Chapter 11

Digital Libraries

by Edward A. Fox and Ohm Sornil

> "The benefits of digital libraries will not be appreciated unless they are easy to use effectively." [LGM95]

11.1 Introduction

Information retrieval (IR) is essential for the success of digital libraries (DLs), so they can achieve high levels of effectiveness while at the same time affording ease of use to a diverse community. Accordingly, a significant portion of the research and development efforts related to DLs has been in the IR area. This chapter reviews some of these efforts, organizing them into a simple framework, and highlights needs for the future.

Those interested in a broader overview of the field are encouraged to refer to the excellent text by Lesk [Les97] and the high quality papers in proceedings of the ACM Digital Libraries Conferences. Those more comfortable with online information should refer to *D-Lib Magazine* [Fri], the publications of the NSF/ARPA/NASA Digital Libraries Initiative (DLI) [Har], or online courseware [Fox]. There also have been journal special issues devoted to the topic [FL93, FAFL95, SC96]. Recently, it has become clear that a global focus is needed [FM98] to extend beyond publications that have a regional [Bar97] or national emphasis [DB94].

Many people's views of DLs build from the foundation of current libraries [Ros96]. Capture and conversion (digitization) are key concerns [CK96], but DLs are more than digital collections [Pet95]. It is very important to understand the assumptions adopted in this movement toward DLs [LM95] and in some cases to relax them [Arm97].

Futuristic perspectives of libraries have been a key part of the science fiction literature [Wel37] as well as rooted in visionary statements that led to much of the work in IR and hypertext [Bus45]. DLs have been envisaged since the earliest days of the IR field. Thus, in *Libraries of the Future*, Licklider lays out many of the challenges, suggests a number of solutions, and clearly calls for IR-related efforts [Lic65]. He describes and predicts a vast expansion of the world of publishing, indicating the critical need to manage the record of knowledge, including search and retrieval and all the related supporting activities. He notes that to handle this problem we have no underlying theory, no coherent representation scheme, no unification of the varied approaches of different computing specialties – and so must tackle it from a number of directions.

After more than 30 years of progress in computing, we still face these challenges and work in this field as a segmented community, viewing DLs from one or another perspective: database management, human-computer interaction (HCI), information science, library science, multimedia information and systems, natural language processing, or networking and communications. As can be seen in the discussion that follows, this practice has led to progress in a large number of separate projects, but also has made interoperability one of the most important problems to face [PCGMW98].

Since one of the threads leading to the current interest in DLs came out of discussions of the future of IR [FFS+93], since people's needs still leave a rich research agenda for the IR community [Cro95], and since the important role of Web search systems demonstrates the potential value of IR in DLs [Sch97], it is appropriate to see how IR might expand its horizons to deal with the key problems of DLs and how it can provide a unifying and integrating framework for the DL field. Unfortunately, there is little agreement even regarding attempts at integrating database management and text processing approaches [GFHR97]. Sometimes, though, it is easier to solve a hard problem if one takes a broader perspective and solves a larger problem. Accordingly we briefly and informally introduce the "4S" model as a candidate solution, and a way to provide some theoretical and practical unification for DLs.

We argue that DLs in particular, as well as many other types of information systems, can be described, modeled, designed, implemented, used, and evaluated if we move to the foreground four key abstractions: streams, structures, spaces, and scenarios. "Streams" have often been used to describe texts, multimedia content, and other sequences of abstract items, including protocols, interactive dialogs, server logs, and human discussions. "Structures" cover data structures, databases, hypertext networks, and all of the IR constructs like inverted files, signature files, MARC records, and thesauri. "Spaces" cover not only 1D, 2D, 3D, virtual reality and other multidimensional forms, some including time, but also vector spaces, probability spaces, concept spaces, and results of multidimensional scaling or latent-semantic indexing. "Scenarios" cover not only stories, HCI designs and specifications, and requirements statements, but also describe processes, procedures, functions, and transformations — the active and time-spanning aspects of DLs. Scenarios have been essential to our understanding of these changes [LGM95], and are particularly important in connection with social issues [Bor96a].

Since the 4S model can be used to describe work on databases, HCI, hyperbases, multimedia systems, and networks, as well as other fields relating to library and information science, we refer to it below to help unify our coverage and make sure it encompasses all aspects of DLs. For example, the 4S model in general, and scenarios in particular, may help us move from a paper-centered framework for publishing and communicating knowledge [CHW97], to one where streams and spaces play a larger role, providing a simple way to organize our thinking and understand the changes that DLs will facilitate:

> "The boundaries between authors, publishers, libraries, and readers evolved partly in response to technology, particularly the difficulty and expense of creating and storing paper documents. New technologies can shift the balance and blur the boundaries." [LGM95]

To ground these and other subsequent discussions, then, we explore a number of definitions of DLs, using 4S to help us see what is missing or emphasized in each.

11.2 Definitions

Since DL is a relatively new field, many workshops and conferences continue to have sessions and discussions to define a "digital library" [Fox93, Har96]. Yet, defining DLs truly should occur in the context of other related entities and practices [Gra97]. Thus, a "digital archive" is like a DL but often suggests a particular combination of space and structure, and emphasizes the scenario of preservation, as in "digital preservation" that is based upon digitization of artifacts. Similarly, "electronic preservation" calls for media migration and format conversions to make DLs immune to degradation and technological obsolescence. Maintaining "integrity" in a DL requires ensuring authenticity, handled by most regular libraries, as well as consistency, which is a concern whenever one must address replication and versioning, as occurs in database systems or in distributed information systems.

While these concerns are important, we argue that "DL" is a broader concept. Because it is true that the "social, economic, and legal questions are too important to be ignored in the research agenda in digital libraries" [LGM95], we really prefer definitions that have communities of users as part of a DL:

> "DLs are constructed – collected and organized – by a community of users. Their functional capabilities support the information needs and uses of that community. DL is an extension, enhancement, and integration of a variety of information institutions as physical places where resources are selected, collected, organized, preserved, and accessed in support of a user community." [Bor96a].

This definition has many aspects relating to 4S, but largely omits streams, and only indirectly deals with spaces by calling for extensions beyond physical places. Its coverage of scenarios is weak, too, only giving vague allusion to user support. In contrast, definitions that emphasize functions and services are of particular importance to the development community [GFA+94], as are definitions concerned with distributed multimedia information systems:

> "The generic name for federated structures that provide humans both intellectual and physical access to the huge and growing worldwide networks of information encoded in multimedia digital formats." [BDMW95]

While brief, this definition does tie closely with 4S, though it is weak on scenarios, only mentioning the vague and limited concept of "access."

To the IR community a DL can be viewed as an extended IR system, in the context of federation and media variations [Bor96a]. Also, DLs must support (large) collections of documents, searching, and cataloging/indexing. They bring together in one place all aspects of 4S, and many of the concerns now faced by IR researchers: multilingual processing, search on multimedia content, information visualization, handling large distributed collections of complex documents, usability, standards, and architectures, all of which are explored in the following sections.

Figure 11.1: Digital objects, handles, and repositories (adapted from [KW95, Arm95, ABO97, Arm98]).

11.3 Architectural Issues

Since DLs are part of the global information infrastructure, many discussions of them focus on high level architectural issues [NFL$^+$95]. On the one hand, DLs can be just part of the "middleware" of the Internet, providing various services that can be embedded in other task-support systems. In this regard they can be treated separately from their content, allowing development to proceed without entanglement in problems of economics, censorship, or other social concerns.

On the other hand, DLs can be independent systems, and so must have an architecture of their own in order to be built. Thus, many current DLs are cobbled together from pre-existing pieces, such as search engines, Web browsers, database management systems, and tools for handling multimedia documents.

From either perspective, it is helpful to extend definitions into more operational forms that can lead to specification of protocols when various components are involved. Such has been one of the goals of efforts at CNRI, as illustrated in Figure 11.1.

Thus, Kahn and Wilensky have proposed one important framework [KW95]. Arms et al. have extended this work into DL architectures [Arm95, ABO97]. One element is that of a digital object, which has content (bits) and a handle (a type of name or identifier) [CNR98], and also may have properties, a signature, and a log of transactions that involve it. Digital objects have associated metadata, that can be managed in sets [Lag96]. Repositories of digital objects can provide security, and respond to an access protocol [Arm98]. Significant progress has been made toward adopting a scheme of digital object identifiers, first illustrated by OCLC's Persistent URLs [OCL98], and agreement seems likely on a standard for DOIs [Fou98].

Other implementation efforts have focused more on services [LE95] and security [LMOY95]. A useful testbed for this work has been computer science reports [DL94], most recently the Networked Computer Science Technical Report Library, NCSTRL [Lag98].

Two large DLI projects have devoted a good deal of attention to architecture, taking radically different approaches. At Stanford, the key concern has been interoperability [PCGMW98]. Their "InfoBus" [PCGM$^+$96] allows a variety of information resources to be connected through suitable mediators, and then used via the shared bus through diverse interfaces. At the University of Michigan, the emphasis has been on agent technologies [BDMW95]. This approach can have a number of classes of entities involved in far-flung distributed processing. It is still unknown how efficiently an agent-based DL can operate.

Ultimately, software to use in DLs will be selected as a result of comparisons. One basis for such comparisons is the underlying conceptual model [Win95]. Another basis is the use of metrics, which is the subject of recent efforts toward definition and consensus building [Lei]. In addition to metrics traditionally used in IR, dealing with efficiency, effectiveness, and usability, a variety of others must be selected, according to agreed-upon scenarios. Also important to understand is the ability of DLs to handle a variety of documents types (combinations of streams and structure), to accurately and economically represent their content and relationships, and to support a range of access approaches and constraints (scenarios).

11.4 Document Models, Representations, and Access

Without documents there would be no IR or DLs. Hence, it is appropriate to consider definitions of "document" [Sch96] and to develop suitable formalizations [LBO88] as well as to articulate research concerns [Lev88]. For efficiency purposes, especially when handling millions of documents and gigabytes of space, compression is crucial [WMB94]. While that is becoming more manageable, converting very large numbers of documents using high quality representations [CG94] can be prohibitively expensive, especially relative to the costs of retrieval, unless items are popular. All of these matters relate to the view of a document as a stream (along with one or more organizing structures); alternatively one can use scenarios to provide focus on the usage of documents. These problems shift, and sometimes partially disappear, when one considers the entire life and social context of a document [BD96, HKB96] and when DLs become an integral part of automation efforts that deal with workflow and task support for one or more document collections.

11.4.1 Multilingual Documents

One social issue with documents relates to culture and language [PP97]. Whereas there are many causes of the movement toward English as a basis for global scientific and technical interchange, DLs may actually lead to an increase in availability of non-English content. Because DLs can be constructed for a particular institution or nation, it is likely that the expansion of DLs will increase access to documents in a variety of languages. Some of that may occur since many users of information desire it from all appropriate sources, regardless of origin, and so will wish to carry out a parallel (federated) search across a (distributed) multilingual collection.

The key aspects of this matter are surveyed in [OD96]. At the foundation, there are issues of character encoding. UNICODE provides a single 16 bit coding scheme suitable for all natural languages [UNI98a]. However, a less costly implementation may result from downloading fonts as needed from a special server or gateway, or from a collection of such gateways, one for each special collection [DMS+97].

The next crucial problem is searching multilingual collections. The simplest approach is to locate words or phrases in dictionaries, and to use the translated terms to search in collections in other languages [HG96]. However, properly serving many users in many languages calls for more sophisticated processing [Oar97]. It is likely that research in this area will continue to be of great importance to both the IR and DL communities.

11.4.2 Multimedia Documents

From the 4S perspective, we see that documents are made up of one or more streams, often with a structure imposed (e.g., a raster organization of a pixel stream represents a color image). Multimedia documents' streams usually must be synchronized in some way, and so it is promising that a new standard for handling this over the Web has been adopted [Hos97].

At the same time, as discussed in Chapters 8 and 9, IR has been applied to various types of multimedia content. Thus, at Columbia University, a large image collection from the Web can be searched on content using visual queries [CSM+97]. IBM developed the Query by Image Content system for images and video [FSN+95] and has generously helped build a number of important image collections to preserve and increase access to key antiquities [GMS+98].

Similarly, the Carnegie Mellon University DLI project, INFORMEDIA [Uni98b], has focused on video content analysis, word spotting, summarization, search, and in-context results presentation [Uni98b]. Better handling of multimedia is at the heart of future research on many types of documents in DLs [Hea96]. Indeed, to properly handle the complexity of multimedia collections, very powerful representation, description, query and retrieval systems, such as those building upon logical inference, may be required [Fuh98].

11.4.3 Structured Documents

While multimedia depends on the stream abstraction, structured documents requires both the abstractions of streams and structures. Indeed, structured documents in their essence are streams with one or more structures imposed, often by the insertion of markup in the stream, but sometimes through a separate external structure, as of pointers in hypertext.

Since Chapter 3 of this book covers many of the key issues of document structure, we focus in this section on issues of particular relevance to DLs [Fur94]. For example, since DLs typically include both documents and metadata describing them, it is important to realize that metadata as in MARC (Machine Readable Catalog) records can be represented as an SGML document, and that SGML (Standard Generalized Markup Language) content can be included in the base document and/or be kept separate [Gay96].

Structure is often important in documents when one wants to add value or make texts "smart" [Che97]. It can help identify important concepts [PJ93]. SGML is often used to describe structure since most documents fall into one or more common logical structures [Sum95], that can be formally described using a Document Type Definition (DTD). Another type of structure that is important in DLs as well as earlier paper forms results from annotation [Mar97]. In this case stream and structure are supplemented by scenarios since annotations result from users interacting with a document collection, as well as collaborating with each other through these shared artifacts [RMW95].

Structure is also important in retrieval. Macleod was one of the first to describe special concerns related to IR involving structured documents [Mac90]. Searching on structure as well as content remains one of the distinguishing advantages of IR systems like OpenText (formerly "PAT"). Ongoing work considers retrieval with structured documents, such as with patterns and hierarchical texts [KM93]. An alternative approach, at the heart of much of the work in the Berkeley DLI project [Teaa], shifts the burden of handling structure in documents to the user, by allowing multiple layers of filters and tools to operate on those so-called "multivalent documents" [UC]. Thus, a page image including a table can be analyzed with a table tool that understands that structure and sorts it by considering the values in a user-selected column.

Structure at the level above documents, that is, of collections of documents, is what makes searching necessary and possible. It also is a defining characteristic of DLs, especially when the collections are distributed.

11.4.4 Distributed Collections

Though our view of DLs encompasses even those that are small, self- contained, and constrained to a personal collection with a suitable system and services, most DLs are spread across computers, that is spread across physical and/or logical space. Dealing with collections of information that are distributed in nature is one of the common requirements for DL technology. Yet, proper handling of such collections is a challenging problem, possibly since we are poorly equipped to think about situations involving spaces as well as the other aspects of 4S.

Of particular concern is working with a number of DLs, each separately constructed, so the information systems are truly heterogeneous. Integration requires support of at least some popular scenarios (often a simple search that is a type of least common denominator) by systems that expect differing types of communication streams (e.g., respond to different protocols and query languages), have varying types of streams and structures, and which combine these two differently in terms of representations of data and metadata. To tackle this problem, one approach has been to develop a description language for each DL, and to build federated search systems that can interpret that description language [CGMH+94].

However, when DL content is highly complex (e.g., when there are "unstructured" collections - meaning that the structure is complex and not well described), there is need for richer description languages and more powerful systems to interpret and support highly expressive queries / operations [Won97b]. An architecture of this type is illustrated in Figure 11.2 about the BioKleisli system [Won97a]. HAVE THE FIGURE AT http://sdmc.krdl.org.sg/kleisli/gif/arch.simple2.gif Figure 11.2. Architecture of the BioKleisli system, adapted from [Won97a]

In addition to these two approaches, namely having reduced functionality for end-users in order to give DL developers more freedom; or increasing functionality by making the federated system smarter and able to use more computational resources on both servers and clients; there is the third approach of making each DL support a powerful protocol aimed at effective retrieval. This latter course is supported by the CIMI effort, wherein a Z39.50 interface exists on a number of museum information servers and clients [Moe98]. While Z39.50 was aimed at the needs of libraries desiring interoperability among library catalogs, it does support many of the needs for DLs. Thus, the CIMI interoperability demonstration, with its support for multimedia content, is of great import, but does leave open further improvement in supporting richer DL interaction scenarios, including more powerful federated searchers.

11.4.5 Federated Search

Federated search work has often been prompted by challenging application requirements. For example, to allow computer science technical reports from around the world to become accessible with minimal investment and maximal local control, the NSF-funded WATERS initiative was launched [MFFS96]. This was then integrated with an effort begun earlier with DARPA funding, the CSTR project [CNR96], leading to a hybrid effort, the Networked CS Technical Reference (previously, Report) Library [Lag98]. At the heart of NCSTRL is a simple search system but a well-thought-out open federated DL protocol and the Dienst reference implementation of that, built at Cornell University [DL94]. While this system was custom-built with little dependence on other software, its type of operation could be constructed more rapidly atop various supports like CORBA [Vin97].

Federated search has had an interesting history, with workers adopting a variety of approaches. First, there are those interested in collecting the required information, often through Web crawling of various sorts [SE95]. Second, there are those focused on intelligent search [ACHK93]. One example is work emphasizing picking the best sites to search [BP94]. These efforts often assume some integrated information organization across the distributed Internet information space [Bre96].

Third, there is work on fusion of results. This can be viewed in the abstract, regardless of whether the various collections are nearby or distributed, with the target of improving retrieval by culling from a number of good sources [BKFS95]. One approach adopts a probabilistic inference net model [CLC95]. Another views the problem as database merging [VT97]. Alternatively one can assume that there are a number of search engines distributed to cover the collection, that must be used intelligently [GWG96].

Fourth, there are commercial solutions, including through special WWW services [Dre98]. Probably the most visible is the patented powerful yet elegant approach by INFOSEEK [Cor].

Finally, there is a new line of work, little explored except in a limited fashion by investigators like Dolan in the UCB DLI project, of developing comprehensive and realistic architectures for federated search. The long-term challenge is to segment the collection and/or its indexes so that most searches only look at a small number of the most useful sources of information, yet recall is kept high. Ultimately, however, there are rich types of use of DL content, once search is carried out.

11.4.6 Access

When priceless objects are described by DL image collections [GMS+98], or when collections are large and/or well organized so as to appear of value to communities of users, or when there are valuable services in information manipulation (searching, ordering, reporting, summarizing, etc.) afforded by a DL, some method of payment is often required [CTS95, CKP+95, BB97, Fer97]. Though historically access to scientific literature was not treated in the same commodity market [Gu8], today DLs clearly must manage intellectual property [MD94]. These services must support agreed-upon principles [All97], copyright practices [Sam97], as well as contracts and other agreements and laws [CON].

Though technology is only part of the picture [Wis98], a key to the implementation of policies for access management [Arm98] is having trusted systems [Ste97]. Security is one topic often ignored by the IR community. However, the many aspects of security can be of fundamental importance in DLs [GL97, Gla97]. Just as encryption is essential to support electronic commerce, watermarking and stronger mechanisms are crucial in DLs to protect intellectual property rights, or to control the types of access afforded to different user groups. Scenarios are important here, to ensure suitable constraints are imposed on processing all the way from input to output. For example, secret documents may not even be made visible in searches through metadata. On the other hand, advertising full documents as well as allowing locating and viewing metadata records is appropriate when the purpose of security is to enforce payment in "pay by the drink" document download systems. Inference systems can be used for complicated rights management situations [ABC+98]. A deeper understanding of these requirements and services can be obtained by considering representative DL projects.

11.5 Prototypes, Projects, and Interfaces

Though numerous efforts in the IR, hypertext, multimedia and library automation areas have been underway for years as precursors of today's DL systems, one of the first new efforts aimed at understanding the requirements for DLs and of constructing a prototype from scratch was the ENVISION project, launched in 1991 [FHH]. Based on discussions with experts in the field, and a careful study of prospective users of the computer science collection being built with the assistance of ACM, the ENVISION system was designed to extend the MARIAN search system [FFS+93] with novel visualization techniques [FHN+93, HHN+95]. Careful analysis has shown its 2-D approach to management of search results is easy to use and effective for a number of DL activities [Now97].

Another early effort, focussed on chemical information, was undertaken by the American Chemical Society, Chemical Abstracts Service, OCLC, Bellcore, and Cornell University, along with other partners [EGL+95]. This project also was concerned with collection building as well as testing of a variety of interfaces that were designed based on user studies.

One of the most visible project efforts is the Digital Libraries Initiative, initially supported by NSF, DARPA and NASA [Har]. Phase 1 provided funding for 6 large projects over the period 1994-1998 [SC96]. Since these have been described elsewhere in depth, it should suffice here to highlight some of the connections of those projects with the IR community. First, each project has included a component dealing with document collections. The Illinois project [Teac] produced SGML versions of a number of journals, while the Berkeley project [Teaa] concentrated on page images and other images classes. Santa Barbara adopted a spatial perspective, including satellite imagery [Teab], while Carnegie Mellon University (CMU) focussed on video [Uni98b]. Standford built no collections, but rather afforded access to a number of information sources to demonstrate interoperability [Gro]. At the University of Michigan, some the emphasis was on having agents dynamically select documents from a distributed set of resources [Tead].

Second, the DLI projects all worked on search. Text retrieval, and using automatically constructed cross-vocabulary thesauri to help find search terms was emphasized in Illinois. Image searching was studied at Berkeley and Santa Barbara while video search was investigated at CMU. Michigan worked with agents for distributed search, while Stanford explored the coupling of a variety of architectures and interfaces for retrieval.

Finally, it is important to note that the DLI efforts all spent time on interface issues. Stanford used animation and data flows to provide flexible manipulation and integration of services [CPW+97]. At Michigan, there were studies of the PAD++ approach to 2-D visualization [BSH94]. Further discussion of interfaces can be found below in the section on usability.

It should be noted that these projects only partially covered the 4S issues. Structure was not well studied, except slightly in connection with the Illinois work on SGML and the Berkeley work on databases. Scenarios were largely ignored, except in some of the interface investigations. Similarly, spaces were not investigated much, except in connection with vocabulary transfer work at Illinois and the spatial collection and browsing work in Santa Barbara. Other projects in the broader international scene may afford more thorough 4S coverage.

11.5.1 International Range of Efforts

DL efforts, accessible over the Internet, now can lead to worldwide access. Since each nation wishes to share the highlights of its history, culture, and accomplishments with the rest of the world, developing a DL can be very helpful [Ber95]. Indeed, we see many nations with active DL programs [FM98] and there are many others underway or emerging.

One of the largest efforts is the European ERCIM program [ERC]. This is enhanced by a large initiative in UK [fLN]. There are good results from activities in New Zealand [Gro98] and Australia [AUD96]. In Singapore, billions are being invested in developing networked connectivity and digital libraries as part of educational innovation programs [SIN]. For information on other nations, see the online table pointing to various national projects associated with a recent special issue on this topic [FM98].

As mentioned briefly above, many nations around the world have priceless antiquities that can be more widely appreciated through DLs [GMS+98]. Whether in pilot mode or as a commercial product, *IBM Digital Library* [Cor98], with its emphasis on rights management, has been designed and used to help in this regard.

These projects all require multimedia and multilingual support, as discussed earlier. Different scenarios of use are appropriate in different cultures, and different structures and spaces are needed for various types of collections. Indeed, many international collections aim for global coverage, but with other criteria defining their focus. Thus, the Networked Digital Library of Theses and Dissertations [NDL] is open to all universities, as well as other supporting organizations, with the aim of providing increased access to scholarly resources as a direct result of improving the skills and education of graduate students, who directly submit their works to the DL.

11.5.2 Usability

Key to the success of DL projects is having usable systems. This is a serious challenge! Simpler library catalog systems were observed in 1986 to be difficult to use [Bor86] and still remain so after a further decade of research and development [Bor96b].

The above mentioned ENVISION project's title began with "User-Centered" and concentrated most of its resources on the interface [HHN+95]. A 1997 study at Virginia Tech of four digital library systems concluded that many have serious usability problems [KSR+97], though the design of the Illinois DLI system seemed promising. The Virginia Tech study uncovered an important aspect of the situation, and suggests it will be years before DL systems are properly understood and used. A pre-test asked about user expectations for a DL, and found that very few have worked with a DL. The post-test showed that user expectations and priorities for various features changed dramatically over the short test period. Thus, it is likely that in general, as DL usage spreads, there will be an increase in understanding, a shift in what capabilities users expect, and a variety of extensions to the interfaces now considered.

Early in the DLI work, DL use was perceived as a research focus [Bis95], and understanding and assessing user needs became a key concern [HLBB96]. For two years a workshop was held at the Allerton conference center of the University of Illinois on this topic. The 1995 event [Gra95] had a diverse group of researchers, so it was necessary to understand the various perspectives and terminologies. There were discussions of fundamental issues, like information, from a human factors perspective [Dil95] as well as specific explorations of tasks like document browsing [Maa95].

The 1996 event was more focussed due to greater progress in building and studying usability of DLs [Gra96]. Thus there was discussion of Stanford's SenseMaker system which supports rapid shifting between contexts that reflect stages of user exploration [Bal97]. Social concerns that broaden the traditional IR perspective were highlighted [Her96]. In addition there was movement toward metrics (see discussion earlier about DL metrics) and factors for adopting DLs [Kan96].

DL interfaces and usability concerns have been central to many efforts at Xerox PARC. Some of the research considers social issues relating to documents [Hea96] while other research bridges the gap between paper and digital documents [HKB96]. There are many issues about documents, especially their stability and how multimedia components as well as active elements affect retrieval, preservation, and other DL activities [Lev94]. Some insight into DL use may result from actual user observation as well as other measures of what (parts of) documents are read [Lev97]. There also has been collaboration between PARC and the UCB DLI team, which has extended Xerox magic filter work into multivalent documents (discussed earlier) as well as developed results visualization methods like TileBars where it is easy to spot the location of term matches in long documents [Hea95].

Further work is clearly needed in DL projects to improve the systems and their usability. But for these systems to work together, there also be some emphasis on standards.

11.6 Standards

Since there are many DL projects worldwide, involving diverse research, development and commercial approaches, it is imperative that standards be employed so as to make interoperability and data exchange possible. Just as any library can buy any book, and any library patron can read anything in the library, DLs must make differences in representation transparent to the users. In real time searching as well, data that can be understand by clients as well as other DLs must be what is transferred from each information source. At the heart of supporting federated DLs is agreement on protocols for computer-computer communication.

11.6.1 Protocols and Federation

In the 1980s it became clear that as library catalog systems proliferated, and library patrons sought support for finding items not locally available through interlibrary loan or remote cataloging search, some protocol was needed for searching remote bibliographic collections. The national standard Z39.50, which later became an international standard as well, led to intensive development of implementations and extensive utilization [Z39]. One example of widespread utilization was the WAIS system, very popular before the WWW emerged, which was based on Z39.50. Ongoing development of Z39.50 has continued, including to apply to DLs, as demonstrated in the CIMI project described earlier, where a number of different clients and server implementations all worked together.

Also mentioned earlier is the NCSTRL effort, starting with CS technical reports, in which the Dienst protocol was developed [DL94]. This is a "lighter" protocol than Z39.50, designed to support federated searching of DLs, but to date the only implementation is from Cornell. It seems suitable for electronic theses and dissertations as well as technical reports, and so has been considered in regard to NDLTD.

These protocols assume that each server and client will be changed to use the protocol. A less intrusive approach, but one harder to implement and enforce, is to have some mechanism to translate from a special server or gateway system to each of the information sources of interest. The STARTS protocol was proposed to move in this direction, but competition among search services on the Internet is so severe that acceptance seems unlikely [STA]. Though this is unfortunate, simple federated schemes have been implemented in the DLI projects at Stanford and Illinois, and a really simple one is in use in NDLTD. Yet, even more important than new protocols for DL federated search is agreement on metadata schemes, which does seem feasible.

11.6.2 Metadata

In the broadest sense, metadata can describe not only documents but also collections or whole DLs along with their services [BCGP97]. In a sense, this reflects movement toward wholistic treatment like 4S. Yet in most DL discussions, metadata just refers to a description of a digital object. This is precisely the role played by library catalog records. Hence, cataloging schemes like MARC are a starting point for many metadata descriptions [oC98].

While MARC has been widely used, it usually involves working with binary records which must be converted for interchange. One alternative is to encode MARC records using some readable coding scheme, like SGML [Gay96]. Another concern with MARC is that there are a number of national versions with slight differences, as well as differences in cataloging practices that yield the MARC records. USMARC is one such version. It is very important in the DL field, and can be encoded using SGML, or easily converted to simpler metadata schemes like the "Dublin Core" [oC97]. Other "crosswalks" exist between Dublin Core (DC), MARC and schemes like GILS, proposed for a Government Information Locator Service [DO97]. A mapping also exists between DC and the Z39.50 protocol discussed in the previous section [LeV98]

DC is a simple scheme, with 15 core elements that can be used to describe any digital object. What is of real import is that it has been widely accepted. That is because there have been several years of discussion and development, focussed around five international workshops [WGMD, OCL, Mil, Woo, Woo]. The core elements include seven to describe content (Title, Subject, Description, Source, Language, Relation, and Coverage). There are four that deal with intellectual property issues (Creator, Publisher, Contributor, and Rights). Finally, to deal with instances of abstract digital objects, there are four other types (Data, Type, Format, and Identifier).

Since digital objects and their metadata often have to be interchanged across systems, the problem of packaging arises. The Warwick Framework, which evolved out of the same type of discussions leading to DC, deals with packages and connections between packages [Lag96]. In general, such discussion about metadata is crucial to allow the move from traditional libraries, with their complex and expensive cataloging, past the WWW, with its general lack of cataloging and metadata, to a reasonable environment wherein metadata is available for all sorts of digital objects, suitable to allow organization of vast collections in DLs [Smi96].

Because the WWW has need of such organization, that has become a concern of its coordinating body, the WWW Consortium [W3C]. In 1996, as concern increased about protecting children from exposure to objectional materials, metadata schemes became connected with censoring and filtering requirements. The problem was renamed for this more general case, in keeping with Harvest's treatment of "resource discovery," to "resource description." The Resource Description Framework (RDF) thus became an area of study for W3C [Swi]. It should be noted that RDF can lead to header information inside digital objects, including those coded in SGML or HTML, as well as XML. In the more general case, however, RDF is essentially a scheme for annotating digital objects, so alternatively the descriptions can be stored separately from those objects. These options bring us back to the Warwick Framework where there may be multiple containers, sometimes connected through indirection, of packages of metadata, like MARC or DC.

We see that DLs can be complex collections with various structuring mechanisms for managing data and descriptions of that data, the so called metadata. However, coding may combine data with metadata, as in the guidelines of the Text Encoding Initiative (TEI) [Ren97]. This reminds us of the complexities that arise when combining streams and structures, where there are many equivalent representations. We also see that for DL standards to be useful, such as appears to be the case for DC, the structures involved must be relatively simple, and have well understand related scenarios of use. While this now appears to work for data interchange, further work is required for interoperability, that is, interchange through the streams involved in protocols.

11.7 Future Challenges

In general, it appears that there are many remaining challenges in the DL field. While TEI provides guidance in complex encoding situations, and has been advocated by the University of Michigan for electronic theses and dissertations, it is unclear how far the rest of the scholarly community will move toward the thorough markup and description of digital objects that characterizes humanistic study [Ren97]. Though such markup is valuable to support context dependent queries as well as electronic document preservation, it will only be generally feasible when there are less expensive tools and more efficient methods for adding in such markup and description. Then too the IR community must provide guidance regarding automatic indexing of marked up documents, metadata, full-text, multimedia streams, and complex hypermedia networks so the rich and varied content of DLs can be searched.

On a grander scale are the problems of handling worldwide DLs, in the context of varying collection principles, enormous difference in response time between local and remote servers, and the needs of users for different views [LFP98]. Thus, one type of scenario might deal with searching all dissertations worldwide, another might be concerned with finding recent results from a particular research group, a third might consider only freely available works in a particular specialty area, a fourth might deal with seeking the new works recently highly rated by a distributed group of close friends, and yet another might involve the most readable overviews in an unknown area.

Other key research challenges have been highlighted in various workshops aimed at establishing an agenda for investigation [LGM95]. Of central concern is covering the range from personal to global DLs, the so-called "scaling" problem. At the same time, the problem of interoperability must be faced [PCGMW98]. As argued earlier, we view the solution to these problems to be the acknowledgement of the role of 4S in the DL arena, and the focus of research and development on treating streams, structures, spaces and scenarios as first class objects and building blocks for DLs. We will continue to explore this approach in future work and believe that to the extent that integrated support for 4S is developed, real progress will be made toward the next generation of digital libraries.

Bibliography

[ABC$^+$98] T. M. Alrashid, J. A. Barker, B. S. Christian, S. C. Cox, M. W. Rabne, E. A. Slotta, and L. R. Upthegrove. Safeguarding Copyrighted Contents, Digital Libraries and Intellectual Property Management, CWRU's Rights Management System. *D-Lib Magazine*, April 1998. http://www.dlib.org/dlib/april98/04barker.html.

[ABO97] W. Y. Arms, C. Blanchi, and E. A. Overly. An Architecture for Information in Digital Libraries. *D-Lib Magazine*, February 1997. http://www.dlib.org/dlib/february97/cnri/02arms1.html.

[ACHK93] Y. Arens, C. Chee, C. Hsu, and C. Knoblock. Retrieving and Integrating Data from Multiple Information Sources. *Journal on Intelligent and Cooperative Information Systems*, 2(2):127–158, 1993.

[All97] National Humanities Alliance. Basic Principles for Managing Intellectual Property in the Digital Environment, March 1997. http://www-ninch.cni.org/ISSUES/COPYRIGHT/PRINCIPLES/NHA_Complete.html.

[Arm95] W. Y. Arms. Key Concepts in the Architecture of the Digital Library. *D-Lib Magazine*, July 1995. http://www.dlib.org/dlib/July95/07arms.html.

[Arm97] W. Y. Arms. Relaxing Assumptions about the Future of Digital Libraries: the Hare and the Tortoise. *D-Lib Magazine*, April 1997. http://www.dlib.org/dlib/april97/04arms.html.

[Arm98] W. Y. Arms. Implementing Policies for Access Management. *D-Lib Magazine*, February 1998. http://www.dlib.org/dlib/february98/arms/02arms.html.

[AUD96] Australian Digital Library Initiatives. *D-Lib Magazine*, December 1996. http://www.dlib.org/dlib/december96/12ian-nella.html.

[Bal97] M. Baldonado. SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests. In *Proceedings of CHI'97*, 1997.

[Bar97] D. Barber. OhioLINK: A Consortial Approach to Digital Library Management. *D-Lib Magazine*, April 1997. http://www.dlib.org/dlib/april97/04barber.html.

[BB97] Y. Bakos and E. Brynjolfsson. Bundling Information Goods: Pricing, Profits, and Efficiency. Technical report, MIT Center for Coordination Science, 1997.

[BCGP97] M. Baldonado, C. K. Chang, L. Gravano, and A. Paepcke. Metadata for digital libraries: architecture and design retionale. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 47–56, 1997.

[BD96] J. S. Brown and P. Duigid. The Social Life of Documents. *First Monday*, May 1996. http://www.firstmon-day.dk/issues/issue1/documents/.

[BDMW95] W. P. Birmingham, E. H. Durfee, T. Mullen, and M. P. Wellman. The distributed agent architecture of the University of Michigan Digital Library. *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995.

[Ber95] J. W. Berry. Digital libraries: new initiatives with world wide implications. In *Proceedings of 61st IFLA General Conference*, August 1995. http://www.nlc-bnc.ca/ifla/IV/ifla61/61-berjo.htm.

[Bis95] A. P. Bishop. Working Towards an Understanding of Digital Library Use: A Report on the User Research Efforts of the NSF/ARPA/NASA DLI Projects. *D-Lib Magazine*, October 1995. http://www.dlib.org/dlib/october95/10bishop.html.

[BKFS95] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing & Management*, 31(3):431–448, May–June 1995.

[Bor86] C. Borgman. "Why Are Online Catalogs Hard to Use? Lessons Learned from Information Retrieval Studies". *Journal of the American Society of Information Science*, 37:387–400, 1986.

[Bor96a] C. Borgman. UCLA-NSF Social Aspects of Digital Libraries Workshop, Feb. 1996. http://ucla.edu.

[Bor96b] C. Borgman. "Why Are Online Catalogs Still Hard to Use?". *Journal of the American Society of Information Science*, 47, July 1996.

[BP94] M. Buckland and C. Plaunt. On the Construction of Selection Systems. *Library Hi Tech*, 12:15–28, 1994.

[Bre96] G. Brett. An Integrated System for Distributed Information Services. *D-Lib Magazine*, December 1996. http://www.dlib.org/dlib/december96/dipps/12brett.html.

[BSH94] B. Bederson, L. Stead, and J. Hollan. Pad++: Advances in Multiscale Interfaces. In *Proc. SIGCHI '94*, 1994. See this and other papers at http://www.cs.umd.edu/hcil/pad++/papers/.

[Bus45] V. Bush. As We May Think. *The Atlantic Monthly*, 176(1):101–108, July 1945.

[CG94] W. B. Cavnar and A. M. Gillies. Data Retrieval and the Realities of Document Conversion. In *Proceedings of Digital Libraries '94*, 1994.

[CGMH+94] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In *Proceedings of IPSJ Conference*, pages 7–18, October 1994.

[Che97] D. R. Chestnutt. The Model Editions Partnership: "Smart Text" and Beyond. *D-Lib Magazine*, July/August 1997. http://www.dlib.org/dlib/july97/07chesnutt.html.

[CHW97] S. Y. Crawford, J. M. Hurd, and A. C. Weller. *From Print to Electronic: the Transformation of Scientific Communication.* New Jersey: Learned Information, 1997.

[CK96] S. Chapman and A. R. Kenney. Digital Conversion of Library Research Materials: a Case for Full Informational Capture. *D-Lib Magazine*, October 1996. http://www.dlib.org/dlib/october96/cornell/10chapman.html.

[CKP+95] S. B. Cousins, S. P. Ketchpel, A. Paepcke, H. Garcia-Molina, S. W. Hassan, and M. Roescheisen. InterPay: Managing Multiple Payment Mechanisms in Digital Libraries. In *Proceedings of Digital Libraries '95*, 1995.

[CLC95] J. P. Callan, Z. Lu, and W. B. Croft. Searching Distributed Collections with Inference Networks. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, 1995.

[CNR96] CNRI. Computer Science Technical Reports Project (CSTR), 1992–1996. http://www.cnri.reston.va.us/home/cstr.html.

[CNR98] CNRI. "Handle Home Page", 1998. http://www.handle.net/.

[CON] CONFU: The Conference on Fair Use. http://www.utsyst-em.edu/ogc/intellectualproperty/confu.htm.

[Cor] Infoseek Corporation. "Distributed Search Patent Home Page". http://software.infoseek.com/patents/dist_search/Default.htm.

[Cor98] IBM Corporation. IBM Digital Library, 1998. http://www.software.ibm.com/is/dig-lib/.

[CPW+97] S. B. Cousins, A. Paepcke, T. Winograd, E. A. Bier, and K. Pier. The digital library integrated task environment (DLITE). In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 142–151, July 1997.

[Cro95] W. B. Croft. What Do People Want from Information Retrieval? (The Top 10 Research Issues for Companies that Use and Sell IR Systems). *D-Lib Magazine*, November 1995. http://www.dlib.org/dlib/november95/11croft.html.

[CSM+97] S. Chang, J. R. Smith, H. J. Meng, H. Wang, and D. Zhong. Finding Images/Video in Large Archives: Columbia's Content-
Based Visual Query Project. *D-Lib Magazine*, February 1997. http://www.dlib.org/dlib/february97/columbia/02chang.html.

[CTS95] B. Cox, J. D. Tygar, and M. Sirbu. NetBill Security and Transaction Protocol. In *Proceedings of the First USENIX Workshop on Electronic Commerce*, 1995.

[DB94] P. Doty and A. P. Bishop. The National Information Infrastructure and Electronic Publishing: A Reflective Essay. *Journal of the American Society for Information Science*, 45(10):785–799, 1994.

[Dil95] A. Dillon. What is the shape of information? Human factors in the development and use of digital libraries. In Allerton Conference, 1995. http://edfu.lis.uiuc.edu/allerton/95/s4/dillon.html.

[DL94] J. R. Davis and C. Lagoze. A protocol and server for a distributed technical report library. Technical Report 94-1418, Cornell University Computer Science Department, June 1994.

[DMS+97] M. Dartois, A. Maeda, T. Sakaguchi, T. Fujita, S. Sugimoto, and K. Tabata. A Multilingual Electronic Text Collection of Folk Tales for Casual Users
Using Off-the-Shelf Browsers. *D-Lib Magazine*, October 1997. http://www.dlib.org/dlib/october97/sugimoto/10sugimoto.html.

[DO97] Network Development and MARC Standards Office. Dublin Core/MARC/GILS Crosswalk, July 1997. gopher://www.loc.gov/marc/dccross.html.

[Dre98] Daniel Dreilinger. "Savvy Seach Home Page", 1998. http://www.cs.colorado.edu/∼dreiling/smartform.html.

[EGL+95] R. Entlich, L. Garson, M. Lesk, L. Normore, J. Olsen, and S. Weibel. Making a Digital Library: The Chemistry Online Retrieval Experiment – A Summary of the CORE Project (1991-1995). *D-Lib Magazine*, December 1995. http://www.dlib.org/dlib/december95/briefings/12core.html.

[ERC] ERCIM. The ERCIM Digital Library Initiative. http://www.area.pi.cnr.it/ErcimDL/.

[FAFL95] E. A. Fox, R. M. Akscyn, R. K. Furuta, and J. J. Leggett. Digital libraries. *Communications of the ACM*, 38(4):22–28, April 1995.

[Fer97] I. A. Ferguson. Paying Their Way: Commercial Digital Libraries for the 21st Century. *D-Lib Magazine*, June 1997. http://www.dlib.orgdliba/june97/zuno/06ferguson.html.

[FFS+93] E. Fox, R. France, E. Sahle, A. Daoud, and B. Cline. Development of a Modern OPAC: From REVTOLC to MARIAN. In *Proc. 16th Annual Intern'l ACM SIGIR Conf. on R & D in Information Retrieval, SIGIR '93, Pittsburgh, PA, June 27 - July 1*, pages 248–259, 1993.

[FHH] E. A. Fox, L. S. Heath, and D. Hix. Project Envision Final Report: A User-Centered Database from the Computer Science Literature. http://ei.cs.vt.edu/papers/ENVreport/final.html.

[FHN+93] E. Fox, D. Hix, L. Nowell, D. Brueni, W. Wake, L. Heath, and D. Rao. Users, User Interfaces, and Objects: Envision, a Digital Library. *Journal of the American Society for Information Science*, 44(8):480–491, Sept. 1993.

[FL93] E. Fox and L. Lunin. Introduction and Overview to Perspectives on Digital Libraries: guest editor's introduction to special issue. *Journal of the American Society for Information Science*, 44(8):441–443, 1993.

[fLN] UK Office for Library and Information Networking. Electronic Libraries Programme. http://www.ukoln.ac.uk/services/elib/.

[FM98] E. A. Fox and G. Marchionini. Toward a Worldwide Digital Library. *Communications of the ACM*, 41(4):29–32, April 1998. http://purl.lib.vt.edu/dlib/pubs/CACM199804.

[Fou98] The International DOI Foundation. The Digital Object Identifier System, 1998. http://www.doi.org/about_the_doi.html.

[Fox] E. A. Fox. Self-study Courseware on Digital Libraries. http://ei.cs.vt.edu/~dlib/front.html.

[Fox93] E. A. Fox. Source Book on Digital Libraries. Technical Report TR-93-35, Virginia Polytechnic Institute and State University, 1993.

[Fri] A. Friedlander. D-lib Program: Research in Digital Libraries. http://www.dlib.org/.

[FSN+95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by Image and Video Content: The QBIC System. *IEEE Computer*, 28(9):23–32, 1995.

[Fuh98] N. Fuhr. DOLORES: A System for Logic-Based Retrieval of Multimedia Objects. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[Fur94] R. Furuta. Defining and Using Structure in Digital Documents. In *Proceedings of Digital Libraries '94*, 1994.

[Gay96] E. Gaynor. From MARC to Markup: SGML and Online Library Systems. *ALCTS Newsletter*, 7, 1996.

[GFA+94] H. Gladney, E. Fox, Z. Ahmed, R. Ashany, N. Belkin, and M. Zemankova. Digital library: Gross structure and requirements: Report from a March 1994 workshop. In *Proceedings of Digital Libraries '94*, pages 101–107, 1994. College Station, TX.

[GFHR97] D. A. Grossman, O. Frieder, D. O. Holmes, and D. C. Roberts. Integrating Structured Data and Text: A

Relational Approach. *Journal of the American Society for Information Science*, 48:122–132, 1997.

[GL97] H. M. Gladney and J. B. Lotspiech. Safeguarding Digital Library Contents and Users: Assuring Convenient Security and Data Quality. *D-Lib Magazine*, May 1997. http://www.dlib.org/dlib/may97/ibm/05gladney.html.

[Gla97] H. M. Gladney. Safeguarding Digital Library Contents and Users: Document Access Control. *D-Lib Magazine*, June 1997. http://www.dlib.org/dlib/june97/ibm/06gladney.html.

[GMS+98] H. Gladney, F. Mintzer, F. Schiattarella, J. Besco's, and M. Treu. Digital access to antiquities. *Communications of the ACM*, 41(4):49–57, April 1998.

[Gra95] Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. Allerton Conference 1995 Homepage, October 1995. http://edfu.lis.uiuc.edu/allerton/95/.

[Gra96] Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign. Allerton Conference 1996 Homepage, October 1996. http://edfu.lis.uiuc.edu/allerton/96/.

[Gra97] P. Graham. Glossary on Digital Library Terminology, 1997. Informal file sent by electronic mail for comments, available with permission of the author, if suitable attribution is made.

[Gro] Stanford DLI Group. Stanford University Digital Libraries Project. http://www-diglib.stanford.edu/diglib/.

[Gro98] New Zealand DL Group. The New Zealand Digital Library Project, 1998. http://www.nzdl.org/.

[Gu8] Jean-Claude Guédon. The virtual library: An oxymoron? NLM and MLA 1998 Leiter Lecture, National Library of Medicine, Bethesda, MD, May 1998.

[GWG96] S. Gauch, G. Wang, and M. Gomez. ProFusion: Intelligent Fusion from Multiple, Distributed Search Engines. *Journal of Universal Computing*, 2(9), September 1996.

[Har] S. Harum. Digital Library Initiative (DLI). http://dli.grainger.uiuc.edu/national.htm.

[Har96] S. P. Harter. What is a Digital Library? Definitions, Content, and Issues. In *Proceedings of KOLISS DL '96: International Conference on Digital Libraries and Information Services for the 21st Century, Seoul, Korea*, September 1996. http://php.indiana.edu/ harter/korea-paper.htm.

[Hea95] M. A. Hearst. TileBar: Visualization of Term Distribution Information in Full Text Information Access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 59–66, Denver, CO, 1995.

[Hea96] M. A. Hearst. Research in Support of Digital Libraries at Xerox PARC, Part I: The Changing Social Roles of Documents. *D-Lib Magazine*, May 1996. http://www.dlib.org/dlib/may96/05hearst.html.

[Her96] C. Hert. Information Retrieval: A Social Informatics Perspective. In Allerton Conference, 1996. http://php.indiana.edu/∼chert/allerton.html.

[HG96] D. A. Hull and G. Grefenstette. Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval. In *Proceeding of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57, 1996.

[HHN+95] L. Heath, D. Hix, L. Nowell, W. Wake, G. Averboch, and E. Fox. Envision: A User-Centered Database from the Computer Science Literature. *Communications of the ACM*, 38(4):52–53, April 1995.

[HKB96] M. A. Hearst, G. Kopec, and D. Brotsky. Research in Support of Digital Libraries at Xerox PARC, Part II: Paper and Digital Documents. *D-Lib Magazine*, June 1996. http://www.dlib.org/dlib/june96/hearst/06hearst.html.

[HLBB96] N. Van House, D. Levy, A. Bishop, and B. Buttenfield. User needs assessment and evaluation: issues and methods (workshop). In *Proceedings of the 1st ACM International Conference on Digital Libraries*, page 186, 1996.

[Hos97] P. Hoschka. Synchronized Multimedia Integration Language, 1997. http://www.w3.org/TR/WD-smil.

[Kan96] P. B. Kantor. Assessing the Factors Leading to Adoption of Digital Libraries, and Growth in Their Impacts: The Goldilocks Principle. In Allerton Conference, 1996. http://http://edfu.lis.uiuc.edu/allerton/96/kantor.html.

[KM93] P. Kilpelainen and H. Mannila. Retrieval from hierarchical texts by partial patterns. In *Proceeding of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214–222, 1993.

[KSR+97] R. Kengeri, C. D. Seals, H. P. Reddy, H. D. Harley, and E. A. Fox. Usability Study of Digital Libraries: ACM, IEEE-CS, NCSTRL, NDLTD, December 1997. http://fox.cs.vt.edu/∼fox/u/Usability.pdf.

[KW95] R. Kahn and R. Wilensky. A Framework for Distributed Digital Object Services. Technical report, CNRI, May 1995. http://www.cnri.reston.va.us/k-w.html.

[Lag96] C. Lagoze. The Warwick Framework: A Container Architecture for Diverse Sets of Metadata. *D-Lib Magazine*, July/August 1996. http://www.dlib.org/dlib/july96/lagoze/07lagoze.html.

[Lag98] Carl Lagoze. "NCSTRL Home Page", 1998. http://www.ncstrl.org.

[LBO88] D. M. Levy, D. C. Brotsky, and K. R. Olson. Formalizing the figural: aspects of a foundation for document manipulation. In *SIGDOC '88. Proceedings of the ACM Conference on Document Processing Systems*, pages 145–151, 1988.

[LE95] C. Lagoze and D. Ely. Implementation Issues in an Open Architecture Framework for Digital Object Services. Technical Report TR95-1540, Cornell Computer Science Department, 1995.

[Lei] B. Leiner. D-Lib Working Group on Digital Library Metrics. http://www.dlib.org/.

[Les97] M. Lesk. *Practical Digital Libraries: Books, Bytes & Bucks.* San Francisco: Morgan Kaufmann, 1997.

[Lev88] D. M. Levy. Topics in document research. In *SIGDOC '88. Proceedings of the ACM Conference on Document Processing Systems*, pages 187–193, 1988.

[Lev94] D. M. Levy. Fixed or fluid?: document stability and new media. In *Proceedings of the 1994 ACM European Conference on Hypermedia Technology*, pages 24–31, 1994.

[Lev97] D. M. Levy. I read the news today, oh boy: reading and attention in digital libraries. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 202–211, July 1997.

[LeV98] R. LeVan. Dublin Core and Z39.50. Draft version 1.2, February 1998. http://cypress.dev.oclc.org:12345/~rrl/docs/dublincoreandz3950.html.

[LFP98] C. Lagoze, D. Fielding, and S. Payette. Making Global Digital Libraries Work: Collection Services, Connectivity Regions, and Collection Views. In *Proceedings of the 3rd ACM International Conference on Digital Libraries*, 1998.

[LGM95] C. Lynch and H. Garcia-Molina. Interoperability, Scaling, and the Digital Libraries Research Agenda: A Report on the May 18-19, 1995, IITA Digital Libraries Workshop, August 1995. http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html.

[Lic65] J. C. R. Licklider. *Libraries of the Future*. Cambridge, Mass.: M.I.T. Press, 1965.

[LM95] D. M. Levy and C. C. Marshall. Going Digital: A Look at Assumptions Underlying Digital Libraries. *Communications of the ACM*, 38:77–84, April 1995.

[LMOY95] C. Lagoze, R. McGrath, E. Overly, and N. Yeager. A Design for Inter-Operable Secure Object Stores (ISOS). Technical Report TR95-1558, Cornell Computer Science Department, 1995.

[Maa95] Y. Maarek. Organizing documents to support browsing in digital libraries. In Allerton Conference, 1995. http://edfu.lis.uiuc.edu/allerton/95/s4/s4index.html.

[Mac90] I. A. Macleod. Storage and retrieval of structured documents. *Information Processing & Management*, 26(2):197–208, 1990.

[Mar97] C. C. Marshall. Annotation: from paper books to the digital library. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 131–141, 1997.

[MD94] F. L. Miksa and P. Doty. Intellectual Realities and the Digital Library. In *Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries*, June 1994.

[MFFS96] K. J. Maly, E. A. Fox, J. C. French, and A. L. Selman. Wide Area Technical Report Server. Published Online, July/August 1996. http://www.cs.odu.edu/WATERS/WATERS-paper.ps.

[Mil] E. J. Miller. CNI/OCLC Metadata Workshop: Workshop on Metadata for Networked Images, September. http://purl.oclc.org/metadata/image.

[Moe98] W. E. Moen. Accessing Distributed Cultural Heritage Information. *Communications of the ACM*, 41(4):45–48, April 1998.

[NDL] NDLTD: Networked Digital Library of Theses and Dissertations. http://www.ndltd.org/.

[NFL+95] P. J. Nuernberg, R. Furuta, J. J. Leggett, C. C. Marshall, and F. M. Shipman III. Digital Libraries: Issues and Architectures. In *Proceedings of Digital Libraries '95*, 1995.

[Now97] L. Nowell. *Graphical Encoding for Information Visualization: Using Icon Color, Shape, and Size to Convey Nominal and Quantitative Data*. Blacksburg, VA: Virginia Tech Dept. of Computer Science Doctoral Dissertation, 1997.

[Oar97] D. W. Oard. Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries. *D-Lib Magazine*, December 1997. http://www.dlib.org/dlib/december97/oard/12oard.html.

[oC97] Library of Congress. Metadata, Dublin Core and USMARC: A Review of Current Efforts. Technical Report MARBI Discussion Paper no. 99, Library of Congress, January 1997. gopher://marvel.loc.gov/00/.listarch/usmarc/dp99.doc.

[oC98] Library of Congress. MARC Home Page (Library of Congress), March 1998. http://lcweb.loc.gov/marc/marc.html.

[OCL] OCLC. Metadata Workshop II, April. http://www.oclc.org:5046/oclc/research/conferences/metadata2/.

[OCL98] OCLC. "PURL Home Page", 1998. http://purl.oclc.org/.

[OD96] D. W. Oard and B. J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, 1996.

[PCGM+96] A. Paepcke, S. B. Cousins, H. Garcia-Molina, S. W. Hassan, and S. P. Ketchpel. Using distributed objects for digital library interoperability. *IEEE Computer*, May 1996.

[PCGMW98] A. Paepcke, C. K. Chang, H. Garcia-Molina, and T. Winograd. Interoperability for Digital Libraries Worldwide. *Communications of the ACM*, 41(4):33–43, April 1998.

[Pet95] P. E. Peters. Digital Libraries Are Much More than Digitized Collections. *EDUCOM Review*, 30(4), 1995.

[PJ93] C. D. Paice and P. A. Jones. The Identification of Important Concepts in Highly Structured Technical Papers. In

Proceeding of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 69–78, 1993.

[PP97] C. Peters and E. Picchi. Across Languages, Across Cultures: Issues in Multilinguality and Digital Libraries. *D-Lib Magazine*, May 1997. http://www.dlib.org/dlib/may97/peters/05peters.html.

[Ren97] A. Renear. The Digital Library Research Agenda: What's Missing – and How Humanities Textbase Projects can Help. *D-Lib Magazine*, July/August 1997. http://www.dlib.org/dlib/july97/07renear.html.

[RMW95] M. Roscheisen, C. Mogensen, and T. Winograd. Interaction Design for Shared World-Wide Web Annotations. Stanford University, 1995. http://walrus.stanford.edu/diglib/pub/reports/brio-chi95.html.

[Ros96] A. Ross. Library Functions, Scholarly Communication, and the Foundation of the Digital Library: Laying Claim to the Control Zone. *The Library Quarterly*, 66:239–265, July 1996.

[Sam97] P. Samuelson. Copyright and digital libraries. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 113–114, 1997.

[SC96] B. Schatz and H. Chen. Building Large-Scale Digital Libraries: Guest editors' introduction to theme issue on the US Digital Library Initiative. *IEEE Computer*, May 1996. http://computer.org/computer/dli/.

[Sch96] L. Schamber. What Is a Document? Rethinking the Concept in Uneasy Times. *Journal of the American Society for Information Science*, 47:669–671, September 1996.

[Sch97] B. R. Schatz. Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science*, 275:327–335, January 1997.

[SE95] E. Selberg and O. Etzioni. Multi-Service Search and Comparison Using the MetaCrawler. *WWW4 Conference*, December 1995.

[SIN] Singapore Advanced Research and Education Network. http://www.singaren.net.sg/.

[Smi96] T. Smith. Meta-Information Environment of Digital Libraries. *D-Lib Magazine*, July/August 1996. http://www.dlib.org/dlib/july96/new/07smith.html.

[STA] STARTS: Stanford Protocol Proposal for Internet Search and Retrieval. http://www-db.stanford.edu/ gravano/starts_home.html.

[Ste97] M. Stefik. Trusted Systems. *Scientific American*, March 1997. http://www.sciam.com/0397issue/0397stefik.html.

[Sum95] K. Summers. Logical Structure Types for Documents. In *DAGS95: Electronic Publishing and the Information Superhighway*, 1995. http://www.cs.dartmouth.edu/~samr/DAGS95/Papers/summers.html.

[Swi] R. Swick. W3C Resource Description Framework Homepage. http://www.w3.org/RDF.

[Teaa] UCB DLI Team. UC Berkeley Digital Library Project. http://elib.cs.berkeley.edu/.

[Teab] UCSB DLI Team. Alexandria Digital Library. http://alexandria.sdc.ucsb.edu/.

[Teac] UIUC DLI Team. University of Illinois at Urbana-Champaign Digital Libraries. http://dli.grainger.uiuc.edu/default.htm.

[Tead] UMICH DLI Team. University of Michigan Digital Library Project. http://www.si.umich.edu/UMDL/.

[UC] UC Berkeley Digital Library Project. About MVD version 1.0alpha3. http://elib.cs.berkeley.edu/java/help/About.html.

[UNI98a] UNICODE.ORG. Unicode, 1998. http://www.unicode.org/.

[Uni98b] Carnegie Mellon University. Informedia, 1998. http://www.inform-edia.cs.cmu.edu/.

[Vin97] S. Vinoski. CORBA: Integrating Diverse Applications Within Distributed Heterogeneous Environments. *IEEE Communications Magazine*, 14(2), February 1997.

[VT97] E. M. Voorhees and R. M. Tong. Multiple search engines in database merging. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 93–102, 1997.

[W3C] W3C. The World Wide Web Consortium Homepage. http://www.w3.org.

[Wel37] H. G. Wells. World Brain: The Idea of a Permanent World Encyclopaedia. Contribution to the New Encyclopedie Francaise, 1937. http://sherlock.berkeley.edu/wells/world_brain.html.

[WGMD] S. Weibel, J. Godby, E. Miller, and R. Daniel. OCLC/NCSA Metadata Workshop Report: The Essential Elements of Network Object Description, June. http://purl.oclc.org/oclc/rsch/metadataI.

[Win95] T. Winograd. Conceptual Models for Comparison of Digital Library Systems and Approaches. Stanford Digital Library Project Working Paper, 1994–1995. http://www-diglib.stanford.edu/diglib/WP/PUBLIC/DOC13.html.

[Wis98] N. Wiseman. Implementing a National Access Management System for Electronic Services: Technology Alone Is Not Enough. *D-Lib Magazine*, March 1998. http://www.dlib.org/dlib/march98/wiseman/03wiseman.html.

[WMB94] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* New York: Van Nostrand Reinhold, 1994.

[Won97a] L. Wong. “BioKleisli architecture”, 1997. http://sdmc.krdl.org.sg/kleisli/kleisli/Architecture.html.

[Won97b] L. Wong. “BioKleisli home page”, 1997. http://corona.iss.nus.sg:8080/biokleisli.html.

[Woo] A. Wood. DC-4: NLA/DSTC/OCLC Dublin Core Down Under / The 4th Dublin Core Metadata Workshop, March. http://www.dstc.edu.au/DC4/.

[Z39] Z39.50 Maintenance Agency. The Library of Congress. http://lcweb.loc.gov/z3950/agency/.