DIGITAL LIBRARY FEDERATION

HARVARD UNIVERSITY

REPORT TO THE DIGITAL LIBRARY FEDERATION FALL, 2003

Table of Contents

I. Collections, Services, and Systems

Harvard Libraries web site

The “Harvard Libraries” site is a comprehensive web interface that presents a single, organized view of web-accessible resources available to the Harvard community. The site also serves as an electronic gateway to Harvard’s union catalogs and to comprehensive information about Harvard’s libraries. On June 25, 2003, the Harvard University Library Office for Information Systems launched a revised "portal" page for the Harvard Libraries site (http://lib.harvard.edu). The major goals of this revision were to improve design and usability, increase flexibility, simplify maintenance, and provide a short-term solution until the planned introduction later in 2004 of a completely new library research portal based on MetaLib software from Ex Libris.
Total number of electronic resources listed as of July 1, 2003: 5,325
http://lib.harvard.edu/

Implementation of SFX

SFX is an exciting new research tool from Ex Libris that was implemented in the Harvard Libraries on January 8, 2003. The tool uses resource-linking technology based on the OpenURL standard to allow users of external research databases to link directly from an article citation or abstract to a variety of related resources determined by the local library or institution. With the click of a button, SFX can provide access to the full text of an article (if available) or to local holdings in the HOLLIS catalog. It permits context-sensitive and dynamic linking between web-based resources in which the actual links are customized to reflect licensed digital resources available to users affiliated with Harvard. During the academic year, usage of SFX quickly neared 2,000 hits per day. In addition to the implementation of SFX, two related products were launched during this past year – *Citation Linker* and *EJ2: supplementary list of e-journals*.
http://hul.harvard.edu/ois/systems/sfx/

The SFX **Citation Linker** was released together with SFX on January 8 and is a web facility that allows the user to enter information directly for a specific article or journal citation so that they may generate an

SFX menu of links for that citation. The Citation Linker is available on the e-resources menu of the Harvard Libraries website and use of this popular tool is increasing at a steady rate; approximately 18,000 hits were recorded from January through June 2003.

EJ2: supplementary list of e-journals was made available on the Harvard Libraries website in June

2003. This list of e-journals is generated from the SFX database and includes hard-to-find titles in aggregated collections. There are approximately 10,000 titles included on the EJ2 list, over half of which currently have no other point of access on the portal or in the OPAC.
http://sfx.hul.harvard.edu:82/sfx_local-e-collection/e-journals-A.html

Electronic Resource Management System

In a collaborative effort with staff from the MIT Libraries and Ex Libris, OIS met extensively during the first half of 2003 to define functional requirements and specific data elements in support of an electronic resource management module to be developed by the Ex Libris Information Services Division and designed to interact heavily with Aleph, SFX and MetaLib. Results from this project were also fed back into related work being done under the auspices of the Digital Library Federation. Participation in the Harvard/MIT work was extended to include members from the North American Aleph Users Group (NAAUG) and the International Consortium of Aleph Users (ICAU). Ex Libris will announce their plans for development of the e-resources management module at the September ICAU meeting in Vienna. Local development at Harvard will supplement the Ex Libris project as appropriate.

Harvard Cross Catalog Search

On November 6, 2002, Cross Catalog Search service was made available from the portal as a demonstration system to gauge the public's reaction to federated searching across multiple Harvard catalogs. It was developed using a subset of an early version of the MetaLib software from Ex Libris. This service is a high-level resource discovery tool which allows the user to search simultaneously across five of Harvard's catalogs, including HOLLIS, Baker, VIA, OASIS and HGL. From November through the end of the academic year, the number of searches totaled over 17,000 with approximately 6,000 sessions recorded.

Feedback from both staff and patrons using Cross Catalog Search indicated a strong desire to be able to search research databases and other external resources together with Harvard library catalogs. With this in mind, the Office for Information Systems began to look seriously at the new version of the MetaLib software which offers federated searching and personalization features not now available on the Harvard Libraries portal. A recommendation to pursue the analysis and implementation of MetaLib as the next generation portal software was approved and a full implementation is planned for mid-2004.
http://crosscatalog.harvard.edu

Digital Audio

Digital audio represents the most complex digital resource that LDI has been asked to support. This year, working in conjunction with David Ackerman, the Audio Preservation Engineer at the Loeb Music Library, LDI staff have developed the specifications for the deposit to DRS of audio works. Deposits consist of multiple versions of digital audio files including high resolution archival and production masters, and lower resolution use copies as well as a wealth of metadata to capture the technical properties of the audio files, the processing history, and the structure and relationships between these

various components. OIS is now developing a new desktop application, Dmart, to automate the complicated packaging of these components. Dmart, along with an upgrade to the DRS data model and new loading procedures to accommodate digital audio works will be available in late summer 2003.

The listening versions of audio works will be delivered via RealAudio through the new Streaming Delivery Service (SDS) developed this year. SDS uses Access Management Service (AMS) to control access to audio materials restricted to the Harvard community and supports usage logs to meet the legal requirements imposed by copyright holders of digitized material. The archival and production master versions of audio files can be retrieved from DRS (by authorized owners) using the WebAdmin interface and Asynchronous Delivery Service (ADS).

Digitizing and Depositing Facilities

Fine Arts Library Digital Imaging Lab (FAL DIL)

The Digital Imaging Lab (DIL) is part of the Fine Arts Library Slides and Digital Images Department in the Harvard College Library (HCL). The lab was established to provide digital images of slides for use with the Instructional Computing Group's (ICG) digital carousel tool currently used by faculty, and to provide study images of slides for VIA. The lab serves the faculty and students of the Department of the History of Art and Architecture, as well as faculty from throughout FAS, including the extension school and Harvard's learning in retirement program.
In FY 2003, FAL DIL scanned and deposited into Harvard's DRS 20,697 images, including some image files from vendors. In addition, the DIL does special project scanning for LDI grants and other projects including an up-coming exhibition and publication detailing the history of the Fogg Art Museum.
http://hcl.harvard.edu/finearts/sdi.html

Harvard College Library Digital Imaging Group (HCL DIG)

HCL DIG, a division of the Preservation & Imaging Department in the Harvard College Library, produces high-quality digital reproductions of library and archival materials, and offers image processing, metadata creation, and DRS deposit services on behalf of HCL and other repositories throughout the University.
During FY03, DIG created and deposited 38,825 digital objects into DRS including 17,950 master archival images with their associated derivatives and 1,228 XML-formatted structural metadata files. This year, HCL DIG work included reformatting for seven LDI-funded grant projects and the scanning and processing of 18,389 page-images for HCL’s Reserves Program.
http://preserve.harvard.edu/dig/

Harvard University Art Museums Digital Imaging and Visual Resources (HUAM DIVR)

HUAM DIVR creates high quality digital images of art objects and ephemera in the collections of the Harvard University Art Museums through direct digital capture and conversion of film surrogates. HUAM DIVR handles internal requests from curators, registrars, and staff in exhibitions, publications and public relations, as well as external requests for scholarly, non-profit and commercial use in research and publications.

During FY 2003, HUAM DIVR created over 70,000 images and deposited 94,719 into DRS. The deposits comprise 2 terra-bytes of space and correspond roughly to 21,864 unique images with their associated derivatives.

Peabody Museum of Archaeology and Ethnology

This year, the Peabody Museum of Archaeology and Ethnology developed the capacity to make batch deposits to DRS. As part of their LDI grant project, the museum outsourced the creation of digital images for photographs from 35mm copy positive reel film and deposited in DRS three digital versions of each image (an archival master, a reference image and a thumbnail). During FY2003, a total of 31,254 image files representing approximately 10,000 photographs were deposited. The images and their associated catalog records are available to the public through VIA and additional information is made available for staff use through the museum's collection management database.

HOLLIS (Harvard Online Library Information System)

The HOLLIS Catalog of the Harvard University Libraries is a database containing over 10 million records for books, journals, electronic resources, manuscripts, government documents, maps, microforms, music scores, sound recordings, visual materials, and data files owned by the Harvard University Libraries. The union catalog is updated continually as material is ordered, received, and cataloged. In FY2003:

- the loading and indexing of 515,300 CJK (Chinese, Japanese, Korean) records was completed, allowing searching in HOLLIS of these records in the native scripts. For Chinese materials, records that were formerly in Wade-Giles Romanization were also converted to pinyin. Additional non-Roman scripts will be added in the coming year.

- Z39.50 access to the HOLLIS Catalog was implemented in March enabling authorized Harvard users to conduct HOLLIS searches using a Z39.50 client such as EndNote, in addition to a web browser. This access is currently limited to members of the Harvard community with valid IDs and PINs.

- Work on the ILS staff functions included the addition of a desktop reporting module which allows users to report on data extracted from Aleph including acquisitions and financial data, circulation history, reserve courses and bibliographic data, and selected fields from the MARC bibliographic and holdings data.

- work began on beta-testing the newest release of the Aleph software, version 16, with full implementation expected by January 2004. The major enhancement included in version 16 is a redesign of the Aleph clients for staff functions, including cataloging and acquisitions. http://nrs.harvard.edu/urn-3:hul.eresource:hollisct

E-reserves

E-reserves is a web-based service that provides students with online access to course reserves reading materials. Through the new HOLLIS Catalog, users have integrated access to both E-reserves and to information about print reserves. In the 2003 academic year, the system supported a total of 136 courses offered by the Faculty of Arts and Sciences and the Harvard Divinity School with links to 2,162 items on reserve.

VIA

VIA (Visual Information Access) is Harvard's web-based union catalog of visual resources in art, architecture, and material culture. VIA records include descriptive information about slides, photographs, drawings, paintings, objects and other artifacts held by the university's libraries, museums, and archives. In FY2003, detailed functional specifications and technical analysis for a new system architecture were completed. The new system with improved functionality will be implemented in FY2004.
Total number of catalog records as of July 1, 2003: 189,225
http://nrs.harvard.edu/urn-3:hul.eresource:viaxxxxx

OLIVIA

OLIVIA is a cataloging system for the creation of descriptive metadata about visual resources that will be exported to VIA for public access. In FY2003 more than 40 catalogers worked in OLIVIA, which served as a primary work environment for approximately half of them. In FY2003, a number of small system enhancements were undertaken to increase cataloging efficiency, including a merge and de-duping function for duplicate catalog records and the capability to link OLIVIA records to restricted images stored in the DRS.
Total number of catalog records as of July 1, 2003: 446,716
http://hul.harvard.edu/ois/systems/olivia/

OASIS

OASIS is an online catalog of electronic finding aids with detailed information about Harvard's archival and manuscript collections. OASIS contributors are increasingly providing links within electronic finding aids to digital content such as correspondence, audio recordings, photographs and other images. During FY2003, detailed functional specifications and technical analysis for a new system with improvements for users was completed.
http://nrs.harvard.edu/urn-3:hul.eresource:oasisxxx

Harvard Geospatial Library

Harvard Geospatial Library (HGL) is both a discovery tool and a data mining environment for geospatial data sets. Unique to the digital library world, HGL provides researches with detailed information about geospatial data and the tools to capture and deliver subsets of the data into their research environment. A major new feature developed in FY2003 allows researchers in other systems to pass information into HGL, combining their data with HGL's to create customized maps. Other enhancements to HGL in FY2003 include metadata and cartographic searching improvements, cataloging and data loading efficiencies, and infrastructure enhancements.
Total number of catalog records as of July 1, 2003: 17 publications representing data sets with 2,500 data layers.
http://nrs.harvard.edu/urn-3:hul.eresource:hgeodesy

TEmplated Database Service (TED)

TED is a powerful new system designed and developed in FY2003 to provide an online home to the myriad of small, specialized collections catalogs which do not fit within the scope of existing Harvard

catalog systems. TED can provide web based access to data that might otherwise be hidden in boxes of cards or desktop computers across campus without requiring an extensive OIS implementation effort or the need for high-level programming skills. Any number of unique databases can be created with TED in order satisfy the needs of many individual projects. With assistance from a metadata analyst, collection managers can create an XML schema, select field names, and define the interface for their own database. Data can be imported from an existing database or created on-line using the TED Maintenance system. Each database is built on the same framework which fosters centralized system support such as software upgrades and data migration. The first collection using TED was launched this year: the Biomedical Image Library (http://nrs.harvard.edu/urn-3:hul.eresource:bioimlib), a set of digital micrographs produced in support of basic biological research. A new collection is scheduled to be available online in the winter of 2003: the Milman Parry Collection of Oral Literature, a text and audio archive relating to South Slavic oral tradition.
Total number of catalog records as of July 1, 2003: 4 publications representing 2,918 images.
http://hul.harvard.edu/ois/systems/ted/index.html

Full-text Search Service

Full-text Search Service (FTS) is a discovery tool that provides researchers with the ability to search full-text associated with scanned image. The FTS server is accessed as an option of Page Delivery Service (see Delivery Services below) for searching the full text (such as OCR) of page-turned objects. The FTS server can also be accessed directly through a web interface, such as the ones used by two Library Digital Initiative projects at: http://hul.harvard.edu/huarc/refshelf/AnnualReportsSearch.htm and http://arboretum.harvard.edu/library/tibet/papers.html.
Total citations as of July 1, 2003: 250

Delivery Services

LDI offers a number of format specific delivery services developed to enable the delivery of digital objects stored in DRS to web browsers. These services include:

- Image Delivery Service (IDS) for delivery of still image files
- Page Delivery Service (PDS) for delivery of scanned page images within the context of logical navigation – in other words, PDS mimics the page-turning functionality of a book. Total number of publications as of July 1, 2003: 723. http://hul.harvard.edu/ois/systems/pds/index.html
- Streaming Delivery Service (SDS) delivers streamed media to web browsers. Currently the service delivers audio files, but it is capable of delivering video as well.

Asynchronous Deliver Service (ADS) allows curators and researchers to request large objects or sets of objects from DRS for downloading upon e-mail notification. Currently, this new service is primarily used to deliver large TIFFs from the Biomedical Image Library for printing or creating image stacks.

In FY2003, significant efforts went into analyzing additional functionality for improvements to IDS that will be implemented next year; the user interface to PDS was redesigned; and SDS and ADS were developed as new services.

Digital Repository Service

Digital Repository Service (DRS) is an integrated set of services to manage, maintain, preserve, and deliver Harvard's digital materials. During FY2003, the system was upgraded to support audio files, and the processes and procedures for auditing all of the copies of each digital object stored in the DRS were established. As a repository, DRS is not visible to researchers and most curators. It is through the DRS delivery services that it is known (see Delivery Services above).
Total number of digital objects stored as of July 1, 2003: 485,963.
http://hul.harvard.edu/ois/systems/drs/

Name Resolution Service

Name Resolution Service (NRS) assigns persistent identifiers to digital objects. Persistent identifiers provide curators and researchers with confidence that the URL they cite will always work.
Total number of persistent identifiers registered as of July 1, 2003: 223,621
http://hul.harvard.edu/ois/systems/nrs/

Access Management Service

Access Management Service (AMS) provides secured access to Harvard's licensed or copyrighted materials. Using the University Personal Identification Number (PIN) and Directory Services, AMS protects the electronic assets of the University from unlawful access and also restricts access to the Harvard Community as required by curators. In FY2003, AMS was upgraded to work with the newest version of the University's Directory Service.

The Harvard–Radcliffe Online Historical Reference Shelf (HROHRS)

A joint venture of the Library Digital Initiative, the Harvard University Archives, and the Radcliffe Archives to provide electronic access to frequently consulted sources on the history of Harvard and Radcliffe including annual reports, narrative histories and founding documents.
http://nrs.harvard.edu/urn-3:hul.eresource:hronhirf

Nineteenth-century American Trade Cards

Descriptions and digital images in VIA of 1,000 advertising trade cards selected from the Historical Collections at the Baker Library. As an indicator of consumer habits, social values, and marketing techniques, trade cards are of interest to scholars of American social, cultural and business history.
http://www.library.hbs.edu/hc/exhibits/tcard

The Hedda Morrison Photographs of China

Descriptions and digital images in VIA of 4,800 photographs made by German photographer Hedda Morrison in the areas of East Asian studies and culture. Taken between 1933 and 1946, this collection from the Harvard-Yenching Library documents the architecture, streetscapes, clothing, religious practices and crafts that in many cases have all but disappeared from modern China
http://hcl.harvard.edu/harvard-yenching/morrison/

Biomedical Image Library (BIL)

A collaboration between the Countway Library, the Biomedical Imaging Laboratory at the Harvard School of Public Health and the Library Digital Initiative to develop a central catalog and collection of biomedical images produced in support of basic biomedical research.
http://nrs.harvard.edu/urn-3:hul.eresource:bioimlib

Maya Archaeological Photographs from the Carnegie Institute of Washington Collection

To view the approximately 10,000 photographs from Peabody Museum of Archaeology and Ethnology this collection that are now available in VIA: enter the search term “Maya” in the first box, select “Anywhere” in the drop-down window, limit to holdings of Peabody Museum of Archaeology, and restrict the search to records that have digital images by checking the box.
http://nrs.harvard.edu/urn-3:hul.eresource:viaxxxxx

South Central China and Tibet: Hotspot of Diversity

A digital collection created by the Arnold Arboretum Library of Harvard University through collaboration with a number of University repositories. Digitized materials include botanical and bird specimens, correspondence, maps and images related to modern and historic botanical expeditions to South China and Tibet, including those of explorer Joseph Rock in the 1920s.
http://arboretum.harvard.edu/library/tibet/expeditions.html

Loeb Design Library Electronic Finding Aid Project

Frances Loeb Library of the Harvard Design School configured the Library's database to enable the export of EAD formatted finding to OASIS. Select the link for *Loeb Design Library* at http://oasis.harvard.edu/ to view the 9 EAD finding aids available online in OASIS as a result of this project

II. Projects and Programs

Library Digital Initiative (LDI)

Harvard University launched the Library Digital Initiative (LDI) in July 1998 to develop the University's capacity to manage digital information by creating a robust technical infrastructure for the acquisition, organization, delivery, and archiving of digital library materials; by providing a team of specialists to advise librarians and others in the University community on key issues in the digital environment; by providing librarians and staff with experience in digital library projects; and by enriching the Harvard University Library system with a significant set of digital resources. Now entering its sixth year, LDI is making it easier for Harvard's libraries to maintain their collections and services in the digital era, without each library having to individually acquire the expertise and systems needed to support digital resources. The development of most of the systems and services documented in this report were funded by LDI.
http://hul.harvard.edu/ldi

Internal Challenge Grant Program

Managers and staff throughout Harvard's libraries, archives, museums and special collections have participated in LDI through the Internal Challenge Grant Program. They have assisted LDI by prioritizing, testing and demonstrating new systems and services while contributing valuable online content for research and education. Projects have had a range of goals including basic digital conversion of a single collection; the creation of a virtual collection by digitizing related material from multiple repositories; and the development of new delivery systems for natively digital material. Many projects have focused on providing access to previously inaccessible collections and making them available online for use by students and scholars at Harvard and around the world. Over the last five years 30 projects were funded through the grant program and nearly 200 Harvard staff members gained experience working with digital projects. In FY 2003, four projects were completed and twelve were newly funded. Completed projects are reported in Section I., Collections, Services, and Systems of this report.
http://hul.harvard.edu/ldi/html/grants.html
http://hul.harvard.edu/ldi/html/funded_projects.html

LDI MAP

LDI Management Assistance and Planning (LDI MAP), is a cost recovery service that provides customized, hands-on assistance to project managers of LDI grant-funded projects (see Internal Challenge grant Program above). In FY2003, the program provided services to four grant projects.
http://hul.harvard.edu/ldi/html/grants.html#ldi-map

Advisory and Technical Services

LDI provides expertise and assistance to the University's libraries, archives, museums, and research projects that are involved in collecting or creating digital resources. These advisory and technical services fall into three main areas: **digital acquisitions ~** for issues of licensing, contracting, and vendor relations; **metadata ~** for standards and best practices related to the creation of data for describing and providing access to digital materials and for managing digital collections; and **reformatting ~** for information about technologies, standards, vendors, and workflow design.
http://hul.harvard.edu/ldi/html/advice.html

Harvard Open Collections Program

In November 2002, the Harvard Open Collections Program was launched as an 18 month pilot project with funding from the Flora and William Hewlett Foundation. The goal of the Open Collections Program (OCP) is to increase the availability and use of Harvard's rich and historically significant collections for teaching, learning, and research by digitizing selected resources in broad topic areas and by providing the larger academic community with access to these resources through Harvard Library catalogs and the World Wide Web. The pilot will focus on women and work in the United States in the late nineteenth and early twentieth centuries. The original source material for the project will include monographs, manuscripts, and visual resources drawn from many of Harvard's libraries, museums and other collections. The resulting digital resources will be added to the appropriate Harvard University Library catalogs (monographs in HOLLIS, manuscripts in OASIS, visual material in VIA) and a subject-specific web site will be created to provide a contextual environment for discovery and exploration of these resources.

Digital library publications and other documents

Digital library information, documentation and publications are generally linked from the following publicly accessible web sites at Harvard University Library:

> The Library Digital Initiative (LDI) site focuses on information about the initiative including technical development, advisory services and the grant program funded through LDI. http://hul.harvard.edu/ldi/
>
> The Office for Information Systems site contains information about available Harvard University Library systems and services including resources for the staff at Harvard's libraries, museums and archive and information technology offices using the systems and services. http://hul.harvard.edu/ois/
>
> The Library Preservation at Harvard site is a collaborative effort of the Weissman Preservation Center in the Harvard University Library and the Preservation & Imaging Department in the Harvard College Library Harvard with information about preservation and imaging services and resources. http://preserve.harvard.edu/

III. Specific Digital Library Challenges

Integration with educational technology at Harvard

The last few years have seen an enormous growth in the use of the web for providing information and tools to students for use in courses. At Harvard, a considerable infrastructure for supporting the electronic delivery of course information has been developed, most notably in the iCommons project (http://icommons.harvard.edu). iCommons gathers instructional software developed at schools throughout Harvard to create an integrated course platform. Libraries of course have had an important traditional role in providing resources for use in instruction, particularly in such areas as undergraduate library collections, course "reserves" services, and collections of teaching slides. With the growth of digital library collections and of course management systems, the mode of providing library support for instructional materials will change. The primary manifestation of that change will be the increasing integration of Harvard's technical infrastructures for digital library content (LDI) and course content (iCommons).

This year, a small step in integrating library and course systems was undertaken with a new facility in VIA, the visual collections catalog. Instructors can make use of an export tool in VIA to download images with descriptive metadata in a way that can be readily imported into a digital carousel tool developed by FAS (and jointly offered by FAS and iCommons) for creating slide shows. At a more general level, discussions are now underway, both in the larger educational environment, and specifically at Harvard, about how digital library systems and various kinds of education and research tools should inter-relate.

There are a growing number of common areas for potential collaboration between iCommons and LDI including:

- Reserves materials are increasingly available in digital formats that could be made directly accessible from course web sites.

- The library's increasing array of digital resources could be presented to students in the context of the courses for which they are most relevant.

- The instructional and reference services already provided by libraries can be made accessible from course web pages.

- The digital collections infrastructure of the libraries can be used to preserve digital materials created specifically for use in courses, and to ensure access to the materials over time.

Defining and implementing modes of interoperation will be a key activity for LDI and iCommons over the next few years.

Digital Preservation

As increasing amounts of digital content are produced at Harvard and stored in the LDI Digital Repository Service (DRS), the importance of ongoing preservation activities cannot be overstated. Digital materials are inherently fragile and completely dependent for long-range viability on technologies that change continuously. To protect Harvard's digital resources into the future, staff are developing expertise in the underlying digital formats of objects accepted into DRS, and requiring extensive technical metadata about these objects. By closely monitoring the technological environment underlying DRS, the various delivery services, and the digital formats stored in DRS, LDI staff will be able to initiate digital preservation activities to ensure the future of the resources.

For LDI and for the University as a whole, digital preservation is a priority that is reflected in several areas of progress in FY 2003:

- A national archiving environment, built upon the distributed activities of independent institutions, requires a formal way of communicating local preservation activities to prevent needless duplication of effort. Various LDI staff are actively participating with the Digital Library Federation in plans for a national digital registry of born-digital materials and digitally reformatted books and journals. http://www.diglib.org/collections/reg/reg.htm

- As a follow-up to last year's Mellon Foundation-funded ejournal archiving planning project, LDI staff have collaborated with the National Library of Medicine (NLM) to produce an open source archiving and interchange XML DTD. The DTD is designed to increase the ease of interchange between publishers and archives for article-level ejournal content. Without this DTD, the structure of ejournal content can vary widely, requiring costly human intervention and multiple parallel workflows within archival repositories. The DTD was designed after extensive document analysis in many subject domains to insure that it does not reflect the bias of any particular academic discipline. Furthermore, it is based on public standards, features a modular structure to allow customization, and should be an easy target of transformation from existing XML or SGML-encoded content. In addition to being used by NLM for the PubMed Central archive, this DTD is well-positioned to become a standard format for the transfer and archival storage of the scholarly literature. http://dtd.nlm.nih.gov/

- DI staff are collaborating with JSTOR to produce an extensible tool, called JHOVE, for automating format-specific validation of digital objects. The tool, which will be made publicly available under an open source license, is particularly useful for the validation of digital objects submitted for deposit

into a digital repository such as DRS. In addition, JHOVE has facilities to extract important technical characteristics of digital objects from the objects themselves. To ensure future use of digital objects, it is important to verify that a format and its characteristics have been correctly identified. The initial deployment of JHOVE will provide validation for the PDF and TIFF formats, including recognition of many specific format profiles, or named constrained subsets. http://hul.harvard.edu/jhove/

- Adobe's Portable Document Format (PDF) has rapidly become a de facto standard for the dissemination and presentation of electronic documents on the web. Unfortunately, the feature-rich nature of PDF permits tremendous variability in the internal structure of documents, and allows documents to be dynamically composed at the time of their display from disparate external resources, which leads to significant difficulties in insuring their long-term viability. In order to address these concerns, a multi-national effort has been established within the ISO standards framework to produce a constrained version of PDF suitable for archival preservation, to be known as PDF/A. Stephen Abrams, Digital Library Program Manager, is the project leader/editor of the ISO Joint Working Group developing PDF/A. http://www.aiim.org/standards.asp?ID=25013
- Most theoretical discussions of archival preservation revolve around three main strategies: migration, emulation, and the newly-proposed Universal Virtual Computer (UVC) approach. However, there is little empirical data by which to evaluate the comparative advantages and disadvantages of these methods. LDI staff will have an opportunity to gain experience in format migration as a result of the implementation of the new LDI Large Image Delivery Service (LIDS). To date, most image objects in the DRS were represented by TIFF master images and one or more JPEG deliverables of various pre-formed sizes and qualities. LIDS will continue to require the TIFF masters, but only a single production master in the JPEG 2000 (ISO 15444-1:2000) format. All deliverables, of arbitrary size and image quality can be dynamically derived from this single JPEG 2000 object. LIDS provides the opportunity to discard the old JPEG deliverables and to convert TIFF images into JPEG 2000 images. This will involve investigating technical problems and formulating new policies in areas such as the appropriate degree of curatorial input, the extent to which the process can be automated, determination of the proper metadata to document the process, and establishing necessary quality assurance procedures.
- As mentioned previously, preservation activities depend upon extensive knowledge of the formats in which digital objects are manifested. Since this same information is useful to all institutions interested in preserving their digital assets, there is great economy of scale to having a central repository for this format information. LDI staff have been instrumental in organizing an ad-hoc international group of interested stakeholders, including representatives of national libraries and archives, and academic research libraries, who have met to discuss the technical and policy issues surrounding the creation and sustainable operation of a global digital format registry. Stephen Abrams, Digital Library Program Manager, co-authored a paper on this topic presented at the 2003 IFLA conference, available at http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf.

Extended Character Set Support

The resources supported by LDI infrastructure encompass many languages and script systems. The encoding of the Unicode character set provides a uniform mechanism for electronically storing living and historic languages and displaying them to online browsers. Most LDI systems provide support for Unicode, including the underlying technologies for HOLLIS, DRS, Full-text Search Service (FTS); TED; OASIS, Page Delivery Service (PDS), and VIA. For ease and efficiency of searching and retrieval in these systems, full text is normalized to a canonical form (devoid of punctuation, case distinction, and diacritic marks) prior to indexing and search operations, while the original form of the text is maintained

for display. All systems share the same set of normalization rules, based on rules from the library community's long-standing Name Authority Cooperative Program (NACO), so that patrons can expect similar search behavior without regard to the particular system in which a search is performed

While this mechanism works well for languages based on the Latin alphabet, challenges remain with properly supporting non-Latin script systems, such as Cyrillic, Hebrew, Arabic, and Asian languages. Content in these languages is easily accommodated in LDI systems using Unicode, but there is no effective way to generate non-Latin Unicode search terms using a standard Latin character-based keyboard. Potential options include transliteration (used, for example, by the HOLLIS OPAC for Chinese) and Input Method Editors (IMEs), desktop applications that provide a visual interface to language-specific "virtual" keyboards. Both of these solutions require significant analysis and present implementation difficulties with regard to providing uniform user interface behavior across LDI systems.

Integration with the larger digital library environment

The digital collections available to Harvard's users are accessed through a large number of highly diverse systems distributed across the entire Internet. Over the past decade this heterogeneous environment has been created by many individual players concentrating on how to best provide access to their own individual set of resources. The current environment is one of enormous richness, and enormous complexity. With the increasing complexity of the digital environment, there is a growing need to begin integrating the many systems that make up our "digital collection" in ways that insulate users from this underlying confusion.

The implementation this year of the SFX system is one step in creating a more integrated environment for Harvard library users. SFX provides a way to navigate easily from a citation in one system to the cited book or article in another. This navigation is tailored to the Harvard information environment, so that users are led only to systems to which they have free access.

During the next year we will implement another tool to help integrate resources, a new portal system that will provide users with the ability to simultaneously search for resources across a variety of systems with a single transaction.

SFX and the new portal are ways for us to bring together diverse resources in systems beyond our borders. There are likewise ways in which Harvard's internal digital resources can be integrated into the larger environment. Making information about our locally produced resources available through outside databases, making our catalogs accessible to other portal systems, and insuring that digital resources we create follow standards so that they are useable in systems beyond Harvard are all steps to increasing interoperation in the wider library environment.

The ability to integrate diverse resources in ways that simplify use is an increasingly important development in the larger information technology environment. Many products are now trying to integrate tools and data into people's working environments in a way that they need not be concerned about *where* those tools or data originate. Digital libraries, with their enormous range of diverse and distributed resources, will benefit greatly from developments of this sort. Integration and the erasing of barriers to the use of distributed resources will be a key theme in digital library developments over the next decade.