Digital Libraries

By Edward A. Fox, Hussein Suleman, Devika Madalli, and Lillian Cassel
Draft 5 / May 2, 2003

For: Practical Handbook of Internet Computing
(to be published by CRC Press)

1. Introduction

Definitions of "digital library" (DL) abound [1] [2], but a consistent characteristic across all definitions is an integration of technology and policy. This integration provides a framework for modern digital library systems to manage and provide mechanisms for access to information resources. This involves a degree of complexity that is evident whether considering: the collection of materials presented through a digital library; the services needed to address requirements of the user community; or the underlying systems needed to store and access the materials, provide the services, and meet the needs of patrons. Technologies that bolster digital library creation and maintenance have appeared over the last decade, yielding increased computational speed and capability, even with modest computing platforms. Thus nearly any organization, and indeed many individuals, may consider establishing and presenting a digital library. The processing power of an average computer allows simultaneous service for multiple users, permits encryption and decryption of restricted materials, and supports complex processes for user identification and enforcement of access rights. Increased availability of high-speed network access allows presentation of digital library contents to a worldwide audience. Reduced cost of storage media removes barriers to putting even large collections online. Commonly available tools for creating and presenting information in many media forms make content widely accessible without expensive special purpose tools. Important among these tools and technologies are coding schemes such as JPEG, MPEG, PDF, and RDF, as well as descriptive languages such as SGML, XML, and HTML.

Standards related to representation, description, and display are critical for widespread availability of DL content [3]; other standards are less visible to the end user, but just as critical to DL operation and availability. HTTP opened the world to information sharing at a new level by allowing any WWW browser to communicate with any information server, and to request and obtain information. The emerging standard for metadata tags is the Dublin Core [4], with a set of 15 elements that can be associated with a resource [5]: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights. Each of the 15 elements is defined using ten attributes specified in ISO/IEC 11179, a standard for the description of data elements. The ten attributes are Name, Identifier, Version, Registration Authority, Language, Definition, Obligation, Datatype, Maximum Occurrence, and Comment. The Dublin Core provides a common set of labels for information to be exchanged between data and service providers.

The technologies are there. The standards are there. The resources are there. What further is needed for the creation of digital libraries? Though the pieces are all available, assembling them into functioning systems remains a complex task requiring expertise unrelated to the subject matter intended for the repository. The field is still in need of comprehensive work on analysis and synthesis, leading to a well-defined science of digital libraries to support the construction of specific libraries for specific purposes. Important issues remain as obstacles to making the creation of digital libraries routine. These have less to do with technology and presentation than with societal concerns and philosophy, and deserve attention from a wider community than the people who provide the technical expertise [6]. Among the most critical of these issues are Intellectual Property Rights, privacy, and preservation.

Concerns related to Intellectual Property Rights (IPR) are not new; nor did they originate with work to afford electronic access to information. Like many other issues, though, they are made more evident and the scale of the need for attention increases in an environment of easy widespread access. The role of IPR in the well-being and economic advance of developing countries is the subject of a report commissioned by the government of the United Kingdom [7]. IPR both serves development, by providing incentives for discovery and invention, and impedes progress, by denying access to new developments to those who could build on the early results to explore new avenues, or could apply the results in new situations. Further issues arise when the author of a work chooses to self archive, i.e., to place the work online in a repository containing or referring to copies of his or her own work, and/or in other publicly accessible repositories. Key questions include the rights retained by the author and the meaning of those rights in an open environment. By placing the work online, the author makes it visible. The traditional role of copyright to protect the economic interest of the author (the ability to sell copies) does not then apply. However, questions remain about the rights to the material that have been assigned to others. How is the assignment of rights communicated to someone who sees the material? Does self-archiving interfere with possible publication of the material in scholarly journals? Does a data provider have responsibility to check and protect the rights of the submitter [8]?

Digital libraries provide opportunities for widespread dissemination of information in a timely fashion. Consequently, the openness of the information in the DL is affected by policy decisions for the developers of the information and those who maintain control of its representations. International laws such as TRIPS (Trade Related aspects of Intellectual Property Standards) determine rights to access information [9]. Digital library enforcement of such laws requires careful control of access rights. Encryption can be a part of the control mechanism, as it provides a concrete barrier to information availability, but adds complexity to digital library implementation.

Privacy issues related to digital libraries involve a tradeoff between competing goals: to provide personalized service [10] on the one hand and to serve users who are hesitant to provide information about themselves on the other. When considering these conflicting goals, it is important also to consider that information about users is useful in determining how well the DL is serving its users, and thus relates to both the practice and evaluation of how well the DL is meeting its goals.

Though the field of digital libraries is evolving into a science, with a body of knowledge, theories, definitions, and models, there remains a need for adequate evaluation of the success of a digital library within a particular context. Evaluation of a digital library requires a clear understanding of the purpose the DL is intended to serve. Who are the target users? What is the extent of the collection to be presented? Are there to be connections to other DLs with related information? Evaluation consists of monitoring the size and characteristics of the collection, the number of users who visit the DL, the number of users who return to the DL after the initial visit, the number of resources that a user accesses on a typical visit, the number of steps a user needs in order to obtain the resource that satisfies an information need, and how often the user goes away (frustrated) without finding something useful. Evaluation of the DL includes matching the properties of the resources to the characteristics of the users. Is the DL attracting users who were not anticipated when the DL was established? Is the DL failing to attract the users who would most benefit from the content and services?

2. Theoretical foundation

To address these many important concerns and to provide a foundation to help the field advance forward vigorously, there is need for a firm theoretical base. While such a base exists in related fields, e.g., the relational database model, the digital library community has relied heretofore only on a diverse set of models for the sub-disciplines that relate. To simplify this situation, we encourage consideration of the unifying theory described in the 5S model [11, 12].

We argue that digital libraries can be understood by considering five distinct aspects: Societies, Scenarios, Spaces, Structures, and Streams. In the next section we focus on Scenarios, related to services, since that is a key concern and distinguishing characteristic in the library world. In this section we summarize issues related to other parts of the model.

With a good theory, we can give librarians interactive, graphical tools to describe the digital libraries they want to develop [13]. This can yield a declarative specification that is fed into a software system that generates a tailored digital library [14]. From a different perspective, digital library use can be logged in a principled fashion, oriented toward semantic analysis [15].

5S aims to support Societies' needs for information. Rather than consider only a user, or even collaborating users, or sets of patrons, digital libraries must be designed with broad social needs in mind. These involve not only humans but also agents and software managers.

In order to address the needs of Societies, and to support a wide variety of Scenarios, digital libraries must address issues regarding Spaces, Structures, and Streams. Spaces cover not only the external world (of 2 or 3 dimensions, plus time, or even virtual environments – all connected with interfaces) but also internal representations using feature vectors and other schemes. Work on geographic information systems, probabilistic retrieval, and content-based image retrieval falls within the ambit of Spaces.

Since digital libraries deal with organization, Structures are crucial. The success of the Web builds upon its use of graph structures. Many descriptions depend on hierarchies (tree structures). Databases work with relations, and there are myriad tools developed as part of the computing field called 'Data Structures'. In libraries, thesauri, taxonomies, ontologies, and many other aids are built upon notions of Structure.

The final 'S', Streams, addresses the content layer. Thus, digital libraries are content management systems. They can support multimedia streams (text, audio, video, and arbitrary bit sequences) that afford an open-ended extensibility. Streams connect computers that send bits over network connections. Storage, compression/decompression, transmission, preservation, and synchronization are all key aspects of working with Streams. This leads us naturally to consider the myriad Scenarios that relate to Streams and the other parts of digital libraries.

2.1 Scenarios

Scenarios "consist of sequences of events or actions that modify the states of a computation in order to accomplish a functional requirement" [12]. Scenarios represent services, as well as the internal operation of the system. Overall, scenarios tell us what goes on in a digital library. Scenarios relate to societies by capturing the type of activity that a user group requires, plus the way in which the system responds to user needs.

An example scenario for a particular digital library might be access by a young student who wishes to learn the basics of a subject area. In addition to searching and matching the content to the search terms, the DL should use information in the user profile and metadata tags in the content to identify material compatible with the user's level of understanding in this topic area. A fifth grader seeking information on animal phyla for a general science report should be treated somewhat differently than a mature researcher investigating arthropoda sub phyla. While both want to know about butterflies, the content should be suited to the need.

Scenarios are not limited to recognizing and serving user requests. Another scenario of interest to the designer of a digital library concerns keeping the collection current. A process for submission of new material, validation, description, indexing, and incorporation into the collection is needed. A digital library may provide links to resources stored in other digital libraries that treat the same topics. The contents of those libraries change and the DL provider must harvest updated metadata in order to have accurate search results. Here the activity is behind the scenes, not directly visible to the user, but important to the quality of service provided.

Other scenarios include purging the digital library of materials that have become obsolete and no longer serve the user community. While it is theoretically possible to retain all content forever, this is not consistent with good library operation. Determining which old materials have value and should be retained is important. If all material is to be kept forever, then there may be a need to move some materials to a different status so that their presence does not interfere with efficient processing of requests for current materials. If an old document is superceded by a new version, the DL must indicate that clearly to a user who accesses the older version.

Services to users go beyond search, retrieval, and presentation of requested information. A user may wish to see what resources he or she viewed on a previous visit to the library. A user may wish to retain some materials in a collection to refer to on later visits, or may simply want to be able to review his or her history and recreate previous result lists. In addition, the library may provide support for the user to do something productive with the results of a search. In the case of the National Science Digital Library, focused on education in the STEM (science, technology, engineering, and mathematics) areas, the aim is to support teaching and learning [16-20]. For example, in the NSDL collection project called CITIDEL (Computing and Information Technology Interactive Digital Educational Library) [21], a service called VIADUCT allows a user to gather materials on a topic and develop a syllabus for use in a class. The syllabus includes educational goals for the activity, information about the time expected for the activity, primary resources and additional reference materials, pre-activity, activity, and post-activity procedures and directions, and assessment notes. The resulting entity can be presented to students, saved for use in future instances of the class, and shared with other faculty with similar interests.

GetSmart, another project within the NSF NSDL program, provides tools for students to use in finding and organizing useful resources and for better learning the material they read [22]. This project provides support for concept maps both for the individual student to use and for teams of students who work together to develop a mutual understanding.

Scenarios also may refer to problem situations that develop within the system. A scenario that would need immediate attention is deterioration of the time to respond to user requests beyond an acceptable threshold [23]. A scenario that presents the specter of a disk crash and loss of data needs to be considered in the design and implementation of the system. An important design scenario is the behavior by a community of users to achieve the level of traffic expected at the site and the possibility of DL usage exceeding that expectation.

3. Interfaces

User interfaces for digital libraries span the spectrum of interface technologies used in computer systems. The ubiquitous nature of the hyper-linked World Wide Web has made that the de facto standard in user interfaces. However, many systems have adopted approaches that either use the WWW in non-traditional ways or use interfaces not reliant on the WWW.

The classical user interface in many systems takes the form of a dynamically generated website. Emerging standards, such as the XSLT transformation language [24], are used to separate the logic and workflow of the system from the user interface. See Figure 1 for a typical system using such an approach. Such techniques make it easier to perform system-wide customization and user-specific personalization of the user interface. Wang's Masters thesis [25] proposes a general solution to connecting digital libraries and visualization systems through a tailored lightweight protocol [26].

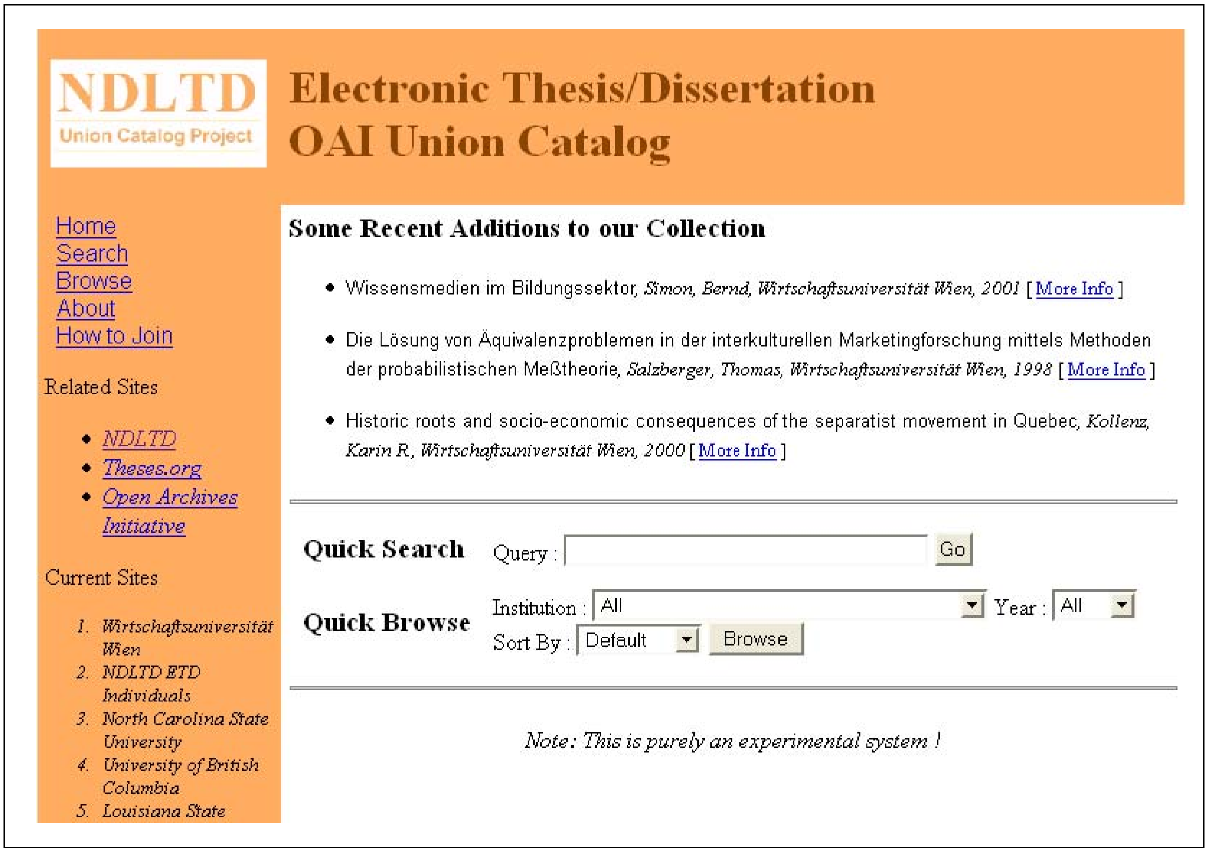

Figure 1. An example of a WWW-based system using a component-based service architecture and XSLT transformations to render metadata in HTML

Portal technology offers the added benefit of a component model for WWW-based user interfaces. The uPortal project [27] defines "channels" to correspond to rectangular portions of the user interface. Each of these channels has functionality that is tied to a particular service on a remote server. This greatly aids development, maintenance, and personalization of the interface.

Some collections of digital objects require interfaces that are specific to the subject domain and nature of the data. Geospatial data, in particular, has the characteristic that users browse by physical proximity in a 2-dimensional space. The Alexandria Digital Earth Prototype [28] allows users to select a geographical region to use as a search constraint when locating digital objects related to that region [29]. Terraserver offers a similar interface to locate and navigate through aerial photographs that are stitched together to give users the impression of a continuous snapshot of the terrain [30]. Both systems offer users the ability to switch between keyword searching and map browsing, where the former can be used for gross estimation and the latter to locate an exact area or feature.

In a different context, multifaceted data can be visualized using 2- and 3-dimensional discovery interfaces, where different facets are mapped to dimensions of the user interface. As a simple example, the horizontal axis is frequently used to indicate year. The Envision interface expands on this notion by mapping different aspects of a data collection or sub-collection to shape, size, and color, in addition to X and Y dimensions [31-33]. Thus, multiple aspects of the data may be seen simultaneously. The Spire project analyzes and transforms a data collection so that similar concepts are physically near each other, thus creating an abstract but easily understandable model of the data [34]. Virtual reality devices can be used to add a third dimension to the visualization. In addition to representing data, collaborative workspaces in virtual worlds can support shared discovery of information in complex spaces [35].

In order to locate audio data such as music, it is sometimes desirable to search by specifying the tune rather than its metadata. Hu and Dannenberg provide an overview of techniques involving such sung queries

[36]. Typically, a user hums a tune into the microphone and the digitized version of that tune is then used as input to a search engine. The results of the search can be either the original audio rendering of the tune or other associated information. In this as well as the other cases mentioned above, it is essential that user needs are met, and that usability is assured [37], along with efficiency.

4. Architecture

Pivotal to digital libraries are software systems that support them; these manage the storage and access to information. To-date, many digital library systems have been constructed, some by loosely connecting applicable and available tools, some by extending existing systems that supported library catalogs and library automation [38]. Most systems are built by following a typical software engineering lifecycle, with an increasing emphasis on architectural models and components to support the process.

Kahn and Wilensky [39] specified a framework for naming digital objects and accessing them through a machine interface. This Repository Access Protocol (RAP) provides an abstract model for the services needed in order to add, modify, or delete records stored in a digital library. Dienst [40, 41] is a distributed digital library based on the RAP model, used initially as the underlying software for the Networked Computer Science Technical Reference Library (NCSTRL) [42]. Multiple services are provided as separate modules, communicating using well-defined protocols both within a single system and among remote systems. A recent approach to supporting a repository is through the DSpace software platform developed at MIT [43].

Other notable pre-packaged systems are E-Prints [44] from the University of Southampton and Greenstone [45] from the University of Waikato. Both provide the ability for users to manage and access collections of digital objects.

Software agents and mobile agents have been applied to digital libraries to mediate with one or more systems on behalf of a user, resulting in an analog to a distributed digital library. In the University of Michigan Digital Library Project [46], DLs were designed as collections of autonomous agents that used protocol-level negotiation to perform collaborative tasks. The Stanford InfoBus project [47-49] not only worked on standards for searching distributed collections [50-52], but also developed an approach for interconnecting systems using distinct protocols for each purpose, with CORBA as the transport layer. Subsequently, CORBA was used as a common layer in the FEDORA project [53], which defined abstract interfaces to structured digital objects.

The myriad of different systems and system architectures has historically been a stumbling block for interoperability attempts [54, 55]. The Open Archives Initiative (OAI) [56], which emerged in 1999 [57], addressed this problem by developing the Protocol for Metadata Harvesting (PMH) [58, 59], a standard mechanism for digital libraries to exchange metadata on a periodic basis. This allows communication involving holders of collections ("data providers") so that the metadata describing the collections can be shared (used by "service providers"). This protocol is widely supported by many current digital library systems.

The Open Digital Library (ODL) framework [60, 61] attempts to unify architecture with interoperability in order to support the construction of componentized digital libraries. ODL builds on the work of the OAI by requiring that every component support an extended version of the PMH. This standardizes the basic communications mechanism by building on the well-understood semantics of the OAI-PMH. The model for a typical ODL-based digital library is illustrated in Figure 2. In this system, data is collected from numerous sources using the OAI-PMH, merged together into a single collection, and subsequently fed into components that support specific services, such as searching. Other efforts have arisen that take up a similar theme, often viewing DLs from a services perspective [62].

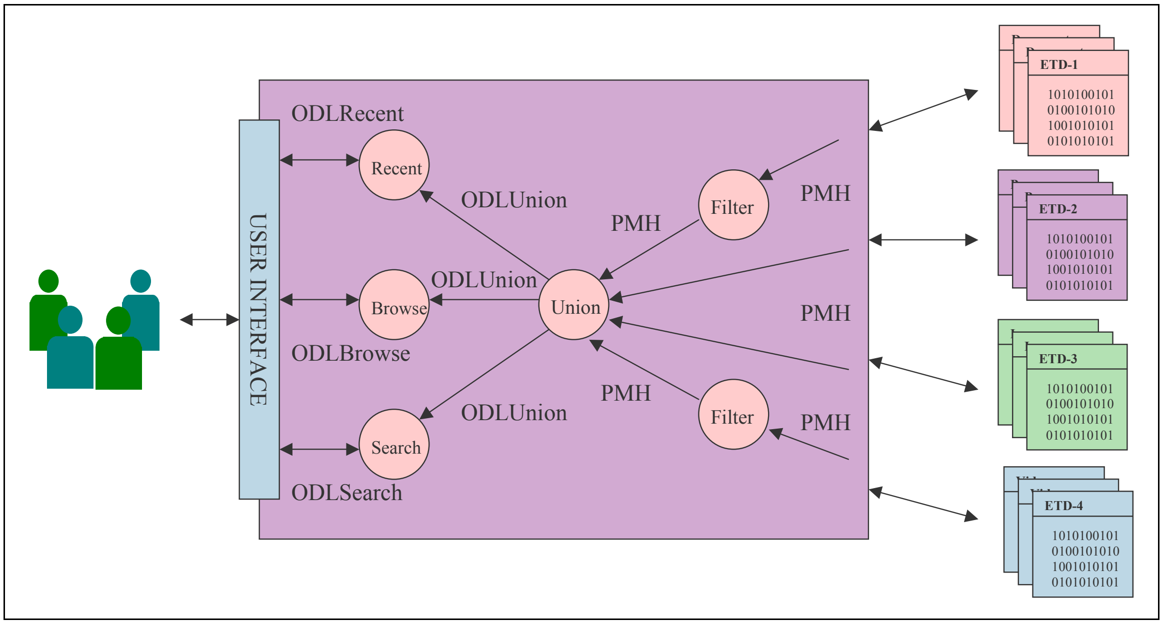

Figure 2. Architecture of digital library based on OAI and ODL components

Preservation of data is being addressed in the Lots of Copies Keeps Stuff Safe (LOCKSS) project [63] which uses transparent mirroring of popular content to localize access and enhance confidence in the availability of the resources. The Internet2 Distributed Storage Initiative [64, 65] has somewhat similar goals and uses network-level redirection to distribute the request load to mirrors that are more accessible [66].

5. Inception

The concept behind digital libraries has its roots in libraries disseminating 'knowledge for all' [67]. Digital libraries break the barrier of physical boundaries and strive to give access to information across varied domains and communities. Though the terms 'Digital Library' and 'Web' both were initially popularized in the early 1990s, they trace back to projects dealing with linking among distributed systems [68], automated storage and retrieval of information [69, 70], library networks, and online resource sharing efforts. Though similar, and mutually supportive in concept and practice, 'Digital Library' and 'Web' differ in emphasis, with the former more focused on quality and organization, and packaged to suit particular sets of users desiring specialized content and services. Accordingly, many digital library projects have helped clarify theory and practice, and must be considered as case studies that illustrate key ideas and developments.

5.1 Digital Library Initiative

The core projects of the US Digital Library Initiative (DLI) Phase I (http://www.dli2.nsf.gov/dlione), started in 1994 as a joint initiative of the National Science Foundation (NSF), Department of Defense Advanced Research Projects Agency (DARPA), and the National Aeronautics and Space Administration (NASA). Phase I involved a total funding of US $24 million for a period of four years from 1994 to 1998. The intent in the first phase was to concentrate on the *investigation and development of underlying technologies for digital libraries*. The Initiative targeted research on information storage, searching, and access. The goals were set as developing technologies related to:

- Capturing, categorizing, and organizing information
- Searching, browsing, filtering, summarizing, and visualization
- Networking protocols and standards

The six projects undertaken in the DLI (1994-1998) were:

- Carnegie Mellon University Project – Informedia Digital Video Library
- Stanford University—Interoperation mechanisms among heterogeneous services
- University of California at Berkeley — Environmental Planning Library and Geographic Information Systems
- University of California at Santa Barbara – Alexandria Digital Library project on spatially-referenced map information
- University of Illinois at Urbana-Champaign – Federating repositories of scientific literature
- University of Michigan Digital Library Project (UMDL) – Intelligent agents for information location

DLI brought focus and direction to developments in the digital libraries arena. Various architectures, models, and practices emerged and precipitated further research. The National Science Foundation announced Phase II in February 1998. In addition to the NSF, the Library of Congress, the Defense Advanced Research Projects Agency (DARPA), the National Library of Medicine (NLM), the National Aeronautics and Space Administration (NASA), and the National Endowment for the Humanities (NEH) served as sponsors. The second phase (1999-2004) went past an emphasis on technologies to focus on applying those technologies and others in real life library situations.

The second phase aimed at intensive study of the architecture and usability issues of digital libraries including research on: a) human-centered DL architecture; b) content and collections-based DL architecture; and c) systems-centered DL architecture.

During this period, many test beds were developed, including at the following universities:

- Carnegie Mellon University (Million books project)
- Columbia University (A Patient Care Digital Library: Personalized Search and Summarization over Multimedia Information)
- Harvard University (Operational Social Science Digital Data Library)
- Stanford University (Stanford Digital Library Technologies Project)
- Tufts University (The Perseus Digital Library Project)
- University of Arizona (High-Performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management)
- University of California Berkeley (Re-inventing Scholarly Information Dissemination and Use)
- University of California Santa Barbara (Alexandria Digital Earth Prototype)

5.2 Networked Digital Libraries

Many DL projects have emerged in the Web environment, where content and users are distributed; some are significant in terms of their collections, techniques, and architecture. For example, NSF partially funded the Networked Computer Science Technical Reference Library (NCSTRL, http://www.ncstrl.org), a digital repository of technical reports and related works. By 2001, however, the Dienst services and software used by NCSTRL no longer fit with needs and practices, so a transition began toward the model advocated by the Open Archives Initiative [57, 59]. OAI ushered in a simple and distributed model for exchange of records. OAI also is used in the Networked Digital library of Theses and Dissertation (NDLTD) [11, 71-73]. NDLTD has over 170 members from over 25 countries sharing electronic theses and dissertations. Eventually this promises to become one of the world's largest digital libraries, with the potential of over 200,000 multilingual hypermedia works being added each year by students to a network of local repositories at universities around the globe. Figure 3 illustrates the typical services offered at one such repository within the larger federation.

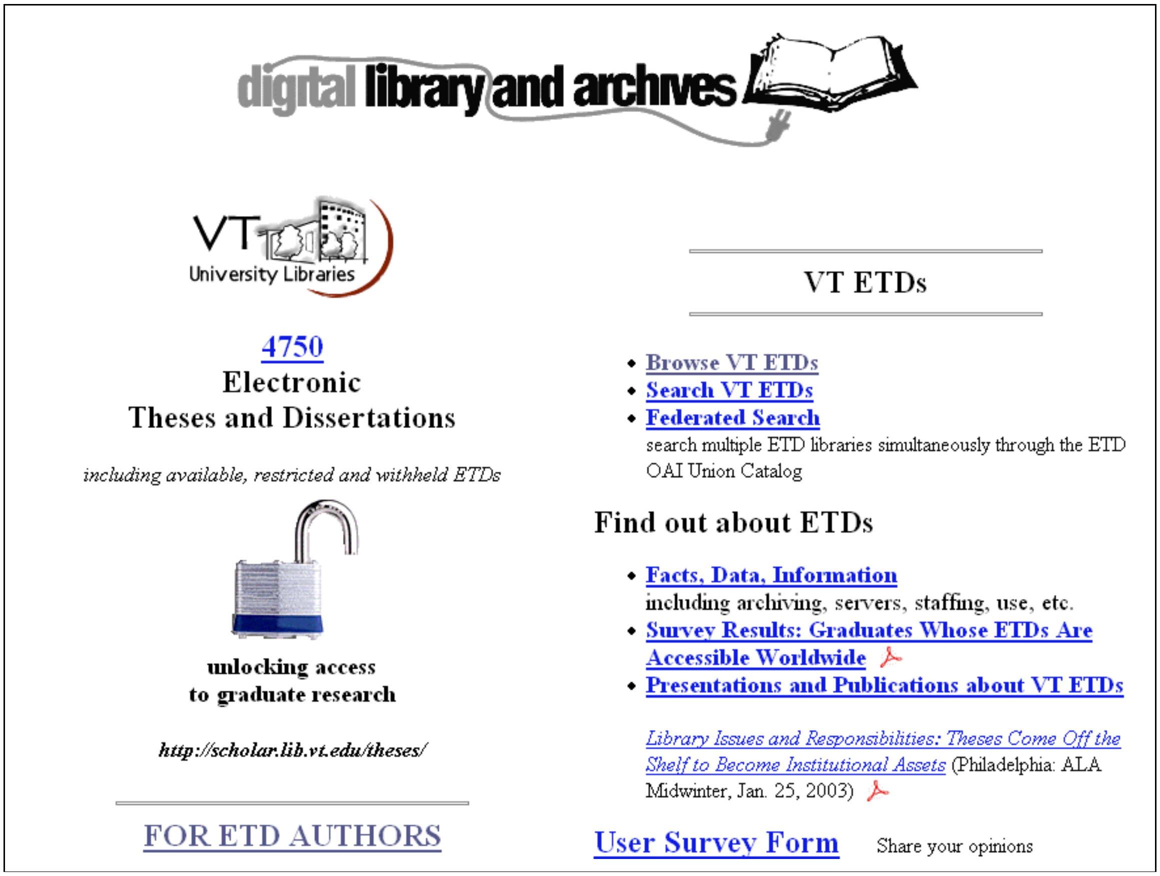

Figure 3. Entry page to the Virginia Tech ETD digital library

Colleges and universities, along with diverse partners interested in education, also are working on a distributed infrastructure for courseware. Building upon DLI, NSF initiated the **N**ational **S**cience, technology, engineering, and mathematics education **D**igital **L**ibrary (NSDL) for the benefit of educators and learners [16-20]. NSDL, already involving over 100 project teams, is projected to have a great impact on education, with the objective of facilitating enhanced communication between and among educators and learners. The basic objective of NSDL is to "catalyze and support continual improvements in the quality of Science, Mathematics, Engineering and Technology education" [74].

5.3 Global DL scenario

Since the early 1990s, work on digital libraries has unfolded all around the globe [75], with many heads of state interested in deploying them to preserve and disseminate the cultural and historic record [76]. There has been some support for research, but more support for development and application, often as extensions to traditional library and publishing efforts.

In Europe there is an annual digital library conference (ECDL), and there have been projects at regional, national, and local levels. The Telematics for Libraries program of the European Commission (EC) aims to facilitate access to knowledge held in libraries throughout the European Union while reducing disparities between national systems and practices. Though not exclusively devoted to digital libraries - the program covers topics such as networking (OSI, Web), cataloging, imaging, multimedia, and copyright - many of the more than 100 projects do cover issues and activities related to digital libraries [77]. In addition there have emerged national digital library initiatives in Denmark, France, Germany, Russia, Spain, and Sweden, among others.

In the UK, noteworthy efforts in Digital libraries include the ELINOR and the elib projects. The Electronic Libraries Programme (eLib, http://www.jisc.ac.uk/elib/projects.html), funded by the Joint Information

Systems Committee (JISC), aims to provide exemplars of good practice and models for well organized, accessible hybrid libraries. The Ariadne magazine (http://www.ariadne.ac.uk/) reports on progress and developments within the Programme and beyond.

The Canadian National Library hosts the Canadian Inventory of Digital Initiatives which provides descriptions of Canadian information resources created for the Web, including general digital collections, resources centered around a particular theme, and reference sources and databases. In Australia, libraries (at the Federal, State, and University levels) together with commercial and research organizations are supporting a diverse set of digital library projects that take on many technical and related issues. The projects deal both with collection building and with services and research, especially related to metadata. Related to this, and focused on retrieval, are the subject gateway projects (http://www.nla.gov.au/initiatives/sg/) which were precursors to the formal DL initiatives.

In Asia the International Conference of Asian Digital Libraries (ICADL, http://www.icadl.org) provides a forum to publish and discuss issues regarding research and developments in the area of digital libraries. In India, awareness of the importance of digital libraries and electronic information services has led to conferences and seminars hosted on these topics. While a national policy on digital libraries is still pending, a number of individual digital library efforts have emerged. In this context, several digital library teams are collaborating with the Carnegie Mellon University Universal Digital Library project. The University of Mysore and University of Hyderabad are among those participating as members in the Networked Digital Library of Theses and Dissertations. In the area of digital library research, Documentation Research and Training (DRTC, www.drtc.isibang.ac.in), at the Indian Statistical Institute, researches and implements the technology and methodologies in digital library architecture, multilingual digital information retrieval, and related tools and techniques. In addition, DRTC also hosts workshops to provide training for information professionals. Other digital library initiatives in Asia are taking shape through national initiatives such as the Indonesian Digital Library Network (http://idln.lib.itb.ac.id), the Malaysian National Digital library (myLib, http://www.mylib.com.my), and the National Digital Library of Korea (http://www.dlibrary.go.kr).

6. Personalization and privacy

Digital libraries allow us to move from the global to the personal. Personalization [10] allows the DL to recognize a returning user and to restore the state of the user relationship with the library to where it was at the time of the last visit. It saves the user time in reconstructing prior work and allows saving the state of the user-DL interaction. It also allows the system to know user preferences and to tailor services to special needs or simple choices. Personalization depends on user information, generally in the form of a user profile and history of prior use. The user profile to some extent identifies an individual and allows the system to recognize when a user returns. In addition to making it possible for the library to provide services, the identification of a user allows evaluation of how well the library is serving that user. If a given user returns frequently, seems to find what was wanted, uses available services, keeps a supply of materials available for later use, and participates in user options such as annotation and discussions, it is reasonable to assume that the user is well served. Thus, an analysis of user characteristics and activities [78] can help determine if the library is serving its intended audience adequately.

There also is a negative side to personalization. Many people are increasingly conscious of diminished privacy, and anxious about sharing data about their personal preferences and contact information. The concerns are real and reasonable and must be addressed in the design of the DL. Privacy statements and a clear commitment to use the information only in the service of the user and for evaluation of the DL can alleviate some of these concerns. Confidence can be enhanced if the information requested is limited to what is actually needed to provide services and if the role of the requested information is clearly explained. For example, asking for an e-mail address is understandable if the user is signing up for a notification service. Similarly, a unique identifier, not necessarily traceable to any particular individual, is necessary to retain state from one visit to another.

With the increasing numbers of digital libraries, repeated entry of user profile data becomes cumbersome. We argue for one way to address these issues - have the users' private profile information kept on their own systems. The user will be recognized at the library because of a unique identifier, but no other information is retained at the library site. In this way, the library can track returns and successes in meeting user needs, and could even accumulate resources that belong to this user. All personal details, however, remain on the user system and under user control. This can include search histories, resource collections, project results such as concept maps, and syllabi. With the growing size of disk storage on personal computers, storing these on the user's system is not a problem. The challenge is to allow the DL to restore state when the user returns.

7. Conclusions

Digital libraries afford many advantages in today's information infrastructure. Technology has enabled diverse distributed collections of content to become integrated at the metadata and/or content levels, for widespread use through powerful interfaces that will become increasingly personalized. Standards, advanced technology, and powerful systems can support a wide variety of types of users, providing a broad range of tailored services, for communities around the globe. Varied architectures have been explored, but approaches like those developed in the Open Archives Initiative, or its extension into Open Digital Libraries, show particular promise. While many challenges remain – such as integration with traditional library collections [79], handling the needs for multilingual access, and long-term preservation – a large research establishment is well connected with development efforts, which should ensure that digital libraries will help carry the traditional library world forward to expand its scope and impact, supporting research, education, and associated endeavors.

References

[1] E. A. Fox and S. Urs, "Digital Libraries," in *Annual Review of Information Science and Technology*, vol. 36, Ch. 12, B. Cronin, Ed., 2002, pp. 503-589.

[2] C. L. Borgman, "What are digital libraries? Competing visions," *Information Processing and Management*, vol. 35, pp. 227-243, 1999.

[3] E. A. Fox and O. Sornil, "Digital Libraries," in *Modern Information Retrieval*, R. Baeza-Yates and B. Ribeiro-Neto, Eds. Harlow, England: ACM Press / Addison-Wesley-Longman, 1999, pp. 415-432, ch. 15.

[4] Dublin-Core-Community, "Dublin Core Metadata Initiative: The Dublin Core: A Simple Content Description Model for Electronic Resources". WWW site. Dublin, Ohio: OCLC, 1999. http://purl.org/dc/

[5] Dublin-Core-Community, "Dublin Core Metadata Element Set". Web pages. Dublin, Ohio: OCLC, 2002. http://dublincore.org/documents/dces/

[6] C. Borgman, "Social Aspects of Digital Libraries," UCLA, Los Angeles, NSF Workshop Report, Feb. 16-17, 1996. http://is.gseis.ucla.edu/research/dl/index.html

[7] IPR-Commission, "IPR Report 2002". Online report. 2002. http://www.iprcommission.org/papers/text/final_report/reportwebfinal.htm

[8] ProjectRoMEO, "Project RoMEO, JISC project 2002-2003: Rights MEtadata for Open Archiving". Web site. UK: Loughborough University, 2002. http://www.lboro.ac.uk/departments/ls/disresearch/romeo/index.html

[9] TRIPS, "TRIPS: Agreement on Trade-Related Aspects of Intellectual Property Rights". Web pages. Geneva, Switzerland: World Trade Organization, 2003. http://www.wto.org/english/tratop_e/trips_e/t_agm1_e.htm

[10] M. A. Gonçalves, A. A. Zafer, N. Ramakrishnan, and E. A. Fox, "Modeling and Building Personalized Digital Libraries with PIPE and 5SL," in *Proceedings of the Joint DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries, June18-20, 2001*. Dublin, Ireland: DELOS, 2001. http://www.ercim.org/publication/ws-proceedings/DelNoe02/Goncalves.pdf

[11] E. A. Fox, "The 5S Framework for Digital Libraries and Two Case Studies: NDLTD and CSTC," in *Proceedings NIT'99, The 11th International Conf. on New Information Technology, Taipei, Taiwan, Aug. 18*. Taipei, Taiwan, 1999. http://www.ndltd.org/pubs/nit99fox.doc

[12] M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp, "Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries," Computer Science, Virginia Tech, Blacksburg, VA, Technical Report, TR-03-04, 2003. http://eprints.cs.vt.edu:8000/archive/00000646/

[13] Q. Zhu, "5Sgraph: A Visual Modeling Tool for Digital Libraries," Virginia Tech Computer Science, Blacksburg, Masters, 2002. http://scholar.lib.vt.edu/theses/available/etd-11272002-210531/

[14] M. Gonçalves and E. Fox, A., "5SL - A Language for Declarative Specification and Generation of Digital Libraries," in *Proc. JCDL'2002, Second ACM / IEEE-CS Joint Conference on Digital Libraries, July 14-18*. Portland, Oregon, USA, 2002, pp. 263-272.

[15] M. A. Gonçalves, G. Panchanathan, U. Ravindranathan, A. Krowne, E. A. Fox, F. Jagodzinski, and L. Cassel., "The XML Log Standard for Digital Libraries: Analysis, Evolution, and Deployment," in *Proc. JCDL'2003, Third ACM / IEEE-CS Joint Conference on Digital Libraries, May 27-31*. Houston, 2003.

[16] NSDL, "NSDL Homepage", *National Science, technology, engineering, and mathematics education Digital Library.* Arlington, VA: NSF, 2002. www.nsdl.org

[17] L. L. Zia, "The NSF National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) Program," *CACM*, vol. 44, 2001.

[18] NSF, "National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL)". Home page with program description, links. Arlington, VA: National Science Foundation, 2003. http://www.ehr.nsf.gov/EHR/DUE/programs/nsdl/

[19] NSF, *National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) - Program Solicitation NSF 00-44*. Arlington, VA: National Science Foundation, 2000. http://www.nsf.gov/cgi-bin/getpub?nsf0044

[20] NSF, *National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) - Program Solicitation NSF 03-530*. Arlington, VA: National Science Foundation, 2003. http://www.nsf.gov/pubsys/ods/getpub.cfm?nsf03530

[21] CITIDEL, "CITIDEL: Computing and Information Technology Interactive Digital Educational Library", D. K. (Edward A.Fox, Lillian Cassel, John A. N. Lee, Manuel Pérez-Quiñones, John Impagliazzo, and C. Lee Giles), Ed. Web site. Blacksburg, VA: Virginia Tech, 2002. http://www.citidel.org

[22] B. Marshall, Y. Zhang, H. Chen, A. Lally, R. Shen, E. A. Fox, and L. N. Cassel, "Convergence of Knowledge Management and E-Learning: the GetSmart Experience.," in *Proc. JCDL'2003, Third ACM / IEEE-CS Joint Conference on Digital Libraries, May 27-31, Houston*, 2003.

[23] E. A. Fox and P. Mather, "Scalable Storage for Digital Libraries," in *Multimedia Information Retrieval and Management*, D. Feng, W. C. Siu, and H. Zhang, Eds.: Springer-Verlag, 2002, pp. chapter 13. http://www.springer.de/cgi/svcat/search_book.pl?isbn=3-540-00244-8

[24] J. Clark, "XSL Transformations Version 1.0, W3C (XSLT),", W3C Recommendation, 1999. http://www.w3.org/TR/xslt

[25] J. Wang, *VIDI: A lightweight protocol between visualization systems and digital libraries*. Blacksburg, VA: Virginia Tech, Department of Computer Science Masters thesis, 2002. http://scholar.lib.vt.edu/theses/available/etd-07012002-145841/

[26] R. Shen, J. Wang, and E. A. Fox, "A Lightweight Protocol between Digital Libraries and Visualization Systems," in *JCDL Workshop on Visual Interfaces to Digital Libraries, Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*. Portland, USA: ACM Press, 2002, pp. 425.

[27] JA-SIG, "uPortal architecture overview". Web site. JA-SIG (The Java in Administration Special Interest Group), 2002. http://mis105.mis.udel.edu/ja-sig/uportal/architecture/uPortal_architecture_overview.pdf

[28] T. R. Smith, G. Janee, J. Frew, and A. Coleman, "The Alexandria digital earth prototype," in*:Proc. First ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2001, 24-28 June*. Roanoke, VA, USA, 2001, pp. 118-199.

[29] G. H. Leazer, A. J. Gilliland-Swetland, and C. L. Borgman, "Evaluating the Use of a Geographic Digital Library in Undergraduate Classrooms: ADEPT," in *Proceedings of the Fifth ACM Conference on Digital Libraries: DL '00, June 2-7, 2000, San Antonio, TX*. New York: ACM Press, 2000, pp. 248-249.

[30] Microsoft, "TerraServer". Web site. Microsoft Corporation, 2002. http://terraserver.microsoft.com
[31] L. Nowell, "Graphical Encoding for Information Visualization: Using Icon Color, Shape and Size to Convey Nominal and Quantitative Data," Virginia Tech Dept. of Computer Science, Blacksburg, VA, Ph.D. Dissertation, 1997. http://scholar.lib.vt.edu/theses/available/etd-111897-163723/
[32] L. Heath, D. Hix, L. Nowell, W. Wake, G. Averboch, and E. A. Fox, "Envision: A User-Centered Database from the Computer Science Literature," *Communications of the ACM*, vol. 38, pp. 52-53, 1995.
[33] J. Wang, A. Agrawal, A. Bazaz, S. Angle, E. A. Fox, and C. North, "Enhancing the ENVISION Interface for Digital Libraries," in *Proc. JCDL'2002 Second ACM / IEEE-CS Joint Conference on Digital Libraries*. Portland, Oregon, USA, 2002, pp. 275-276.
[34] J. Thomas, K. Cook, V. Crow, B. Hetzler, R. May, D. McQuerry, R. McVeety, N. Miller, G. Nakamura , L. Nowell, P. Whitney, and P. Chung Wong, "Human Computer Interaction with Global Information Spaces - Beyond Data Mining," Pacific Northwest National Laboratory, Richland, WA 1998. http://www.pnl.gov/infoviz/papers.html
[35] K. Börner and C. Chen, "Visual Interfaces to Digital Libraries", in *LNCS 2539*: Springer Verlag, 2002.
[36] N. Hu and R. B. Dannenberg, "A Comparison of Melodic Database Retrieval Techniques Using Sung Queries," in *Second ACM/IEEE-CS Joint Conference on Digital Libraries, 14-18 July*. Portland, OR, USA, 2002, pp. 301-307.
[37] R. Kengeri, C. D. Seals, H. D. Harley, H. P. Reddy, and E. A. Fox, "Usability study of digital libraries: ACM, IEEE-CS, NCSTRL, NDLTD," *International Journal on Digital Libraries*, vol. 2, pp. 157-169, 1999. http://link.springer.de/link/service/journals/00799/bibs/9002002/90020157.htm
[38] M. A. Gonçalves, P. Mather, J. Wang, Y. Zhou, M. Luo, R. Richardson, R. Shen, X. Liang, and E. A. Fox, "Java MARIAN: From an OPAC to a Modern Digital Library System," in *9th String Processing and Information retrieval Symposium (SPIRE 2002), September*. Lisbon, Portugal, 2002.
[39] R. Kahn and R. Wilensky, "A Framework for Distributed Digital Object Services". Technical report. Reston, VA: CNRI, 1995. http://www.cnri.reston.va.us/k-w.html
[40] C. Lagoze and J. R. Davis, "Dienst: An Architecture for Distributed Document Libraries," *Communications of the ACM*, vol. 38, pp. 47, 1995.
[41] J. R. Davis, D. Krafft, and C. Lagoze, "Dienst: Building a Production Technical Report Server," in *Advances in Digital Libraries '95*: Springer Verlag, 1995, pp. 211-222.
[42] J. R. Davis and C. Lagoze, "NCSTRL: Design and Deployment of a Globally Distributed Digital Library," *J. American Society for Information Science*, vol. 51, pp. 273-280, 2000.
[43] MIT, "DSpace: Durable Digital Depository". Web site. Cambridge, MA: MIT, 2003. http://dspace.org
[44] EPrints.org, "E-Prints". Web site. 2002. http://www.eprints.org/
[45] I. H. Witten, R. J. McNab, S. J. Boddie, and D. Bainbridge, "Greenstone: A Comprehensive Open-Source Digital Library Software System," in *Proceedings of the Fifth ACM Conference on Digital Libraries: DL '00, June 2-7, 2000, San Antonio, TX*. New York: ACM Press, 2000, pp. 113-121.
[46] W. P. Birmingham, "An Agent-Based Architecture for Digital Libraries," *D-Lib Magazine*, vol. 1, 1995. http://www.dlib.org/dlib/July95/07birmingham.html
[47] M. Roscheisen, M. Baldonado, C. Chang, L. Gravano, S. Ketchpel, and A. Paepcke, "The Stanford InfoBus and Its Service Layers: Augmenting the Internet with Higher-Level Information Management Protocols," in *Digital Libraries in Computer Science: The MeDoc Approach, Lecture Notes in Computer Science, No. 1392,*: Springer, 1998. http://dbpubs.stanford.edu:8090/pub/1998-25
[48] M. Baldonado, C.-C. K. Chang, L. Gravano, and A. Paepcke, "The Stanford Digital Library Metadata Architecture," *International Journal on Digital Libraries*, vol. 1, pp. 108-121, 1997. http://www-diglib.stanford.edu/cgi-bin/-WP/get/SIDL-WP-1996-0051
[49] A. Paepcke, "Using the InfoBus". Web site. Palo Alto: Stanford University Digital Libraries Project, 1999. http://www-diglib.stanford.edu/diglib/pub/userinfo.html
[50] A. Paepcke, R. Brandriff, G. Janee, R. Larson, B. Ludaescher, S. Melnik, and S. Raghavan, "Search Middleware and the Simple Digital Library Interoperability Protocol," *D-Lib Magazine*, vol. 6, 2000. http://www.dlib.org/dlib/march00/paepcke/03paepcke.html

[51] L. Gravano, C.-C. K. Chang, H. Garca-Molina, and A. Paepcke, "STARTS: Stanford Proposal for Internet Meta-Searching," in *Proceedings 1997 ACM SIGMOD Conference*. Tucson, 1997, pp. 207-218.

[52] L. Gravano, C.-C. K. Chang, H. Garcia-Molina, and A. Paepcke, "STARTS: Stanford protocol proposal for Internet retrieval and search," Stanford University, Stanford, Technical Report, SIDL-WP-19960043, August, 1996. http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1996-0043

[53] S. Payette and C. Lagoze, "Flexible and Extensible Digital Object and Repository Architecture," presented at Second European Conference on Research and Advanced Technology for Digital Libraries, Heraklion, Crete, Greece, 1998.

[54] A. Paepcke, C.-C. K. Chang, H. Garcia-Molina, and T. Winograd, "Interoperability for Digital Libraries Worldwide," *Communications of the ACM*, vol. 41, pp. 33-43, 1998.

[55] S. Payette, C. Blanchi, C. Lagoze, and E. A. Overly, "Interoperability for Digital Objects and Repositories: The Cornell/CNRI Experiments," *D-Lib Magazine*, vol. 5, 1999. http://www.dlib.org/dlib/may99/payette/05payette.html

[56] H. Van de Sompel and C. Lagoze, "Open Archives Initiative". WWW site. Ithaca, NY: Cornell University, 2000. http://www.openarchives.org

[57] H. Van de Sompel and C. Lagoze, "The Santa Fe Convention of the Open Archives Initiative," *D-Lib Magazine*, vol. 6, 2000. http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

[58] C. Lagoze, H. Van de Sompel, M. Nelson, and S. Warner, "The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0, Open Archives Initiative,". Technical report. thaca, NY: Cornell University, 2002. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[59] H. Van de Sompel and C. Lagoze, "The Open Archives Initiative Protocol for Metadata Harvesting: Protocol Version 1.0, Document Version 2001-01-21". Technical report. Ithaca, NY: Cornell University, 2001. http://www.openarchives.org/OAI/1.0/openarchivesprotocol.htm

[60] H. Suleman and E. A. Fox, "A Framework for Building Open Digital Libraries," *D-Lib Magazine*, vol. 7, 2001. http://www.dlib.org/dlib/december01/suleman/12suleman.html

[61] H. Suleman and E. A. Fox, "Designing Protocols in Support of Digital Library Componentization," in *Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002, Rome, Italy, September, 16-18 2002*, M. Agosti and C. Thanos, Eds., 2002, pp. 568-582.

[62] D. Castelli and P. P., "OpenDLib: A Digital Library Service System," in *Research and Advanced Technology for Digital Libraries, Proceedings of the 6th European Conference, ECDL 2002, Rome, Italy, September 2002*, 2002, pp. 292-308.

[63] V. Reich and D. S. H. Rosenthal, "LOCKSS: A Permanent Web Publishing and Access System," *D-Lib Magazine*, vol. 7, 2001. http://www.dlib.org/dlib/june01/reich/06reich.html

[64] M. Beck and T. Moore, "The I-2 DSI Project: An Architecture for Internet Content Channels," *Computer Networking and ISDN Systems*, vol. 30, pp. 2141-2148., 1998.

[65] M. Beck, "Internet2 Distributed Storage Infrastructure (I2-DSI) home page": UTK, UNCCH, and Internet2, 2000. http://dsi.internet2.edu

[66] A. Pande, M. Kothapalli, R. Richardson, and E. A. Fox, "Mirroring an OAI archive on the I2-DSI channel," presented at JCDL'2002, Second ACM / IEEE-CS Joint Conference on Digital Libraries, Portland, Oregon, USA, 2002.

[67] H. G. Wells, *World brain*. Garden City, New York: Doubleday, 1938.

[68] D. C. Englebart, "Conceptual framework for the augmentation of man's intellect," in *Vistas in Information Handling*, P. W. H. D. C. W. (Eds), Ed. Washington, D.C: Spartan Books, 1963, pp. 1-20.

[69] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[70] G. Salton, *Automatic Information Organization and Retrieval*. New York: McGraw-Hill, 1968.

[71] E. Fox, "Networked Digital Library of Theses and Dissertations: An International Collaboration Promoting Scholarship," *ICSTI Forum, Quarterly Newsletter of the International Council for Scientific and Technical Information*, vol. 26, pp. 8-9, 1997. http://www.icsti.org/icsti/forum/fo9711.html#ndltd

[72] E. Fox, "NDLTD: Networked Digital Library of Theses and Dissertations". Web site. 1997. http://www.ndltd.org

[73] E. A. Fox, "Networked Digital Library of Theses and Dissertations," in *Proceedings DLW15*. Japan: ULIS, 1999. http://www.ndltd.org/pubs/dlw15.doc

[74] C. A. Manduca, F. P. McMartin, and D. W. Mogk, "Pathways to Progress: Vision and Plans for Developing the NSDL," NSDL March 20, 2001, 2001. http://doclib.comm.nsdlib.org/PathwaysToProgress.pdf (retrieved on 11/16/2002)

[75] C. L. Borgman, *From Gutenberg to the global information infrastructure: Access to information in the networked world*. Cambridge, MA: MIT Press, 2000.

[76] E. Fox, R. Moore, R. Larsen, S. Myaeng, and S. Kim, "Toward a Global Digital Library: Generalizing US-Korea Collaboration on Digital Libraries," *D-Lib Magazine*, vol. 8, 2002. http://www.dlib.org/dlib/october02/fox/10fox.html

[77] T. Kuny, "Digital Library Projects: European Commission Telematics for Libraries Program," *Network Notes*, vol. 46, 1997.

[78] M. A. Gonçalves, G. Panchanathan, U. Ravindranathan, A. Krowne, E. A. Fox, F. Jagodzinski, and L. Cassel, "The XML Log Standard for Digital Libraries: Analysis, Evolution, and Deployment," in *Proc. JCDL'2003, Third ACM / IEEE-CS Joint Conference on Digital Libraries, May 27-31, Houston*, 2003.

[79] B. Wang, "A hybrid system approach for supporting digital libraries," *International Journal on Digital Libraries*, vol. 2, pp. 91-110, 1999.