DLF

DIGITAL LIBRARY FEDERATION

FALL FORUM 2004

Hyatt Regency Baltimore, Baltimore Inner Harbor, Maryland
Monday, October 25th - Wednesday, October 27th, 2004

****

DLF DEVELOPERS' FORUM

New Discovery Tools and Metasearch Derivatives -- What's Coming?

Monday, October 25
8:30am-12:45pm
The Chesapeake AB room.
Meeting includes continental breakfast, break food, and lunch.

This Developers' Forum will focus on interesting and innovative projects related to discovery (aka "search") going on at a number of institutions, ranging from data mining to faceted browsing to semantic web applications, in a variety of relevant domains. Examples of these efforts include D2K, a data mining suite emerging from NCSA; Grokker, a visual and dynamic categorization tool under investigation at Stanford; Sentient Discover, a tool facilitating metasearch integration with learning management systems, developed initially in partnership with the LSE; SIMILE, a RDF and Semantic Web-based project at MIT, and more. **Co-chaired by Peter Brantley (CDL) and MacKenzie Smith (MIT).**

****

PROGRAM

Monday October 25

12.00pm-1.00pm Registration *(Foyer F Side)*

1.00pm-2.10pm Keynote Address: ***Cyberinfrastructure for the Humanities and Social Sciences.*** **John Unsworth, University of Illinois, Urbana-Champaign** *(Constellation E/F)*

2.10pm-2.30pm Break

2.30pm-4.00pm Session 1: THE GREENSTONE DIGITAL LIBRARY *(Constellation E)*

Digital Libraries in New Zealand: An Update. David Seaman, Digital Library Federation

A brief update on a range of digital library initiatives currently underway in New Zealand (excluding Greenstone), including institutional repository planning at the University of Auckland, electronic texts work at the University of Victoria at Wellington, and a range of digital preservation metadata and tools building work at the National Library, along with their recently-announced NZ Online national plans.

The Greenstone Digital Library and G3: A Demonstration of Current Capabilities and Future Developments. Ian Witten, University of Waikato, New Zealand

The Greenstone digital library software is a comprehensive, open-source system for constructing, presenting, and maintaining information collections. It is widely used internationally, and collections exist in many of the world's languages (the interface itself has been translated into thirty languages). Greenstone runs under Unix, Windows and Mac (OS/X) and is issued under the GNU general public license. Greenstone's "librarian" interface lets users gather together sets of documents, import or assign metadata, build them into a collection, and serve it from their web site (or write it to CD-ROM or DVD). Such collections automatically include effective full-text searching and metadata-based browsing facilities that are attractive and easy to use. The facilities that a collection provides, including the user interface for searching and browsing, can be customized at many different levels based on whatever document formats and metadata are available. Documents can include text, pictures, audio, and video. The interface explicitly supports four levels of user: Library Assistants, who can add documents and metadata to collections, and create new ones whose structure mirrors that of existing collections; Librarians, who can, in addition, design new collections, but cannot use specialist IT features (e.g. regular expressions); Library Systems Specialists, who can use all design features, but cannot perform troubleshooting tasks (e.g. interpreting debugging output from Perl scripts); and Experts, who can perform all functions.

This talk will introduce Greenstone and demonstrate the use of the librarian interface to build a multimedia collection about the Beatles, including discographies (.html), guitar tablature (.txt), album covers (jpeg), audio recordings (.mp3), midi versions (.midi), supplementary material (.doc and .pdf), and library records (in MARC format). It will also describe current work on Greenstone3, a complete reimplementation (downloadable in prototype form) structured as a network of modules that communicate in terms of XML messages. All modules characterize the functionality they implement in response to a "describe yourself" message, and can transform messages using XSLT to support different levels of configurability. Traditional library values such as backwards compatibility and multiplatform operation are combined with the ability to add new collections and services adaptively.

Note: Greenstone, out of the box, supports many standard formats. It can ingest full-text documents in Word, PDF, PostScript, HTML, LaTex, E-mail, PowerPoint, Excel, and many others; metadata in MARC, XML-based Dublin Core, CDS/ISIS, ProCite, BibTex, Refer; image and multimedia documents in JPEG, GIF. TIFF, MPEG, MP3, MIDI, etc. It can also ingest data from OAI sites. Current projects include an OAI server for Greenstone collections and the use of METS as an alternative internal representation for documents (both almost complete).

2.30pm-4.00pm Session 2: REPOSITORY DEVELOPMENT AND MANAGEMENT I *(Constellation F)*

Digital Library Repository Development at the UVa Library. Leslie Johnston, University of Virginia

This fall, the UVa Library made a first version Digital Library Repository based on the Fedora open source architecture available to its community. At the time of launch, the Repository contained TEI and image collections produced locally since mid-2003, selected licensed image collections, and with all UVa Library Special Collections EAD finding aids added more recently.

This presentation will generally discuss the development of Fedora to date, but primarily focuses on the UVa Library implementation of Fedora. Fedora is the underlying architecture for the Repository, but not the complete management, indexing, discovery, and delivery application. The UVa Library's implementation required a large-scale effort to define a local architectural and service overlays specific to UVa's collections and functional requirements. The development process encompassed:

- the creation, documentation, and adoption of new holistic standards for production;
- a detailed analysis of the formats and configuration of media files and metadata for legacy collections and comparison of the results to the new production standards;
- functional requirements for discovery and delivery services;
- unified interface design across collections;

- the implementation of new software tools and scripts for all aspects of production and delivery.

The Repository also includes the first release of the "Collector Tool" that allows users to create personal portfolios of objects. The current tool includes the ability to collect images into personal portfolios and generate slide shows or electronic reserve web sites that include pointers to the images and metadata in the Repository. Later releases will be generalized to support the collection of other object types, a sort of combination shopping cart and basic authoring tool for the entire Repository.

The next steps for the Repository infrastructure development process are an evaluation of the production workflows, usability testing of the interface with groups of faculty, undergraduates, and graduate students, and migration to Fedora 2.0 with its improved support for the updating of objects. The next steps for Repository content development are a review of additional legacy image and text collections for migration to current production standards and ingestion, and planning for additional content types, including data sets and digital video.

Architecture, Design and Creation of Software to Support an Institutional Research Repository: The ARROW Experience. Geoff Payne and Andrew Treloar, Monash University

This presentation will describe the development of the software to support ARROW - Australian Research Repositories Online to the World (an Australian Commonwealth Government funded project under the Research Information Infrastructure Framework for Australian Higher Education). The ARROW initiative aims to identify, develop and test software to support best-practice institutional digital repositories at the ARROW Consortium member sites to manage e-prints, digital theses and electronic publishing, and to develop and test a national resource discovery service using metadata harvested from the institutional repositories by the National Library of Australia. Refer http://arrow.edu.au/ for more information.

The presentation therefore begins by describing the global and Australian context in which the project came about. It then moves on to the design brief for the software, based on the list of functional requirements that were encapsulated in the successful funding bid. Next it describes the process of defining the overall software architecture and list of functions to be provided. In order to turn this into architecture into reality, a series of software decisions needed to be made. The presentation discusses the decision about the foundation repository, the application development framework and the search and exposure layer for the repository contents. It also explains why the project decided to sub-contract out the software development, rather than doing it in-house (as originally planned). Finally it will describe the current state of the project (including the first phase of the delivered software), and when the project expects to be ready for widespread deployment.

Plans and Early Results from Ensuring Access to Mathematics over Time.

Marcy E. Rosenkrantz and William R. Kehoe, Cornell University;
Marcus Enders, University of Gottingen

The growth of digital serial literature has presented librarians with many complex problems in fulfilling their familiar archival and preservation functions. To address the questions surrounding how best to develop and maintain reliable digital archives, our project focuses on the complex discipline of mathematics. With the cooperation of selected publishers, we plan to develop an archive of serial mathematics literature that will serve as a model for similar cooperative efforts in other disciplines within the library and publishing communities. To do so, we will develop and implement a system that adheres to the principles put forth in the Open Archival Information System (OAIS) Reference Model. Our institutions will establish requirements and procedures that will enable separately administered repositories to function as a single digital archiving system.

The National Science Foundation and the Deutsche Forschungsgemeinschaft have funded this project, which we affectionately call EATMOT. This panel will include project participants from the Cornell University Library and the State and University Library at Göttingen, Germany. We will discuss our plans and present an early progress report.

4.00pm-4.30pm Break

4.30pm-6.00pm Session 3: PANEL: DLF AND THE CYBERINFRASTRUCTURE FOR HUMANITIES AND SOCIAL SCIENCES *(Constellation E)*

Abby Smith, Council on Library and Information Resources

New digital information environments in the humanities and social sciences are changing the intellectual landscape. Computational power and new digital tools are being applied to such age-old problems as deciphering ancient languages through text mining and pattern recognition. Digital documentation of archeological sites enables new interrogations of data while protecting the authenticity of multiple layers of evidence. Applications of Geographic Information Systems (GIS) are giving rise to place-based research projects in collections as diverse as historical maps of land ownership and biological specimen collections. Roman fora and Romanesque cathedrals are being recreated in virtual spaces for study and testing of hypotheses about past cultural practices and belief systems.

What are the particular needs of the humanities and social science domains for digital information environments, and what infrastructure is needed to support these environments? This is the subject of a year-long study undertaken by the American Council of Learned Societies (ACLS), a group representing 68 scholarly organizations. Recognizing the need for the humanities and social sciences to help shape this emerging infrastructure, ACLS appointed a commission in early 2004, comprising nine nationally prominent individuals engaged in digital scholarship and chaired by Dean John Unsworth of the University of Illinois. They are charged to define the current state of digital scholarship in the humanities and social sciences, to demonstrate the potential of systematic capabilities inherent in cyberinfrastructure to catalyze innovation in these domains, and to change the terms of reference within which the digital future is understood and imagined.

The commission is gathering information through public sessions, surveys, and targeted interviews with leaders in humanities and social science disciplines, including scholars, academic administrators, librarians and archivists, museum curators, technologists, publishers, and funders, among others. They anticipate their report will be available for public review and commentary in early 2005. The report will address in detail the emerging trends in digital humanities and social sciences, new opportunities for global-scale collaboration among domain specialists, the potential of cyberinfrastructure for advancing both research and teaching, and the numerous technology and policy issues that demand resolution in order for cyberinfrastructure to support knowledge communities across the globe. For more information about work to date, see **http://www.acls.org/cyberinfrastructure/cyber.htm**.

Digital Library Federation libraries are positioned to become essential elements of the cyberinfrastructure, and this panel explores the work of Commission with respect to DLF activities. Chaired by Abby Smith (CLIR), senior editor of the Commission, panelists include Daniel Greenstein (CDL), Beth Sandore (UIUC), and Margaret Hedstrom (U-Mich), a member of the Commission.

4.30pm-6.00pm Session 4: DIGITAL PUBLISHING *(Constellation F)*

Cornell's Digital Publishing System (DPubS) as an Open Source Electronic Publishing Solution. David Ruddy, Cornell University

There is a strong and growing need for flexible, cost-effective alternatives to traditional scholarly communications models. Cornell University Library has been developing content management and delivery software since 2000 that offers an adaptable electronic publishing solution to university publishing initiatives. Developed at Cornell, DPubS (Digital Publishing System) provides the technical infrastructure that supports Project Euclid (**http://projecteuclid.org**), a library-based electronic publishing initiative aimed at providing affordable alternatives for independent publishers of mathematics and statistics serial literature. The Library has recently received generous support from The Andrew W. Mellon Foundation to extend and enhance DPubS. This work will create a general-purpose publishing platform to support the dissemination of scholarly literature in diverse fields. DPubS will support peer review, have extensive administrative functionality, and will provide interoperability with other open source repository systems such as Fedora and DSpace.

Cornell University Library is collaborating with the University Libraries and the University Press at Pennsylvania State University to test and refine DPubS. The resulting software will be released under an Open Source license, making it available to libraries, university presses, and other independent publishers, thereby expanding opportunities for creative and affordable communication among scholars around the world.

The presentation will describe the history and functionality of DPubS, the proposed extensions and enhancements for the current two-year project, and the challenges that university electronic publishing efforts face.

XTF: Building a Digital Library Publishing Framework. Kirk V. Hastings, California Digital Library

The California Digital Library has evaluated, adapted, put into production and eventually abandoned a number of SGML/XML publishing frameworks over its short history. Rigid architectures, proprietary query and stylesheet languages, and expense have made most enterprise level solutions inappropriate for our particular needs. In 2002, we decided that our rapidly growing digital object collections warranted an investment in the development of a search and presentation architecture that would meet all of our requirements, be completely under our control, and allow for the future expansion of functionality and collection types.

The eXtensible Text Framework (XTF) organizes and searches collections of large documents in multiple formats, providing sophisticated query capabilities and flexible navigation with hits marked in context. It is a deceptively simple marriage of the Lucene full-text indexing package with XSLT-driven configuration and display via a JAVA servlet. XTF has been moved into production and we continue to migrate our services from proprietary applications.

Goals for the system are:

1. Support diverse document formats, flexible collection organization, and various display and functional requirements;
2. Promote small incremental feature development without high-level programming skills by using XSLT to add intelligence to the data pipeline;
3. Build on existing open source tools and standards efforts to provide a no-cost, non-proprietary solution to the digital library community.

In pursuit of this final goal, we plan to announce the distribution of XTF through sourceForge and will be handing out CDs at the presentation.

Engaging the User: The "Encyclopedia of Diderot and d'Alembert: Collaborative Translation Project" and New Scholarly Paradigms. **Kevin Hawkins, University of Michigan, Ann Arbor; Jason Kuznicki, Johns Hopkins University**

The Encyclopedia of Diderot and d'Alembert: Collaborative Translation Project is building an online, freely available English-language edition of the Encyclopédie, originally published from 1751 to 1777 and containing more than 70,000 articles. The project directors seek copyright permission to republish articles that have been previously translated and assign articles to volunteer translators based on their expressed interests. Submissions are reviewed by the project directors, but translators retain copyright to their text. The Scholarly Publishing Office of the University of Michigan University Library publishes the articles using the DLXS Middleware with the XPAT search engine.

While the project has broken with the traditional publishing model by making the content freely available and allowing individual contributors to retain copyright, it has yet to adopt Creative Commons-type licenses or establish a framework for truly collaborative translation and editing, such as fostered by Wikimedia Foundation projects. As will be explained in more detail, limited resources and institutional caution have slowed the adoption of an alternative intellectual property scheme and delivery system, respectively.

Nevertheless, the Diderot project grows based on its community of users and encourages them to contribute content. The project blurs the line between user and creator of content (reader and writer): for example, people use the website and then decide to contribute, and in the classroom it can be used not only as a source of assigned texts but also as the basis of an assignment ("do your own translation"). While not quite the NPR model gaining popularity among libraries and similar projects (such as the Stanford Encyclopedia of Philosophy), the project directors are confident that they can continue to encourage contributions from faculty, grad students, undergrads, and amateurs while not endangering the academic integrity of the project.

As time allows, we will also discuss recent cooperation with the ARTFL Encyclopédie Project (which presents the entire French original online through a subscription model) to create a bilingual controlled vocabulary for use by both projects.

6.00pm-10.00pm **Tour and Reception** *(National Aquarium in Baltimore)*

6.00pm – 7.00pm
National Aquarium in Baltimore
Self-guided tours

Prior to the Reception, all Forum attendees are invited to take a self-guided tour of the National Aquarium free of charge from 6.00pm – 7.00pm. The Aquarium asks that you arrive no later than 6.15pm to ensure you have adequate time to tour the building before it closes at 7.00pm.

Directions to the building entry for touring
When facing the Main Aquarium Building, walk down the pier, past the blue construction wall, straight past the main entrance escalator and the stroller check tent, to the Security Entrance. Enter through the sliding glass doors and identify yourself at the security desk as a member of the event. You will then take the elevator to the 1st floor to begin the self-guided tour.

7.00pm – 10.00pm
DLF Fall Forum 2004 Reception
Marine Mammal Pavilion

Please join us for cocktails and hors d'oeuvres in the National Aquarium's Marine Mammal Pavilion. Entry begins at 7.00pm.

Directions to the Marine Mammal Pavilion

From the Hyatt, walk towards the Power Plant Building (the building with the large red guitar at the top which houses ESPN Zone, Hard Rock Café, and Barnes and Noble). The Marine Mammal Pavilion will be on your right. Turn right and walk down the pier. At the edge of the building (at the restaurant Chipotle), turn left. The Marine Mammal Pavilion is located on your right. Go up the stairs before the Dolphin Fountain to enter.

Tuesday October 26

8.00am-9.00am Breakfast *(Foyer F Side)*

9.00am-10.30am Session 5: PANEL ON DEVELOPMENTS IN PRODUCTION OUTSOURCING *(Constellation E)*

Panel organized by Perry Willett (Michigan) and Stephen Rhind-Tutt (Alexander Street Press), with Gurvinder Batra (TechBooks), Mark Gross (Data Conversion Laboratory), and Peter B. Kaufman (Innodata Isogen), and Joel Poznansky (Apex CoVantage ePublishing Solutions).

This panel will discuss the latest trends in mark-up, re-keying, indexing and other services for digital libraries as offered by 4 of the leading vendors. Each vendor will present his view of the market, how it is evolving, what they expect for the future and then discuss how they're developing services to respond. The panel will touch on issues such as the increasing effectiveness of OCR, the changing needs of mark-up, meta-data creation and other new services. Each vendor will be invited to answer questions like; "What developments in rekeying/offshore services/mark-up are you most excited about? How do you think these developments will impact the creation of digital libraries?"

9.00am-10.30am Session 6: ART AND IMAGES *(Constellation F)*

Exploring METS: A Case Study Using Architectural Images. Eileen Llona, Diana Brooking, and Marsha Maguire, University of Washington Libraries

Architectural images have many layers of metadata, pertaining both to the content and the image formats. While we are currently using a flat system for expressing metadata relationships in architectural image collections, we are looking at other metadata schemas that can provide a mechanism for expressing these complex relationships.

Starting with a flat file set of metadata, we have been exploring mapping these elements into a METS structure. In the process, we have been evaluating other descriptive schemas that can be incorporated into METS, such as VRA and MODS.

Using display tools available from the Digital Library program at New York University, we have started developing some prototypes of how an architectural image collection might look in METS. Promising aspects of the METS structure include the ability to link various levels of image formats to one content-based record, which reduces redundancies in cataloging, and the ability to hierarchically structure content-based and technical metadata.

Questions that have come up along the way include:

- Which is the best descriptive metadata scheme to use?

- Does the cataloging have to be modified from an image-based system to a work-based system?

We will describe the format of the original metadata, methods for mapping from the home-grown system to METS, and techniques we used for comparing other descriptive metadata schemas.

Finding and Mixing Licensed and Local Content: ARTstor's Approaches and Challenges. James Shulman and Bill Ying, ARTstor

ARTstor is a non-profit initiative, founded by The Andrew W. Mellon Foundation, with a mission to use digital technology to enhance scholarship, teaching and learning in the arts and associated fields. As recent research by Penn State and CDL has confirmed, users will always need more images than any one resource can to provide. Hence, one of our key challenges is to enable users to have a range of ways to integrate their individual and institutional content with content from ARTstor, either within ARTstor's Digital Library or in existing institutional environments.

In this session, we will describe and seek feedback concerning:

- The possible ways that we can expose ARTstor metadata for federated searching;
- The various ways that institutions and end users can integrate local content into ARTstor;
- Lessons learned in creating APIs for importing content and integrating with other repositories.

We will demonstrate software tools we have developed to encourage flexible use of ARTstor and discuss upcoming technical and service developments. Since ARTstor takes a system-wide perspective that attempts to balance the changing needs of both users and content owners, we believe that a conversation with the DLF that explores evolving methods of serving users while respecting the intellectual property concerns of content contributors would be productive for both ARTstor and the user community.

Access to Cultural Heritage Materials in the Digital Realm. Ann Whiteside, University of Virginia; Trish Rose, UC, San Diego

In the art and cultural heritage communities, the most fully developed type of data standards are those that enumerate a set of categories or data elements that can be used to create a structure for a fielded format in a database. Categories for the Description of Works of Art (CDWA) and the VRA Core Categories, Version 3.0 (VRA Core) are examples of data structures or metadata element sets. Although a data structure is the logical first step in the development of standards, a structure alone will achieve neither a high rate of descriptive consistency on the part of cataloguers, nor a high rate of retrieval on the part of end-users.

Unlike the library and archival communities, which have well-established rules for data content in the form of the Anglo-American Cataloguing Rules (AACR), the cultural heritage community in the United States has never had published guidelines similar to AACR that meet the unique and often idiosyncratic descriptive requirements of one-of-a-kind cultural objects.

Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images (CCO) is the first standard within the cultural heritage community to comprehensively address issues of data content for cultural heritage works and their images. This guide helps the cataloger choose terms and define the order, syntax, and form in which those data terms should be entered into a data structure.

This session will discuss why CCO is an important advancement for the cultural heritage community and what the implications will be for access to cultural materials; the challenges of integrating metadata records from heterogeneous/non-standard data and how CCO will help address these challenges; and the development of an XML schema for the VRA Core.

10.30am-11.00am Break

11.00am-12.30pm Session 7: STANDARDS FOR DIGITAL LIBRARY ACCESS AND CONTROL *(Constellation E)*

Shibboleth. Rick Ochoa and Tom Cunningham, New York University

Shibboleth is a powerful way to integrate access control amongst distributed sets of resources. Content can be protected based on user roles as defined by their parent institution. Multiple authentication mechanisms integrate with Shibboleth, from simple authentication to systems using LDAP. The framework provides access to these resources not only on a particular campus, but anywhere a user can connect, thus removing a significant barrier to legitimate use: the need for IP-based authentication.

New York University has been testing Shibboleth as a single-sign-on mechanism for access to multiple resources on campus, from Library resources to NYU's Home portal application. This presentation will explore the modification and integration of Shibboleth code with various systems including Darwin Streaming Server 5 and Apache 1.3, NYUHome, and the Database of Recorded American Music, a digital collection of American Music. It will cover the planning, testing, implementation, and deployment issues across multiple systems on campus and beyond.

Applications of the Digital Object Identifier. Norman Paskin, International DOI Foundation

This presentation will review the Digital Object Identifier (DOI), a persistent identifier system initiated from the publishing industry and now becoming more widely used, characterizing it as an implemented identifier system providing persistent identification, semantic interoperability, and practical implementation mechanisms. The four components of the DOI system are described in outline:

1. Numbering of an entity
2. Resolution of an identifier to current state data (using the Handle System)
3. Interoperable description of the identified resource
4. Governance, policy and implementation including infrastructure maintenance

Where existing identifiers or metadata schemes are present, the DOI system is able to re-use these. The presentation will describe the model for cost recovery and social infrastructure, and show examples of some of the 14 million+ DOIs which are currently assigned. Initial implementations have been simple applications which are now being supplemented by more advanced applications making use of more complex metadata and application interfaces

This presentation will also provide an update on recent development and implementations of the Digital Object Identifier as a persistent identifier. Initially and widely implemented in text publishing (both commercial and non-commercial), the DOI is now being taken up widely as a tool for government documents (through adoption in the UK, European Union, OECD, etc) and for scientific data.

Among issues to be discussed will be:

- The origins, aim and role of the DOI
- The functionality required of a persistent identifier
- The essential role of metadata and how existing metadata schemes can be used
- The operational costs, social infrastructure and governance
- Issues of relevance to libraries such as the "appropriate copy" problem and the role of national libraries and archives
- The role of DOI and related technologies in tools for information access management such as licenses
- The importance of non-commercial usage of, and participation in, DOI

11.00am-12.30pm Session 8: MANAGING METADATA *(Constellation F)*

Designing a Metadata Management Repository. Nathan Rupp, Cornell University; Michael Pelikan, Penn State University; Jeff Young, OCLC

Over the past decade, organizations responsible for digital library projects have moved from individual projects to programs in which a number of projects have been developed. These projects may use different metadata schemes, different schemes to transform one type of metadata into another, and different scripts to facilitate those transformations. Continuing digital library efforts in these organizations may involve metadata, transformation schemes, and facilitating scripts that have been used in the past and it is useful for organizations to record these metadata objects so they can be reused. This is especially true in large, complex, distributed organizations in which practitioners working with metadata may not be in close contact with one another on a daily basis. One way to record these objects is by developing what we have come to call a metadata repository. Such a repository would record the metadata schema used in a particular project, schema used to transform one type of metadata to another, and the scripts used to facilitate those transformations.

This presentation will discuss Cornell University's plan for a repository used to record the different metadata objects used in digital library projects at Cornell, tools that could be added to such a repository beyond just the metadata objects themselves that would further their use, and an OCLC project involving an SRW/U-accessible catalog of XSLT stylesheets that implements some of these concepts.

Preservation Metadata for Digital Repositories. Rebecca Guenther, Library of Congress

PREservation Metadata: Implementation Strategies, or PREMIS, is an activity jointly sponsored by OCLC and RLG, focusing on issues associated with implementing preservation metadata in digital archiving systems. PREMIS is composed of nearly 30 international experts representing national and university libraries, museums, archives, government agencies, and the private sector. The objectives of PREMIS are two-fold:

1. Define a core set of preservation metadata elements, applicable across a broad range of digital preservation activities, and supported by a data dictionary;
2. Identify and evaluate alternative strategies for encoding, storing, managing, and exchanging the core elements within a digital archiving system.

Initiated in June 2003, PREMIS will conclude its activities by the end of 2004. This presentation will report on the initial findings and conclusions of the PREMIS working group. A draft of the recommended core preservation metadata and its supporting data dictionary will be presented as well as a review of some key issues that arose in the consensus-building process. The presentation will also discuss the results of a survey conducted by the PREMIS group, addressing the mission, policy, economic, and technical aspects of digital repositories, as well as current practices for creating, managing, and maintaining preservation metadata within the repository environment. A set of emerging best practices for implementing preservation metadata, culled from the survey responses, will also be summarized.

Automatic Exposure? Capturing Technical Metadata for Digital Still Images. Robin Dale, RLG

This RLG-led initiative seeks to minimize the cost of technical metadata acquisition and maximize the cultural heritage community's capability of ensuring long-term access to digital assets. The goal of the initiative is to lower the barrier for institutions to capture the data elements specified in NISO Z39.87: Technical Metadata for Digital Still Images [currently a Draft Standard for Trial Use]. NISO Z39.87 defines a standard, comprehensive set of data elements key to an institution's ability to manage and preserve its digital images.

The project has engaged device manufacturers to determine what technical metadata their products currently capture and to encourage greater capture of Z39.87-defined technical metadata. It also engages cultural heritage professionals to determine how digital repositories and asset management systems can be supplied with technical metadata that is automatically captured by high-end scanners and digital cameras.

This session will discuss the work of the project, including: identified tools for harvesting technical metadata; upcoming "Automatic Exposure Scorecards" which profile and review the available technologies for capturing technical metadata; progress in influencing technical metadata changes in technology standards followed by device manufacturers; and the development of a Z39.87-Adobe Extensible Metadata Platform (XMP) panel to allow the extension of the metadata handling capabilities of Adobe Photoshop, a commonly used tool in the cultural heritage digitization process.

12.30pm-2.30pm Break for Lunch

2.30pm-4.00pm Session 9: TOOLS AND TRAINING *(Constellation E)*

Interactions of Emerging Gather/Create/Share End-User Tools with Digital Libraries. Raymond Yee, Interactive University Project, University of California, Berkeley

As the quality, quantity, and diversity of scholarly information grow, end-users tools to access and manage this bewildering array of information have been rapidly evolving. In this talk, we will summarize the range of current strategies and tools that enable users to effectively "gather, create, and share" digital information: next generation web browser technology (e.g., Mozilla FireFox and its extensions); personal information managers such as Chandler, web-services enabled- and XML-aware office suites (such as Microsoft Office 2003 and OpenOffice.org); academic projects such as the Scholar's Box, a tool we are building that enables users to gather resources from multiple digital repositories in order to create personal and themed collections and other reusable materials that can be shared with others for teaching and research; high profile open source "Collaboration and Learning Environment (CLE) software" such as Sakai; evolving next generation operating systems, such Microsoft Longhorn. We will analyze the implications of such end-user tools and environments on digital library infrastructure and technology.

New Approaches to Digitization Training for Cultural Heritage Institutions. Amy Lynn Maroso, University of Illinois at Urbana-Champaign

The benefits of digitizing library collections are important and diverse. Patrons outside traditional geographic boundaries can be served alongside local patrons. Access to area history, a vibrant part of many library holdings, and original, rare, and/or valuable materials can be greatly expanded. However, the practice of digitization can fall short of its promise due to poor planning and a lack of digitization skills. Giving cultural heritage institutions the opportunity to learn the skills they need to carry out successful projects is paramount. But how does one make training budget- and time-friendly?

The Basics and Beyond digitization training program, funded by an IMLS National Leadership grant and administered by the University of Illinois at Urbana-Champaign Library, the Illinois State Library, and the Illinois Heritage Association offers an innovative solution. Three training options are available to cultural heritage institutions: one-day on-site workshops, three-week online training, and three-week training followed by a hands-on workshop. The online courses are accessible to anyone with a Web connection, providing institutions across the country and around the world a unique learning opportunity. The online training is affordable to most organizations and its asynchronous nature allows librarians and staff to easily fit the course into their work schedules. As surveys, quiz results, and other data from the courses have consistently shown, the objective of the training is being accomplished: to present cultural heritage institutions with different types of digitization training to suit their time constrains, budgets, and education needs and produce a new set of digitizers who will create successful and long-lasting projects.

2.30pm-4.00pm Session 10: REPOSITORY DEVELOPMENT AND MANAGEMENT II *(Constellation F)*

Economic Growth Center Digital Library: Creating Access to Statistical Sources Not Born Digital. Ann Green, Sandra Peterson, and Julie Linden, Yale University

With funding from the Andrew W. Mellon Foundation, two units at Yale (one from the Library and one from Information Technology Services) are building a statistical digital archive as an extension of Yale's Economic Growth Center Library Collection. In a departure from most digital library conversion projects, which concentrate on images or texts, this project focuses on statistical tables in print.

The digital collection is comprised of the annual statistical abstracts for Mexico's 31 states, from 1994 to 2000. The data include population, industrial, service and commercial censuses, annual and quarterly economic indicators, and trade, financial, and production statistics. The system is being built so that additional series in other languages and from other countries may be added over time. (ssrs.yale.edu/egcdl/)

The goals of the project are to: build a prototype archive of statistics not born digital; implement standard digitization practices and emerging metadata standards for statistical tables; document the costs and processes of creating a statistical digital library from print; build the collection based upon long-range digital life cycle requirements; and present the prototype digital library to the scholarly community for evaluation.

The research questions addressed by the project include:

- Are common digitization practices and standards suited to statistically-intensive documents?
- What are the costs of producing high-quality statistical tables with OCR and editing?
- How scalable is this process, for what kinds of collections, and for what purposes?
- Do faculty members see value in a collections-based digitization effort or prefer on-demand digitization services? How best to integrate the materials into existing resources?
- What effect does online access to the statistical information have on scholarly use of the materials?

The project team is currently evaluating the processes, standards, and best practices emerging from the grant activity and how best to build on the success of the project. It would be our pleasure to discuss the project as well as its findings and challenges at the Fall DLF Forum.

Dynamic de-duplication of bibliographic data for user services. Thorsten Schwander and Herbert Van de Sompel, Los Alamos National Laboratory

De-duplication of bibliographic data is a challenging problem that presents itself in a variety of information aggregating applications including union catalogues, FRBR-oriented catalogues, portals to search distributed databases, and OAI service providers. These challenges are largely caused by idiosyncrasies in the creation of metadata by various parties, and by general data quality problems.

In the context of the user services of the Research Library of the Los Alamos National Laboratory, the de-duplication problem is especially challenging due to the size and heterogeneity of the dataset that needs to be de-duplicated:

1. approximately 60,000,000 bibliographic records originating in over 10 locally stored abstracting and indexing databases need to be de-duplicated to allow presenting users with de-duplicated search results
2. over 500,000,000 citations -- from the ISI citation databases -- must be de-duplicated in order to be able to present citation information for search results.

The Digital Library Research and Prototyping Team has been conducting research aimed at handling the de-duplication challenge in an on-the-fly manner, in an attempt to shift away from the batch-computational approach that is used in current operational services. A solution has been prototyped based on commercial software running on a cluster of blade computers. The software implements a unique fuzzy matching algorithm to search for duplicates of a given bibliographic record (or citation). The system can be trained by librarians to optimize matching results to a specific dataset. The prototyped de-duplication component is integrated in the service environment using a standard-based approach:

1. bibliographic data is added to the database hosted by the blade cluster by harvesting bibliographic data from the Los Alamos Repository (see Abuquerque DLF Forum) using the OAI-PMH
2. queries are performed using NISO OpenURL.

The results of this research are very promising and will lead to moving the component in production. It will remove the burden of recurrent batch de-duplication, significantly increase the flexibility to adopt matching algorithms to the ever growing dataset, and provide better de-duplication results.

4.00pm-4.30pm Break

4.30pm-6.00pm BIRDS OF A FEATHER SESSIONS

1) BOF: Web Services Interoperability and the DLF OCKHAM Reference Model. Martin Halbert, Emory University *(Camden Room)*

This "birds of a feather" session will provide an opportunity for discussion of interoperability using web services, with an initial introduction to the reference model developed by the DLF OCKHAM task force.

The OCKHAM task force was convened two years ago by Dan Greenstein to study interoperability architecture issues (see http://www.diglib.org/architectures/ockham.htm). The task force continued intermittently over a period of months, eventually resulting in a successful research grant application to the NSF NSDL.

This research project is now in the first year of work to develop a simple and broadly applicable framework for interoperability of DL infrastructures using web services, as well as a set of testbed services to try out the framework in practical ways.

The OCKHAM reference model now under development reflects input from the NSDL Core Integration team, as well as input from the DLF OAI Best Practices committee. This BOF will provide an opportunity for interested DLF institutions who have not participated in the OCKHAM task force to hear about and discuss the draft reference model, its approach to interoperability using web services and JXTA, and the testbed services under development as part of the NSDL. The results of this discussion will be fed into the OCKHAM project development, as well as forwarded to the DLF DODL development committee for comment.

2) BOF: Integrating Personal Collection Services. Daniel Chudnov, Yale Center for Medical Informatics; Jeffrey Barnett, Yale University Library; Jeremy Frumkin, Oregon State University Libraries *(Calvert Room)*

A number of recent initiatives address the intersection of personal information management and formal collection development with services for building, publishing, transforming, and sharing "personal collections." BoF attendees involved with or interested in this work will be invited to report on their activities, and to consider approaches for integrating the functions personal collection services provide with digital library services and learning management systems.

3) BOF: OAI Best Practices. Sarah L. Shreeves, University of Illinois Library at Urbana-Champaign *(Douglass Room)*

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been widely adopted since its inception in 2001; there are currently over 500 active data providers from a wide variety of domains and institution types. The protocol has demonstrated its usefulness as a tool to move and aggregate metadata from diverse institutions. However, as the protocol has become more widely adopted, several areas of concern have surfaced that would benefit from documentation of best practices. This session will report on a DLF-convened effort to develop a set of best practices for OAI data and service providers. We encourage participants to share their concerns, questions, and ideas.

4) BOF: Digital Video. Jennifer Vinopal, New York University *(Pratt Room)*

This BOF is for those just thinking about or already involved in digital video projects.

More and more academic institutions are taking on digital video projects, large and small, and are discovering the challenges of working with this media in digital form: the steep learning curve and expense of hardware and software, complex workflows, and practical issues such as storage, preservation, retrieval and presentation, metadata, video standards and file formats, integration with the OPAC, etc.

Based on their experience with New York University's Hemispheric Institute's video project, members of NYU's Digital Library Team will:

1. Outline the basic workflow they've developed for this project;
2. Highlight some of the sticking points in realizing a viable digital video collection available online to the public;
3. Lead a discussion in which participants can share experiences, knowledge, and, especially, solutions to challenges.

We envision this as the first of a series of conversations that may lead to the identification of "good practices" for working with digital video.

Wednesday October 27

8.00am-9.00am Breakfast *(Foyer F Side)*

9.00am-10.30am Session 11: STANDARDS AND SERVICES *(Constellation E)*

EAD 2002: Finding Aid Delivery Using Native XML Technologies. Charles Blair, University of Chicago

The Digital Library Development Center and the Special Collections Research Center of the University of Chicago Library have partnered to deliver searchable finding aids over the Web using native XML technologies. Challenges have included:

- Conversion from SGML to XML
- Migration from EAD 1.0 to EAD 2002
- Choosing markup tools and strategies
- Revising workflows
- Promoting discovery of the finding aids

Questions have included:

- Should we use high-tech or low-tech tools?
- Should we maintain the finding aids directly in the XML database?
- How might we best produce both electronic and printed versions?
- Should we derive descriptive metadata for OAI from MARC or from EAD?

Factors influencing decision-making have included:

- Training costs
- Ease of maintenance
- Technical considerations

Virtual Browsing Via Deep-linked Catalog Searches. Scott Warren, North Carolina State University

This project facilitates virtual browsing by means of deep-linked catalog searches based on Library of Congress (LC) subject headings and call numbers. Collocating these searches by discipline in pull-down menus opens up the print collection in a seamless online fashion to users lacking LC subject heading knowledge, but possessing subject knowledge.

In effect, these menus become interactive thesauri. LC headings are rewritten to patron-recognizable terms. Such rewriting is possible because the thesaurus taps into the catalog, but is a completely separate digital construction that subsets the catalog.

Virtual browsing is important because it may aid awareness of and increased use of print collections in large libraries, especially by researchers who rarely visit the library physically. This technique is amenable to subject areas that strongly rely on print collections (mathematics, history, etc.) that will not be digitized in the near future because of their scale and attendant costs.

Exploring how and if LC virtual browsing could be automated is an open question. At present the technique involves simple JavaScript and HTML, but relies upon direct librarian production. If LC headings and call number ranges could be harvested based on discipline and then automatically added to a browsing menu, all that would be required of the subject specialist is to construct a synonym for each LC heading. Such a technique would more fully utilize the power of LC cataloging in a digital environment and aid in the discovery and use of possibly hidden print collections within a large research collection.

For current working examples, see:

- http://www.lib.ncsu.edu/risd/guides/mathematics/mathbooks.html
- http://www.lib.ncsu.edu/risd/guides/mathematics/mathjournals.html
- http://www.lib.ncsu.edu/risd/guides/history/maps.html (see the section at the bottom, Regional Historical Atlases)
- http://www.lib.ncsu.edu/risd/guides/USRole.html

NISO: Current Work and Strategic Directions. Pat Stevens, OCLC

"Technical standards provide the infrastructure that makes information systems and databases less expensive to develop, easier to use, and universal in value." Do they really?

This quote from the NISO website would be accepted by some and challenged by others. The talk will focus on what NISO is doing today to make the statement a living reality.

The talk begins with a discussion of what NISO means today by the word standards. Guidelines, best practices and rigorous specifications that define how to create a product or a service that conforms with the standard so that it provides an agreed upon level of functionality or quality are one form of standard. Another type of standard defines practices for exchange or commerce. Each type of standard will emerge at a different point in the market, business or service cycles. And, each type requires a different standards development approach. The rest of the presentation will focus on the development of exchange formats; examples from NISO's current work will illustrate the stages of the process.

The talk concludes with a review of NISO's strategic review process, comments heard to date from the DLF community and upcoming opportunities for feedback and involvement.

9.00am-10.30am **Session 12: FILE FORMATS** *(Constellation F)*

Format Dependencies in Repository Operation. **Stephen L. Abrams and Gary McGath, Harvard University**

Almost all aspects of repository operation are conditioned by the format of the objects in the repository. The Harvard University Library Digital Repository Service (DRS) has been in production operation for four years and has over 1.5 million digital objects (7 TB) under managed storage.

A recent comparison of internal technical metadata extracted from these objects with the external metadata supplied in the objects' Submission Information Packages (SIPs) revealed some troubling inconsistencies. Additionally, a small percentage of objects were found to be invalid or malformed with respect to their formats. A post- mortem investigation determined the cause of these problems to include both human and system failures, in some instances on a systemic basis with regard to format.

We report on the findings of this effort and discuss systems that are now in place and under development for automated SIP construction and pre-ingest validation intended to mitigate such problems in the future. We also present an update on JHOVE, the JSTOR/Harvard Object Validation Environment, useful for format-specific object identification, validation, and characterization.

Long Server: A Collaborative Web Site towards Universal File Format Conversion. **Kurt D. Bollacker, The Long Now Foundation**

We believe that in order to survive in the long term, digital data must be able to move around easily, both in hardware (to make copies for redundancy) and in software (converted between encodings and formats to stay comprehensible). We are starting the Long Server project to create tools to promote this notion of high "data mobility". One of the first of these tools is a collaborative Web site for digital file format converters. This site is intended to be a central place for information about all known file format converters. Users will be able to use search and browsing tools to figure out which converters they need and where to get them.

Discussion and documentation may be contributed by users and will be made searchable. To encourage good archival practices, the site will recommend formats that are likely to be long lived and well documented. For converter developers and other contributors, converter records can be added, augmented, or updated collaboratively.

This file format converter Web site is intended to be built and grown quickly. Our goal is to accumulate information that is as complete and rich as possible, rather than be vetted against a high standard of consistency. It is intended to be complementary to the Global Digital Format Registry project by both leveraging GDFR standards and practices, while serving as a "reconnaissance" project to discover where user needs and interests exist. We hope to have the first public services available within 6 months of starting.

Assessing the Durability of Formats in a Digital Preservation Environment: The INFORM Methodology. **Andreas Stanescu, OCLC**

INFORM (INvestigation of FOrmats based on Risk Management) is a methodology for investigating and measuring the risk factors of digital formats and providing guidelines for preservation action plans. In other words, this methodology attempts to discover specific threats to preservation and measure their possible impact on preservation decisions. Moreover, by repeating the process, involved parties can detect changes in the threat model over time, to which digital archives can act accordingly. A comprehensive approach to the format assessment must include the following considerations: (1) risk assessment; (2) significant properties of the format under consideration; (3) the features of the format as defined in the format specification. This report concentrates on the first aspect; the method of incorporating the latter two aspects, those which reflect the quality of the digital format specification, in a preservation decision will be detailed at a later time. Digital archives, institutional repositories and digital libraries can take advantage of the measurements offered by the INFORM method to select digital formats most apt for long term viability. While individuals are biased and subjective in their aversion to risk, collating the assessments of many individuals should generate group-consensus or group-averaged objective results. Hence, preservation plans can be based on objective analysis of risk trends instead of individuals' opinion developed in the relative isolation of their institutions.

10.30am-11.00am Break

11.00am-12.30pm Session 13: METADATA AND DATA HARVESTING
(Constellation E)

Archive Ingest and Handling Test: Interim Report. **Martha Anderson, Library of Congress; Clay Shirky, New York University**

Abstract: In-progress report from a multi-institution test of ingest of a bit-identical archive into different digital preservation repositories.

Description: In May, the Library of Congress, as part of its NDIIPP program, partnered with 4 institutions working on digital preservation frameworks -- Harvard, Johns Hopkins, Old Dominion, and Stanford -- to conduct an Archive Ingest and Handling Test (AIHT). Each institution is taking a bit-identical copy of an existing, moderately complex archive. They are then creating necessary meta-data, and ingesting it into their respective systems. The test provides an opportunity for the participating institutions to understand one another's approaches to digital preservation and, ideally, to find commonalities or areas of complementary work.

The archive itself is a collection of materials gathered by GMU in the aftermath of the September 11th, 2001 attacks, including written accounts, photos, video, audio, and even complete websites. It holds roughly 57,000 files, of nearly a hundred different file types. The archive is accompanied by only minimal metadata. Running the test with this archive allows us to observe how various digital preservation regimes deal with archives too large for human examination of each object before ingest, and where the markup is poor and the file types are varied.

Our presentation will focus on three areas: what we have learned from the completion of Phase I of the test: examination of the archive, meta-data creation and ingest; our progress to date on Phase II, where the participating archives share data with one another; and proposed approaches from the four participants for Phase III: strategies for migrating or emulating content whose native format (e.g. JPEG) may become obsolete in the future.

mod_oai - Metadata Harvesting for Everyone. Herbert Van de Sompel and Xiaoming Liu, Los Alamos National Laboratory; Michael L. Nelson and Aravind Elango, Old Dominion University

We describe the development of an Apache module, mod_oai, which allows for the easy proliferation and adoption of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). While OAI-PMH has significantly impacted digital libraries (DLs), it has yet to make an impact in the general web community despite recent studies that show OAI-PMH is applicable for a variety of purposes. Apache is an open-source web server that is used by 63% of the websites in the world. Apache defines an extensible module format to allow specific functionality to be incorporated directly into the web server. Building an Apache module that "automatically does" OAI-PMH would make the power and flexibility of OAI-PMH available to the general web community.

The OAI-PMH is a simple protocol that defines six "verbs" to facilitate the incremental harvesting of "metadata", or more generally, XML-expressible content. Typically, an OAI-PMH repository is attached to an existing digital library, database, or some other pre-existing content management system. While this allows for the harvesting of the "deep" or "hidden" web, commercial web robots (e.g. Google) do not yet implement the OAI-PMH. This is partially due to the fact that the OAI-PMH community is not yet large enough to appear on their radar.

However, if the installed base of OAI-PMH repositories grew significantly, OAI-PMH enabled web robots could be much more efficient than current web robots, which traverse an entire web site to locate new and updated web pages. Using the OAI-PMH, the new and updated pages would be immediately visible. Eliminating the unnecessary accesses to unchanged pages would result in quicker harvesting for web robots as well as significantly reduced network traffic for web sites.

An introduction to Heritrix, an Open Source Archival-Quality Web Crawler. Dan Avery and the Internet Archive Web Archive Technical Team

Heritrix is the Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project. The Internet Archive started Heritrix development in the early part of 2003. The intention was to develop a crawler for the specific purpose of archiving websites and to support multiple different use cases including focused and broad crawling. The software is open source to encourage collaboration and joint development across institutions with similar needs. A pluggable, extensible architecture facilitates customization and outside contribution. Now, after over a year of development, the Internet Archive and other institutions are using Heritrix to perform focused and increasingly broad crawls. The crawler has been adopted by the IIPC (International Internet Preservation Consortium) as the "official crawler" supported by this group. It is also of particular interest to universities trying to figure out how to do web archiving.

11.00am-12.30pm Session 14: ANNOTATING AND INDEXING DIGITAL RESOURCES *(Constellation F)*

Vivo: A Case Study for the Collaborative Life Sciences Library. Jon Corson-Rikert and Medha Devare, Cornell University

Project Team: Helen-Ann Brown, Kathy Chiang, Phil Davis, Zsuzsa Koltay, Marty Schlabach, Leah Solla, and Susanne Whitaker

As the field of life sciences rapidly expands to embrace aspects of other natural sciences, engineering, philosophy, and computer science, academics and librarians need help identifying life sciences resources and collaborators across Cornell's complex institutional structure and multiple campuses. In response to these problems, the life sciences working group of the Cornell Libraries has been developing a unified library presence to support research and instruction for the university-wide New Life Science Initiative, and will soon release a unique web-based index, called Vivo (http://vivo.library.cornell.edu).

Vivo indexes faculty, courses, events, departments, research groups, services, recent publications, online tools and databases, as well as library resources and services representing 8 libraries on 3 campuses. Life science librarians supervise content selection and curate additional external resources including bibliographic and non-bibliographic life sciences databases, unique services, software tools, and image collections. Vivo uses open-source software (Java, Java Server Pages, and MySQL) and a flexible ontology structure to organize entries by type and cross-index them by an extensible set of relationships that provides users a consistent interface for browsing and searching, with links from each entry directly to the original resource.

Feedback from demonstrations to numerous stakeholders in Cornell's libraries and among faculty and students has been very positive, to the extent that Vivo will be featured in Cornell's upcoming life science faculty recruitment advertisement in Science. These responses indicate that Vivo will fulfill an important need for cross-disciplinary and cross-institutional resource discovery, helping to generate a vibrant virtual community for the life sciences at Cornell.

EVIA: Creating a Digital Archive for Annotated Video. **William Cowan and Jon Dunn, Indiana University, Bloomington**

The Ethnomusicological Video for Instruction and Analysis (EVIA) Digital Archive is a multi-year collaborative project between Indiana University and the University of Michigan, supported in part by a grant from the Andrew W. Mellon Foundation, to create a digital archive for field video recordings captured over the years by ethnomusicology researchers. This digital archive will serve both to preserve this content for future generations of scholars and also to provide a resource to support teaching and learning in ethnomusicology, anthropology, and related disciplines. EVIA has involved a unique collaboration between ethnomusicologists, librarians, archivists, and technologists in carrying out all stages of the project, including video digitization, metadata creation, and system and user interface design.

In this presentation, we will discuss the project's goals, our experiences in working in this collaboration, and lessons learned to date. We will focus particularly on the video segmentation and annotation process and will demonstrate a software tool developed by the project for this purpose. This tool, created using Java Swing and Apple's QuickTime for Java, provides the ethnomusicologist with the ability to segment video, add commentary, annotations, and controlled vocabulary descriptions, and to output this data as a METS document containing MODS descriptive metadata for deposit into a digital repository system and for use in video searching, browsing, and presentation tools currently being developed by the project.