The plan projected for 6 years was essentially achieved within half the time. The update provides details.
- **Digital Library Plan, 2001-2007: Report to Carnegie Mellon President's Council (August 2004). Available: http://www.library.cmu.edu/Libraries/DigiPlanSHOW.pdf**
  Gloriana St. Clair reported digital library successes to university administration.
- **Project Updates: Posner and Million Book Projects (July 2004). Available: http://www.library.cmu.edu/Libraries/Staff_UpdateMtg7_04rev21.ppt**
  Denise Troll Covey, to University Libraries' faculty and staff, July 26 and July 27, 2004.
- **Million Book Project FAQ (updated August 2004). Available: http://www.library.cmu.edu/Libraries/MBP_FAQ.html**
  Project tackles key research issues.
- **Million Book Project: Poster Session Report to NSF re ITR Grant (June 2004). Available: http://www.library.cmu.edu/Libraries/NSFitr.pdf**
  Principal Investigators Gloriana St. Clair and Raj Reddy summarized project status.