Design Architecture: An Introduction and Overview

Edward A. Fox
Department of Computer Science
Virginia Tech, Blacksburg, VA 24061 USA

Hussein Suleman
Department of Computer Science
University of Cape Town
Cape Town, SOUTH AFRICA

Ramesh C. Gaur
Tata Institute of Fundamental Research
Mumbai 400 005 INDIA

Devika P. Madalli
Documentation Research and Training Centre
Indian Statistical Institute
Bangalore 560 059, Karnataka, INDIA

Abstract

Digital libraries evolved in response to the need to manage the vast quantities of electronic information that we produce, collect, and consume. Architects of such systems have adopted a variety of design approaches, which are summarized and illustrated in this chapter. We also introduce the following three chapters, and provide suitable background. From a historical perspective, we note that early systems were designed independently to afford services to specific communities. Since then, systems that store and mediate access to information have become commonplace and are scattered all over the Internet. Consequently, information retrieval also has to contend with distributed/networked systems, in a transparent and scalable fashion. In this context, digital library architects have adopted various interoperability standards and practices to provide users with seamless access to highly distributed information sources. This chapter looks at current research and emerging best practices adopted in designing digital libraries, whether individual or distributed.

1. Introduction

Every Digital Library (DL) is constructed according to some design and architecture. These DLs are built upon suitable technology, and must support operations, as well as function as integrated systems, to support a target user community. While there are generic needs common for most DLs, such as searching and browsing, specific communities often require specialized services, and prefer particular types of user interfaces and display formats. In accordance with one of S.R. Ranganathan's Five Laws of Library Science, i.e., "Every Reader His / Her Book", a user-centred design is desired. DLs should be designed so that user information needs are met.

Other DL requirements must be satisfied by suitable DL architectures, with regard to conformance to standards, digital preservation, indexing styles, logging, security, and tuning. These are addressed by DL architectures of individual systems, like those discussed in Section 2 of this chapter. In the case of distributed DLs, discussed in Section 3 of this chapter, there also are requirements for data federation, interoperability, scalability, service federation, and Web Services.

This portion (Part 2) of the book presents an overview of DL design architecture. While this chapter provides an introduction and overview, the remaining chapters focus on particular issues in design and architecture, presenting focused results from the research and development arena. In sequence, they address:

1. The architecture of the Indonesian DL Network
2. Management of metadata
3. An architecture for information filtering and personalization

Chapter 2 describes the Indonesian Digital Library Network (IndonesiaDLN), a nation-wide DL initiative, from inception to its current status. The basic objective of IndonesiaDLN is to collect, manage, share, and reuse the nation's intellectual capital towards the development of a knowledge based society. Early work was guided in part by the evolution of the Networked Digital Library of Theses and Dissertations (NDLTD). The overall initiative is explained using the 5S (Societies, Scenarios, Spaces, Structures, Streams) Framework. One result is the Ganesha Digital Library (GDL) open source software, which has been made available to all member libraries in the network.

Chapter 3 describes dynamic metadata management for digital archives. It discusses the development, features, structure, functions, and use of Metalogy - an XML metadata framework system developed in the context of the Digital Museum Project funded by the National Science Council, Taiwan. Beginning with the basic design concept of a metadata system with the requirement of supporting various metadata formats, a solution based on multi-XML schema is presented in terms of information organization, schema construction, and importing/exporting/conversion of metadata.

Chapter 4 addresses information filtering and personalized services. Content-based filtering addresses the difficult task of delivering relevant information resources to diverse users, through tracking, studying, and representing users' interests. Theoretical and experimental results of the advantages of a probabilistic model over the vector space model are presented. The discussion first looks at representational issues. Then, to improve the accuracy of the collaborative filtering algorithm, a matrix conversion method is proposed. Results are given regarding training set size and improving performance and prediction. A client-server architecture is described, along with prototype personalized searching and recommending services, in the TH-PASS system, which supports management of personal profiles, covering users' interests and bookmarks.

The sections below provide an introduction and overview to DL design and architecture, helping prepare the reader for subsequent chapters in this part of the book.

2. Individual Systems

Before any single digital library (DL) can address the questions of scalability and interoperability with other systems, it has to meet the needs of its local user population. These needs may include submission workflows, peer review, the ability to browse through resources, subscription-based "push" services, or a myriad of other popular DL services. The most obvious, however, is the ability to search through a local collection. Thus, much of this section is devoted to recent ideas in information retrieval, largely from the perspective of a single DL system.

The development and configuration of DL services requires the skills of both computer scientists and library scientists, where the former address technical issues while the latter focus on information seeking and the behavioural aspects of users. There is agreement, however, that the design and architecture of DL services should be user-centred.

DL systems are usually complex, with many components, handling authentication and authorization, user interaction, searching and browsing, retrieval and presentation, analysis and indexing, multimedia management, logging, preservation, link management, and other functions. To reinforce the need for modularity in the system as a whole, we argue next for modularity in a single portion, focusing on query expansion.

Provision of a simple and efficient DL retrieval interface is imperative, though the content may be complex and varied. General purpose search engines incorporating merely technical mechanisms for information search and access cannot guarantee precision. Providing retrieval tools for a collection is more complex, involving mapping user needs and collection

concepts in the proper context and order. Accordingly, facet analysis is regarded as a powerful methodology for the creation of structures appropriate to specific retrieval requirements in a range of contexts (Broughton, 2001). The emphasis is on the problems of complex subject description and representation of multidimensionality in the domain to aid retrieval. While in the traditional environment information organising tools such as library classification systems have proven to be indispensable, the potential of such facet analytical knowledge structures for the management of digital materials has been demonstrated in several systems (Egan, et al., 1989; Allen, 1994). The Subject Based Domain Search System (SBDSS, see Figure 1) (Madalli, et al., 2003), a system for supporting query modification and reformulation using knowledge structures, demonstrates the significance of the subject approach to digital library retrieval. SBDSS accepts a user query and retrieves thesaurus entries that contain at least one of the query keywords at every level of the hierarchy. It ranks and displays the retrieved entries. The input module accepts a natural language query. Stop words are identified and removed. After the standard stemming procedure, a set of keywords remains; these are sent to the lookup module.

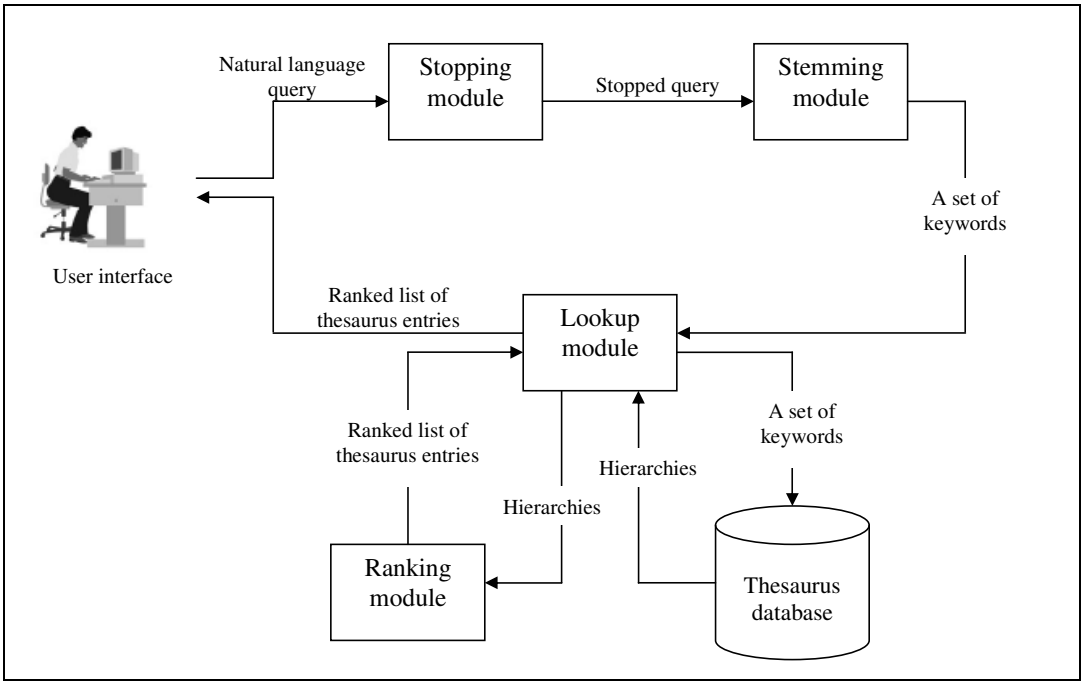

Figure 1. Architecture of SBDSS

The lookup module, augmented with a thesaurus database, searches for the keywords; each resulting context specifies the hierarchies for a term occurrence. The user may interactively choose the term in the context best suited and issue the final search query. The system design (Figure 1) is modular; thus, the thesaurus may be replaced by other formal knowledge structures such as classification schemes.

Shifting to another aspect of the user interface, we note that the structured display and visualisation of result sets is a long-standing topic for retrieval systems (Tudhope & Cunliffe, 2001). Result set displays should carry sufficient semantic information about the retrieved resources; this can be enhanced through the use of metadata, when available. Thus, a display with bibliographic element descriptions, using widely-used standards such as Dublin Core (DC), can add semantic value to retrieval.

Many types of standalone DLs have been built. Some are derived from information retrieval systems. Several evolved from library catalog systems. Other origins include: archive management systems, computer-supported cooperative work systems, database management

systems, directory systems, educational technology or courseware management systems, geographic information systems, hypertext systems, multimedia information systems, and text processing systems. Because of strong current interest, we focus on institutional repositories in the next subsection; we provide examples of other approaches in the following subsection.

2.1. Institutional Archives

The information revolution is leading to a 'flood of publications' or 'information glut'. Users have more journals to read but prices have increased faster than library budgets, so it is becoming difficult for libraries even to maintain their current subscriptions. To balance budgets, libraries have been forced to cancel some journal subscriptions. Yet, scholarly communications are a must for any institution. In part to help cope with this situation, 'institutional repositories' or 'institutional archives' have emerged - representing one of the potentially major components in the evolution of scholarly communications towards digital collections. At the same time, the World Wide Web has opened up a new way for individual scientists and their institutions to preserve and leverage their intellectual assets directly and freely online. So, institutional archives can provide an immediate and valuable complement to the existing scholarly model, and at the same time help institutions in developing their own resource bases and subsequently new areas of resource sharing with other institutions (as discussed in Section 3).

Contents and Services

Institutional archives may contain a wide range of intellectual assets such as pre-prints, working papers, articles, course materials, handouts, theses and dissertations, monographs,

institute journals, standards, reports, meeting proceedings, and notes. Data formats include text documents, research data, and multimedia.

Services are needed to support each of the types of assets, which vary regarding workflow, user community, size and granularity, half-life of use, method of aggregation, support for quality control and refinement, and degree of linking with other types of works. In addition, there are generic services that are popular in any institutional repository, and which tend to distinguish such from general purpose DLs. Such services include registration, (self-) archiving, certification, preservation, and awareness.

Benefits and Problems

Institutional archives are beneficial to all researchers, scholarly institutions, and the entire research community. Major benefits include cost saving, avoiding duplication of effort, broadening of the communication process, reduction in time in announcing findings, expansion of audience, and, above all, preserving information assets for the use of future generations. Since local electronic storage is employed, many large multimedia works may be included, eliminating the page limit constraints of traditional journals. Institutional archives may help an institution to improve its prestige, as well as visibility, worldwide. Thus it may lead indirectly to additional revenues. Institutional archives are especially beneficial in the developing world, which is little represented in the journal literature. Institutions may immediately become visible if their works are indexed by popular search engines. Institutional archives can help in bridging the digital divide and also may help in enriching education, by sharing learning resources among rich and poor nations.

Use of such archives, however, presupposes a pervasive telecommunications infrastructure, so works can be created, uploaded, checked/edited, and later accessed. In developing countries such infrastructure may not exist, or may be very expensive. In such cases,

institutional repositories are not practical, and their lack can lead to further disenfranchising of less wealthy institutions. Fortunately, for some countries (e.g., Portugal), the national library provides such repository and depository services, yielding many of the key benefits, even if small institutions connect only as needed.

2.2. Individual DLs and DL Toolkits: Case Studies

arXiv: http://www.arxiv.org/

arXiv is an electronic pre-print service in the fields of physics, mathematics, non-linear science, computer science, and quantitative biology. arXiv is owned, operated and funded by Cornell University, a private not-for-profit educational institution. arXiv is also partially funded by the National Science Foundation and was formerly supported by the Los Alamos National Laboratory as the XXX service. arXiv has enjoyed wide support from its community of users and the steady or increasing rate of access to articles in arXiv is an indicator of the importance and usefulness of subject-based pre-print archives (Ginsparg, 2003).

CogPrints: http://cogprints.ecs.soton.ac.uk/

CogPrints is an electronic self-archiving service for papers in any area of psychology, neuroscience, and linguistics, and many areas of computer science (e.g., artificial intelligence, robotics, vision, learning, speech, neural networks), philosophy (e.g., mind, language, knowledge, science, logic), biology (e.g., ethology, behavioral ecology, sociobiology, behaviour genetics, evolutionary theory), medicine (e.g., psychiatry, neurology, human genetics, imaging), anthropology (e.g., primatology, cognitive ethnology, archeology, paleontology), as well as any other portions of the physical, social and mathematical sciences

that are pertinent to the study of cognition. Cogprints is based at the University of Southampton.

Eprints: http://www.eprints.org/

Based in University of Southampton, this open source software system supports e-print and pre-print services. It is dedicated to opening access to the refereed research literature online through author/institution self-archiving. At Virginia Tech, it has been used in the Department of Computer Science for local technical reports. In addition, in support of the Networked Digital Library of Theses and Dissertations, it is used to run an open service for archiving of individual electronic theses and dissertations. In a different institutional context, it has been used to run an e-print archive for the Indian Institute of Science, Bangalore (http://eprints.iisc.ernet.in/), that contains research papers, pre-prints, book chapters, technical reports, unpublished findings, conference papers, magazine articles, etc.

DSpace: http://www.dspace.org/

DSpace is an open source software platform that enables institutions to capture and describe digital works using a submission workflow module, distribute an institution's documents over the Web, support search and retrieval, and preserve digital works. Its advent greatly helped expand interest in, and utilization of, institutional repositories.

India: DRTC and other DLs

A specialist Digital Library for Library and Information Science is hosted by the Documentation Research and Training Center, Indian Statistical Institute, (https://drtc.isibang.ac.in/). The DRTC DL is open to worldwide participation for both access

and submission. The collection is structured into seminars and conference proceedings, student theses and dissertations, multilingual resources, and pre-prints. The DRTC DL is powered by DSpace and is compliant with OAI-PMH v2.0 (see Section 3.1). It uses CNRI's Handle server to generate unique URIs for independent access to each resource. As a Unicode implementation, DRTC DL offers multilingual support and hosts resources (and accepts submissions) in Indian and other languages. Besides the main communities of Indian library and information science professionals and academics, the DL has membership from other countries like the UK, USA, and France, among others. The DRTC DL differs from other examples in that it is a broad-based central collection of resources rather than a collection of resources from a broad collection of sites.

There are a variety of other systems of interest in India, including the following non-exhaustive list:

- Developing Libraries Network (DELNET, http://www.delnet.nic.in/): National network of libraries in India. There are several databases related to books, journals, theses, and dissertations.
- Indian Academy of Sciences (IAS) Journals (http://www.ias.ac.in/): 11 journals available online at no charge, covering key scientific results from India and beyond.
- IndMed Database (http://indmed.delhi.nic.in/): It is the first Web-based Indian biomedical database covering 75 Indian Journals. Lead responsibility is with the Indian MEDLARS Centre, New Delhi.
- National Centre of Biodiversity Informatics (http://www.ncbi.org.in/): A range of related efforts dealing with biology, ecology, geography, and related areas.
- National Collection of Industrial Micro-organisms (NCIM, http://www.ncl-india.org/ncim/)

- The Digital Library of India Site (http://*www.dli.gov.in/*): The Indian Institute of Science is the lead player in the University Digital Library Project in association with CMU. The collection developed under this project, in distributed scanning centers of India, has led to the Digital Library of India.

- Indian Institute of Information Technology, Bangalore, (*http://www.iiitb.ac.in/ digital_library.htm*), Hyderabad (*http://www.iiit.net/infrastructure.htm*), and Indian Institute of Technology, Delhi (*http://www.iitd.ernet.in/acad/library/ project.html):* These initiatives provide access to existing international collections (such as those of ACM and IEEE) as well as local collections of books, articles, technical reports, courseware, etc.

- Indian National Science Academy (INSA): INSA, with membership of leading Indian scientists, makes available to scientific organizations and institutions access to its key Digital Library projects: digital records of its fellows and online journals of the academy.

- Network of Automated Library and Archives (Nalanda, *http://www.nalanda.nitc.ac.in/*): The NITC Library has both modernized its traditional collections and began a move towards born-digital resources such as ETDs and electronic databases.

- NCSI, Indian Institute of Science: NCSI has designed, developed and is hosting the SciGate Science Journal Gateway. In addition it is involved with many DL related activities such as beta testing of Greenstone DL software for UNESCO and providing local institutional archiving using the EPrints software (http://*www.eprints.iisc.ernet.in/).*

- University of Hyderabad Central Library: The Indira Gandhi Memorial Library (IGML) of the University of Hyderabad hosts one of biggest Digital Library Projects with external funding from Sun Microsystems and the University Grants Commission (UGC, India) and partnerships with VTLS Inc., USA, Compaq and IBM India. It provides access to key databases like EBSCO and ScienceDirect, a pioneering effort in bringing these services to the Indian academic world. The DL effort has earned recognition as a "Centre of Excellence for Digital Libraries" by Sun Microsystems.

In addition to such individual DLs, there are many distributed systems, as discussed in the following sections.

3. Distributed Digital Libraries

3.1 Federation and Harvesting

While the provision of discovery and management services has largely driven the design of digital libraries, an increasingly important requirement is the need for interoperability with peer systems. Most of the early attempts have focused on the sharing of data among systems, as this replication improved system reliability, reduced latency, and made more effective use of precious network bandwidth by bringing resources closer to users.

FTP mirrors were among the earliest forms of data federation. Simple non-interactive clients can easily obtain a listing of files in a single directory, thereafter fetching the files that have changed and recursing through subsequent directories lower down in the file system hierarchy. Mirroring of websites is not as simple, however, because Web servers provide only an abstract view of the source data – it is quite often the case that two different URLs can access the same website. Website mirroring software therefore must rely on hyperlinks

to “crawl” through a website and produce a mirror of the view constructed by the Web server rather than a mirror of the source data, though, even then, unconnected components and pages can be omitted. This is further exacerbated by the introduction of dynamically-generated websites where timestamps are not valid indicators of change and there is no one-to-one correspondence between URLs and Web pages. Most digital libraries use technology that falls into this category and, as a result, crawling/mirroring is usually ineffective. Thus, an early prototype of the Networked Digital Library of Theses and Dissertations (NDLTD, www.ndltd.org) Union Catalog service provided users with the ability to search and then navigate through related material with a simple hyperlink, which was inadvertently used by a popular search engine, in its attempt to crawl the site automatically, to perform a search with every document as a query. The net effect was similar to a denial-of-service attack and Web crawlers had to be explicitly disallowed in the future. Much effort has been expended in the digital library community on solving precisely this problem: how to efficiently transfer a collection of data from point to point without requiring best-effort heuristics at the destination and without repeated transfer of the same data from the source.

Designers of digital library systems have had to contend with a shift in emphasis of Internet use from file-based systems such as FTP to service-based systems such as the WWW. Early projects such as RePEc (Krichel, 2000) are based on a strong foundation of FTP-able data collections. Modern online information systems, however, exploit the increased capabilities of the WWW to provide advanced services that mediate between users and data. In this context, interoperability can be viewed as a service provided by a system. In keeping with this view, various experiments were conducted in the 90’s, e.g., Cornell/CNRI Experiments (Payette, Blanchi, Lagoze, & Overly, 1999), and Stanford InfoBus (Roscheisen, et al., 1998), to connect together disparate systems. These have shown that it is feasible but not simple to implement, manage, and maintain such configurations over time.

In searching for a simple interoperability solution, the Open Archives Initiative (OAI) was launched at a meeting of representatives of large-scale digital libraries in Santa Fe in October 1999. This loosely-formed alliance had a single purpose in life – to develop a low cost interoperability solution (Lagoze & Van de Sompel, 2001). Initially, the process was driven by the need to exchange metadata about electronic pre-prints but a number of initial discussions and workshops expanded the scope of this mandate well beyond e-prints. In June 2002, after more than 2 years of experimentation, development and testing, the OAI released a stable second version of its Protocol for Metadata Harvesting (PMH) (Lagoze, Van de Sompel, Nelson & Warner, 2002), a simple high-level network protocol to incrementally transfer metadata from one system to another.

The OAI-PMH is a client-server network protocol that facilitates the transfer of metadata from a provider of data to a provider of services. A data provider is defined as that network-accessible entity that owns a collection of metadata which is to be made available to others. A service provider is defined as the entity that obtains metadata from the data provider, usually with the intention of providing user-directed services. This functional split was largely motivated by the realisation that data providers do not often have the best suite of services available and, conversely, that service providers do not often make available the best data sets. In conventional client-server terminology, a data provider is a server while a service provider is a client. As such, a data provider runs a Web application that listens for and processes requests for data. A service provider runs an application, known as a harvester (Web client application that conforms to a harvesting protocol), that periodically obtains new and updated metadata from a data provider. The data thus obtained then either is merged into the local metadata collection or passed on to a service provider component, such as a search engine, for further processing.

In keeping with the notion of simplicity, the OAI-PMH is a stateless protocol. Requests are sent using the HTTP URL encoding that is popular among CGI-based Web applications. Responses are encoded in XML, adhering to best practices in the use of namespaces and XML Schema, so as to be both generalisable and simple to understand/interpret. Data providers must be able to understand and generate responses to a set of 6 service requests, which may be submitted by service providers attempting to harvest metadata. These service requests and their semantics, forming the core of the protocol, are described in Table 1.

Table 1. OAI-PMH service requests and semantics of responses

Service Request	Semantics of Response
Identify	Archive-level information: name, contact details, policies, optional protocol features supported, etc.
ListSets	List of all subsets of the archive that may be harvested selectively.
ListMetadataFormats	List of all metadata formats supported by the archive or that are available for a single item.
ListIdentifiers	List of identifiers of items in the archive or in the subset identified by optional date range and/or set parameters.
GetRecord	Metadata record corresponding to the specified identifier and metadata format.
ListRecords	List of metadata records for all items in the archive or for a subset identified by optional date range and/or set parameters.

The OAI-PMH has proven to be popular because of its unobtrusive nature and loosely-connected mode of operation. Data providers act in an entirely passive manner where they do not do any work until a request arrives, thus concentrating on their primary local functions. Service providers, similarly, do not contact data providers unless actively harvesting data. For greater control, data providers can moderate the flow of data by sending back truncated lists of data along with resumption tokens that service providers may return in order to resume transferring the list (of metadata records, identifiers, etc.). In addition, data providers

may utilise HTTP headers to deny, delay, or postpone requests. In practice, these mechanisms contribute to avoiding denial-of-service attacks and ensure that providers of data are always in full control of the process of metadata harvesting, albeit that it is initiated by service providers.

Ultimately, the simplicity, adherence to standards, clean separation of responsibilities and preservation of autonomy of individual systems rank high on the list of reasons for why existing digital libraries are comfortable with using OAI-PMH in order to interoperate with other systems. Also, new systems that are designed to be highly distributed can readily adopt OAI-PMH, because of the user community and toolsets that have already been and are being developed. Other ongoing OAI work is looking at evaluation of individual design decisions, rights management in the framework of the OAI-PMH, and the application of Web Services standards and practices to the OAI and OAI-PMH.

3.2 Integration of Services

Besides data transfer, a large-scale digital library could conceivably include multiple user interfaces or multiple variants of the same service. Additionally, these interfaces and services could be provided at different physical locations, connected only by a network. In order to present users with a coherent view of the collection of data and a sensible set of services, there is a need for remote access mechanisms for digital library services. The simplest of these is the concept of remote searching, where the machine interface to a search engine can be accessed remotely in addition to, or in place of the human interface, thus allowing integration of remote search engines into local user interfaces and workflows.

The Z39.50 protocol (ANSI/NISO, 1995) was specifically designed for such remote retrieval operations and is popular in library systems. However, it uses older technology (pre-XML)

and is not popular among designers of smaller individual system because of a high degree of complexity. Given that large-scale digital libraries are frequently aggregations of small projects, the added complexity can be problematic. In addition, as the scale of a digital library network increases, remote searching becomes less viable as a basic interoperability mechanism because of the increase in points of failure and an increased possibility of network latency effects. A solution that builds on the best of both worlds is to use data harvesting to create one or more central collections of data, then provide one or more remote search interfaces on these data collections. This is the approach adopted in NDLTD, as illustrated in Figure 2. A single service provider at OCLC harvests data from many remote sites and then exposes an OAI-PMH interface as well as a remote search interface. A second service provider mirrors the metadata and provides yet another remote search interface as well as a user interface that makes use of it.

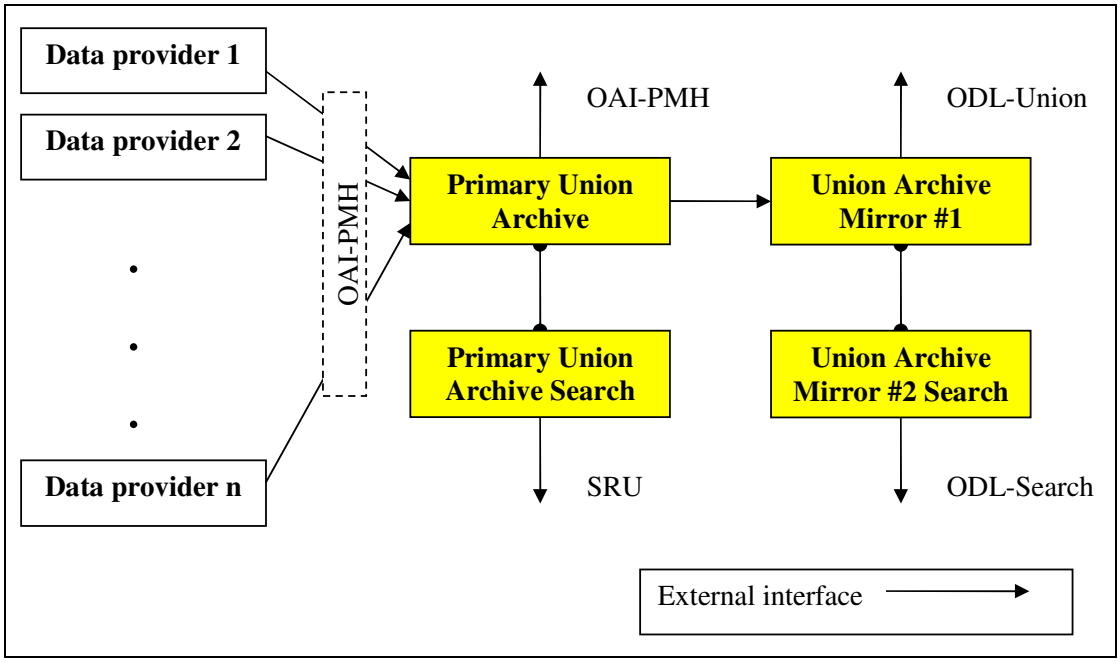

Figure 2. Architecture of NDLTD digital library network

The figure makes reference to SRU, which is part of the "Z39.50 International: Next Generation" (ZING) project (Library of Congress, 2003) to upgrade Z39.50 so that it is

simpler and more accessible as a Web-based service. ODL-Union and ODL-Search are experimental protocols that were developed as part of the Open Digital Library (ODL) project (Suleman & Fox, 2001), where popular services were cast as extensions of the OAI-PMH. Although ZING and ODL have been motivated by different needs, they are both based on the concept of location-independence or federation of digital library services.

3.3 DLs and Web Services

The basic idea behind federated services is, however, not at all specific to digital libraries and is more widely known as the "service-oriented computing" paradigm (Papazoglou & Georgakopoulos, 2003). In this model, the components of a system are analogous to service providers with well-defined machine interfaces. One realisation of this model is the Web Services project of the World Wide Web Consortium (W3C, http://www.w3.org/), which is defining a framework whereby service-oriented components can interact over the WWW.

At the core of this framework is the SOAP protocol (Gudgin, Hadley, Mendelsohn, Moreau, & Nielsen, 2003) that defines how to encapsulate an XML-formatted message for delivery between components of a distributed system. The SOAP specification concentrates on genericity so that the payload, sequence of data transfer, and transport protocols have as much flexibility as possible. For example, SOAP messages may be sent and received by email communication, thus enabling the use of email for directly requesting services from a system.

However, SOAP defines only the encoding and transportation parameters for messages. The syntax and semantics are specified and enforced using different mechanisms. The Web Services Description Language (WSDL) (Christensen, Curbera, Meredith, & Weerawarana, 2001), as the next logical step, defines syntactic elements such as the interfaces and

parameters associated with specific services. The Universal Description, Discovery and Integration of Web Services (UDDI) registries then provide public access to these descriptions of interfaces. A popular example is the Google search engine's free Web Service (Google, 2003), a WSDL definition of which can be discovered through the UDDI registries. This allows any software developer to incorporate machine access to Google using SOAP messages sent over the WWW. Based on this SOAP, WSDL, and UDDI foundation, newer standards such as the Web Services Flow Language (WSFL) (Leymann, 2001) are being developed to coordinate workflows and enable aggregation and composition of service components.

The SRU service, introduced in the previous section, adheres to the Web Service standards and is a prime example of how Web Service technology pervades the networked digital library community. Newer digital library standards for highly connected systems may build on the Web Services initiative – a prototype SOAP-based version of the OAI-PMH has already been developed and tested (Merchant, Gaylord & Congia, 2003) to illustrate the effectiveness and efficiency of Web Services applied to digital library standards. Other standards will likely follow as more Web Services specifications are standardised.

3.4 DLs and Internet Technology

The explicit adoption of Web Services standards by projects such as ZING and implicit adoption of the service-oriented computing architecture by initiatives such as OAI point to an impending convergence between the fields of digital libraries and Internet-based information systems. Just as the Web is becoming a highly-connected system of semantically-rich services, so too are digital libraries becoming a network of service-based components and systems.

The boundaries between digital libraries and other Web-based systems are no longer well-defined. In the domain of content management systems, the popular RSS standard (Winer, 2002) allows for syndication of content such as newsfeeds among dynamically-generated websites. The principles and operation of RSS are very similar to OAI-PMH, except that there are fewer parameters in the former protocol. In contrast, the IMS Digital Repositories specification (IMS Global Learning Consortium, 2003) explicitly avoids overlapping standards and recommends different externally-defined protocols for different scenarios, including the OAI-PMH for metadata harvesting.

From a security perspective, digital library projects may opt to use existing standards defined for rights management and authentication/authorisation. The RoMEO project (Gadd, Oppenheim, & Probets, 2003) studied issues in rights management and recommended use of the existing Open Digital Rights Language (ODRL) and Creative Commons licences for rights specification in an interoperable environment. The Shibboleth project defines a trust relationship framework for authentication and authorisation in a distributed system, and is a prime technology enabler for remote data access and service invocation for digital libraries (Gourley, 2003).

As we move towards larger-scale service-oriented networked information systems, the projects discussed above indicate an increasing degree to which digital libraries rely on emerging Internet standards and vice versa. In the context of this convergence of technologies, it is crucial to design modern distributed digital libraries taking into account current best practices in digital libraries as well as Internet standards to enable the broadest possible spectrum of use cases.

3.5 Distributed DLs: Case Studies

As explained in the last subsection, case studies are helpful to illustrate the key architectures and approaches to distributed digital libraries. Those following provide a representative sample of research and development efforts in the area.

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 (Lagoze and Davis, 1995) or supporting a lightweight FTP-based protocol for metadata transfer (Davis and Lagoze, 2000). Since the introduction of the OAI-PMH, the old system has been replaced by components adhering to newer interoperability standards. Figure 3 shows the architecture of the current system.

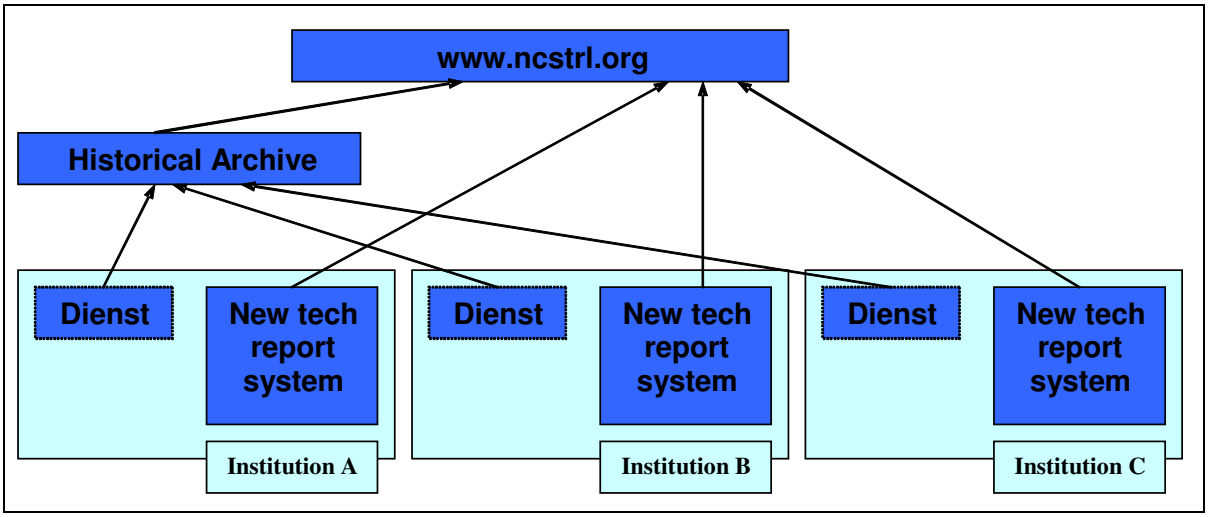

Figure 3. Architecture of NCSTRL

The central *www.ncstrl.org* website provides users with search and browse facilities, using the ARC software (Liu, et al., 2001). The repository stores metadata for the documents in the

Dublin Core (DC) format. The actual documents are stored independently in the providers' archives and URLs are provided in the metadata records. The metadata fields are stored in an indexed database that provides fast search capabilities through the metadata fields. The central NCSTRL site 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.

NDLTD (http://www.ndltd.org/)

The Networked Digital Library of Theses and Dissertations (Fox, E. A., 1998) is a collaborative effort of universities around the world to promote creating, archiving, distributing and accessing Electronic Theses and Dissertations (ETDs), as well as to encourage local advancement and adoption of digital library technologies. Initially, when numbers were small, a federated search approach was adopted. As membership expanded, NDLTD shifted to maintain a union catalog that provides a means to search and retrieve ETDs from the combined collections of NDLTD member institutions. To gather metadata in the ETD metadata standard (ETDMS) format and then to make it accessible at a central portal, the system uses the OAI-PMH. NDLTD has more than 200 international members from over twenty countries sharing electronic theses and dissertations.

The concept of electronic theses and dissertations (ETDs) was first openly discussed at a 1987 meeting in Ann Arbor arranged by UMI, and attended by representatives of Virginia Tech (Fox from Computer Science and Bright from the Computing Center), University of Michigan, SoftQuad, and ArborText. As followup, Virginia Tech funded development of the

first SGML Document Type Definition (DTD) for this purpose, by Yuri Rubinski of SoftQuad. Later work shifted to using XML for metadata, and XML, PDF, and multimedia formats for the works themselves.

Initially, Fox and Eaton (Dean of the Graduate School) at Virginia Tech investigated problems associated with production, archiving, and access - initially with a local faculty committee. Starting in 1992 they worked with the Coalition for Networked Information (CNI), the Council of Graduate Schools (CGS), UMI and other interested organizations, helping run a series of design and discussion meetings. Additionally, the University Library's Scholarly Communications Project developed the procedures and systems for processing, archiving, and providing public access to Virginia Tech's graduate research works. Subsequently, projects in Australia, Brazil, Germany, Portugal, and other nations, along with support from international organizations like UNESCO, led to a number of software systems for local, regional, and national initiatives and services, reinforced by an annual international ETD conference to promote international collaboration. NDLTD, a non-profit educational and charitable corporation open to membership from around the globe, continues to guide the initiative and to promote advances in digital libraries to ensure worldwide scholarly communication.

NSDL (http://www.nsdl.org/)

The National Science, Technology, Engineering, and Mathematics (STEM) Education Digital Library is an effort initiated by the US National Science Foundation to organise and make easily accessible electronic resources for teaching and learning in the STEM areas. There are over 100 projects that have made up this initiative, engaged in targeted research, services development and deployment, and in supporting varied communities with specialized collections. NSDL offers interoperability at three levels: federation, harvesting, and

gathering (Arms et al., 2002). Federation enables interaction with collections that are compliant with standards such as Z39.50. However, since the limitations and challenges of federated searching are widely known, NSDL provides the facility of harvesting from OAI-compliant repositories and building a central searchable database. Over and above these, crawlers and community based activities are deployed to gather resources, similar to the method used by general-purpose Web search engines. NSDL uses the latest technology and best practices outlined in this document in supporting both a broad base of remote sites and a varied and configurable set of user services at central locations. NSDL and NDLTD are both featured in a UNESCO report that advocates a digital library for education in every nation (Kalinichenko, 2003).

RePec (http://repec.org/)

RePEc (Research Papers in Economics) is a collaborative effort of over 100 volunteers in 41 countries to enhance the dissemination of research in economics. The heart of the project is a decentralized database of working papers, journal articles and software components. Any institution is welcome to join in contributing its research materials. All RePEc material is freely available.

Universal Digital Library (http://disc.iisc.ernet.in/unidiglib.html)

The aim of this project is to digitize around a million books in the next three years. This joint initiative is planned to synergistically capitalise on the availability of state-of-the-art hardware and software in the US for digitizing, storing, and accessing information, and the high quality manpower available in India. This would act as a forerunner for many such initiatives with other countries, particularly in China and Korea, and would culminate in the

grand vision of digitizing all formal knowledge and making it available in a location- and time-independent way for the benefit of mankind.

4. Summary and Future Directions

Digital libraries (DLs) have been under development since the early 1990s. There have been centralized and distributed architectures, wherein content and/or services have been in one place, or distributed according to a variety of design principles. Some DLs serve an individual, and use light-weight methods (Maly, Zubair, & Liu, 2001). Many DLs serve an institution, including its various communities. Larger DLs serve a regional or national or worldwide community, typically in a particularly disciplinary area (e.g., economics) or with regard to a particular genre (e.g., theses).

Distributed systems, early on, supported federated search. Later systems shifted to OAI-PMH. Newer systems are moving toward a Web Services paradigm. Continuing research is needed to ensure interoperability, efficiency, robustness, and reliability across the global Internet.

As these systems become more widely used, including for commercial activities, further work is needed with regard to authentication, digital rights management, and security. In addition, as a DL industry emerges, the separation of data from services promulgated in OAI is likely to be extended, with rapid growth not only of specialized collections, but also of integrated services. In academic settings these already are called for with regard to learning management systems. More broadly, they fit into the move towards a Semantic Web.

This introduction should help readers explore other design and architectural issues in the DL context, such as those discussed in the next three chapters. More broadly, it is hoped that it will provide some foundation for considering the spread of DLs throughout Asia and beyond.

5. References

Allen, R.B. (1994). Navigating and searching in hierarchical digital library catalogs. *Proceedings of Digital Libraries `94*, College Station, TX, USA, pp. 95-100, June 1994.

Arms, W., et al. (2002). A Spectrum of Interoperability: The Site for Science Prototype for the NSDL. *D-Lib Magazine*, Vol. 8 No. 1, January 2002. Retrieved July 11, 2003, from http://www.dlib.org/dlib/january02/arms/01arms.html

ANSI/NISO (1995). *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification (ANSI/NISO Z39.50-1995)*. Bethesda, MD: NISO Press.

Broughton, V. (2001). Faceted Classification as a Basis for Knowledge Organisation in a Digital Environment: The Bliss Bibliographic Classification as a Model for Vocabulary Management and the Creation of Multidimensional Knowledge Structures. The New Review of Hypermedia and Multimedia, Vol. 7, 2001, pp. 67-102.

Christensen, E., Curbera, F., Meredith, G., & Weerawarana, S. (2001) *Web Services Description Language (WSDL) 1.1*. W3C. Retrieved November 7, 2003, from http://www.w3.org/TR/wsdl

Davis, J., R., & Lagoze, C. (2000). NCSTRL: Design and Deployment of a Globally Distributed Digital Library, in *JASIS*, Vol. 51, No. 3, pp. 273-280.

Egan, D., Remde, J.R., Gomez, L.M., Landauer, T.K., Eberhardt, J., & Lochbaum, C.C. (1989). Formative Design and Evaluation of SuperBook. *ACM Transactions on Information Systems*, Vol. 7, , pp. 30-57.

Fox, E.A. (1999). Networked Digital Library of Theses and Dissertations (NDLTD). *Nature Web Matters*, August 1999. Retrieved February 28, 2003, from http://www.nature.com/nature/webmatters/library/library.html

Gadd, E., Oppenheim, C., & Probets, S. (2003). The Intellectual Property Rights Issues Facing Self-archiving: Key Findings of the RoMEO Project. *D-Lib Magazine*, Vol. 9, No. 9. Retrieved November 5, 2003, from http://www.dlib.org/dlib/september03/gadd/09gadd.html

Ginsparg, P. (2003). *Can Peer Review be better Focused?*. Cornell University. Retrieved March 4, 2003, from http://arxiv.org/blurb/pg02pr.html

Google (2003). *Google Web APIs*. Retrieved November 7, 2003, from http://www.google.com/apis/

Gourley, D. (2003). *Library Portal Roles in a Shibboleth Framework*. Shibboleth Project. Retrieved November 5, 2003, from http://shibboleth.internet2.edu/docs/gourley-shibboleth-library-portals-200310.html

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J., & Nielsen, H. F. (2003). *SOAP Version 1.2 Part 1: Messaging Framework* and *SOAP Version 1.2 Part 2: Adjuncts*. W3C. Retrieved November 5, 2003, from http://www.w3.org/TR/2003/REC-soap12-part1-20030624/ and http://www.w3.org/TR/2003/REC-soap12-part2-20030624/

IMS Global Learning Consortium, Inc. (2003). *IMS Digital Repositories Interoperability - Core Functions Information Model*. Retrieved November 5, 2003, from http://www.imsglobal.org/digitalrepositories/driv1p0/imsdri_infov1p0.html

Kalinichenko, L. (coordinating author) (2003). *Digital Libraries in Education: Analytical Survey*. UNESCO Institute for Information Technologies in Education, Moscow.

Krichel, T. (2000). Working towards an Open Library for Economics: The RePEc project. *Proceedings of The Economics and Usage of Digital Library Collections*, Ann Arbor, MI, USA, 23-24 November 2000. Retrieved November 7, 2003, from http://openlib.org/home/krichel/myers.html

Lagoze, C., and Davis, J. R. (1995). Dienst - An Architecture for Distributed Document Libraries. *Communications of the ACM*, Vol. 38, No. 4, ACM, p. 47.

Lagoze, C. & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. *Proceedings of the ACM-IEEE Joint Conference on Digital Libraries*, Roanoke, VA, 24-28 June 2001, pp. 54-62.

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). *The Open Archives Initiative Protocol for Metadata Harvesting – Version 2.0*. Open Archives Initiative, June 2002. Retrieved November 5, 2003, from http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Leymann, F. (2001). *Web Services Flow Language (WSFL 1.0)*. IBM, May 2001.

Library of Congress (2003). *ZING Z39.50 International: Next Generation*. Retrieved November 7, 2003, from http://www.loc.gov/z3950/agency/zing/zing-home.html

Liu, X., Maly, K., Zubair, M., & Nelson, M. L. (2001). Arc: an OAI service provider for cross-archive searching. *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, VA, USA, 24-28 June 2001, pp. 65-66.

Madalli, D. P., et al. (2003). Subject Based Domain Search System. Paper based on internal project report of CS department, Virginia Tech, Blacksburg, VA (unpublished), USA.

Maly, K., Zubair, M., & Liu, X. (2001) Kepler - An OAI Data/Service Provider for the Individual. *D-Lib Magazine*, Vol. 7, No. 4, April 2001. Retrieved March 4, 2004, from http://www.dlib.org/dlib/april01/maly/04maly.html

Merchant, B., Gaylord, M., & Congia, S. (2003). *SOAPifying the Open Archives*. Technical Report CS03-13-00, Department of Computer Science, University of Cape Town.

Papazoglou, M. P. & Georgakopoulos, D. (2003). Service-Oriented Computing. *Communications of the ACM,* Vol. 46, No. 10, pp. 25-28.

Payette, S., Blanchi, C., Lagoze, C., & Overly, E. A. (1999). Interoperability for Digital Objects and Repositories: The Cornell/CNRI Experiments. *D-Lib Magazine,* Vol 5, No. 5, May 1999. Retrieved November 7, 2003, from http://www.dlib.org/dlib/may99/payette/05payette.html

Roscheisen, M., Baldonado, M., Chang, C., Gravano, L., Ketchpel, S., & Paepcke, A. (1998). The Stanford InfoBus and Its Service Layers: Augmenting the Internet with Higher-Level Information Management Protocols. *Digital Libraries in Computer Science: The MeDoc Approach*, Lecture Notes in Computer Science, No. 1392, Springer, 8 August 1998. Retrieved November 5, 2003, from http://dbpubs.stanford.edu:8090/pub/1998-25

Suleman, H. & Fox, E. A. (2001). A Framework for Building Open Digital Libraries. *D-Lib Magazine*, Vol. 7, No. 12, December 2001. Retrieved November 5, 2003, http://www.dlib.org/dlib/december01/suleman/12suleman.html

Tudhope, D., & Cunliffe, D. (2001). Editorial: Introduction to theme on Digital Libraries. *The New Review of Hypermedia and Multimedia*, Vol. 7, 2001. Retrieved June 11, 2003, from http://www.comp.glam.ac.uk/~NRHM/volume7/e-2001.htm

Winer, D. (2002). *RSS 2.0 Specification*. Retrieved November 5, 2003, from http://blogs.law.harvard.edu/tech/rss