DIGITAL LIBRARY
FEDERATION

SPRING FORUM 2004
PROGRAM

Winners
DLF Forum Fellowships For Librarians New To The Profession

Hannah Frost, Stanford University
Kevin Hawkins, University of Michigan
Alison Morin, Library of Congress
Jacqueline Samples, North Carolina State University
Jewel Ward, University of Southern California

Program Committee

Martin Halbert: Emory University
Leslie Johnson: University of Virginia
Jerome McDonough: New York University
John Ober: California Digital Library
David Reynolds: Johns Hopkins University
David Seaman: Digital Library Federation
Jennifer Vinopal: New York University

Pre-Forum

Monday April 19

9:00am-12:30pm: **DLF Developers Forum**. *(Cabildo Room)* [Open to DLF Developer representative]

9:00am-12:30pm: **Catalog Browse Team**. *(Pontalba Room)* [Closed meeting]

DIGITAL LIBRARY
F E D E R A T I O N

Spring 2004 Forum

Monday April 19

12.00pm-1.00pm Registration *(Le Foyer)*

1.00pm-2.00pm The DLF Today. David Seaman, Digital Library Federation *(Vieux Carré A & B)*

2.00pm-2.30pm Break

2.30pm-4.00pm Session 1: DISTRIBUTED SEARCHING *(Pontalba Room)*

Experiences from an NSF-funded Distributed Search Project ("A Distributed Digital Library of Mathematical Monographs"): Technical and Social Perspectives on Interoperability. David Ruddy, Cornell University; John P. Wilkin, University of Michigan

> The university libraries of Cornell, Göttingen, and Michigan have made available a significant body of mathematical monographs with access provided through a distributed full text search protocol. The virtual collection, comprising more than 2,000 volumes of significant historical mathematical material (nearly 600,000 pages), resides at the three separate institutions and is provided through interfaces to the three entirely different software systems. Two distinct public interfaces to the collection are currently available, both based on the common protocol but reflecting different development efforts at Michigan and Cornell and different perspectives on how to best mediate searches.
>
> The protocol for this distributed search was developed by the three participating institutions over the last three years, with generous support provided by the National Science Foundation. Working from the roots of the DIENST and the then-emergent OAI protocols, the project team focused on creating a new protocol--dubbed CGM, for "Cornell, Göttingen, Michigan"-- that was consistent with OAI, borrowed from DIENST, and added mechanisms for full text searching.
>
> David Ruddy will give a technical overview of the CGM protocol and how it has been implemented in this project, describing in particular how the protocol conveys information about document structure. He will focus attention primarily on the Search verb and several of its challenges: improving search precision through structured queries; identifying common search regions across documents with incongruent structures; and handling large results sets with scaffolding techniques.

John P. Wilkin will suggest that the CGM experience demonstrates that interoperability is not a technical problem, but rather a social one. The CGM protocol demonstrates a solid beginning for full text interoperability; nevertheless, he argues, the problem of content silos has more to do with "dysfunction" among developing digital libraries. Ultimately, what we should hope to accomplish is large, shared repositories. At this point, "interoperability" will cease to be the excuse for not sharing, but will instead be the glue that binds together large multi-institutional efforts.

2.30pm-4.00pm **Session 2: VIRTUAL COLLECTIONS** *(Cabildo Room)*

Digital Asset Management System (DAMS) Infrastructure: a collaborative metadata pilot. **Yong-Mi Kim, University of Michigan**

An effort is underway at the University of Michigan to pilot a digital asset management system for the diverse digital artifacts created by individual academic units. One of its aims is to build an environment where assets are easily searched, shared, edited and repurposed in the academic model. DAMS utilizes IBM's Content Manager and Ancept Media Server, along with other software for access, manipulation and control of digital content such as audio, video, and images. Users will be able to ingest digital assets (image, audio, video, metadata), perform searches, preview assets and retrieve them.

A major challenge has been the metadata to be used in such a system, given the diversity of academic disciplines as well as types of assets to be ingested. In general, academic units over many years have developed and grown materials for academic use, but without a systematic effort to catalog them. Thus these materials are not easily found for reuse by faculty and students. The approach taken in this project has been to:

- Identify and define a core set of metadata for all digital assets, based on existing metadata standards, in particular Dublin Core;
- Identify and define discipline-specific metadata, drawing on existing controlled vocabularies

The processes followed for two academic units, and resulting metadata schemas, will be presented in detail, along with lessons learned.

Toward a User-Centered Digital Library. **Curtis Fornadley and Howard Batchelor, University of California, Los Angeles**

How can a digital library be personal and yet remain connected to authentic primary source material? Recent DLF presentations and papers have suggested a growing interest among system designers in creating tools to support scholarship within digital libraries, and in integrating digital library content with on-line learning environments. Digital libraries of the future may

well be hybrid forms that can respond to, and track, the needs and actions of those who use primary sources for research and teaching.

UCLA Digital Library has created an application called Virtual Collections that offers an environment to support the effective discovery, fusion, and reuse of content from digital collections:

- It allows users to create private or public collections with personal annotations.
- It enables workgroup collaborations on selections from the entire range of digital objects within the collection.
- It supports dynamic collection of metadata from a community of scholars or learners.
- It allows export of content from the Digital Library to the desktop, course web site, or course management system, encouraging the re-purposing of primary content by teachers or students.

The technical architecture leverages lightweight XML and Oracle database technologies to provide a framework for a service that integrates the collections of digital libraries into applications for teaching and research, while also allowing research outcomes to remain within the source collection so that commentary can be shared within collaborative groups.

The presentation will provide an overview of the application, discuss feedback received from early users within a Mellon-funded usability study of the OAI Sheet Music project, and describe in more detail facilities provided for creating and exporting the content of Virtual Collections.

Implementing a Digital Library Architecture at the University of Virginia.
Thornton Staples, University of Virginia Library

In 1999 the University of Virginia Library started working to create an integrated digital library that could serve a broad-based university community in the year 2020. The assumption was that eventually the term "digital library" would refer to a federated effort on the part of somewhere between 10 and 1000 research libraries, working with a variety of other information providers, to provide a seamless integrated network of information for K-12 and college classrooms, scholarly research and for life-long learners. UVA intends to be one of those libraries, building a broad collection of digital resources in all content types and all media.

This presentation will discuss and demonstrate the first phase of a digital library system that presents information as an integrated network of content. This network can be seen as a graph of content nodes delivered and managed as digital objects using the Fedora system. The graph allows for arbitrary levels of aggregation of content, for overlapping sub-graphs of content to represent any number of contexts for a given resource and could be seamlessly distributed across a federation of Fedora repositories.

A first implementation of the system that includes modern English texts, descriptions of art objects and architectural sites and finding aids, all with associated images, will be described. All three collections are searchable from one discovery index that is designed to be integrated with the on-line catalog of traditional resources. Each of the collections has a full-text index of its textual content. Work towards the integration of complex born-digital scholarly resources as sub-graphs will also be discussed. Future plans that include the integration of quantitative datasets, and video and audio collections will be outlined.

4.00pm-4.30pm Break

4.30pm-6.00pm Session 3: DIGITIZING AND ACCESS *(Pontalba Room)*

Copyright Permission for Open Access: Costs, Strategies, and Success Rates. Denise Troll Covey, Carnegie Mellon University Libraries

This presentation will describe three studies conducted by Carnegie Mellon University Libraries to acquire permission to provide open Internet access to copyrighted books. The first study, a random sample feasibility study in 2000-2001, secured an overall success rate of 22%, though the success rate varied significantly by publisher type. The second study, in 2003, sought to acquire permission to digitize and provide open access to a collection of fine and rare books and accompanying archival documents. Different strategies for negotiating with publishers in this study yielded an overall success rate of 44% with a transaction cost of $78 per title. It also showed that authors and estates are as likely to grant permission as university presses. The current and largest study is an attempt to acquire copyright permission to provide open access to 500,000 copyrighted books. Strategies being tested in this study are designed to increase the success rate and decrease the transaction cost per title. Using items in selected collections as an approval plan for publishers, educating these publishers about user behaviors and preferences, offering them incentives to participate in the project, and doing prompt personal follow-up to the initial request letter have already yielded permission to digitize thousands of out-of-print, in-copyright books -- at a cost of $1.50 per title. In addition to the copyrighted books made available on the web, this study will produce best practices for acquiring copyright permission, a database of publisher contact information, and ultimately an outcomes assessment of participating publisher attitudes towards open access.

Update on ACLS History E-Book Project. Nancy Lin, ACLS History E-Book Project; Maria Bonn, University of Michigan

The ACLS History E-Book Project is a cooperative publishing venture among the American Council of Learned Societies, eight scholarly societies, and ten university presses to publish high-quality history books in electronic format.

The project launched as a library subscription product in September 2002 and currently includes 790 previously published books ("backlist") selected by historians from participating learned societies, as well as 15 new titles ("frontlist") developed with participating presses. Each year, 250 backlist titles will be added to the collection. Over the next few years, a total of 85 new frontlist titles will be published, ranging from "print-first" to completely new, "born-digital" titles. The technology back-end for the project is provided by the Scholarly Publishing Office (SPO) at the University of Michigan.

Our presentation will include project updates, demos, and discussions on production processes, workflow, and technology development. We will also discuss challenges and issues, including the need to improve interoperability among digital collections. We will review technologies used (scanned page/OCR, XML, DTDs, XSLT, etc.), design and structural issues (text-chunk size, paragraph numbering, linking, etc.), and production workflow (working with publishers, vendors, print/electronic composition, etc.). We will also discuss SPO's development of the technology back-end using Michigan's DLXS system, and some of the ways in which the needs of electronic publishing are distinct from the needs of digital library projects.

We are seeing that our authors are incorporating materials from digital collections such as APIS, Perseus, Making of America, and other online collections. We also link to book reviews in JSTOR, Project Muse, and the History Cooperative. With this increased interlinking among online resources, collections must adopt use of persistent IDs and clearly identify how one can cite and permanently link to an online object (beyond basic location URL). To ensure that scholars can effectively use and create digital material, it is critical for our new cyberinfrastructure to coordinate use of standards and protocols to facilitate navigation, linking, and scholarly citation. We hope to begin discussions on this and other issues with the digital library community.

For details on XML development for the project, see white paper "Report on Technology Development and Production Workflow for XML Encoded E-Books" http://www.historyebook.org/heb-whitepaper-1.html.

Lessons Learned from RedLightGreen. Merrillee Proffitt, Research Libraries Group

One year ago, RLG was preparing to launch RedLightGreen, a free online service aimed at college undergraduates and optimized to provide access to a wealth of high quality, trusted, print resources through a simple, easy-to-use interface. At the Spring 2003 DLF Forum, we gave a presentation that highlighted use of FRBR, MARC in XML, data mining, user studies, and future directions. Now, with a full semester and more of academic trial use, and with continued funding from the Andrew W. Mellon Foundation, RLG can report:

- Who's using the system, and how?

- Further findings from extended user studies, and how user studies have specifically influenced interface design and helped dictate future directions for the service
- Planned future directions for RedLightGreen
- How institutions can join an expanded partnership for RedLightGreen -- for free.

4.30pm-6.00pm **Session 4: OAI** *(Cabildo Room)*

Enabling Better Collaboration between OAI Metadata and Service Providers: Report from the Third International Workshop on the Open Archives Initiative (OAI3). Kat Hagedorn, University of Michigan Libraries; Jeff Young, OCLC Online Computer Library Center, Inc.; Thomas Habing, Grainger Engineering Library Information Center at the University of Illinois, Urbana-Champaign

Use of the Open Archives Initiative Protocol for Metadata Harvesting has reached critical mass. Of central importance now is making the transition from experimental protocol to a robust, reliable infrastructure component. This will require the building and regularization of collaborative relationships between OAI metadata providers and service providers. Means and ways to enable and facilitate closer, more productive relationships and interactions between metadata providers and service providers was the focus of a breakout session at the recent OAI3 meeting in Geneva. Proposed panel members were leaders of that breakout session and will report in this panel on issues, discussions, and consensuses that emerged from that breakout session. DLF members interested in OAI will be able to gain better insights into tools and resources available and also get a better sense of where OAI is going next.

6.00pm-9.00pm **Reception** *(Les Continents)*

Tuesday April 20

9.00am-10.30am **Session 5: METASEARCHING** *(Pontalba Room)*

What You Need to Know About Metasearching: Lessons and Questions from Metasearch Pioneers. Roy Tennant, California Digital Library; Marty Kurth, Cornell University; Kristin Antelman, North Carolina State University

> Metasearching, also known as cross-database searching or federated searching, is fast becoming a hot new tool for unifying access to disparate databases through one search box. But metasearching is not without its pitfalls. Software is at a very early stage of development, and the success of metasearch services greatly depends on how such services are configured and deployed. This panel of early adopters will focus in on specific lessons learned and provide hard-won advice based on those lessons. Unanswered questions and issues will also be raised to stimulate audience discussion.

> Marty Kurth: Cornell University was an early implementer of, and development partner for, Endeavor's ENCompass metasearch product. Cornell will share the lessons learned from the metasearch services they have released since 2003.

> Kristin Antelman: North Carolina State University constructed its own metasearch service, called "MultiSearch," which has been available since Fall 2002.

> Roy Tennant: The California Digital Library has been involved with metasearching since the release of its SearchLight service in January 2000. CDL's experience with this service has led to a new model of metasearch deployment -- multiple portals tailored to specific audiences and needs. CDL will discuss this model and why it may be a useful model for other research institutions to consider.

9.00am-10.30am **Session 6: ARCHIVING WEB SITES** *(Cabildo Room)*

METS and MODS in the MINERVA project: standards used at LC for archiving Web sites. Allene Hayes, Library of Congress; Leslie Myrick, New York University; Rebecca Guenther, Library of Congress

> This program will review the utilization of developing metadata standards through a review of the purpose and scope of the MINERVA project, LC's Web archiving project. In collaboration with other institutions, it will discuss a METS application profile for Web sites that is under development. The session will review the various collections that have been archived and the progression to fuller use of metadata. The Election 2002 collection was the first to have rich metadata for each Web site, using the Metadata Object

Description Schema (MODS). MODS will be introduced and its particular use for Minerva sites will be reviewed. With MINERVA's 107th Congress collection LC is experimenting with making METS objects for Web sites with richer MODS descriptive metadata.

Virtual Remote Control (VRC). Nancy Y. McGovern, Cornell University Library

Virtual Remote Control (VRC) is Cornell's risk management approach for Web resources. *Virtual* because the approach uses web tools to develop baseline data models representing essential features of selected sites that enable ongoing monitoring. *Remote* because the approach is intended for use by cultural heritage institutions interested in the longevity of web resources residing on remote servers, i.e., not owned or managed by the institution itself. *Control* because at the most proactive end of the approach a monitoring organization may act to protect another organization's resources by agreement or implicit consent through notification and/or action. The VRC approach includes but does not presume the capture of Web sites -- a monitoring organization may not have the means, authority or desire to capture all or some iterations of a Web site. Ongoing monitoring and evaluation allows monitoring organizations to intelligently manage Web resources over time. In conducting our research, we have learned a lot about good Web site management and about promulgating good practice to encourage Web longevity. We will present a review of the model and our findings.

10.30am-11.00am Break

11.00am-12.30pm Session 7: ELECTRONIC ARCHIVES *(Pontalba Room)*

NARA's Electronic Records Archives (ERA) - The Electronic Records Challenge. Fynnette Eaton, National Archives and Records Administration

NARA's position as a public trust requires the preservation and maintenance of records that ensure the accountability and credibility of America's national institutions and document the American national experience. While NARA holds vast amounts of material in many formats, it is the fastest growing recordkeeping medium, electronic records, which provide the largest challenge to maintain and store into the future. NARA's bold initiative, the Electronic Records Archives (ERA), is meeting this challenge. When operational, ERA will be a comprehensive, systematic, and dynamic means for preserving any kind of electronic record, free from dependence on any specific hardware or software. It also will make it possible for Federal agencies to transfer any type or format of electronic record to NARA, as well as allow citizens to find records of interest and obtain them in the formats they want.

After providing an overview of both the current status of the ERA program and how it has been informed by its key partnerships, this presentation will highlight current approaches for preserving the content, context, structure, behavior, and authenticity of documents so as to allow access over time. ERA's involves a fusion of different technologies, such as distributed computing, large scale object storage and access methods, secure infrastructure, and forward-thinking record preservation strategies. Also under discussion will be the ERA system architecture and design, which must not only guard against obsolescence of hardware, software and original record format, but it also must accommodate the ingesting of an immense volume of heterogeneous records.

Implementing OAIS Reference Model at OCLC. Leah Houser and Andreas Stanescu, OCLC Online Computer Library Center, Inc.

This case study examines the development of the OCLC Digital Archive, a third-party service that provides (1) tools for the capture of individual online resources and offline collections; (2) a repository in which those resources and collections can be stored for preservation purposes; and (3) an administration module, which allows depositors to manage their archived resources after submission.

The OCLC Digital Archive complies with the Reference Model for an Open Archival Information System (OAIS). OAIS is a framework, implementations of which vary. The case study focuses on OCLC's development of requirements based on the OAIS and member input, highlighting factors that influenced our decisions.

Several categories of factors influenced the three-year development project. These factors include the nature of OCLC, the institution developing the archive; the local depositor community; and the global digital archiving community. Implementation decisions affected include object types and formats accepted into the archive, access methods, preservation metadata creation, types of tools developed, rights management capabilities, and preservation planning.

Current developments in preservation planning and the upcoming OCLC Digital Archive Preservation Policy are outlined.

Kickin' It Up a Notch: Cooking with the Digital Registry. Taylor Surface, OCLC Online Computer Library Center, Inc.; Erika Linke, Carnegie Mellon University; Rebecca Guenther, Library of Congress; Robin Wendler, Harvard University

Over the past year the Digital Registry Working Group has been formulating MARC/AACR2 guidelines for the Registry of Digital Masters (http://www.diglib.org/collections/reg/reg.htm) to be hosted at OCLC. The guidelines, affectionately known amongst working group participants as "The Cookbook", are now complete and ready for use in practice for registering

collections of digital master objects. At this presentation you'll learn the background and purpose of the Digital Registry, a key piece of future digital preservation infrastructure. In addition, you'll hear several mini-case studies from practitioners in the working group about preparing their metadata for use in the registry.

11.00am-12-30pm Session 8: COLLECTING AND PRESERVING MULTIMEDIA *(Cabildo Room)*

Audio for the Digital Age: the National Recording Preservation Act and the Future of Sound. **Samuel Brylawski, Library of Congress; Abby Smith, Council on Library and Information Resources**

The National Recording Preservation Act of 2000 calls for a study of the current state of recorded sound preservation and a national plan to ensure future access to audio through digital networks. Under the aegis of the Library of Congress, a national board is now at work on key elements of the plan: addressing technical challenges in capturing audio on analog formats and migrating them to digital output; assessing the legal environment for the preservation of recorded sound; and identifying impediments to the fair use of audio for educational purposes. The Library has hired CLIR to help in the implementation of the study and development of the plan.

Details of the plan will be discussed, including progress to date and next steps, as well as the ongoing development of the National Audio Visual Preservation Center at Culpeper, VA.

Collecting Digital Video. **Judith Thomas and Michael Tuite, Robertson Media Center, University of Virginia Library**

Issues of digital video collection building are only slowly entering the digital library discourse: text and images have held sway for more than a decade. The reasons are easy to understand: digital video is tremendously demanding of system and staff resources; there are no clear technical or metadata standards; technological developments are driven by forces outside the world of academia. The primary reason, though, is simply this: libraries generally do not assign the same importance to their motion media collections as they do to their text.

However, we live in a world saturated with motion media. Over the course of the last few years, digital video technology has advanced to the point that production-level creation and delivery are a real possibility for academic libraries, and we are now looking at our video collections with new eyes. At the University of Virginia, our forays into this realm are being driven by the demands of our user community, faculty who are eager for access to digital video for teaching and research.

This presentation will focus on several issues relating to digital video collection-building and delivery. Using three case studies, we will discuss

technical and metadata decision-making and describe the workflow currently in place at UVa. The cases will feature three types of content: videos purchased from a media vendor; unique film footage from our Special Collections; field-based documentation created by a faculty member. We will also present two new "homegrown" tools created to manage metadata and facilitate access to our digital video collections.

12.30pm-2.30pm Break for Lunch

2.30pm-4.00pm Session 9: DIGITAL IMAGES *(Pontalba Room)*

Dumbing up or Dumbing Down? Developing a flexible information architecture for image/metadata retrieval and display in a Digital Library context. Joseph B. Dalton, The New York Public Library

> The challenges in developing a flexible information architecture for searching and examining 200,000+ images from across The New York Public Library's research collections are many. In light of The Library's mandate to "provide free and open online access" to its immense physical collections, one of the Digital Library Program's initial challenges has been to develop a set of consistently "user-friendly" search and retrieval functions appropriate for a wide audience. The NYPL Research Libraries' traditional user-base has included curators, academic librarians, professors, authors and other researchers, but it is anticipated that NYPL Digital Gallery's future audience will likely include a majority of other users (K-12 students, post-secondary students, hobbyists, the intellectually curious, the causal browser referred from an external source, etc.). How do we ensure that the site's functionality is largely transparent for a wide variety of users, while providing context, metadata and access points appropriate for imaged material from The Research Libraries?

Digital Image Services Come of Age (But Will They Ever Grow Up?). Laine Farley, California Digital Library; Henry Pisciotta, Pennsylvania State University Libraries

> The introduction of digital image services has revealed complexities in service creation and delivery and the need for a deeper understanding of users' personal image collections. Based upon Penn State's Visual Image User Study (VIUS) and LionShare projects and UC's Image Demonstrator Project the presenters will draw upon surveys, focus groups, and related data, as well as experience in service prototyping to explore image services from the user perspective. Through the continuum of users' efforts to create, discover, use and reuse images for research and instruction, the presenters will discuss what has been learned about user needs and institutional capabilities to meet them, what is puzzling, and what areas still need investigation. Penn State's VIUS documented the importance of content and one-stop-shopping to potential

system users. The UC's work pinpoints critical areas (metadata, software, workflow, and others) in the complex process of coordinating multiple collections. The research of both institutions underscores the importance of personal collections (44% of pictures users maintain one.) UC is working with LUNA's Insight to test a personal collection manager that is fully coordinated with institutional collections. Penn State's LionShare project proposes a peer-to-peer model that would enhance access to personal collections and facilitate interaction with institutional collections.

Reviving DIDO: Using Contextual Inquiry to Inform the Redesign of an Art Image Resource. Michelle Dalmau, Indiana University Digital Library Program

Indiana University's Digital Library Program (DLP) and the Fine Arts Slide Library have begun to re-assess the Digital Images Delivered Online (DIDO) system as it is straining to meet the needs of art history faculty and students. DIDO, originally developed in 1996, was intended as a resource to supplement the traditional 35mm slides lecture format, but has now become a primary source for art history faculty who wish to present lectures in a digital format. DIDO needs to evolve from a basic search and display tool to one that supports digital content creation for courses. In order to provide meaningful design recommendations for the next generation of the system, the processes of Contextual Design, especially Contextual Inquiry, have been applied to better understand how faculty create and present lectures.

The Contextual Design approach supplies the user-centered tools and techniques designers and usability professionals require to create innovative software and hardware systems that truly do support the work practices of the targeted user group. It provides a framework for designers and usability professionals to evolve design ideas based on a shared understanding of how people work in various contexts. This talk will introduce the framework, with a focus on Contextual Inquiry, the first of seven major steps of Contextual Design, and explain why it is a valuable data gathering method for designing digital libraries with pedagogic and didactic purposes.

By illustrating Contextual Inquiry along with Work Modeling and Consolidation, the two subsequent steps of Contextual Design, with example data collected from recent DIDO studies, it will become apparent how a design team can easily appropriate the approach towards the vision and development of an intuitive and useful system.

2.00pm-3.30pm Session 10: OPEN SOURCE SOFTWARE IN DIGITAL INITIATIVES *(Cabildo Room)*

From Creation to Dissemination: A Case Study in the Library of Congress's use of Open Source Software. Corey Keith, Library of Congress

The Library of Congress's use of open source software tools has enabled the rapid and flexible development and management of multiple digital projects.

In our environment, the mantra is to get data into XML as early in the production process as possible thus enabling the flexible nature of XML and the use of common solutions subsequently in the production process.

We will show how LC makes this initial conversion of data from disparate sources into XML. Then we will show the aggregation of these XML streams to produce complex digital objects, using METS for standards support. On the delivery side we will show the pipelined approach to dissemination of complex digital objects which allows user interface development to be separate from application logic.

During this presentation we will also highlight other open source tools not directly involved in this flow of digital object data. LC is adopting open source tools for the management of digital projects also. We are using defect tracking applications, version control, and other tools to better manage digital projects from small to large.

SRW: the Search and Retrieve Web Service. Robert Sanderson, University of Liverpool

SRW, the Search/Retrieve Webservice, is an XML oriented protocol designed to be a low- barrier-to-entry solution to searching and other information retrieval operations across the internet. It uses existing, well tested and easily available technologies such as SOAP and XPath to perform what has been done in the past using proprietary solutions.

The design has been informed by 20 years of experience with Z39.50, and is both robust and easy to understand while still retaining the important aspects of its predecessor. Building on Z39.50 semantics enables the creation of gateways to existing Z39.50 systems; web technologies reduce the barriers to new information providers allowing them to make their resources available via a standard search and retrieve service.

After an initial discussion of the protocol and the changes between the experimental version 1.0 and the stable 1.1 (released just this past February), the presentation will look briefly at open source implementation details from several independant developers, including the LC's gateway.

4.00pm-6.00pm BIRDS OF A FEATHER SESSIONS

Open Source Software in Digital Initiatives. Corey Keith, Library of Congress; Rob Sanderson, University of Liverpool *(Cabildo Room)*

This session will follow up from the presentation in Session 10 (above) and allow more detailed sharing and discussion of presenters' and participants' experiences with Open Source software in the digital development environment.

ARTstor. James Shulman, ARTStor *(Pontalba Room)*

> ARTstor is a non-profit service that provides useful collections of art images for non-commercial educational use. The ARTstor Charter Collections (available July 1 on a site-licensed basis) will include 300,000 images, tools that allow users to make active use of the collections, and an intellectual property environment that has community-wide support. At DLF, we will also report on ARTstor's policies and procedures concerning interoperating, recognizing that ARTstor will need to "land" very differently at different institutions, including those that have already made substantial investments (and progress) in building, managing, and making use of digital images.

Between the Sheets: Enriching the Catalog. Roy Tennant, California Digital Library *(Vieux Carré A)*

> For almost three decades librarians have advocated the enhancement of online library catalog records with book tables of contents, sample text, indexes, reviews, cover images, etc. We believe that deployed technologies, user expectations, and emerging standards such as METS, OAI-PMH, and ONIX make this a propitious time for libraries to aggressively pursue bibliographic record enhancement strategies. This session will briefly report on an ad hoc collaborative effort begun at ALA Midwinter 2004 to build an infrastructure to enable distributed, non-duplicative input of record-enriching content using standards and practices currently available and proved effective. We will invite BOF attendees to share their concerns, ideas, and comments. As this effort is an informal collaborative, anyone is welcome to participate in advancing the future of the library catalog. Come join us!

Database-driven approaches to EAD. Stephen Davis, Columbia University *(Vieux Carré B)*

> Database-driven approaches to EAD and archival management information, including an EAD / SQL data model.
> http://www.columbia.edu/cu/libraries/inside/projects/findingaids/planning/considerations_2002-08-27.html

Wednesday April 21

8.00am-9.00am Breakfast *(Le Foyer)*

9.00am-10.30am Session 11: PRESERVATION REPOSITORIES *(Pontalba Room)*

Building a robust knowledge base for digital formats. John Mark Ockerbloom, University of Pennsylvania

> Long term preservation and reuse of digital information requires detailed knowledge of the formats used by this information. Several major libraries and archiving institutions have proposed a Global Digital Format Registry to collect format information and make it available for digital library needs.
>
> Building a knowledge base that is authoritative, comprehensive, and widely used, however, is easier said than done. Many details concerning the information the registry should collect, how the information should be managed, and how the registry will interact with users and other systems are still uncertain. These details may prove crucial to the long-term success of a global format registry.
>
> At the University of Pennsylvania, we are developing a prototype registry service to test some design hypotheses for a format registry. Fred, our Format Registry Demonstration, first went online in late March, and allows interested parties to contribute, view, and maintain format information. Fred is not itself intended to be the global format registry, but rather a testbed for ideas on how to design, build, and maintain such a registry.
>
> In my presentation, I will discuss the initial design and implementation of Fred, and how it relates to existing format information systems like MIME, TOM, and PRONOM, as well as to shared information resources like authority control systems and Wikipedia. I'll also discuss our initial experiences with the system, what we hope to learn from it, and how the DLF community can participate in building a better format registry.
>
> For more information about Fred, see http://tom.library.upenn.edu/fred/.

A Repository of Metadata Crosswalks. Carol Jean Godby, Devon Smith, Eric Childress, and Jeff Young, OCLC Online Computer Library Center, Inc.

> In "Two Paths to Interoperable Metadata ," we argued that XSLT scripts are an appropriate tool for processing crosswalks when the metadata translation task is straightforward. In response to interest from the metadata community, we have created an OAI repository of XSLT-encoded crosswalks, which we will demonstrate. We will also discuss some of the conceptual problems that arise when we try to make the XSLT scripts more usable by documenting the meaning behind the transforms. Our demo associates three pieces of information: the crosswalk, the source metadata standard, and the target metadata standard, each of which may have a machine-readable encoding and human-readable description. This representation brings together all of the

information required to access and interpret crosswalks. But it raises questions about how best to describe these complex objects and exposes gaps that must eventually be filled in by practitioners.

This exercise also forces us to assess the theoretical significance of crosswalks. On the one hand, crosswalks may simply represent a stopgap solution to the problem of heterogeneous data. This view implies that the metadata translation problem is local and temporary and that crosswalks are not meant to be reusable. A more hopeful view is that crosswalks are persistent and represent an attempt to identify interoperable elements among metadata standards that have been developed in different communities of practice. A well-designed repository of metadata crosswalks enables us to see how far we have come toward resolving this important issue for stewards of digital libraries.

Digital Repository Interoperability with Learning Systems. David Greenbaum, University of California at Berkeley; Leslie Johnston, University of Virginia Library.

To make the most effective use of digital content in teaching, learning applications need to be able to easily interoperate with digital repositories so that teachers and students can discover, access, view, quote, adapt, and evaluate appropriate learning material. Unfortunately, many data sources have not been designed to interoperate with other repositories or with learning applications. A working group, supported by the Mellon Foundation and DLF, has developed a set of use-case scenarios and a report that present a checklist and discussion of digital repository services that are needed to make digital content usable by learning applications. An overview of the use-case scenarios and checklist of interoperability guidelines will be presented in this session.

9.00am-10.30am **Session 12: SPECIAL COLLECTIONS** *(Cabildo Room)*

The OpenEmblem Portal at the University of Illinois at Urbana-Champaign. Nuala Koetter, University of Illinois at Urbana-Champaign

The OpenEmblem Portal aims to be a resource for emblem book researchers from around the world, helping them share resources and discussions with others in the emblem scholarly community. The University of Illinois holds an internationally renowned collection of emblem books that is among the most highly utilized primary source materials of its type worldwide. Nationally and internationally known emblem scholars regularly consult our collections and the collections have been the topic of numerous publications about the emblems themselves and their bibliographic environments. Emblem books can possibly be looked upon as the multi-medial publications of the 17th and 18th centuries. They are books that link together three constitutive elements-a motto, a woodcut or engraving and an explanatory poem. An emblem is more

than the sum of its parts, because the interplay between text and image produces a great meaning than any of the individual components can provide.

UIUC has just recently set up a new portal for the world-wide emblem scholarly community, using the Internet Scout Portal Toolkit, developed at the University of Wisconsin. In this presentation, we will showcase the digitized emblem books from the University of Illinois which have been cataloged using a metadata schema developed specifically for emblem books and which we map to the Dublin Core schema. We will discuss how the digitized materials have been integrated, using the OAI protocol, into the OpenEmblem Portal, together with love emblems from the University of Utrecht and other future plans for the emblem portal.

Opportunities for Collaboration: The HEARTH Project. Joy Paulson and Nathan Rupp, Cornell University

The Home Economics Archive: Research, Tradition, and History (HEARTH) project at Cornell University's Albert R. Mann Library, a core electronic collection of monographs and serials in home economics and related disciplines published between 1850 and 1950, is a prime example of a successful, collaborative digital library project. The HEARTH project has multiple components: metadata, content, a system and user interface for storing and accessing the metadata and content, and a front end on the World Wide Web to provide some context to the overall project. Rather than concentrating the work on all these components in one particular library unit, the work was dispersed throughout all sections of the library. The workflow for this project fell across various units within the library, including the information technology section, collection development and preservation, technical services and public services. Staff in the information technology section created the systems that were used to create structural metadata and tie it together with the other metadata components. Metadata was created by three different project groups: structural metadata by preservation staff, descriptive metadata by technical services staff, and administrative metadata by the scanning vendor. Staff in the public services section created the web-based front end used to access the content in HEARTH. We will discuss the workflow and connections between the departments that were associated with this project. We will show how the cooperation between these groups resulted in a successful digital library system.

The Usability of Electronic Finding Aids During Directed Searches. Christopher J. Prom, University of Illinois at Urbana-Champaign

This presentation presents findings from a major research project conducted to measure the usability of on-line archival finding aids. The study measured responses for users interacting with eight finding aids, including interfaces at DLF members Illinois, Yale, Princeton, and U-C Berkeley/CDL. The study provides specific insights regarding how users navigate archival descriptive

information and how archivists and digital librarians might design interfaces which facilitate effective search strategies. Both the methodology employed and the conclusions will likely be of broad interest to conference attendees. The study juxtaposed different interfaces in a ASP-driven search portal. (See http://web.library.uiuc.edu/ahx/survey/usab-test/ for a non-functional version.) In addition, all users took a survey and thirty-five of the 89 participants were observed by the project director or his assistant. Both statistical and qualitative findings are provided and correlated to demographic data such as archival/library experience and self-reported computer expertise.

The study found that system (i.e. computer) expertise was a more salient predictor of quick finding aid usage than was domain (i.e. archival) expertise. Experienced archival users and novices utilize very different methods of searching for archival information. Nevertheless, certain finding aid features (including alphabetical lists, page-top tables of contents, Google-like search algorithms, and single-page search options) enabled both sets of users to use some interfaces much more efficiently than alternate designs. The study provides baseline data and conclusions which will assist in reengineering access to archival finding aids and by implication digital libraries.

10.30am-11.00am Break

11.00am-12.30pm Session 13: SEARCH ENGINE TECHNOLOGY AND DIGITAL LIBRARIES *(Pontalba Room)*

Beyond Digital Libraries - The Use of Search Engine Technology to Create Next Generation Scholarly Portals. Norbert Lossau and Friedrich Summann, University of Bielefeld; Dr. Bjoern Ohlstadt, FAST Search

Current Portal solutions (incl. the Digital Library North Rhine-Westphalia, iPort, Electra, Metalib/DigiTool, EnCOMPASS) respond to the need for integrated access to the increasing number of electronic resources that reached the market over the last ten years. Their technology and concepts are often based in the first hand on searching metadata (bibliographic descriptions, keywords, abstracts). Full text search features for e-journals have only been introduced over the last years.

How should next generation portals be designed and what should be our strategy forward? Bielefeld UL has taken a pragmatic approach that builds on existing state-of-the-art search and content matching technology and develops on top of it where necessary. The main focus is not on generic research but on improvements or adoptions by development of add-ons. Instead of developing a new system or spending resources on rebuilding a powerful search architecture, efforts and resources should better be focusing on improving user-interfaces, adding intelligent browsing and navigation features to search boxes or developing and introducing more generic connectors to integrate the "deep" web resources.

The paper will report on the activities at Bielefeld University Library in evaluating and testing search engine technology. An early implementation of search engine technology will be presented that integrates distributed digitised collections (incl. Cornell, Michigan, Göttingen, and Bielefeld University Library's resources), an online library catalogue, preprint servers, subject databases, electronic journals and institutional repositories.

11.00am-12.30pm **Session 14: E-RESOURCES** *(Cabildo Room)*

Digital Library at Dartmouth: Evolution of a New Service. Mary M. LaMarca, Dartmouth College Library

Faced with the task of creating a "Digital Library at Dartmouth", the designated working group decided to create a digital directory of all web-based digital resources owned or licensed by the library. We named this digital directory, eResources.

This new service allows generation of lists of digital resources by type; these include: subject guides, encyclopedias/dictionaries, article indexes, research databases, electronic journals, electronic books, electronic news sources and manuscript finding aids. Users can search or browse by type of electronic resource, or by subject. Users can limit their search, and have access to an advanced search with Boolean capability.

During the past year, eResources has gone through a number of modifications and enhancements based on librarian and user feedback. This talk will outline the evolution of this new service and its current use at the Dartmouth College Library.

XML Schema for E-Resource Licenses. Nathan Robertson, Johns Hopkins University; Tim Jewell, University of Washington

An important focus of the DLF-sponsored Electronic Resource Management Initiative is to foster appropriate metadata standards to allow parties to exchange information about e-resources, e-resource packages, and licenses. While an early goal of the Initiative was to present a draft XML schema that would encompass most relevant functions and data elements, time constraints and the rapid emergence of proprietary Digital Rights Management and Rights Expression Language initiatives have led the project's Steering Group to refocus its XML work on the area in which libraries have the greatest immediate stake: how license data is defined, structured and expressed.

Consistent with its effort to utilize existing standards wherever possible, the Initiative has explored the possibility of expressing e-resource licensing through an existing DRM standard. The result of that exploration is a prototype ERMI license expression in an extended version of the Open Digital Rights Language (ODRL). This presentation will discuss the prototype and describe the advantages, disadvantages, and difficulties of this attempt to extend an existing standard.

Post-Forum

Wednesday April 21

2.00pm-6.00pm: METS Editorial Board. *Vieux Carré A* [Closed meeting]

2.00pm-6.00pm: Fedora/ARROW meeting. *Vieux Carré B* [Closed meeting]

Thursday April 22

9.00am-1.00pm: METS Editorial Board. *Vieux Carré A* [Closed meeting]