DESIDOC Bulletin of Information Technology, Vol. 17, No. 6, November 1997, pp. 15-21

Digital Library Projects in the United States

Padmini Srinivasan*

Abstract

This paper surveys select Digital Library projects in the United States. Representative projects such as those from academia, publishing agencies, and the Library of Congress are included. The projects indicate a variety of strategies and emphases. It is clear that this is a very active and opportune time for Digital Library research and development. For every project surveyed there are at least 10 other up and coming projects that are not surveyed. A new entrant in this area has a number of models and approaches from which to select. At the same time, the field of Digital Libraries is young enough to support entirely new approaches.

1. INTRODUCTION

In recent years we have seen rapidly increasing interest in Digital Libraries (DLs). Academia, libraries, publishers, system developers and diversive user populations are collaborating in exciting new ways to create DL models, technologies, test collections and prototype systems. The subject area of DLs offers many theoretical and practical challenges that are best addressed by interdisciplinary team work. The aim in this article is to present a brief survey of representative DL efforts in the United States. A few years ago the area became highly energised when the US Government invested almost 25 million dollars into several DL initiatives. However DL research in the US predates this event. In addition to the aforementioned initiatives, this paper surveys sample projects conducted by universities and organisations such as Xerox, the Library of Congress, OCLC and RLG. The plan is to give the reader a flavour of the different directions of investigations in the area of Digital Libraries.

2. NSF/ARPA/NASA JOINT DL INITIATIVE

In 1994, the Joint Digital Libraries Initiative (DLI) sponsored by the National Science Foundation (NSF), Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA) was established as a project of the US Federal Internet Infrastructure Technology and Applications Working Group of the High Performance Computing and Communications Initiative. The goal of this four year initiative is to explore the technical requirements for future digital libraries. The joint initiative funded DL proposals from six universities. Each proposal though initiated by a University involved several external collaborators.

Both the University of Michigan at Ann Arbor and the University of Illinois at Urbana-Champaign are funded to augment scientific and engineering education by placing journals and other media on the Web. While the former takes an image-based approach, the

*School of Library and Information Science
The University of Iowa
Iowa City, Iowa 52242, USA
padmini-sril~ivasan~uiowa.edu

latter explores SGML-based techniques. Stanford University is funded to explore inter operability challenges between various third-party vendors of information services. Carnegie-Mellon University is funded to explore search and discovery from video collections using speech recognition and image analysis technologies. Finally, the Universities of California at Berkeley and at Santa Barbara are funded for geographic/spatial systems and their interfaces to environmental databases. Each project is ongoing and has far more complex features than can be fully described in this paper. Instead we present a few examples of significant features regarding their goals, their test collections and their external collaborators.

2.1 Joint DLI: University of Michigan

This effort derives from previous collaborative work on a 1991 DL project called TULIP, described later in this article. The current goal is to study a complex array of technical and socio-economic issues through the design, implementation and evaluation of a DL test bed. Their test bed collection focuses on earth and space sciences and is intended to serve a variety of users. Key contributions of the Michigan DL project include an agent based DL architecture founded on a 'conspectus', i.e., formalised description methods for information objects and services in the DL test bed[4]. An agent represents an element of the DL, i.e., a collection or a service.

Agents are autonomous and negotiate with each other for access to resources and services using a common communications protocol. Three classes of agents with subtypes are defined: user interface, mediator and collection interface agents. Agents negotiate to build a collaborative team in response to a query goal. Agents communicate using messages containing performatives such as ASK and TELL. This type of agent communication via performatives is modelled after efforts such as KQML[11]. As of July 1996, more than 2000 titles were included in their prototype[31]. Examples of external collaborators are IBM, Elsevier Science. UMI International and Kodak.

2.2 Joint DLI: University of Illinois at Urbana-Champaign

This emphasis is on developing a professional quality information infrastructure to effectively search technical documents on the Internet[9,21]. They are building repositories of indexed multiple-source collections and federating them into a single virtual collection. More than 20 technical journals provided by a variety of publishers are made available in full text, SGML format. Open Text Corporation's Open Text Search engine is employed to index and search the collection. Their prototype allows comprehensive search and display of articles, including text, figures, equations and tables. For example, a user could execute an equation while browsing through a document. UI UC is collaborating with the University of Michigan DL team in interface technology, with the UC Santa Barbara DL team in semantic concept space technology and with NCSA on designing appropriate software. Examples of partners are IEEE Computer Society, Hewlett-Packard and John Wiley and Sons.

2.3 Joint DLI: Carnegie-Mellon University

The focus here is on search and discovery in the video medium[25]. Called the Informedia Digital Video Library Project, its goal is to integrate speech, language and image understanding technologies to support both the creation of and retrieval from the digital library. They use Sphinx-II, a speech recogniser developed at CMU to automatically transcribe narratives and dialogues from each video or broadcast into a technical transcript. Close-captioning information is used for this when available. video sequences are partitioned into segments at CMU's Image Understanding Systems Laboratory to support retrieval and manipulation. Machine vision techniques are used to determine the boundary points of scenes. The initial focus is on several thousand hours of raw and edited video from public television programs with test beds in K-12 schools and CMU. Industrial partners include DEC, Bell Atlantic, Intel Corporation and MicroSoft, Inc.

2.4 Joint DLI: University of California at Berkeley

The aim here is to provide a large test bed of data pertaining to the environment, including documents, reports, image collections, maps, sensor data and related collections.[18,20] Most of the sensor data is unnumbered. As of July 1996, they had 268 GB, with an expected size of 3.5 TB. The goal is to offer WWW based access to this data with CGI used to provide interactions between WWW clients and other software systems. A significant contribution of the Berkeley initiative is the concept of 'Multivalent Documents'. This new concept allows one to view a document with complex content, as having multiple 'layers' of distinct but intimately related content. Moreover, small, dynamically loaded program objects (also termed 'behaviours') activate the content and work in concert to present the user with a unified document. Thus, documents are perceived as interacting layers of content and functional behaviours.

This scheme allows the later (and optionally local) addition of appropriate content as well as behaviour. Each content layer and behaviour unit is kept simple. The multivalent document concept has lead to a number of interesting applications. For example, an image based 'cut and paste' mechanism is implemented by linking a page image to a layer containing the corresponding ASCII characters and another layer representing the characters' geometric positions on the image. (Note that these ASCII and positional layers are generated automatically).

Other behaviours may be associated with the same content such as a search function where the user interacts with the page image. Interestingly, the multivalent document concept naturally supports a distributed document model where the document parts may be geographically and temporarily distributed. In an application employing the distributed model, for example, a user may select a word from a page image which is then sent to Encyclopaedia Britannica's web site. The definition, returned in HTML format, is filtered and sent to the user. Partners include the California Resources Agency.

2.5 Joint DLI: Stanford University

The Stanford focus is on inter operability among third party repositories and services [19]. This is a challenging objective since resources and services such as selective monitoring and dissemination of information and information summarisation are being independently developed. Moreover, users are faced with many variations in communication methods. Besides telnet and http, some services require login id, others are authenticated by IP number while still others offer open access. In addition many services charge a fee which may depend upon the services used. Stanford's goal includes providing a mediating service that is aware of such variations. They employ an architecture termed the 'InfoBus'. Services, collections and clients plug into this and communicate via built-in inter operability mechanisms that include a common language to describe information management tasks and objects. Implementation is with CORBA (Common Object Request Broker Architecture) distributed object technology implemented through Xerox's Inter-Language Unification System. The Stanford test bed provides access to Internet search services such as Lycos and Alta Vista as well as commercial services such as Dialog. Services offered include text summarisation and online payment options. Sample industrial partners are Xerox PARC, Bell Communications Research, Hewlett-Packard and Knight-Ridder Information Services.

2.6 Joint DLI: University of California at Santa Barbara

The goal here is to research problems related to a 'distributed digital library for geographically-referenced information', where the phrase geographically-referenced implies indexing of information by geographic locations in addition to indexing by other attributes[14]. Their Web based prototype called the Alexandria Digital Library supports both text based and visual query languages to identify and work with relevant data sets. Users may visualise and browse data sets and transfer subsets to other

locations. The test bed focuses on populating the test bed, maintaining compatibility with various standards (example: spatial and library standards); scalability aspects and user evaluation goals. Their test bed collection includes digitised maps images and other graphical information related to Santa Barbara, Ventura and Los Angeles counties. Example partners are the National Centre for Geographic Information and Analysis, Digital Equipment Corporation and Xerox. Further details on these NSF/ARPA/NASA DL initiatives may be found at the Web site http://www.clse.nsf.gov/iris/DLHome.html. These projects are not only interesting for their research contributions but also for their test beds which should be available for further research.

3. LIBRARY OF CONGRESS

The Library of Congress (LC) has been a key leader in the use of automation to promote inter-library collaboration and in developing standards such as MARC and Z39.50. The LC is conducting a five-year DL program to assemble a core digital collection of American historical and cultural primary source material[1,2]. These collections, collectively referred to as 'American Memory', derive from heterogeneous resources and are intended to serve a range of users including scholarly researchers and school children. Most of the books and papers have been encoded in SGML with embedded links to the images of original pages. However these are also converted to HTML to support a wider user group. The LC has recognised a number of special challenges. For example digitising historical collections poses special requirements since it is necessary to minimise the handling of fragile or deteriorating originals. Similarly significant staff time and technical expertise are required to create work flow plans and process materials through to successful digitalisation. Other conclusions made include recognition of the need for a naming scheme and the impracticality of expecting uniformity in description. The American memory collections are also linked to the INQUERY search system[5] to offer both browsing and full text searches. The digital archive expected to be 50 TB by the year 2000 requires management and preservation. Other related LC projects are THOMAS to provide public access to legislative information[7] and CORDS to accept materials in digital form for copyright registration and deposit. LC is also a participant in the National Digital Libraries Federation (NDLF) established in 1995[15] a consortium of the country's fifteen largest research libraries and archives. In addition LC participates in the G-7 program on the Global Information Society. This program emphasises international cooperation in projects such as the Electronic Libraries Project which aims to build a prototype global library.

4. XEROX

Vendors such as Xerox, have had a strong role in the development of digital libraries. Xerox in collaboration with Indiana University Purdue University at Indianapolis explored the viability of a WWW based 'electronic reserve section'. In collaboration with Cornell University they investigated digital image scanning as an alternative to microfilm and photocopy archival of brittle library materials. These projects indicate that the interaction between technology and the social and political context of the development effort is a major factor determining success[6]. This paper includes some thought provoking discussions examining a number of assumptions typically made. These include, assumptions regarding technology, users, copyright and document revision management. For example, the task of scanning brittle materials, initially regarded as a low-skill job proved to be relatively highly skilled.

5. ONLINE COMPUTER LIBRARY CENTER

OCLC has taken a strong leadership position with its research and development efforts on Persistent URLS (PURLS)[22] and more generally on Resource Metadata. Functionally a PURL is a URL. However, it points to a resolution service that returns the actual URL of the selected resource. The OCLC PURL Service has been functioning since January 1996. In addition, the Dublin Core Metadata Element Set resulted from a 1995 Metadata workshop jointly hosted by OCLC and NCSA[26]. A follow up workshop in

1996 yielded the 'Warwick Container Architecture' that allows individual metadata sets (including Dublin Core) to co-exist[15]. This architecture allows designers of individual metadata sets to focus on their specific requirements. Moreover, the same object may be accessed via separate metadata sets representing different user communities. This flexibility accommodates future metadata sets without requiring changes to existing ones. These and other related efforts also emphasise the importance of metadata standards for effective resource discovery.

6. TULIP

TULIP (The University Licensing Program, 1991-1995) a collaboration between Elsevier Science and nine US universities is definitely a significant part of US DL history. The collaboration tested systems for the networked delivery to, and use of electronic journals at the user's desktop[23]. The scanned pages, the bibliographic data and unedited, OCR generated, ASCII full text of 43 Elsevier and Pergamon journals were made available to the collaborating universities which in turn developed systems for desktop delivery to users. This project yielded a number of important conclusions. An example technical conclusion is that scalability of TULIP-like projects will be hindered by bandwidth and the limits of massive storage technology. An end-user related conclusion is the importance of coverage in attracting users. An organisational conclusion made is the need for a reliable infrastructure: systems, networking printing facilities, etc. Project participants also see the role of libraries increasing and not decreasing with a major role in providing coherence to the vast array of delivery, cost and interface models that users face. Finally, for the publisher, electronic publishing offers challenges very different from paper publishing, challenges that typically slow such expansions.

7. INTER-UNIVERSITY COLLABORATION

NCSTRL (Networked CS Technical Report Library) is an excellent example of a collaborative DL project involving a number of universities. It is a distributed digital library offering access to Computer Science Technical Reports. It has the advantage that the underlying social context of computer scientists exchanging technical reports is an established one. The development of NCSTRL from its predecessors UCSTRI[24], then WATERS[13] and DIENST[17] is a good example of system evolution[8]. Both UCSTRI and WATERS adopt a centralised search model while Dienst involves a distributed set of servers for both indexing and searching. NCSTRL uses the Dienst protocol for distributed storage and retrieval developed by Davis and Lagoze at Cornell University. Future versions will be based on the distributed digital object architecture of Kahn (CNRI) and Wilensky (tic Berkeley). A NCSTRL user views the technical report collection as a single entity although the collection and search engines are distributed over the net. The search interface is web form based. Participating institutions may collaborate in standard or in lite mode. The former implies the installation of the NCSTRL server software. This allows the ability to provide customised user interfaces; store data locally and maintain own search engines and the ability to store documents in multiple formats. The lite version requires only the availability of an anonymous ftp site. As far as the user is concerned, there is no distinction between NCSTRL server documents and documents on an NCSTRL-lite ftp site. Inter-operability of NCSTRL sites is through an open protocol and three component services: repository, index and interface services. One hopes to see similar applications built for other disciplines.

A similar distributed effort to build a national electronic collection of theses and dissertations (ETD) has been initiated at Virginia Tech in collaboration with organisations such as Cornell, UMI, IBM, Virginia Tech, Adobe. The idea is to build a unified national collection to support higher education with each university responsible for its local collections. Two document formats are involved. First Adobe's PDF and second SGML. In the later a special DTD (Document Type Definition) tailored to electronic theses and dissertations has been developed. The SGML approach simplifies the

creation of derivative objects such as HTML versions, MARC records and Dissertation Abstracts entries. Moreover, author submission and verification may be done through the WWW. This application, since it encompasses the global TD domain, requires work flow redesign at the local end which in return requires strong local consensus and training.

8. RESEARCH LIBRARIES GROUP

Research Library Group (RLG), a not-for-profit cooperation of universities, archives, libraries, etc., produces and maintains RLIN a union catalog of over one-hundred-million bibliographic records in more than 365 languages[10]. RLG promotes collaborative efforts such as resource-sharing agreements and thus the RLG environment is naturally supportive of collaborative digital conversion of materials. 'Studies in Scarlet' is the first in a series of RLG Digital Collections projects. The goal is to digitise legal research materials on marriage and sexuality in the US and UK between 1815 and 1914. Participating institutions develop and contribute resources. Scanning, text conversion and SGML encoding are part of the overall process. This and other collaboratively established collections will form the basis of user studies, metadata archival and finding aids research. RLG efforts will be interesting to watch as they slowly become more global with increasing international partnerships.

9. CONCLUSIONS

We make a few conclusions that are potentially relevant to groups or organisations considering research and development in Digital Libraries. First, there appears to be no single model or method to follow. For example, despite the collective experience gathered thus far the choice between Adobe's PDF format and SGML format for digital documents is not a simple one. Needles to say more complex decisions such as the choice of architectural model and the design of collaborative connections remain highly challenging. Each project surveyed appears to have taken an almost unique approach determined by its goals, resources and the extent to which creativity and novelty are encouraged. The newcomer to the DL R&D world has at one extreme the option of emulating an existing project and at the other extreme the option of carving out an independent path. Either way, it is necessary to pay close attention to the more frequently made observations such as the importance of strong organisational commitment, a critical mass of information and a well defined collaborative approach. DL projects continue to be exciting ventures with many open challenges and opportunities.

10. BIBLIOGRAPHY

1. Arms, CR. Historical collections for the national digital library: Lessons and challenges at the Library of Congress. Part-1. *D-Lib Magazine*, April 1996. http://lcweb2.loc.gov/ammem/ammemhe.html
2. Arms, CR. Historical collections for the national digital library: Lessons and challenges at the Library of Congress. Part-2. *D-Lib Magazine*, May 1996.
3. Atkins, D. The University of Michigan digital library project: The test bed. *D-Lib Magazine*, July/August 1996.
4. Birmingham, WP. An agent-based architecture for digital libraries. *D-Lib Magazine*, July 1995.
5. Callan, JB, Croft, WB, and Harding, SM. The inquery retrieval system. *In* Proceedings of the Third International Conference on Database and Expert Systems Applications, Valencia, Spain, 1992. Spring-Verlag, pp. 78-83.
6. Crocca, W and Anderson, W. Delivering technology for digital libraries: Experiences as vendors. *Digital Libraries*, 1995.
7. Croft, W, Cook, R, and Wilder, D. Providing government information on the internet: Experiences with THOMAS. *Digital Libraries*, 1995.
8. Davis, J. Creating a networked computer science technical report library. *D-Lib Magazine*, September 1995.

9. Digital Libraries Initiative. University of Illinois at Urbana-Champaign, October 31, 1996.

10. Erway, R. Digital initiatives of the research libraries group. *D-Lib Magazine,* December 1996.

11. Fin in, T, Labrou, Y and Mayfield, J. KQML as an agent communication language. http://www.cs.umbc.edu/agents/introduction.

12. Fox, E, *et al.* National digital library of theses and dissertation: A scalable and sustainable approach to unlock university resources. *D-Lib Magazine,* September 1996. http://etd.vt.edu/etd/

13. French, JC, *et al.* Wide Area Technical Report Service: Technical Reports Online. *CACM,* 1995, **38**(4), 45.

14. Frew, J. The Alexandria digital library test bed. *D-Lib Magazine,* July/August 1996.

15. http: //lcweb . lsc. gov/lsc/ndcf .

16. Lagoze, C. The Warwick framework: A container architecture for diverse sets of metada. *D-Lib Magazine,* July/August 1996.

17. Lagoze, C, and Davis, JR. Dienst: An Architecture for Distributed Document Libraries. *CACM,* April 1995, **28**(4), 47.

18. Ogle, V. and Wilensky, R. Test bed development for the Berkeley digital library project. *D-Lib Magazine,* July/August 1996.

19. Paepcke, A. Summary of Stanford's digital library testbed design and status. *D-Lib Magazine,* July/August 1996.

20. Phelps, T and Wilensky, R. Toward active, extensible, networked documents: multivalent architecture and applications. Proceedings of ACM Digital Libraries, March 20-23, 1996.

21. Schatz, B and Chen H. *IEEE Computer magazine,* May 1996.

22. Shafer, K., *et al.* Introduction to persistent uniform resource locators. OCLC Online Computer Library Center, Inc. http://purl.oclc.org .

23. Smith, T. The Meta-information environment of digital libraries. *D-Lib Magazine,* July/August 1996.

24. TULIP: Final report. Elsevier Science, NY. 1996.

25. Van Heyningen, M. The unified computer science technical report index. *In* Proceedings of the Second International WWW '94: Mosaic and the Web, WWW '94, 1994. 535-43.

26. Wactlar, H. Informedia digital video library: Technology outreach. *D-Lib Magazine,* July/August 1996.

27. Weibel, S. Metadata: The foundations of resource description. *D-Lib Magazine,* July 1995.