Building Digital Libraries Made Easy: Toward Open Digital Libraries

Edward A. Fox, Hussein Suleman and Ming Luo

Digital Library Research Laboratory, Virginia Tech
Blacksburg, VA 24060 USA
{fox, hussein, lming}@vt.edu
http://www.dlib.vt.edu

Abstract. Digital libraries (DLs) promote a sharing culture among those who contribute and those who use resources. This same approach works when building Open Digital Libraries (ODLs). Leveraging the intellectual and practical investment made in the Open Archives Initiative through an eXtended Protocol for Metadata Harvesting (XPMH), one can build lightweight protocols to tie together key components that together make up the core of a DL. DL developers in various settings have learned how to apply this framework in a few hours. The ODL approach has been effective with the Computer Science Teaching Center (www.cstc.org), the Networked Digital Library of Theses and Dissertations (www.ndltd.org), and AmericanSouth.org. Hence, to support our Computing and Information Technology Interactive Digital Educational Library (www.citidel.org) and to provide a generic capability for other parts of the US National Science, technology, engineering, and mathematics education Digital Library (www.nsdl.org), we are developing a "DL-in-a-box" toolkit. When lightweight protocols, pools of components, and open standard reference models are combined carefully, as suggested in the OCKHAM discussions, both the DL user and developer communities can benefit from the principle of sharing.

1 Introduction

"Digital libraries" has many definitions and can be viewed from many perspectives [4, 13, 14]. Here we consider it as referring, in different contexts, to two related modern constructs. The first has been called a digital library service system (DLSS, see [6]); it typically is a large, monolithic software package. The second, a type of institution, which is the target of the 5S framework [11], integrates at least: community, services, and content, supported by a DL system. In this paper we focus on the former case, though we are fully aware that the second approach is essential if the DL field to become a science [17].

Our focus is on supporting the large number of people – in libraries, documentation centers, computing centers, and research centers – who deal with the first construct on the way to satisfying requirements implicit in the second type of construct. They face serious problems, so a thoughtful approach is essential.

1.1 Problem

However, instead of building upon the work of others, most DL developers continue to "reinvent the wheel". Why? Here are some of the top reasons given:

1. The library budget won't allow purchase of a commercial DL system.
2. Unless the development effort is local, there won't be any control.
3. DLs are extensions of DBMSs, so they are simple applications to develop.
4. Since DLs operate on the Web, one must adopt the newest W3C proposal.
5. Since technology moves so quickly, it is essential to follow the latest fad.
6. CS students always develop from scratch.
7. This team knows it can do it better.
8. This system must have more capabilities than any other system.
9. This DL has to be more flexible and extensible.
10. This is *the* right system architecture – at last!

Note: Lest we be accused of falling prey to the last myth above, it bears stating that our goal is not to develop the best architecture, but rather one that really is simple and easy to use.

1.2 Approach

Simplicity is the driving principle in our approach. This is reflected in our involvement in the OCKHAM initiative [18, 20, 21] as is discussed further in Section 5 below. It is exemplified by our building upon the work of the Open Archives Initiative [24, 28, 37], with its emphasis on low barriers to entry and support of interoperability.

Interoperability is a key goal in the field of DLs [29]. It has led to many investigations of DL architecture. Thus, at Stanford, the InfoBus is the mechanism (building on CORBA) that allows modules to function cooperatively [1]. On the other hand, at the University of Michigan, an agent approach was employed [2].

Since many DLs are built as distributed systems, the parts of such DLs must be able to communicate with each other. Further, since many DLs are in reality federations of independent DLs, these separate systems must speak a common protocol. Clearly we see that a simple protocol must be an essential element of our approach.

In Section 2 we explain that approach: building Open Digital Libraries. To make the idea concrete, we consider in Section 3 its application in the National Science Digital Library (NSDL). Section 4 gives additional examples of ODL's adoption and use in other applications. Then, Section 5 explains how ODL fits into the OCKHAM activities, while Section 6 concludes this paper.

2 Open Digital Libraries

In [38] we sketched the key ideas of the Open Digital Libraries (ODL) approach, and invited the DL community to comment as well as work with us. We set up a web site to provide current information about the freely available software we have been de-

veloping, along with related documentation and publications [39]. The following subsections summarize the key ideas.

2.1 Definition

The ODL approach calls for lightweight protocols that allow DL development to proceed simply by interconnecting components. Because of the success of the Open Archives Initiative [28], we build upon the OAI Protocol for Metadata Harvesting (PMH, see [41]). In this regard, we adopt a 2 step approach [35].

First, we developed a new protocol that is an extended form of PMH: XOAI-PMH. Since this was undertaken when PMH was at version 1, we were able to argue for these extensions, and some were incorporated in PMH version 2. The other extensions aimed to support general component-component communication inside a DL. In particular, these allow:

- Response-level containers
- Submission (using PutRecord)
- Ignoring the requirement to support DC (inside the DL)

Second, we developed specialized versions of XOAI-PMH for particular types of components. Examples include:

Annotate (with PutRecord to add annotations for items whose ID is supplied using the set parameter)

- Browse (with the set parameter encoding the categories and sort order)
- Rate (with a metadata record encapsulating numerical rating and item ID)
- Search (with the keyword list, query language, and bounds for range of returned results all encoded in the set parameter)

Building on this 2 step approach of protocol extension, we then were able to develop components that satisfied these protocols, and thus allowed key DL functionality to emerge, as is explained in the next subsection.

2.2 Components

From our perspective [14], DLs can be thought of as powerful, high-end information systems that integrate a variety of multimedia, database, information retrieval, and human-computer interface technologies. They encompass creation, discovery, retrieval, and use of information. They support electronic publishing and content management [22]. Thus, a broad range of basic components are needed, and it is essential that they can be composed so that larger and larger systems can be developed. This is possible since an ODL component can be either an OAI data provider, and OAI service provider, or both.

Figures 1-3 illustrate both some of the components developed, and their composition to build a variety of digital libraries. In Figure 1 we see that a small group of individuals, each with a set of suitable XML files, can easily make these available as an OAI data provider. In addition, they can become an OAI service provider, supporting both searching and browsing. All together, this can be thought of as a basic DL.

Figure 2 illustrates a more complex DL. There is one new type of output supported, by a "what's new" service provider. And there are 3 more types of input. One supports harvesting from open archives. The second allows submission of content, such as by authors or data entry personnel. The third, developed by our partners at NCSA, turns a relational database management system into an OAI data provider, which can be filtered first to ensure that only selected information is passed on.

Finally, Figure 3 illustrates fairly rich services. Composition also is illustrated, such as where IRDB-2 supports searching of annotations. Further, both the Recommend and Rate components can be accessed by a single service provider / interface.

For real-life DLs, however, even more complex systems may be needed. Sections 3 and 4 illustrate this point by way of exploring a variety of DL applications.

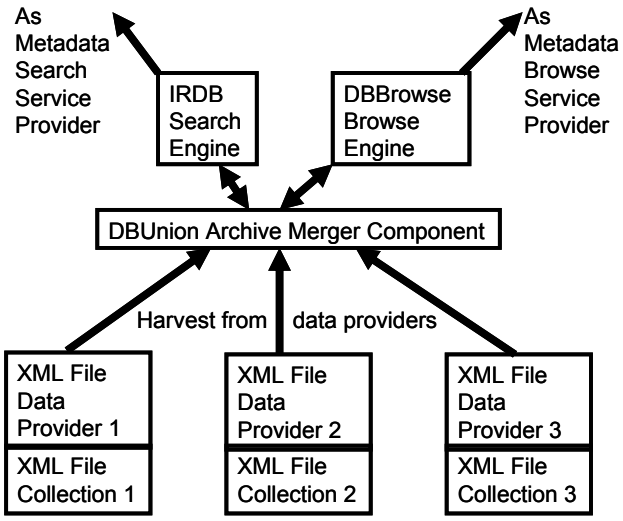

Fig. 1. Simple DL built from 4 basic types of components.

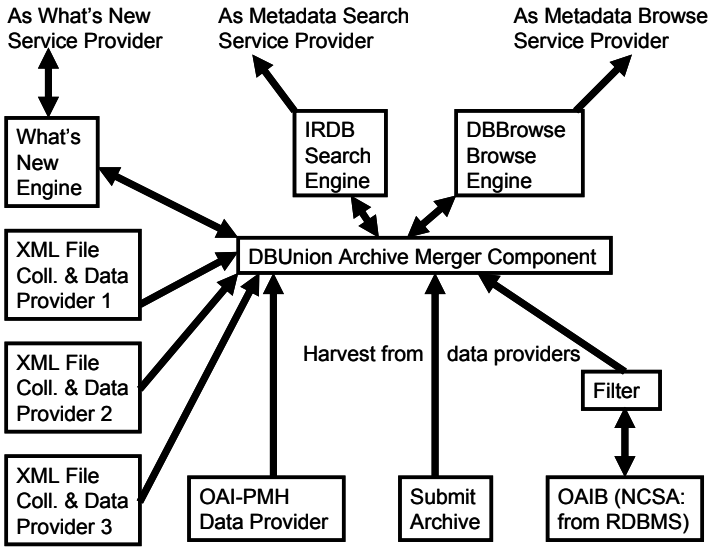

Fig. 2. Intermediate DL built using 9 types of components.

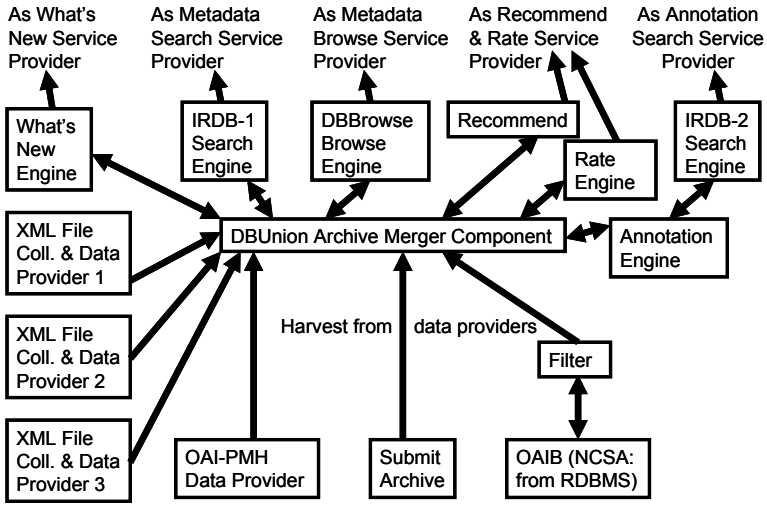

Fig. 3. More complex DL built using 12 types of components.

3 NSDL Applications

One of the largest DL activities currently underway is the National STEM education Digital Library – NSDL for short [31]. Sometimes, for simplicity, "STEM", which stands for Science, Technology, Engineering and Mathematics (replacing the old form, SMET), is expressed as "Science".

NSDL has 4 tracks. One deals with the Core Integration efforts. A second involves support for key Collections. The third focuses on Services. The fourth, and smallest, involves specialized Research, including evaluation.

By the end of 2002, when an initial version of NSDL will be open for first large-scale testing and deployment, there should be about 90 projects that the US NSF is supporting, in the 4 above mentioned tracks. Clearly, interoperability is essential, so there is widespread use of OAI by the Core Integration and Collection projects. However, while metadata can be harvested easily from a wide variety of sources, integrating a diversity of separately developed services is not planned for 2002.

Fortunately, ODL has been tried in a number of educational settings [8]. We believe that it can be effective with regard to integrating both collections and services, as is explained in the next subsections.

3.1 CITIDEL

Virginia Tech has primary responsibility for the CITIDEL part of NSDL [9]. This Collection project covers the topical areas of computing and information technology. Figure 4 explains both collection and services that are under development.

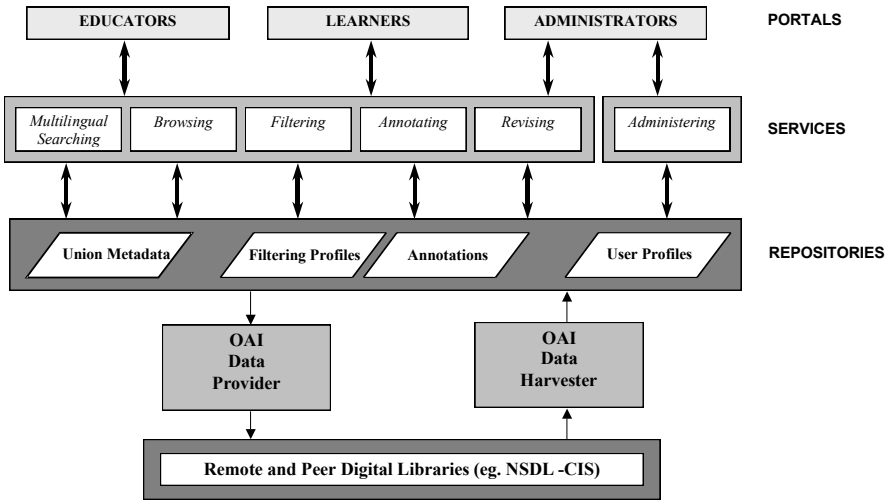

Fig. 4. CITIDEL schematic from original proposal showing collections and services.

Many of the requirements for CITIDEL can be met using existing ODL components, at least for an initial prototype. But when CITIDEL has a union collection with on the order of a million records, and becomes widely used by undergraduate and graduate students, as well as teachers/trainers and other learners (both younger, in public schools) and older (some as lifelong learners), performance may become an issue. Consequently, one of the research activities being explored with regard to ODL is the matter of performance.

Based in part on this experience with CITIDEL, we are working, along with NCSA, on a project awarded to University of Florida, in the NSDL Services track, as is explained in the next subsection.

3.2 DL-in-a-box

As is explained in Section 1.1, the tendency in NSDL is for each newly funded project to start from scratch in developing software. Consequently, there is a real opportunity to reduce overall costs if new projects can instead begin with a basic but extensible digital library. Our web site for such a digital library in a box [26] aims to support such an approach. We are working to provide additional documentation of components and subsystems (compositions of components), as well as to develop additional components. We are open to requirements statements from others, comments on enhancements and extensions, and will provide full support as funding permits. We hope that gradually others will provide components as well, both those engaged in other NSDL activities, as well as those working on other projects, such as the ones discussed in the next section.

4 Other Applications

Digital libraries can be used in many application domains, and can extend traditional approaches [13]. In the subsections below we explore four other types of applications.

4.1 CSTC

The Computer Science Teaching Center [23] has been one of our test domains for DL development over the last 4 years. It covers the full spectrum of services from author submission to support of end-user searching and browsing. In addition, it supports peer review and editorial control, along with notification through email to editors and reviewers. Further, thanks to support from ACM, CSTC is connected with the ACM Journal of Educational Resources in Computing [5]. This means that submissions to CSTC may be considered for JERIC, and then may appear there if editorial concerns are all addressed. This interconnection is further complicated in that both CSTC and JERIC are collections that are part of CITIDEL, in each case with both their metadata and the full text covered. Fortunately, ODL allows modular development, so parts of CSTC have been replaced by components (e.g., Browse) while the rest of the system has stayed as-is. Clearly such incremental testing and development, and such flexible

interconnection, bode well for ODL being deployed in legacy contexts as well as in new situations. A similar situation is considered in the next subsection too.

4.2 NDLTD

The Networked Digital Library of Theses and Dissertations, NDLTD [10], which supports graduate education [8], also has benefited from the ODL approach. This is fortunate, since NDLTD aims to support change [12] and hence must remain agile. The plan in NDLTD is for as many members as possible to become OAI data providers, so metadata can be easily harvested. Already, harvesting into a union catalog [36] occurs, and a set of services is provided (i.e., browse, recent, and search) [33, 34]. In addition, the union catalog feeds into the Virtua DL system developed by VTLS, Inc. Thus, we have services provided both through ODL and through a commercially available monolithic DL – allowing us to undertake scientific comparisons over the next year.

4.3 AmericanSouth.org

While Virginia Tech has lead responsibility in CITIDEL and NDLTD, it only provides technical assistance in the AmericanSouth.org effort, which is led by Emory University [19]. Thus, this activity demonstrates that others can deploy ODL. While we provide assistance, a number of universities around the Southeast, that are willing to employ OAI to make available metadata about local history and culture, can use components to build up their local services as well as their support for interoperability. Further, this effort has in part led to the OCKHAM effort, discussed in the next section.

4.4 Classes

To demonstrate further that ODL can be deployed easily, it was explained to learners, in several class and tutorial settings. These included at library and digital library conferences, as well as in the Virginia Tech course "Information Storage and Retrieval", wherein almost 70 students (in 2 sections, so that each student could work on their own computer):

- Learned about OAI and ODL
- Installed components on their computer
- Configured the components
- Ran the set of components as a small DL

It is clear that all this can take place in less than 3 hours, and that students can both learn a great deal and gain confidence in their understanding of DL practice. But for students focused in this area, it is helpful to set this in a broader context, as explained in the next section.

5 OCKHAM

In the summer of 2002, the Open Community Knowledge Hypermedia Applications & Metadata initiative was launched. A web site was developed [18], and discussion proceeded through a listserv [20]. The concept was disseminated at ECDL'2002 [21] and was well received; feedback suggests that further meetings will gain support.

There are four main ideas:

1. Components
2. Lightweight protocols
3. Open reference models
4. Community perspective and involvement

The first two have already been discussed above, and are the basis for ODL. The other points are explained in the next two subsections.

5.1 Open reference models

While it is clear that components and lightweight protocols are helpful when building DLs, more is needed. In the case of the applications discussed above, the context and assumed general architecture / reference model has been well understood. However, this is not always the case! Fortunately, however, the library and information science world has invested considerable time in preparing open reference models [3, 7, 15, 16, 25, 27, 30, 40]. Of particular interest are architectures like DNER, meetings and work encouraged by UKOLN, and efforts related to the archival community. In short, it is important that development of components and lightweight protocols takes place in a suitable framework, where modularity has been carefully thought through.

5.2 Community perspective and effort

Such frameworks, however, are in turn based on community activity. Thus, fundamentally, OCKHAM depends on effort to achieve consensus by a group with a common aim. Only with a unified perspective can a community develop a reference model that in turn allows efficient and effective development of components and protocols. Fortunately, in cases such as NSDL and NDLTD, years of discussion and prototyping have led to clear understanding by a broad community.

6 Conclusions

As explained above, when the right conditions exist, it is possible to build DLs easily. We argue for the ODL approach, with components and lightweight protocols. Those work best when there is an open reference model, which has arisen to reflect community perspective, and where community effort helps carry the project forward. Yet, we live in a world where other forces also apply. In some cases we have existing subsys-

tems. Thus, for example, in some cases we may want to simply achieve interoperability at the level of interconnecting DL and information visualization systems [42, 43]. Or, we may need to build upon a particular software infrastructure like Web Services [42]. Such situations may occur, and yet the ODL approach may apply, as long as key concepts, and the essential principle of simplicity, are carefully considered.

7 Acknowledgements

Portions of this work were funded by the US National Science Foundation through grants DUE-9752190, 9752408, 0121679, 0121741, and 0136690; and IIS-0002935, 0080748 and 0086227. Among these are subcontracts with original funding to UNC Wilmington, U. of Arizona, and U. of Florida.

Other aspects of this work were funded in part by the Mellon Foundation, through a subcontract, with original funding to SOLINET for AmericanSouth.org.

References

1. Baldonado, M., Chang, C.K., Gravano, L., and Paepcke, A. The Stanford Digital Library Metadata Architecture. International J. on Digital Libraries 1(2): 108-121, 1997.
2. Birmingham, W.P. An Agent-Based Architecture for Digital Libraries. D-Lib Magazine 1(7), July 1995
3. Blinco, K. Modeling Hybrid Information Environments: The Librarian and the Super Model. PowerPoint presentation for 9th MODELS workshop, 13-14 Oct 1999, UKOLN, http://www.ukoln.ac.uk/dlis/models/models9/presentations/kb-m9.ppt
4. Borgman, C.L. What are digital libraries? Competing visions. Information Processing and Management 35: 227-243, 1999.
5. Cassel, L., Fox, E.A. Introducing the ACM Journal of Educational Resources in Computing (JERIC), editor-in-chiefs' introduction. 1(1), March 2001, http://doi.acm.org/10.1145/376697.382399
6. Castelli, D., Pagano, P. OpenDLib: A Digital Library Service System. In "Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002, Proceedings", eds. Maristella Agosti and Constantino Thanos, pp. 292-308.
7. CCSDS (Consultative Committee for Space Data Systems). Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-R-1, Red Book, May 1999, e Model http://www.ccsds.org/documents/p2/CCSDS-650.0-R-1.pdf
8. Fox, E. Advancing Education through Digital Libraries: NSDL, CITIDEL, and NDLTD. In the Proceedings of Digital Library: IT Opportunities and Challenges in the New Millennium, ed. Sun Jiazheng, Beijing, China: Beijing Library Press, July 9-11, 2002, pp. 107-117.
9. Fox, E. CITIDEL: Computing and Information Technology Interactive Digital Educational Library, 2002, http://www.citidel.org
10. Fox, E. NDLTD: Networked Digital Library of Theses and Dissertations, 2002, http://www.ndltd.org
11. Fox, E. The 5S Framework for Digital Libraries and Two Case Studies: NDLTD and CSTC. In Proceedings NIT'99. Taipei, Taiwan, 1999. http://www.ndltd.org/pubs/nit99fox.doc

12. Fox, E., Gonçalves, M., McMillan, G., Eaton, J., Atkins, A., Kipp, N. The Networked Digital Library of Theses and Dissertations: Changes in the University Community. Journal of Computing in Higher Education, 13(2): 3-24, Spring 2002.
13. Fox, E., Marchionini, G. Digital Libraries: Extending Traditional Values. Guest Editors' Introduction to special section on Digital Libraries. Commun. of the ACM, 44(5): 30-32, May 2001, http://doi.acm.org/10.1145/374308.374329
14. Fox, E. and Urs, S. Digital Libraries. In Annual Review of Information Science and Technology (ARIST), v. 36, B. Cronin, Ed.: American Society for Information Science, 2001.
15. Gardner, T. The MIA Logical Architecture: MODELS Information Architecture (MIA) Requirements Analysis Architecture, UKOLN, 1999, http://www.ukoln.ac.uk/dlis/models/requirements/arch/
16. Garrett, J. ISO Archiving Standards – Overview. Last Revised: 29 July 2002. http://ssdoo.gsfc.nasa.gov/nost/isoas
17. Gonçalves, M.A., Fox, E.A. 5SL – A Language for Declarative Specification and Generation of Digital Libraries. In Proc. JCDL'2002, Second Joint ACM / IEEE-CS, Joint Conference on Digital Libraries, July 14-18, 2002, Portland, pp. 263-272.
18. Halbert, M., ed. OCKHAM: Open Community Knowledge Hypermedia Applications & Metadata. 2002. http://ockham.library.emory.edu
19. Halbert, M. D., ed. AmericanSouth.org: A joint project of Emory University and ASERL, sponsored by the Andrew W. Mellon Foundation. 2002. http://AmericanSouth.org
20. Halbert, M. D., ed. Archives of OCKHAM-SYS@LISTSERV.CC.EMORY.EDU: OCKHAM System Framework Listserv, 2002, http://www.listserv.emory.edu/archives/ockham-sys.html
21. Halbert, M. D., Morgan, E. L., Fox, E. A. OCKHAM: Coordinating Digital Library Development with Lightweight Reference Models. Panel at "Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002", Rome, Italy, September 16-18, 2002
22. Hunter, P. "The Management of Content: Universities and the Electronic Publishing Revolution". Ariadne Issue 28, 22-June-2001. http://www.ariadne.ac.uk/issue28/cms/
23. Knox, D., Fox, E.A., Suleman, H., eds. CSTC: Computer Science Teaching Center. 2002. http://www.cstc.org
24. Lagoze, C., and Van de Sompel, H. The Open Archives Initiative: Building a low-barrier interoperability framework. In Proceedings of JCDL 2001, Roanoke VA, June 2001, ACM Press, pp. 54-62.
25. Library of Congress. METS: Metadata Encoding & Transmission Standard, Official Web Site, Last Revised: February 19, 2002, http://www.loc.gov/standards/mets/
26. Luo, Ming. DL-in-a-box: digital library in a box, website. 2002. http://dlbox.nudl.org
27. Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W., and Gupta, A. Collection-Based Persistent Digital Archives - Part 1. D-Lib Magazine, March 2000, 6(3), http://www.dlib.org/dlib/march00/moore/03moore-pt1.html
28. OAI. Open Archive Initiative. 2002. http://www.openarchive.org/.
29. Paepcke, A., Chag, C.-C. K., Garcia-Molina, H., Winograd, T. Interoperability for Digital Libraries Worldwide. Communications of the ACM, vol. 41, pp. 33-43, 1998.
30. Powell, A. JISC Information Environment Architecture, on "DNER Architecture" for the JISC Distributed National Electronic Resource, JISC, 2002 http://www.ukoln.ac.uk/distributed-systems/dner/arch/
31. NSDL. NSDL: The National Science Digital Library. 2002. http://www.nsdl.org
32. Shen, R., Jun Wang, Edward A. Fox. A Lightweight Protocol between Digital Libraries and Visualization Systems. JCDL Workshop on Visual Interfaces to Digital Libraries (see p. 425 of Proc. JCDL 2002), July 18, 2002, Portland.
33. Suleman, H., Atkins, A., Gonçalves, M.A., France, R.K., Fox, E.A., Chachra, V., Crowder, M., Young, J. Networked Digital Library of Theses and Dissertations: Bridging the

Gaps for Global Access - Part 1: Mission and Progress. D-Lib Magazine, 7(9), Sept. 2001, http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html
34. Suleman, H., Atkins, A., Gonçalves, M.A., France, R.K., Fox, E.A., Chachra, V., Crowder, M., Young, J. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 2: Services and Research. D-Lib Magazine, 7(9), Sept. 2001, http://www.dlib.org/dlib/september01/suleman/09suleman-pt2.html
35. Suleman, H., Fox, E.A. 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, Proceedings", eds. Maristella Agosti and Constantino Thanos, pp. 568-582.
36. Suleman, H., Fox, E.A. Towards Universal Accessibility of ETDs: Building the NDLTD Union Archive. In Proc. ETD'2002, BYU, Provo, Utah, May 30 - June 1, 2002, preprint at http://rocky.dlib.vt.edu/~hussein/etd_2002/etd_2002_paper_final.pdf
37. Suleman, H., Fox, E.A. The Open Archives Initiative: Realizing Simple and Effective Digital Library Interoperability. J. Library Automation, 35(1/2):125-145, 2002.
38. Suleman, H., Fox, E.A. A Framework for Building Open Digital Libraries. D-Lib Magazine, 7(12), Dec. 2001, http://www.dlib.org/dlib/december01/suleman/12suleman.html
39. Suleman, H. Open Digital Libraries. 2002. http://oai.dlib.vt.edu/odl/
40. Thibodeau, K. Building the Archives of the Future: Advances in Preserving Electronic Records at the National Archives and Records Administration. D-Lib Magazine February 2001, 7(2), http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html
41. Van de Sompel, H., Lagoze, C. The Open Archives Initiative Protocol for Metadata Harvesting. Open Archives Initiative, 2001. http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html
42. Vasudevan, V. A Web Services Primer, 2001. http://www.xml.com/pub/a/2001/04/04/webservices/index.html
43. Wang, J. VIDI: A Lightweight Protocol Between Visualization Tools and Digital Libraries. Master's Thesis, Virginia Tech, May 2002.