DRTC Workshop on
Digital Libraries: Theory and Practice
March, 2003
DRTC, Bangalore

Paper: G

Design and Implementation of Networked Digital Libraries: Best Practices

Dr.Hussein Suleman
Lecturer

Prof. Edward A. Fox
Virginia Tech, Blacksburg, VA – 240 61, USA
email: *fox@vt.edu*

and

Dr. Devika P. Madalli
Lecturer, Documentation Research and Training Centre
Indian Statistical Institute, Bangalore - 560 059
email: *devika@drtc.isibang.ac.in*

Abstract

This paper introduces the concept of Networked Digital Libraries (NDLs). Discusses the important components of NDLs. Also presents case studies of a few NDLs. Summarizes the best practices for design and implementation effective NDLs.

1. INTRODUCTION

Libraries have come a long way from being just physical collections of books and bound volumes to latest digital libraries which are essentially integrated system of information services. Digital libraries by their content and function are complex systems and require specialized skills for their management. As a discipline, Digital library is interdisciplinary combining theory and practices in library science, computer sciences and information systems, networking among other areas. Digital libraries are meant to serve as a one stop shop for information requirement of the users through personalized services rather than the traditional environment where such integration was not possible due to physical collection being housed at different locations. An extension of the digital library environment is the Networked Digital Library (NDL) where digital libraries with similar content, clientele and services form a network to give integrated services to users at all nodes thus resulting in making the benefits and impact of the digital library multifold. This paper discusses the components of such networked digital library and also presents case studies and summarizes the efforts in certain guidelines that may be followed for the design and implementation of Networked Digital Libraries.

2. COMPONENTS OF NETWORKED DIGITAL LIBRARIES

The concept Networked digital library is intended for resource sharing among digital libraries of similar interests and content. The National Science Digital Library (NSDL) and Networked Digital Library of Theses and Dissertations (NDLTD) are examples of NDLs which have integrated several smaller digital library projects to holistically serve communities across geographical and institutional boundaries.

NDLs are basically combined efforts of a few member institutions or projects bringing together their resources and hence the basic component is the collection resulting from such co-operative efforts. The functional aspect encompasses sharing the work involved for making such a network successful. This component is about the members, policy makers, administration and management. The core comprises of the actual data, formats, hardware solutions, software solutions, interfaces and managing protocols and upgrades as and when required. However it is to be noted that Digital Libraries in themselves represent complex systems. Networked Digital Libraries face additional challenges like maintaining standards in data representation, retrieval and transfer. In addition, they have to gear up to the technological disparities and work with changing network architectures and protocols. We would like to emphasize here that since NDLs are complicated systems and involve all sorts of physical and logical components, there should enough efforts and time dedicated to planning and implementation of NDLs.

We present in the sections that follow a few NDLs with an objective of summarizing the best practices for the design and implementation of NDLs.

3. CASE STUDIES OF DIGITAL LIBRARY PROJECTS

3.1. NSDL

The **N**ational **S**cience, Mathematics, Engineering and Technology (SMET) Education **D**igital **L**ibrary (NSDL) (www.nsdl.org/) is a project initiated by National Science Foundation. A project of significant importance and magnitude, NSDL is projected to have a great impact in education with the objective of facilitating enhanced communication between 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*. (1).

NSDL is a digital library of *"exemplary resource collections and services*, organized in support of science education at all levels. In addition, NSDL also aims at being a *center of innovation* in digital libraries as applied to education, and a *community center* for groups focused on digital-library-enabled science education. NSDL is a comprehensive, online source for science, technology, engineering and mathematics education. The NSDL mission is to both deepen and extend science literacy through access to materials and methods that reveal the nature of the physical universe and the intellectual means by which we discover and understand it. (www.nsdl.org)

3.2. NDLTD

The Networked Digital Library of Theses and Dissertations (NDLTD)(2) is a collaborative effort of universities around the world to promote creating, archiving, distributing and accessing Electronic Theses and Dissertations (ETDs). Since its inception in 1996, over a hundred universities have joined the initiative, underscoring the importance institutions place on training their graduates in the emerging forms of digital publishing and information access. The outreach and training mission of NDLTD is an ongoing project. Recent research has focused on creating a union database that will provide a means to search and retrieve ETDs from the combined collections of NDLTD member institutions. In response to the need for a focused and accessible catalog with a low barrier to participation, NDLTD has adopted a solution that uses the Open Archives Initiative's Metadata Harvesting Protocol (4) to gather metadata in the ETDMS format and then to make it accessible at a central portal. NDLTD project has international members from over a dozen countries sharing electronic theses and dissertations. Eventually this will become one of the world's largest digital libraries, with the potential of 200,000 multilingual hypermedia works being added each year.

3.3. NCSTRL

The Networked Computer Science Technical Reference Library (NCSTRL) is a distributed digital library of technical reports published by computer science departments internationally. Originally, the system was made up of a central site and multiple remote sites either running Dienst (3) or supporting a lightweight FTP-based protocol for metadata transfer (5). Since the introduction of the Open Archives Initiative's (6). Protocol for Metadata Harvesting (OAI-PMH) (4), the old system has been replaced by components adhering to newer interoperability standards. Figure 1 shows the architecture of the current system.

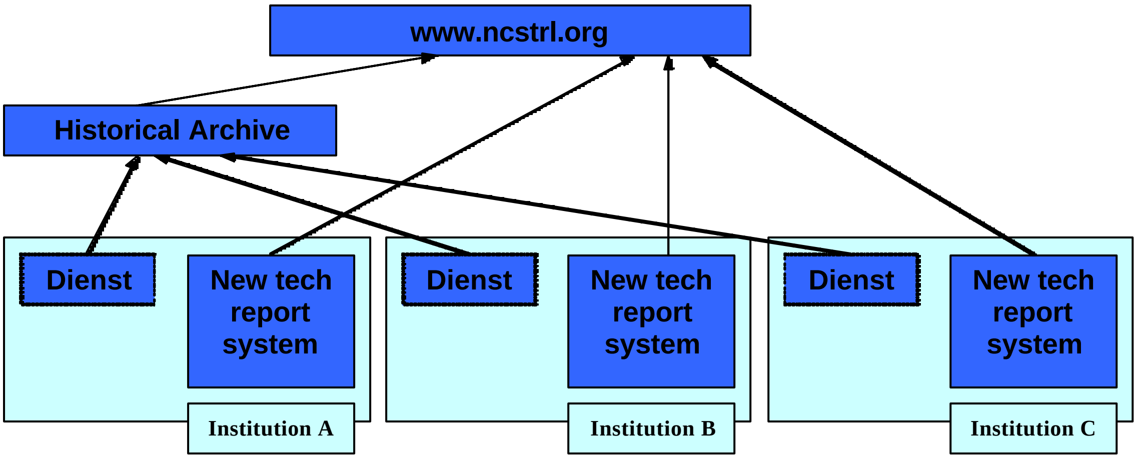

Fig.1: Architecture of NCSTRL

The central *www.ncstrl.org* site provides users with search and browse facilities, using the ARC software (7) over the metadata collected from all partner sites.

The computer science department at each institution runs any OAI-compliant DL software. The central NCSTRL site then harvests metadata from each partner site on a periodic basis using the OAI protocol. In addition, in order to maintain continuity between the old and new systems, a historical archive was set up to store a snapshot of the entire NCSTRL system before the transition. This historical archive stores a copy of all metadata and documents from partner sites, exposing them to the central site using the OAI protocol. The central site then uses data from the historical archive whenever such data is not available directly from each partner site.

4. BEST PRACTICES FOR NETWORKED DIGITAL LIBRARIES

4.1 Systems Planning

4.1.1. System/Network Resources

Individual digital library systems are based loosely on the client-server model where the data and metadata are stored on one or more centrally-located servers which are typically accessed by Web clients. The hardware required to set up such a system is influenced by the DL software, system architecture, projected system use and projected content size. A networked digital library requires hardware resources at each node of the distributed network architecture, as discussed in later sections.

For a single DL instance, a single low-end desktop PC has sufficient computing cycles to cater for the requirements of open source software such as Eprints (8) or Repository-in-a-Box (9). Both of these software packages, among others, allow for multiple installations on a single server so a dedicated machine is not required. On the other hand, some packages, such as the ARC system (7), require the use of higher-end commercial databases, which may impose additional hardware requirements. The lowest common denominator – homegrown software – varies in requirements depending on the development platform.

System architecture depends on the software used and the expected popularity of the system. Dedicated machines for search engines and/or databases are a possibility in the long term and should be planned for if needed. However, the limiting factor in most developing countries is network bandwidth rather then compute cycles so optimizing the use of the network is most critical. Due consideration should be given to packages such as Greenstone (10, 11)[Witten, et al., 2000] [Witten, 2003], which has the ability to work off a standard CDROM with no network connectivity, a feature especially aimed at disseminating information in areas with poor network connectivity.

A 24/7 (T1, DSL, fixed line, leased line, Ethernet) network connection is required for any online system that is available to the general public, especially those aimed at an international audience with users in different timezones. The bandwidth of this connection is often the determining factor but this may be upgraded as needed. The need for stable network capacity also suggests the possibility of outsourcing the web hosting or opting for a co-location solution with a network services vendor. The philosophy of taking the data to the users is also applicable: if the majority of users of a system are in India, a server should be located on a local network; if the majority of users are in Europe, a server should be located in Europe. This will avoid network delays transferring data to/from a developing country with little bandwidth to spare.

4.1.2. Software and OS

The software suite chosen in many instances dictates the operating system (or type of operating system) to be used. If not, there is a compromise between cost and support. Free operating systems, such as Linux and FreeBSD, are easily obtained and have a low acquisition cost but may incur support costs if expertise is not readily available. On the other hand, commercial server vendors such as Microsoft provide an international support network for their products. Service-oriented companies such as RedHat provide support for Linux, thus attempting to provide the advantage of stable service without incurring high software costs.

Beyond the operating system, a DL requires software to manage the data and arbitrate the requests from users. This software can be developed in-house or a pre-packaged system may be used. Popular packages used in production environments include Eprints (8), Greenstone(10) and the newly released DSpace (12, 21) [MIT Libraries-HP, 2003]. All offer the ability to submit items, access the items, manage them internally by means of a review mechanism and various levels of customisation.

4.2. Backups and Replication

Like any production system, a DL should be backed up regularly, preferably off-site. With a server, this can be accomplished through automated procedures. Another alternative to safeguard data is the use of RAID, which is possible without additional hardware on some operating systems.

Replication on a distributed basis is also an option. The LOCKSS project (13) is setting up distributed replicates of critical DL-like Web systems. In this approach, content is preserved at multiple remote locations, which are used to automatically reconstruct the content in the event of a catastrophic failure.

4.3. System Architecture

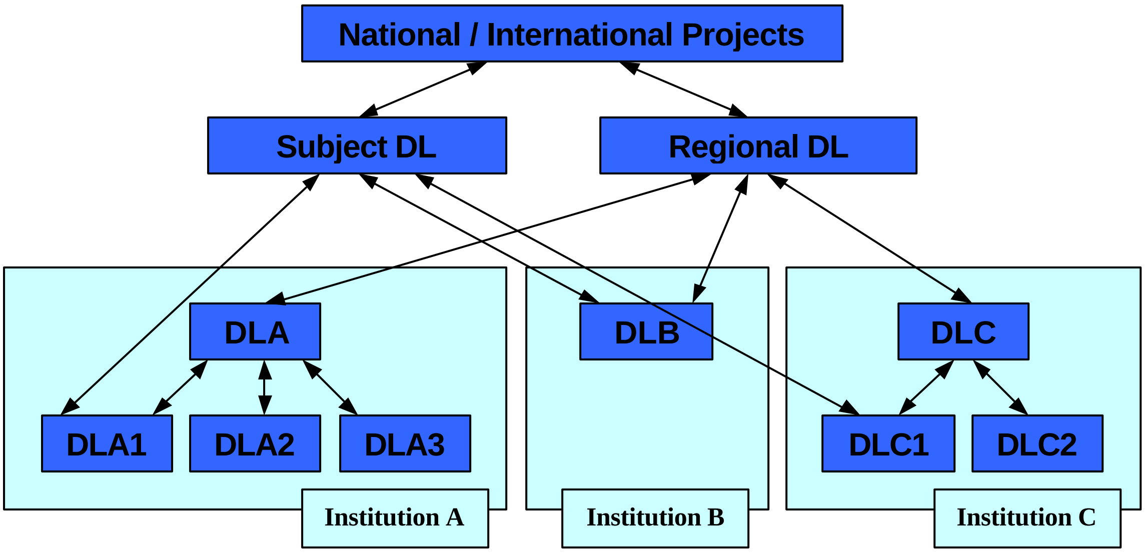

Fig. 2: Generalised architecture of networked digital library

Figure 2 shows a generalised view of the architecture of a typical networked digital library. Data collection and basic services are provided at each institution or organisation, which manages its own DL or DLs. If there are multiple DLs for specific subject areas or subsections of the organisation, these can be merged into institution-wide archives (e.g., DLA). Regional archives can then interoperate with the institution-wide archives of each organisation while subject archives can interoperate with the subject archives (e.g., DLA1) and extract subject-specific data where individual subject archives do not exist (e.g., DLB).

In general, collection of data is done as close to the source as possible, thereby giving the creators of data control over the management of the data. Services, however, are provided at a sufficiently high level of aggregation so that the data is interesting to users. This is the model followed by most current large-scale networked DL projects, including NDLTD, NCSTRL and CITIDEL.

4.4. Protocols

4.4.1. Metadata Harvesting

One of the simplest forms of interoperability among individuals DL systems is the harvesting of metadata. Any DL that wishes to share its data with other systems can implement the Open Archives Initiative's Protocol for Metadata Harvesting (4). This protocol allows a remote system to obtain a copy of all the metadata in a specific format, with incremental updates at discrete intervals. The protocol is designed for simplicity so it can easily be implemented in a custom-built DL (14). In addition, many popular DL (e.g., EPrints, DSpace) and ILS systems (e.g., SIRSE) have built-in support for the OAI protocol.

4.4.2. Remote Search

Interoperability can also be accomplished at the higher level of federated searching, if all parties agree to a common protocol or wrappers can be created to access different remote systems. Library systems traditionally use Z39.50 (19) [ANSI/NISO], while newer DLs prefer simple protocols such as SDLIP (15). The recent ZING (20) project is an attempt to incorporate best practices from Web development, including simplicity, while retaining the essence of Z39.50.

4.5. Component-based Services

As digital libraries become more commonplace, system designers are moving towards the use of standard components to support reusability and repeatability of the process of building digital libraries (16, 17). Emerging component-based DL systems, such as OpenDLib (18) and ODL (19), use well-defined APIs and external Web interfaces to facilitate inter-component interaction. This type of DL allows replacement of individual components, interconnection of DLs and extension of functionality without impacting the existing system. Some service components are available for interconnection into any existing systems solely by means of the OAI-PMH.

5. CONCLUSION

Digital libraries emerged with increased availability of digital information and user demand for services in digital format. They became widely used in the research and academic world with the feasibility of data dissemination speedily over networks. Just as library networks emerged for resource sharing networked digital libraries also share the same aims and objectives. However, the planning and implementation of networked digital libraries poses new challenges and involves policy making regarding the members, content, content management, governance, maintenance and the technical know-how. However in the networked world where more and more information is made available online and through distributed systems, properly implemented NDLs would be very impactful in promoting research and education.

6. REFERENCES

1. Manduca, C. A., etal (2001). Pathways to progress: Vision and plans for developing the NSDL. Retrieved on November 16th 2002, from World Wide Web: http://doclib.comm.nsdlib.org/PathwaysToProgress.pdf
2. Fox, E.A. (1998). Networked digital Lirbary of theses and dissertations (NDLTD) at http://www.nature.com/nature/webmatters/library/library.html retrieved on 02/28/2003
3. Lagoze, C., and J. R. Davis (1995), “Dienst - An Architecture for Distributed Document Libraries”, in *Communications of the ACM*, Vol. 38, No. 4, ACM, p. 47.
4. Lagoze, Carl, Herbert Van de Sompel, Michael Nelson and Simeon Warner (2002), *The Open Archives Initiative Protocol for Metadata Harvesting – Version 2.0,* Open Archives Initiative, June 2002. Available http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
5. Davis, James R., and Carl Lagoze (2000), “NCSTRL: Design and Deployment of a Globally Distributed Digital Library”, in *JASIS*, Vol. 51, No. 3, pp. 273-280.
6. OAI (2003), *Open Archives Initiative*. Website http://www.openarchives.org
7. Liu, Xiaoming, Kurt Maly, Mohammad Zubair, and Michael L. Nelson (2001), “Arc: an OAI service provider for cross-archive searching”, in *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, VA, USA, 24-28 June 2001, pp. 65-66.
8. OpCit (2003), *E-Prints*. Website http://www.eprints.org/
9. NHSE (2003), *Repository-in-a-Box*. Website http://www.nhse.org/RIB/
10. Witten, I. H. (2003) *New Zealand Digital Library*. Website http://www.nzdl.org
11. Witten, I. H., R. J. McNab, S. J. Boddie, and D. Bainbridge (2000), “Greenstone: A Comprehensive Open-Source Digital Library Software System”, in *Proceedings of Fifth ACM Conference of Digital Libraries*, San Antonio, Texas, USA, 2-7 June 2000, pp. 113-121
12. MIT Libraries (2003), *DSpace: Durable Digital Repository*. Website http://dspace.org
13. Reich, Vicky, and David S. H. Rosenthal (2001), “LOCKSS: A Permanent Web Publishing and Access System”, in D-Lib Magazine, Vol. 7, No. 6, June 2001. Available http://www.dlib.org/dlib/june01/reich/06reich.html
14. Dobratz, Susanne, and Birgit Matthaei (2003), “Open Archives Activities and Experiences in Europe: An Overview by the Open Archives Forum”, in D-Lib Magazine, Vol. 9, No. 1, January 2003. Available http://www.dlib.org/dlib/january03/dobratz/01dobratz.html
15. Paepcke, A., R. Brandriff, G. Janee, R. Larson, B. Ludaescher, S. Melnik and S. Raghavan (2000), “Search Middleware and the Simple Digital Library Interoperability Protocol”, in D-Lib Magazine, Vol. 6, No. 3, March 2000. Available http://www.dlib.org/dlib/march00/paepcke/03paepcke.html
16. Gladney, H., Z. Ahmed, R. Ashany, N. J. Belkin, E. A. Fox and M. Zemankova (1994), “Digital Library: Gross Structure and Requirements”, Workshop on On-line Access to Digital Libraries, June 1994.
17. DELOS (2001) Digital Libraries: Future Directions for a European Research Programme, San Cassiano, Alta Badia, Italy, 13-15 June 2001. Available http://delos-noe.iei.pi.cnr.it/activities/researchforum/Brainstorming/brainstorming-report.pdf
18. Castelli, Donatella, and Pasquale Pagano (2002), “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, pp. 292-308.
19. Suleman, Hussein, and Edward A. Fox (2001), “A Framework for Building Open Digital *D-Lib Magazine*, Vol. 7, No. 12, December 2001. Available http://www.dlib.org/dlib/december01/suleman/12suleman.html

20. ANSI/NISO (1995), Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995), Bethesda, MD: NISO Press.
21. Library of Congress (2003), *ZING – Z39.50 International: Next Generation*. Website http://www.loc.gov/z3950/agency/zing/zing-home.html
22. Smith, MacKenzie, Mary Barton, Margret Branschofsky, Greg McClellan, Julie Harford Walker, Mick Bass, Dave Stuve and Robert Tansley (2003), "DSpace: An Open Source Dynamic Digital Repository" in D-Lib Magazine, Vol. 9, No. 1, January 2003. Available http://www.dlib.org/dlib/january03/smith/01smith.html