Interoperability, Scaling, and the Digital Libraries Research Agenda:

A Report on the May 18-19, 1995
IITA Digital Libraries Workshop
August 22, 1995

Clifford Lynch
Hector Garcia-Molina

Introduction

This report summarizes the results of a workshop on Digital Libraries held under the auspices of the U.S. Government's Information Infrastructure Technology and Applications (IITA) Working Group in Reston, Virginia on May 18-19, 1995. The objective of the workshop was to refine the research agenda for digital libraries with

specific emphasis on issues of scaling and interoperability, and to identify the infrastructure developments needed to make progress on these issues.

While there have been a number of workshops and other meetings examining the broader questions of support for applications in the National Information Infrastructure (NII), we believe this was the first workshop that focused specifically on Digital Libraries in this context. In the past year, Digital Libraries have emerged as one of the central and most compelling applications enabled by the NII; numerous digital library research projects are underway, including six large-scale pilot projects that have been funded jointly by ARPA, NASA, and NSF. While Digital Libraries are now a vibrant research area, and also a field in which considerable commercial development is taking place (presaging the future economic importance of Digital Library technology to the United States), many new questions are emerging as a result of this flowering of research activity. Informed by insights gained from current research, this workshop offered an opportunity to consider questions such as interoperability objectives that might be defined among projects now underway.

The workshop was organized by Hector Garcia-Molina of Stanford University and Clifford Lynch of the University of California Office of the President. The IITA working group, which sponsored the meeting, reports to the National Science and Technology Council (NSTC) through the High Performance Computing, Communications, and Information Technology subcommittee of the Committee on Information and Communication. The workshop was attended by some 60 leading digital library researchers and developers and by representatives from a wide range of federal government organizations concerned with research and development and policy formulation related to digital libraries (see Appendix 1 for a roster of attendees).

Workshop attendees were asked to consider the following questions as a point of departure in developing the research agenda:

1. What is a Digital Library? How does it differ from an information repository or from today's World Wide Web? How many Digital Libraries will there be, and how will they interlink? How might this look to users?
2. What Digital Library infrastructure is needed? What does "infrastructure" consist of in this context and how does it differ from the broader applications support infrastructure for the emerging NII? What is the relationship between infrastructure and standards? Who will use this infrastructure? When must it be defined, and what parts are most urgently needed? How does the infrastructure relate to intellectual property management and publisher concerns?
3. How can a Digital Library be evaluated? How will we know in three to four years if current research projects have been successful in developing effective digital library services for their user communities?

To further frame and stimulate discussion, Hector Garcia-Molina prepared a position paper discussing the issues and distributed it prior to the workshop (see Appendix 2).

Participants spent the majority of the workshop in one of five groups; unlike many workshops, in which each group is assigned a different set of issues, here each group approached the full spectrum of questions from a specific, unique viewpoint and generated a summary of their discussions that reflected that viewpoint. After a presentation from the five group leaders representing each group's approach to the

issues, each participant selected his or her group. The five groups and their leaders were

Bill Arms,
Corporation for National Research Initiatives:
The Publishing Perspective

Michael Lesk,
Bellcore:
The Commercial Perspective

Bruce Schatz,
University of Illinois Urbana Champaign:
The Library Perspective

Mike Schwartz,
University of Colorado:
The Internet Perspective

Terry Smith,
University of California, Santa Barbara:
The Multimedia Perspective

The reports of these five groups appear in Appendix 3. This summary of the workshop extracts common themes and also key points of disagreement from the work of the five groups and places them in broader context. The report is not a consensus document; while it draws heavily on the five group reports and has also benefited greatly from comments from attendees, it does not attempt to reflect completely any of the five group reports.

This report addresses responses to the first two questions posed to the attendees (the definition of a digital library and infrastructure needs to support digital librariesand discusses the research agenda. The third question posed to the attendees -- how to evaluate Digital Library projects - -- did not receive much attention from most of the groups; it is to be the subject of a separate workshop on User Evaluation Methods to be held October 29-31 at the Allerton center under the auspices of The University of Illinois Urbana-Champaign and NSF. Some groups did identify the need for consistent instrumentation and data gathering across projects to facilitate evaluation. In addition, several groups stressed the need to make the transition from a systems technology framework to one driven by user access and collection organization in developing future digital library technology and systems. This view is perhaps most eloquently stated in the reports of the Internet working group and the Library working group.

Definitions and Roles of Digital Libraries

Considerable work has already been done on operational definitions of Digital Libraries and their relationship to traditional library institutions, as well as to the broader systems of scholarly and commercial publishing (see, for example, Communications of the ACM, April 1995). Much of the discussion in this workshop was motivated by questions of scaling, interoperability and needed support infrastructure.

Digital libraries were viewed as systems providing a community of users with coherent access to a large, organized repository of information and knowledge. One group made the provocative proposal that this organization of information was characterized by the absence of prior detailed knowledge of the uses of the information. The ability of the user to access, reorganize, and utilize this repository is enriched by the capabilities of digital technology; the Multimedia group provided particularly vivid examples of these possibilities.

Several groups pointed out that, in fact, digital libraries would, for the foreseeable future need to span both print and digital materials and that the central issue was to provide a coherent view of a very large collection of information. In this sense, an emphasis on content solely in digital format is too limiting. Really, the objective is to develop information systems providing access to a coherent collection of material, more and more of which will be in digital format as time goes on, and to fully exploit the opportunities that are offered by the materials that are in digital formats. Additionally, the comprehensiveness and value of the collection accessible through a digital library system can be strengthened by the ability to integrate materials in digital formats that have not been well-represented, easy to access, or effectively usable in traditional library collections, such as multimedia, geospatial data, or numerical datasets. There is, in reality, a very strong continuity between traditional library roles and missions and the objectives of digital library systems.

Participants in the workshop repeatedly underscored this continuity, and emphasized that the traditional library institutional missions of collection development, collection organization, access, and preservation must extend to the digital library environment. Digital libraries will be a component in the broader range of future library services, and librarians will play a central role in developing and managing digital libraries.

While there would be many digital repositories, a given digital library system should provide a coherent, consistent view of as many of these repositories as possible. From the user's perspective, to the extent possible, there should appear to be a single digital library system. Users increasingly have access to various types of digital collections and information systems: personal information resources, workgroup and organizational information collections and collaboration environments, and more public digital libraries. Defining the boundaries and characteristics of these information spaces and exploring ways in which they can be fused into a coherent whole is a central problem that cuts

across all aspects of the research agenda. From the user's perspective, the digital library system needs to extend smoothly from personal information resources, workgroup and organizational systems, and out to personal views of the content of more public digital libraries.

Some groups raised, but did not resolve, the question of the extent to which the digital library system should incorporate support for publishing, annotation, and integration of new information, and the extent to which additions to repositories within the digital library system should be mediated by librarians. It is clear that the development of digital libraries is closely linked to the changes that are occurring in modes of scientific and scholarly communication; the extent to which the digital library should actively embrace -- and perhaps even drive -- these changes remains to be fully explored.

Libraries -- digital or traditional -- exist to serve diverse purposes and constituencies. To some extent, each discipline, constituency, and collection creates its own organization of information. In the digital library world this differentiation among library collections, organization, and services may become more visible. One of the key challenges is to retain this diversity, which is responsive to unique constituencies, and at the same time permit information to be effectively shared across disciplines and constituencies. This is an essential component of the interoperability questions that formed a major focus for the workshop. Workshop participants represented many of these diverse perspectives: university research libraries, archives, libraries supporting teaching, public libraries, and libraries of the performing arts.

Defining Interoperability in the Digital Library Environment

Defining interoperability proved difficult. It is clear that this is still a central research problem in its own right, and one that merits continued attention. Discussions of infrastructure focused on common tools, enabling technologies and standards that would provide a basis for further exploration of interoperability issues, particularly by encouraging and facilitating the growth of digital libraries on the Internet. Considerable effort was spent on identifying infrastructure that was either unique or particularly critical to progress in digital libraries, as opposed to more general-purpose infrastructure that a range of NII applications, including digital libraries, might share. One clear theme was that an understanding of interoperability issues required operational experience which could only be gained by large-scale deployment of digital library systems. Speculation about interoperability in the abstract is of very limited value.

Participants expressed a full spectrum of views on interoperability. At one end of the spectrum is the use of common tools and interfaces that provide a superficial uniformity for navigation and access but rely almost entirely on human intelligence to provide any coherence of content. At the opposite end of the spectrum is deep semantic interoperability. The precise definition of deep semantic interoperability was the subject

of some debate, but deals with the ability of a user to access, consistently and coherently, similar (though autonomously defined and managed) classes of digital objects and services, distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations. It also extends beyond passive digital objects to actual services offered by specific digital library systems. Deep semantic interoperability is a "grand challenge" research problem; it is extraordinarily difficult, but of transcendent importance, if digital libraries are to live up to their long-term potential. An intermediate position between these two extremes advocates primarily syntactic interoperability (the interchange of metadata and the use of digital object transmission protocols and formats based on this metadata rather than simply common navigation, query, and viewing interfaces) as a means of providing limited coherence of content, supplemented by human interpretation.

Note that the term "digital object" here is intended only to describe, in the broadest sense, the type of information objects that may comprise a digital library -- textual, audio, video, numeric, computer programs, or multimedia composites of such components. It is not intended either to endorse or preclude an object-oriented architectural framework for digital library systems (in the sense of object-oriented programming or object-oriented databases, for example).

Infrastructure Requirements for Digital Library Research

The most urgent infrastructure need is to establish common schemes for the naming of digital objects, and the linking of these schemes to protocols for object transmission, metadata, and object type classifications. The consensus of the groups was that naming schemes for digital objects that allow global unique reference represented perhaps the most immediate infrastructure deployment priority in order to facilitate resource sharing, linkages, and interoperation among digital library systems and to facilitate scale-up of digital library prototypes. It was recognized that the design of large-scale naming systems and their integration into the larger digital library framework will continue to be an important research area, but that infrastructure support needs to be put in place quickly for at least an interim system, and that in fact experience with such an interim system would inform further research.

The deployment of a public key cryptosystem infrastructure -- including the development of a system of key servers and the definition of standards and protocols -- was also identified as essential to progress in digital libraries; this is necessary to support digital library needs in areas such as security and authentication, privacy, rights management, and payments for the use of intellectual property. While the need for public key cryptosystem infrastructure is hardly unique to digital libraries, the importance of the digital library services and components which depend on this infrastructure mean that its absence represents a significant barrier. In particular, until these problems are addressed, it seems unlikely that we will see commercial publishers and other information suppliers making large amounts of high-value copyrighted

information broadly available to digital library users. This in turn will constrain the development of research prototypes and may be a distorting factor in studies of user behavior.

Research Issues and Priorities

The working groups outlined a wide range of important research issues; most groups were less successful at prioritizing them, beyond the immediate infrastructure needs already discussed. The five key research areas that emerged from the workshop are described below; arguably, the first three are of most central and immediate importance, specifically to the development of digital libraries, though the long-term importance of research in the fifth area (economic, social, and legal issues) cannot be overemphasized. The distinctions among the five areas are to some extent arbitrary; for example, progress on interoperability (the first area) depends critically on progress in our ability to describe successfully objects and repositories (the second area).

1. Interoperability

The difficulty in defining the objectives for interoperability have already been discussed; clarifying these objectives, mapping the spectrum of interoperability, and establishing the key challenges at points along this specturm are key research issues in their own right.

The more technical interoperability research involve protocol design that supports a broad range of interaction types, inter-repository protocols, distributed search protocols and technologies (including the ability to search across heterogeneous databases with some level of semantic consistency), and object interchange protocols. Interoperability is not simply a matter of providing coherence among passive object repositories. Digital library systems offer a range of services, and these services must be projected in an interoperable fashion as well. One particular issue that emerged was that existing Internet protocols (such as HTTP, the basis of the World Wide Web) are clearly inadequate. Research must move beyond the current base of deployed protocols and systems. This raises complex questions about how to deploy prototype systems and the tradeoffs between advanced capabilities and ubiquity of access.

The practical question of the nature of the installed technology base and the need to support this installed base will increasingly frame and influence interoperability research. Access to digital libraries is not an end in itself for most users, but rather a support service; many will be willing to sacrifice advanced functionality for consistency, stability, and ability to use familiar, common access tools. Just as the installed base has become the greatest barrier to meaningful large-scale trials of new approaches that improve existing services (as opposed to providing entirely new services which do not compete with an installed base) in the overall Internet environment, user expectations and the installed base will ultimately impede progress in fundamental technology

research within the large-scale experiments necessary to gain insights into interoperability among digital libraries. Managing this tension will be a critical element in the continued development of the community's research agenda.

It should be noted that, at this relatively early stage in the evolution of digital library technology, it is of vital importance that projects strive for approaches that incorporate high functionality and extensibility. A high level of functionality in the standards and protocols used, even if not fully exploited initially, will postpone the time when the inertia of the installed base begins to confine research opportunities. Careful design of extensibility in digital library systems will facilitate continued research progress and understanding of the impact of new approaches on the user community without the need to attempt to displace an installed base.

2. Description of Objects and Repositories

In order to provide a coherent view of collections of digital objects, they must be described in a consistent fashion which can facilitate the use of mechanisms such as protocols that support distributed search and retrieval from disparate sources. Research in description of objects and collections of objects provides the foundation for effective interoperability. Interoperability at the level of deep semantics will require breakthroughs in description as well as retrieval, object interchange, and object retrieval protocols.

Issues here include the definition and use of metadata and its capture or computation from objects, the use of computed descriptions of objects, federation and integration of heterogeneous repositories with disparate semantics, clustering and automatic hierarchical organization of information, and algorithms for automatic rating, ranking, and evaluation of information quality, genre, and other properties. Other key issues involved knowledge representation and interchange, and the definition and interchange of ontologies for information context. The idea of active "information matchmaking" emerged in several group reports.

Research is also needed to understand the strengths and limitations of purely computer-based technologies for describing objects and repositories, and the appropriate roles for the efforts of human librarians and subject experts in the digital library context as a complement to these technology-based approaches.

3. Collection Management and Organization

Collection management and organization research is the area where traditional library missions and practices are reinterpreted for the digital library environment. Progress in this area is essential if digital library collections are to meet successfully the needs of their user communities.

Policies and methods for incorporating information resources on the network into managed collections, rights management, payment, and control issues were all identified as central problems in the management of digital collections. Approaches to replication and caching of information and their relationship to collection management in a distributed environment need careful examination. The authority and quality of content in digital libraries is of central concern to the user community; ensuring and identifying these attributes of content calls for research that spans both technical and organizational issues. Research is also needed to clarify the roles of librarians and institutions in defining and managing collections in the networked environment.

With the enhanced potential to support nontextual content effectively in the digital library environment, issues in nontextual and multimedia information capture, organization, and storage, indexing and retrieval are clearly key research areas. However, textual digital documents remain a vitally important research area in their own right, and are far from fully understood. The role of knowledge bases in digital libraries remains a poorly explored but potentially important question.

The preservation of digital content for long periods of time, across multiple generations of hardware and software technologies and standards is essential in the creation of effective digital libraries. This is an extraordinarily difficult research problem which has not received sufficient attention.

4. User Interfaces and Human-Computer Interaction

While user interfaces and human-computer interaction issues are an extensive field of research in their own right, there are some specific problems that are central to progress in digital libraries.

Display of information, visualization and navigation of large information collections, and linkages to information manipulation/analysis tools were identified as key areas for research. The use of more sophisticated models of user behavior and needs in long-term interactions with digital library systems is a potentially fruitful area for research. The necessity for a more comprehensive understanding of user needs, objectives, and behavior in employing digital library systems was stressed repeatedly as a basis for designing effective systems. Finally, it was observed that digital library systems must become far more effective in adapting to variations in the capabilities of user workstations and network connections (bandwidth) in presenting appropriate user interfaces; new technologies such as personal digital assistants and nomadic computing models will emphasize this need.

5. Economic, Social, and Legal Issues

Digital libraries are not simply technological constructs; they exist within a rich legal, social, and economic context, and will succeed only to the extent that they meet these

broader needs. Rights management, economic models for the use of electronic information, and billing systems to support these economic models will be needed. User privacy needs to be carefully considered. There are complex policy issues related to collection development and management, and preservation and archiving. Existing library practice may shed some light on these questions. The social context of digital documents, including authorship, ownership, the act of publication, versions, authenticity, and integrity require a better understanding. Research in all of these areas will also be needed if digital libraries are to be successful.

Conclusions

This workshop has made substantial progress in refining and focusing a research agenda for digital libraries, as well as in developing insights into questions about interoperability among digital libraries and the infrastructure necessary to support such interoperability. Interoperability is likely to continue to be a useful organizing theme in refining this agenda in the coming years. The outcomes of the workshop also suggest that a focus on broad architectural issues in digital libraries will be fruitful. Several working groups commented on the need to develop component software strategies that would facilitate the transfer of technology among the current digital library pilot projects and from these projects to other new digital library research efforts. The Internet working group went further in suggesting that the development of a broadly available software base for the digital library community would contribute to rapid progress, and we believe that this suggestion deserves careful consideration.

Scaling was identified as a major area of concern. The common vision is one of tens of thousands of repositories of digital information that are autonomously managed yet integrated into what users view as a coherent digital library system. Accommodating this very large number of repositories -- a very different environment than that in which today's handful of pilot projects operate -- will clearly have major implications for infrastructure definition and design. We must move rapidly towards an infrastructure that can support and facilitate research towards this common vision. The full range of issues here are unclear. Some immediate needs are evident; these are reflected in the emphasis on establishing naming systems for digital objects as a high priority, for example.

We don't know how to approach scaling as a research question other than to build upon experience with the Internet. However, attention to scaling as a research theme is essential and may help in further clarifying infrastructure needs and priorities, as well as informing work in all areas of the research agenda outlined above. For example, reliability questions are poorly understood; in a sufficiently large system, some components will inevitably be out of service during the processing of any given query. The need to support large-scale deployment projects (in terms of size of user community, number of objects, and number of repositories) and to study subsequently the effectiveness and use of such systems was emphasized repeatedly. It is clear that

limited deployment of prototype systems will not suffice if we are to understand understand the research questions involved in digital libraries.

Research in scale-up is very difficult to perform except by building and deploying a large-scale digital library system. Establishing infrastructure and tools to facilitate experimentation with large-scale systems is essential, as is funding to study use and behavior of large-scale systems once deployed through this infrastructure. The Internet as a context for deploying digital library systems offers an unprecedented opportunity -- not only technically by providing connectivity to an enormous potential user base but also culturally, given the Internet community's models and traditions of technology diffusion through the distribution of publicly available prototype software -- to move ahead large-scale experiments. Research efforts should exploit these opportunities.

Finally, it seems clear that the inevitable presence of large amounts of commercially valuable, proprietary information in the future -- which can be viewed as another form of scale-up in digital libraries -- will also shape the research agenda in new ways. The near-term focus is on overcoming the infrastructural barriers to supporting proprietary information (such as authentication, billing, and rights management). There are research issues in the design of such an infrastructure, but also operational and policy problems impeding deployment. While some of the research issues are complex and will require ongoing exploration, putting at least the first steps towards the necessary infrastructure in place to accommodate such commercially valuable information is a high priority in advancing the research agenda and addressing scale-up issues. It will also stimulate commercial developments that will complement existing research initiatives. The development of an increasingly rich marketplace of information resources under a wide range of economic and legal constraints will create new opportunities in all areas of the research agenda presented above, and will allow us to explore vital new research questions in the development of description, navigation, access, and resource discovery technologies and systems that can function in this broader environment.

Appendix 1: Participants at the IITA Workshop

Name	Organization
Allen, Robert B.	Bellcore
Anderson, Greg	MIT
Arms, William Y.	CNRI
Becker, Herbert	Library of Congress
Behrens, Cliff	Bellcore
Bergman, Larry	IBM
Brett, George H. II	CNIDR-ITD-MCNC
Browne, Shirley	U Tennessee, Knoxville
Chen, Su-shing	NSF
Chien, Y.T.	NSF
Crowder, Grace	UMBC
Daddio, Ernest	NOAA
Davis, James R.	Xerox: Des. Res. Inst.
Denning, Peter J.	George Mason Univ.
Finin, Timothy	UMBC

Fischer, Christoph	UC, Alexandria
Fordham, Brad	UMBC
Fox, Edward A.	Virginia Tech
Frank, Randy	U Michigan
French, James	U Virginia
Funk, Steven	NSF
Garcia-Molina, Hector	Stanford Univ.
Garrett, John R.	CNRI
Gifford, David	MIT
Glick, Norm	NSA
Griffin, Steve	NSF
Howe, Sally	NCO
Hurley, Bernard	UC Berkeley Library
Hylton, Jeremy	MIT
Jacobson, Robert	Chronicle of Higher Ed.
Jefferson, David	NIST
Kahn, Robert E.	CNRI
Ketchpel, Steven	Stanford Univ.
Klensin, John	MCI
Krafft, Dean B.	Cornell Univ.

Lal, Nand	NASA
Lannom, Larry	DynCorp - ATS
Lasher, Rebecca	Stanford Univ.
Leiner, Barry	ARPA
Lesk, Michael	Bellcore
Likens, William	NASA Ames
Lynch, Clifford	UCOP
Michelson, Avra	The Mitre Corporation
Miya, Eugene	NASA
Neches, Bob	ARPA
Overman, Ronald	NSF
Preston, Cecilia	Self
Rundensteiner, Elke	U Michigan
Schatz, Bruce	NCSA/U Illinois
Schwartz, Mike	University of Colorado
Shackelford, Walter	US EPA/RTP, NC
Sirbu, Marvin	CMU
Smith, Terence	UC, Santa Barbara
Sunkara, Anil	GMU
Terstriep, Jeff	NCSA/U Illinois

Thomas, Mary Augusta	Smithonsian Institution
Turek, John J.	IBM
Valluri, Sven	GMU
Vassallo, Paul	NIST
Wactlar, Howard D.	CMU
Wake, William	Virginia Tech
Weibel, Stuart	OCLC
Weider, Chris	Bunyip
Westermeyer, Beverly	Smithsonian Institution
Wilensky, Robert	UC Berkeley
Winograd, Terry	Stanford Univ.
Zijlstra, Jaco	Elsevier

Appendix 2: A "Strawman" Report

for the

IITA Digital Libraries Workshop

Hector Garcia-Molina

This document was developed -- in the form of a workshop report -- prior to the Workshop, as a means of focusing discussion and providing some positions to which the attendees might react.

Starting Point: NII Report

On February 28 and March 1, 1994, various organizations (including the Computing Research Association, the Council on Competitiveness, and the Cross Industry Working Team) sponsored a comprehensive workshop on the research and infrastructure needs for the emerging National Information Infrastructure (NII). The report from that workshop (available from EDUCOM at nii-forum@educom.edu) discusses in great detail the principal challenges and makes extensive research and development recommendations. Since digital libraries are an important component of the NII and share many of the same challenges, we do not wish here to redevelop the same type of comprehensive list of research issues. Instead, we will simply summarize the main findings of the NII report, refer the reader to the full report for additional information, and focus here on highlighting the differences and the needed library-specific priorities.

According to the NII report, it will support advanced applications by providing: (a) thousands of information repositories, (b) wide bandwidth data networks and information appliances, and (c) advanced communications and information access services.

The report identifies critical technical challenges in the following areas:

(1) Network components that can handle voice, video, and text simultaneously, and can operate seamlessly.

(2) Information appliances and services that can provide access and services in a scalable, efficient, and interoperable way.

(3) Information access techniques that can enable efficient searches of large distributed information repositories, making the myriad of information resources understandable.

(4) Multimedia information technologies that can, for example, synchronize and integrate real-time delivery of voice and video, and can support search and retrieval based on image content.

(5) Infrastructure for application development that can provide common solutions.

(6) Technologies that are dependable and manageable.

(7) Technologies that are easy to use and services that are accessible by users with widely varying skills, experiences, abilities, and backgrounds.

(8) Interoperability among heterogeneous systems will be required on an unprecedented scale.

(9) Security and privacy technologies that are easy to use and provide appropriate levels of security to suit the requirements, cost constraints, and convenience of the end user.

(10) Technologies and services that provide portability, mobility, and ubiquity .

The NII Report goes on to say that the Federal Government has at least two roles to play in the development of the NII:

- Devise and implement effective policies that enable the development of a coherent infrastructure, while allowing competitive market forces to drive the creation of products and services;
- Foster and support a long-range research program that addresses the many technical problems.

In addition, the report states that NII research should be guided by pilot projects to develop appropriate technologies, evaluate them, and get them into the hands of users.

What are Digital Libraries?

Digital libraries provide the critical information management technology for the NII, and at the same time represent its primary information and knowledge repositories. In other words, digital libraries are the core of the NII. The information services, search facilities, and multimedia technologies of items (2), (3), and (4) above constitute the digital libraries technologies. Like other NII technologies, they must provide for dependability, manageability, ease of use, interoperability, and security and privacy (items 6, 7, 8, 9). The information repositories mentioned in (a) at the beginning of the previous section constitute the contents of the digital libraries.

Notice that this new notion of a "library" is broader than the traditional view. In particular, information does not have to be processed by a human (e.g., catalogued,

approved, edited) before it can become part of the library. Nevertheless, we expect that there will be some repositories with controlled collections.

Also notice we are using the plural term "digital libraries." We do not expect to see a single digital library in the NII. Each information repository will be managed separately, possibly with different technologies, and hence each will constitute a digital library. However, it will be possible to (and actually critical) to integrate "virtually" separate libraries into a single one, by providing a software layer on top of the libraries.

Digital Libraries Research Agenda

Given our definition, the research agenda for digital libraries is the research agenda of the NII. However, it is useful to reorganize the research topics listed above to highlight the critical library specific issues. Thus, we propose the following classification of digital libraries research and development problems.

Conceptually, there is a single problem to be addressed by digital libraries -- that of information discovery. How do we put a user "in touch" with the information that is of interest to him? All other problems are subsets. We may have to pay to get the user his information (item DL8 below); we may have to scan in the information before we can provide it (item DL7 below); or we may have to search through heterogeneous repositories (item DL4 below). And, incidentally, the information that a user wants may not be in a repository, but may have to be created by a service. But at the highest level, our goal is "information matchmaking."

To provide users with the information they want, we need to address the following interrelated research:

(DL1) Understanding user needs. Before we can find the information, we need to know what the user wants. We need to develop expressive query languages and user interfaces that allow a user to describe, naturally and accurately, his information needs.

(DL2) Resource discovery. An initial step in the actual matchmaking process is to find out what digital libraries are available and may have relevant information or services. The challenge is to characterize the information contents (e.g., meta-information) and service capabilities of libraries in a compact and meaningful way.

(DL3) Information retrieval. If the desired information is in one or more repositories, we have to find it efficiently, without also retrieving irrelevant information, and without missing anything relevant. To do this, we need mechanisms for identifying the relevance of information to a given user request, as well as access structures to perform the identification and retrieval efficiently.

(DL4) Heterogeneity. Information will be stored or provided by digital libraries using different commands, and will be returned using different representations. Standardized

commands, protocols, and models will help, but we expect that there will always be a significant level of heterogeneity. Thus, we need to develop technology for interoperation between digital libraries, that will allow searches and interactions to span multiple libraries.

(DL5) Scale and distribution. We expect dramatic growth in the number of digital libraries, the volume of information in them, the number of users, and the number of requests. We need to develop technologies (for the rest of the items addressed in this list) that will scale and that will work efficiently in spite of the wide geographic distribution (and possible temporary disconnection) of the information resources.

(DL6) Information input and collection building. Clearly, the information in the digital libraries must enter the system somehow. Some of it will come from conventional media such as printed documents or videotape. We need to develop mechanisms for digitizing this information accurately and efficiently. Other sources of information will be the library users themselves, and hence we also need natural and easy-to-use mechanisms for generating new information, as well as for annotating or modifying existing information.

(DL7) Preservation. Some of the digital information needs to be preserved for future generations. The key challenge is to ensure that at least one copy of the medium that holds the information (e.g., the tape or CD-ROM) physically survives, that that medium can be read in the future, and that the digital information can be interpreted properly.

(DL8) Security, privacy, and charging. Before providers will make their information available, they need to be assured that they will be compensated economically, if so desired, and that the information will not be accessible to unauthorized users. We need to develop schemes for protecting information without unduly interfering desired information sharing. We also need mechanisms for tracking access and charging for it, in a way that encourages providers to make even more useful information available.

Research Priorities

We believe that all of the research problems of the previous section are important and must be addressed. However, from a short- to medium-term perspective (e.g., 5 years), we believe there are some problems that are more critical to the potential of digital libraries, and to the continued support and interest from funding agencies and the public in general.

Currently, commercial information vendors such as Knight-Ridder's Dialog and Mead Data provide a basic but very useful level of functionality over significant collections. For example, Dialog provides access to over 400 databases, many of which full text. They also provide some simple but powerful resource discovery tools for identifying relevant databases in their collections.

Most critical. We believe that it is critical to provide at least this level of functionality (e.g., boolean queries) over heterogeneous collections, geographically distributed at several organizations, with some level of resource discovery. This will let us demonstrate that investments in digital libraries will be sharable across organizations. At the same time, it will let us offer a useful service. Even though the offered user interfaces, query languages, and so on, would be limited, we know from experience with the commercial vendors that they can provide an extremely useful service. For this work, the critical research problems that need to be addressed are those of items DL2 and DL4.

Next most critical. Next in priority (short to medium term) are scalability (DL5) and security and charging (DL8) problems. Again, in terms of rapidly demonstrating the potential of digital libraries, it is important to provide access to valuable and useful information, which requires security and charging mechanisms. Similarly, we need to demonstrate access to significant volumes of information. Current commercial systems are already struggling with scalability problems, so if we want to go beyond the sizes of their collections, it is important to develop scalable mechanisms.

Not critical. The rest of the research topics are not on our critical list. This does not mean they are not important. They are, and it is essential to continue research on those issues. However, we believe that the short-term payoff from being able to, say, scan more documents (beyond what is already being done commercially), or to provide slightly better precision and recall (over what current information retrieval techniques achieve) is lower than the payoff from solving the other problems.

Infrastructure

A single organization can build a single digital library. However, to share information across libraries, it is important to have a common infrastructure that facilitates such sharing. Furthermore, this same infrastructure can also support sharing of technologies used to build the digital libraries.

In our view, the infrastructure for digital libraries should have the following components:

(IN1) Shared information representation models, service representation models, and access protocols. These will facilitate the sharing of information and services across digital libraries.

(IN2) Information "content" sharing agreements. This will take the form of communities of organizations that agree to share their collections. Initially, the sharing may be free, but eventually the communities will institute common charging schemes. The communities will also provide rules for having additional members join.

(IN3) Resource directories. To facilitate resource discovery, the infrastructure should provide "directories" that describe available information resources and the models and protocols they follow, and characterize their contents. Similarly, technology directories could be provided to help in sharing of developed technologies.

(IN4) Coordination forum. The goal of this forum is to coordinate national research and development activities. It could provide help in organizing workshops, conferences, and newsletters, whose goal would be to define further the national digital libraries infrastructure. It could also provide a mechanism for circulating and commenting on proposed "standards," similar to the RFC mechanism.

It is important to understand that the digital libraries infrastructure is neither centralized nor a single entity. It is a collection of agreements and distributed (or replicated) meta-knowledge repositories that support digital libraries research and development.

The infrastructure is also not a "pilot project" as described in the NII Report. The pilot projects will build some of the initial digital libraries (or will integrate collections of libraries) aided by the libraries infrastructure. Clearly, it is very important to have such pilot projects to demonstrate potential and achieved results.

Since the digital libraries infrastructure plays such an important role, we believe it is essential to put its initial components in place as soon as possible, say within the next two years.

Evaluation

Many of the national challenges have clearly quantifiable goals, and this makes it easy to evaluate progress. For example, one can measure the number of instructions per second required in a new processor design, or the plasma temperature needed for nuclear fusion. With digital libraries, we are not as fortunate.

There are of course some metrics that can be used, but we think their use is limited. For instance, one can count the number of documents scanned into a library, or the number of queries run at a server. But the raw number of documents or queries does not reflect the quality of the information. Furthermore, it is not easy to come up with an overall meaningful target for those metrics.

Traditional information retrieval metrics such as precision and recall also are limited when applied to huge heterogeneous collections of information. In particular, it is difficult to say what the representative queries are, and it is hard to know how many documents were missed by a particular search. (It could take a lifetime to examine manually all reachable documents to see which were actually relevant.)

So, even though the traditional metrics will be useful in some cases, we will not be able to rely on them to evaluate progress in digital libraries. In a way, the situation is analogous to the World Wide Web (WWW) a few years ago. At that time, it would have been impossible to predict how successful the WWW would be, and how it would be used. A formal evaluation of the WWW a few years back could have easily concluded that it was not useful, for example, in terms of number of queries run, or in terms of how useful the documents were to a particular search.

Given the limitations of traditional metrics, we believe that the most promising approach is to evaluate informally what users are doing, and what services they want. This could involve informal interviews or feedback sessions with researchers implementing new services or libraries.

Appendix 3: Report of the Library Working Groups

Report of the Publishing Perspective Working Group

IITA Digital Libraries Workshop

William Arms

A. Introduction

The working group considered research in digital libraries from the perspective of all creators, originators, editors, rights-holders, and publishers of material in the digital library. This first section describes some underlying issues behind the group's recommendations in later sections.

1. Organizations and publishing. 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.

Publishers and libraries perform many functions that go far beyond the creation and management of physical items. Examples range from editing and refereeing, to abstracting and indexing. We believe, therefore, that the roles of libraries and publishers will continue even as their specific practices change with the technology. The new forms of publishing and library organizations that will emerge are open to speculation, but we believe they will be shaped by natural market forces and are not a topic for research.

2. The social, economic, and legal frameworks. The research agenda in digital libraries should not be restricted to technical areas. The social, economic, and legal questions are too important ignore.

Publishing and libraries exist in a social and economic framework, where the operating rules are codified by a network of laws and business relationships. One of the greatest forces inhibiting the rapid deployment of digital libraries is the need to modify this framework. Two key topics are understanding how copyright functions in digital libraries and how the various costs will be covered.

3. Ease of use. The benefits of digital libraries will not be appreciated unless they are easy to use effectively. Some experienced people already meet a high percentage of their library needs with networked information. These individuals have development heuristics for finding and evaluating information, but their practices are difficult to describe and teach to less-experienced users.

The ease of use will develop naturally when the rate of change slows down, conventions develop, and less successful systems are withdrawn. However, natural progression alone is unlikely to be sufficient by itself.

B. The Need for Research in Digital Libraries

1. What is a digital library? A library is a system in which large volumes of information from many sources are assembled, organized, and made accessible without detailed prior knowledge of that information's use.

A digital library is a library where the information is stored and processed in digital formats. (The World Wide Web is a simple example.) The digital library system will contain many components, with different technical underpinnings, managed by many organizations.

2. Why is research in digital libraries important? Libraries are important because: (1) they retain the social, scientific, legal and other records of our culture; (2) they provide wide, inexpensive access; and (3) they provide access to this record supporting economic and cultural development.

Digital Libraries are important because: (1) they have the potential to provide library services more effectively; (2) they can store information that exists only in digital form; and (3) they provide new opportunities to organize and disseminate information.

Research in digital libraries is needed to tackle the hard, technical questions that must be resolved for the essential functions of libraries to continue into the digital age and to realize their new potential. It will be essential, for example, to develop ways for independently developed digital libraries to interoperate.

In addition to supporting the development of digital libraries, research in this area will necessarily address core problems in network computing, that are key to the development of many other areas of national interest, such as electronic commerce.

C. Needs of Originators, Creators, and Publishers

In this section, the value and potential of digital libraries is explored through the needs of "originators." This word is used to describe all forms of creators and publishers -- people or organizations who generate, organize, or otherwise create material that they wish to distribute in digital form.

1. **Dissemination**. The basic need of originators is an infrastructure that supports widespread distribution of digital library objects within a simple framework.
2. **Access**. The second need is a library system that provides access to these objects. This requires tools for finding material, such as catalogs and indexes, and systems for managing access, such as authentication and payment tools.
3. **Archiving**. Originators usually expect that their material will be preserved over long periods of time. They require systems that will ensure access despite changes in organizations and technology.
4. **Control**. When originators distribute their material, they usually require some control over how it is used. This control varies from placing the material in the public domain to tight restrictions on access. It includes decisions about who can alter the material and other considerations of integrity.
5. **Legal and social**. A society that enables orderly dissemination is crucial. Legal areas include copyright and other intellectual property, privacy, obscenity, and libel. Business practices include acceptable use policies, codes of practice, and standard contracts.
6. **Tools**. Originators need computers, networks, and software tools for the straightforward and orderly creation, distribution, and access to all types of information.

D. The Top Research Topics

This section lists key topics where research can contribute to the development of digital libraries, satisfying the needs described in Section C.

1. General

a) Scale and complexity. Many of the problems faced by digital libraries are already solved on a small scale, but deployment on a large scale is more difficult. It is a deep research problem to create widely dispersed, distributed systems, developed by many organizations, across national boundaries, with technology from many sources.

b) Integration of digital and conventional libraries. Digital libraries and conventional libraries will coexist indefinitely. Two major research topics are: (1) how to build integrated libraries where some of the material is in conventional formats, notably paper, and some is digital; and (2) how to build indexing and abstracting service that combine the effectiveness of human and computer systems.

c) Measurement of effectiveness. Research in libraries, including digital libraries, lacks measures of effectiveness. For example, the classical measures of recall and precision are widely disliked, yet no alternatives exist.

d) Tools for creating and managing digital libraries. At present, digital libraries are very labor intensive, as are traditional libraries. Tools are needed to simplify the tasks of creating, managing, and using them.

2. Content

a) Text. Text retains a special place in the digital library, because it is the primary medium of human communication. There are many complex research questions about creating digital text, organizing it for retrieval and display, and combining it with other material.

b) Active library objects. The digital medium allows for new types of library objects such as software, simulations, animations, movies, slide shows, and sound tracks, with new ways to structure material, such as hypertext. Active library objects enable the form of an object with which the user interacts to be very different from the stored form.

c) Integration of mixed media. Much of the development of multi-media and mixed media is happening independently. Digital library research will integrate these materials and develop systems to provide access to them.

3. The long term

a) Preservation. To preserve material in the digital library is to retain its content over long periods, without necessarily retaining the media, the format, or other methods of representing the content. Preservation of material over very long periods is one of the defining characteristics of libraries, archives, and museums.

To preserve digital library material, more than bits must be retained. The library must be able to recognize formats and have the technical ability to display, perform, or otherwise interact with materials originally developed for long-dead computer systems written in forgotten programming languages.

b) Naming. Naming systems are a key component of libraries. They need to support the access to materials long after their creators cease to exist. The problems divide into two sections-- naming individual digital objects and naming works -- which may be composed of many digital objects. Each part of the problem has to deal with both static and dynamic objects and to resolve the issues of equivalence.

4. Computer systems

The research problems here concern library and publishing functions in distributed systems.

a) Repository access protocols. No existing protocol for communication between library client and the various types of repositories and archives is adequate.

b) Security and authentication. Security and authentication are essential. General developments for the National Information Infrastructure (NII) may create services that digital libraries can use, but protocols that deal with the specific practices of publishers and libraries must be developed.

c) Mixed environments. Users of the digital library will have a huge variety of computers, connected over widely differing communications channels, operating in different social and legal frameworks. The digital library must adapt to these mixed environments, providing suitable services with good performance.

5. Social

The social aspects of digital libraries are some of the most difficult. Here are two vital topics:

a) Human-computer interaction. The problem in human-computer interaction lies in the structuring of information sources and services. Users must not be obliged to serve a long apprenticeship before they can make effective use of the digital library.

b) Rights management. Rights management is a key part of control. Rights in intellectual property must be identified and tracked. Rights management can be linked with questions of payment.

E. Areas Not Recommended for Research.

The list of research topics in Section D is long, but many important topics have been omitted.

1. Scope. Some important areas can be left to normal developments. In these areas, the main concern is that the interests of existing organizations should not inhibit entrepreneurs and innovation.

Some topics fall cleanly within other research fields. For example, we do not recommend specific research in networks or multimedia, except in areas where digital libraries have special needs.

2. Difficulty. Some areas are important but so complex that we see little hope of successful research. Some of the social and economic questions fall in this area. Other areas are so straightforward that they do not justify specific research.

3. Unserved public. Libraries have been a great contributor to the continuing openness of society. Although we do not recommend specific research in this area, digital libraries must serve the nation and the world as broadly as possible and not be confined to people who have advanced equipment and resources.

4. Transfer from research. The research topics proposed have a bias in favor of long-term, fundamental research. This must be combined with more effective methods of technology transfer.

F. Fragmentation and Coordination of Research

Research into digital libraries is poorly served by the existing and planned conferences and journals. The community needs a small number of high-quality methods of exchanging research ideas and develop standards.

Each of the federal funding agencies sponsoring digital library research has its own mission. These do not always map cleanly onto the needs for research and advanced development. Interagency cooperation will be as valuable as it has been in the overall HPCC program.

We encourage the continuing development of a framework for organizing, coordinating, and communicating of digital library research. This will act as a bridge with organizations engaged in implementation.

Report of the Commercial Perspective Working Group

IITA Digital Libraries Workshop

Rashomon Meets Digital Libraries

Michael Lesk

Commercial exploitation of the Internet, the Web, and digital libraries is rushing towards us. What important limits on economic use of digital libraries could be alleviated by new research? This was the theme of our group discussion.

We discussed briefly the definitions of "digital library" and "infrastructure." Key ingredients in a "library" are organization and access tools, in addition to piles of bits. Publishers must be involved in the infrastructure debate, and we note approvingly that the existing digital library projects all involve publishers in cooperative roles. But a key question is what new research can do to assist the broad range of economic impacts from digital libraries, not just the effects on publishers or libraries.

Commercial organizations can exploit digital libraries in many ways. They can obtain information from libraries to help their operations; they can use library software to manage their own information; and they can sell services based on the delivery of information. The information industry is in a state of flux and rapid development, taking advantage of the quick progress in computer networks. However, there are still obstacles that new research might overcome, and opportunities that new research

might provide, which would both facilitate the development of the industry. Many of these issues are targeted at opening the markets for information services and digital libraries, assuring all companies an equal chance at participating in these efforts.

We assume a familiar context: Digital libraries are collections of byte streams, stored in ways that permit users to retrieve and view information that they want. There will be many such collections, distributed around the United States and the world. Some will be freshly created material, and some will be converted from other forms. The various repositories of digital information will be connected, and users will be able to view the multiple repositories as if they were one big library, even though they will not be owned by one organization. There are many technical advances needed to bring about this world, and many are being developed right now, funded by various governmental and private organizations. This report points to some new areas where additional research is of highest priority.

The most important issues are about basic infrastructure support. In the context of digital libraries, these issues include preservation, collection, location, and mapping of information names ("handles") to locations. Preservation includes technology for long-term stable storage, techniques for managing archiving practice and refreshing, ways of verifying the integrity of stored files, and methods of tracking the number of copies of files in distributed repositories. For example, it should be possible to design file repositories to keep a count of the number of copies of each file, even though they are geographically spread, and to maintain a threshold so that a file can not be deleted if this would reduce the number of copies below the stated threshold. Collection involves technology to select material for long-term storage, to provide assistance in cataloging and storing the material, and to describe the collection for use in information navigation and retrieval. Location and location mapping require technologies for rapid retrieval of items of known locations and mapping of semantically meaningful names to locations.

Query handling must also be improved to facilitate information services. Queries in new and larger information systems often retrieve a great many documents, and techniques for visualizing and summarizing answer sets are required. Query negotiation is going to be necessary as well, since queries will often be sent to libraries under circumstances that only allow constrained processing. The constraints may arise out of limited bandwidth, access restrictions based on charging, or other circumstances. More general browsing interfaces are also needed.

Interoperability is also a key requirement for digital libraries. It must be possible to send queries to multiple index servers and retrieve documents from multiple repositories without human intervention. This will require either standardization or automatic translation. Research into both areas is necessary.

Remaining topics we identified as important include:

User modeling. The maintenance of state from one session to another, and the acquisition of information about each user's goals and intents. This can be used to form models of what kind of query processing will be advantageous to each user, and to improve the performance of systems in a world in which many queries are too short and need supplementary context.

Automatic methods of assessing quality, genre, and other properties of documents. Traditional library classification systems address subject content only, and do not deal with other aspects of documents that are often important for user needs. Given the costs of manual evaluation, we need fully automatic methods, perhaps involving language processing, to extract these properties of digital documents (or other digital objects).

Economic and social models and alternate structures for publishing. One of the key bottlenecks in the development of digital libraries has been uncertainty about which organizations would perform which tasks, and how they would be able to recover their costs. Technology research is needed to simplify these tasks and to provide systems which can be used for collecting revenue. Since the entire structure is uncertain, techniques for economic evaluation, perhaps including simulation, may be needed to suggest the best organizational roles for digital library administration.

Access control and economic charging structures. This topic is related to the previous one, but deals more directly with possible charging algorithms. Authors or readers might pay for information, and they might pay by the month, byte, page, article, minute, or other measures. Practical methods for administration of cost recovery in digital libraries, both for individual users and in the context of site licenses to institutions, are necessary. One very important technical issue is downstream protection: We need technological ways to make it difficult people to resell copies of purchased information.

Multimedia authoring and querying. Although simple text retrieval is now well understood, searching sounds, images, and video is still a difficult task. We need research on indexing, matching, and clustering of all kinds of media. In addition, there is a danger that the rise of multimedia will decrease the diversity of information sources available because of the increased cost of developing this material. Technology to improve authoring would alleviate this problem.

Individual tools to support use of combined personal and public files in a workstation library for use by one researcher. Individual information systems are going to become commonplace, and the methods by which they are connected to distributed national digital libraries are not certain. Research to improve an individual's use of information is needed to help people make the best use of digital library information.

Publicly managed cryptosystems so that businesses can use a standard form of cryptographic protection while avoiding monopolistic practices. The need for trust in

cryptographic software and key servers makes it unlikely that many small vendors can serve this market, and having only one vendor will raise monopolistic risks. Public management of keys will avert these risks.

Evaluation criteria. Basically we wish to have commerce in electronic information continue to grow and thrive, and establish the US as a world leader. We need user acceptance, including institutional and individual reliance on digital libraries, and public acceptance (e.g., when use of the word "library" no longer conjures up an idea of paper books any more than use of the word "watch" implies a circular dial today). Instrumentation of our programs is also important so that we can tell how often and perhaps even how effectively they are being used.

We also thought a few mileposts along the road to acceptability should be noted. Some have already been passed: There exist libraries today which spend more on electronics than on paper, for example (typically in pharmaceutical companies). We suggested:

1. A single on-line source sells electronic articles from a variety of publishers.
2. An electronic information company makes it into the Fortune 500 (almost true today for American Online).
3. A major library devotes more space to people than to paper.
4. People throw away books with the same ease and personal comfort that they have when the type "rm".
5. A faculty member at a top-rate university gets tenure for papers published only electronically.
6. An ARPA grant is given to a proposal citing only electronic references.

Report of the Library Perspective Working Group

IITA Digital Libraries Workshop

Bruce Schatz

Appendix 3-3 Library Perspectives

Introduction

1. Philosophical Issues: What is a Digital Library and How Does it Relate to Traditional Libraries?

2. Research Issues.

3. Priorities and Recommendations.

Introduction

This report briefly summarizes the discussions from the "Library" group. The perspective of this group was of "librarians" who might be seen as custodians of large digital repositories in the future. Thus, there was a strong concentration on retrieval from existing collections, rather than on generation of new materials. This focus

provided a significant discussion on problems of search, and largely omitted problems of publishing. The library perspective might be summarized as building and maintaining large distributed repositories. Repositories are the technology to support users in search sessions across collections. In information infrastructure such as digital libraries, the technology and systems cannot exist alone in the absence of users and collections.

Most of the discussion in this group centered around the important research issues for repositories, with the hopes of encouraging research and funding into solutions of these issues. The issues are discussed below, preceded by a brief philosophical discussion, and followed by a prioritization of the most pressing research issues. The issues in the short-term concentrate on syntax, while the more important ones in the long-term concentrate on semantics.

1. Philosophical Issues: What is a Digital Library and How Does it Relate to Traditional Libraries?

An extended discussion of digital libraries made it clear that the defining factor is searching large collections. The key to a digital library is not yhr digitization of physical materials, but the organization of an electronic collection for better access. The organization provides coherence to a massive amount of shared knowledge, while the access provides convenient retrieval for a wide range of users distributed across a network. Therefore,

a digital library deals with organization and access of a large information repository.

The issues in digital libraries are thus quite similar to those in traditional libraries. Most of the major problems are the same, with some change in orientation due to electronic rather than physical materials. For the foreseeable future, digital libraries are likely to augment physical libraries, much as an on-line card catalog augments, rather than strictly replaces, a book collection. A user's information needs will nearly always be satisfied by some combination of digital and physical materials, each relying on the availability of collections and appropriateness of access. As a rule of thumb, the digital medium tends to be better for searching, and the physical medium tends to be better for reading. Remote items are more easily available, and relationships between items can be more easily followed.

2. Research Issues.

After much discussion, the research issues for digital libraries were divided into four major categories. The first two deal with the technologies and systems at the syntactic (interoperability) and the semantic (description) levels. The second two deal with the users and the collections support.

2.1. Interoperability. These issues deal with the global architecture necessary to deploy digital libraries widely. They are primarily at the syntactic level, dealing with the mechanisms for passing digital objects and operations around the network between collections and users. Thus, these issues concentrate mostly on access:

Naming of digital objects. Giving a unique and invariant name to information objects.

Protocols for object transmission. Executing operations across the network (e.g., issuing a search query to multiple collections).

Types of digital objects. Keeping track of the class definitions for information objects.

Metadata (syntax-level). Registering and reconciling the object schema.

2.2. Descriptions. These issues deal with the resources necessary to retrieve objects adequately from digital libraries. They are primarily at the semantic level, dealing with the mechanisms for describing the meaning of the objects in the collections. Thus these issues concentrate mostly on organization.

Metadata (semantics-level). Defining the value and meaning of the object substructure.

Computed descriptions. Extracting meaning deduced from object content (rather than recorded in static metadata fields).

Unification. Merging the semantics of the metadata across descriptions (e.g., interpreting an author search "properly" across multiple collections with different definitions).

Organization. Clustering the descriptions to facilitate navigation (e.g., building indexes at multiple levels to categorize the networked information).

2.3. Users. These issues deal with the interaction required for users to adequately access a digital collection.

Needs. Understanding what the users need and how to provide it (user assessment, user interface, and new information types).

Contributions. Enabling the users to organize the digital collections for better personal access (annotation, groupwork, and authoring).

2.4 Collections. These issues deal with the management required for collections to be adequately organized.

Archiving. Insuring that access to the digital collection is possible on a "permanent" basis (preservation of objects, conservation of operations).

Virtual collection development. Providing tools to organize a collection consisting of objects distributed across the network.

Repository management. Providing tools to update and maintain a digital collection.

3. Priorities and Recommendations.

From the library perspective, all of these issues are important. Users and collections must be served, and the underlying technology must be available for organization and access. However, some of the issues are pre-requisites to the others. In particular, digital objects must be available for the collections to be generated. So object naming (to reference the objects) and object archiving (to preserve them) are of immediate importance. Once the collections can exist, the adequacy of organization and access holds the most immediate importance. Metadata (both syntactic and semantic descriptions) and needs (understanding the users) are the next most immediate issues. After these critical topics, the rest of the issues were judged to be roughly of the same immediacy. The largest technology pay-off is in the semantic description issues, notably unification and organization; but much research remains to be done for a comprehensive solution to mapping semantics across repositories. A final recommendation is a plea for information systems research, instead of computer technology research. Digital libraries need to be tested with large collections and users since the value of the technology cannot be evaluated in isolation. Large testbeds of systems with new functionality are necessary to prototype new digital libraries. Deployment monies become as important as development monies for digital library funding.

Interoperability, Scaling, and the Digital Libraries Research Agenda:

A Report on the May 18-19, 1995
IITA Digital Libraries Workshop
August 22, 1995

Clifford Lynch
Hector Garcia-Molina

Introduction

This report summarizes the results of a workshop on Digital Libraries held under the auspices of the U.S. Government's Information Infrastructure Technology and Applications (IITA) Working Group in Reston, Virginia on May 18-19, 1995. The objective of the workshop was to refine the research agenda for digital libraries with specific emphasis on issues of scaling and interoperability, and to identify the infrastructure developments needed to make progress on these issues.

While there have been a number of workshops and other meetings examining the broader questions of support for applications in the National Information Infrastructure (NII), we believe this was the first workshop that focused specifically on Digital Libraries in this context. In the past year, Digital Libraries have emerged as one of the central and most compelling applications enabled by the NII; numerous digital library research projects are underway, including six large-scale pilot projects that have been funded jointly by ARPA, NASA, and NSF. While Digital Libraries are now a vibrant research area, and also a field in which considerable commercial development is taking place (presaging the future economic importance of Digital Library technology to the United States), many new questions are emerging as a result of this flowering of research activity. Informed by insights gained from current research, this workshop

offered an opportunity to consider questions such as interoperability objectives that might be defined among projects now underway.

The workshop was organized by Hector Garcia-Molina of Stanford University and Clifford Lynch of the University of California Office of the President. The IITA working group, which sponsored the meeting, reports to the National Science and Technology Council (NSTC) through the High Performance Computing, Communications, and Information Technology subcommittee of the Committee on Information and Communication. The workshop was attended by some 60 leading digital library researchers and developers and by representatives from a wide range of federal government organizations concerned with research and development and policy formulation related to digital libraries (see Appendix 1 for a roster of attendees).

Workshop attendees were asked to consider the following questions as a point of departure in developing the research agenda:

1. What is a Digital Library? How does it differ from an information repository or from today's World Wide Web? How many Digital Libraries will there be, and how will they interlink? How might this look to users?
2. What Digital Library infrastructure is needed? What does "infrastructure" consist of in this context and how does it differ from the broader applications support infrastructure for the emerging NII? What is the relationship between infrastructure and standards? Who will use this infrastructure? When must it be defined, and what parts are most urgently needed? How does the infrastructure relate to intellectual property management and publisher concerns?
3. How can a Digital Library be evaluated? How will we know in three to four years if current research projects have been successful in developing effective digital library services for their user communities?

To further frame and stimulate discussion, Hector Garcia-Molina prepared a position paper discussing the issues and distributed it prior to the workshop (see Appendix 2).

Participants spent the majority of the workshop in one of five groups; unlike many workshops, in which each group is assigned a different set of issues, here each group approached the full spectrum of questions from a specific, unique viewpoint and generated a summary of their discussions that reflected that viewpoint. After a presentation from the five group leaders representing each group's approach to the issues, each participant selected his or her group. The five groups and their leaders were

Bill Arms,
Corporation for National Research Initiatives:
The Publishing Perspective

Michael Lesk,
Bellcore:
The Commercial Perspective

Bruce Schatz,
University of Illinois Urbana Champaign:
The Library Perspective

Mike Schwartz,
University of Colorado:
The Internet Perspective

Terry Smith,
University of California, Santa Barbara:
The Multimedia Perspective

The reports of these five groups appear in Appendix 3. This summary of the workshop extracts common themes and also key points of disagreement from the work of the five groups and places them in broader context. The report is not a consensus document; while it draws heavily on the five group reports and has also benefited greatly from comments from attendees, it does not attempt to reflect completely any of the five group reports.

This report addresses responses to the first two questions posed to the attendees (the definition of a digital library and infrastructure needs to support digital librariesand discusses the research agenda. The third question posed to the attendees -- how to evaluate Digital Library projects - -- did not receive much attention from most of the groups; it is to be the subject of a separate workshop on User Evaluation Methods to be held October 29-31 at the Allerton center under the auspices of The University of Illinois Urbana-Champaign and NSF. Some groups did identify the need for consistent instrumentation and data gathering across projects to facilitate evaluation. In addition, several groups stressed the need to make the transition from a systems technology framework to one driven by user access and collection organization in developing future digital library technology and systems. This view is perhaps most eloquently stated in the reports of the Internet working group and the Library working group.

Definitions and Roles of Digital Libraries

Considerable work has already been done on operational definitions of Digital Libraries and their relationship to traditional library institutions, as well as to the broader systems of scholarly and commercial publishing (see, for example, Communications of the ACM, April 1995). Much of the discussion in this workshop was motivated by questions of scaling, interoperability and needed support infrastructure.

Digital libraries were viewed as systems providing a community of users with coherent access to a large, organized repository of information and knowledge. One group made the provocative proposal that this organization of information was characterized by the absence of prior detailed knowledge of the uses of the information. The ability of the user to access, reorganize, and utilize this repository is enriched by the capabilities of digital technology; the Multimedia group provided particularly vivid examples of these possibilities.

Several groups pointed out that, in fact, digital libraries would, for the foreseeable future need to span both print and digital materials and that the central issue was to provide a coherent view of a very large collection of information. In this sense, an emphasis on content solely in digital format is too limiting. Really, the objective is to develop information systems providing access to a coherent collection of material, more and more of which will be in digital format as time goes on, and to fully exploit the opportunities that are offered by the materials that are in digital formats. Additionally, the comprehensiveness and value of the collection accessible through a digital library system can be strengthened by the ability to integrate materials in digital formats that have not been well-represented, easy to access, or effectively usable in traditional library collections, such as multimedia, geospatial data, or numerical datasets. There is, in reality, a very strong continuity between traditional library roles and missions and the objectives of digital library systems.

Participants in the workshop repeatedly underscored this continuity, and emphasized that the traditional library institutional missions of collection development, collection organization, access, and preservation must extend to the digital library environment. Digital libraries will be a component in the broader range of future library services, and librarians will play a central role in developing and managing digital libraries.

While there would be many digital repositories, a given digital library system should provide a coherent, consistent view of as many of these repositories as possible. From the user's perspective, to the extent possible, there should appear to be a single digital library system. Users increasingly have access to various types of digital collections and information systems: personal information resources, workgroup and organizational information collections and collaboration environments, and more public digital libraries. Defining the boundaries and characteristics of these information spaces and exploring ways in which they can be fused into a coherent whole is a central problem that cuts across all aspects of the research agenda. From the user's perspective, the digital library system needs to extend smoothly from personal information resources, workgroup and organizational systems, and out to personal views of the content of more public digital libraries.

Some groups raised, but did not resolve, the question of the extent to which the digital library system should incorporate support for publishing, annotation, and integration of new information, and the extent to which additions to repositories within the digital library system should be mediated by librarians. It is clear that the development of digital libraries is closely linked to the changes that are occurring in modes of scientific and scholarly communication; the extent to which the digital library should actively embrace -- and perhaps even drive -- these changes remains to be fully explored.

Libraries -- digital or traditional -- exist to serve diverse purposes and constituencies. To some extent, each discipline, constituency, and collection creates its own organization of information. In the digital library world this differentiation among library collections, organization, and services may become more visible. One of the key

challenges is to retain this diversity, which is responsive to unique constituencies, and at the same time permit information to be effectively shared across disciplines and constituencies. This is an essential component of the interoperability questions that formed a major focus for the workshop. Workshop participants represented many of these diverse perspectives: university research libraries, archives, libraries supporting teaching, public libraries, and libraries of the performing arts.

Defining Interoperability in the Digital Library Environment

Defining interoperability proved difficult. It is clear that this is still a central research problem in its own right, and one that merits continued attention. Discussions of infrastructure focused on common tools, enabling technologies and standards that would provide a basis for further exploration of interoperability issues, particularly by encouraging and facilitating the growth of digital libraries on the Internet. Considerable effort was spent on identifying infrastructure that was either unique or particularly critical to progress in digital libraries, as opposed to more general-purpose infrastructure that a range of NII applications, including digital libraries, might share. One clear theme was that an understanding of interoperability issues required operational experience which could only be gained by large-scale deployment of digital library systems. Speculation about interoperability in the abstract is of very limited value.

Participants expressed a full spectrum of views on interoperability. At one end of the spectrum is the use of common tools and interfaces that provide a superficial uniformity for navigation and access but rely almost entirely on human intelligence to provide any coherence of content. At the opposite end of the spectrum is deep semantic interoperability. The precise definition of deep semantic interoperability was the subject of some debate, but deals with the ability of a user to access, consistently and coherently, similar (though autonomously defined and managed) classes of digital objects and services, distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations. It also extends beyond passive digital objects to actual services offered by specific digital library systems. Deep semantic interoperability is a "grand challenge" research problem; it is extraordinarily difficult, but of transcendent importance, if digital libraries are to live up to their long-term potential. An intermediate position between these two extremes advocates primarily syntactic interoperability (the interchange of metadata and the use of digital object transmission protocols and formats based on this metadata rather than simply common navigation, query, and viewing interfaces) as a means of providing limited coherence of content, supplemented by human interpretation.

Note that the term "digital object" here is intended only to describe, in the broadest sense, the type of information objects that may comprise a digital library -- textual, audio, video, numeric, computer programs, or multimedia composites of such components. It is not intended either to endorse or preclude an object-oriented

architectural framework for digital library systems (in the sense of object-oriented programming or object-oriented databases, for example).

Infrastructure Requirements for Digital Library Research

The most urgent infrastructure need is to establish common schemes for the naming of digital objects, and the linking of these schemes to protocols for object transmission, metadata, and object type classifications. The consensus of the groups was that naming schemes for digital objects that allow global unique reference represented perhaps the most immediate infrastructure deployment priority in order to facilitate resource sharing, linkages, and interoperation among digital library systems and to facilitate scale-up of digital library prototypes. It was recognized that the design of large-scale naming systems and their integration into the larger digital library framework will continue to be an important research area, but that infrastructure support needs to be put in place quickly for at least an interim system, and that in fact experience with such an interim system would inform further research.

The deployment of a public key cryptosystem infrastructure -- including the development of a system of key servers and the definition of standards and protocols -- was also identified as essential to progress in digital libraries; this is necessary to support digital library needs in areas such as security and authentication, privacy, rights management, and payments for the use of intellectual property. While the need for public key cryptosystem infrastructure is hardly unique to digital libraries, the importance of the digital library services and components which depend on this infrastructure mean that its absence represents a significant barrier. In particular, until these problems are addressed, it seems unlikely that we will see commercial publishers and other information suppliers making large amounts of high-value copyrighted information broadly available to digital library users. This in turn will constrain the development of research prototypes and may be a distorting factor in studies of user behavior.

Research Issues and Priorities

The working groups outlined a wide range of important research issues; most groups were less successful at prioritizing them, beyond the immediate infrastructure needs already discussed. The five key research areas that emerged from the workshop are described below; arguably, the first three are of most central and immediate importance, specifically to the development of digital libraries, though the long-term importance of research in the fifth area (economic, social, and legal issues) cannot be overemphasized. The distinctions among the five areas are to some extent arbitrary; for example, progress on interoperability (the first area) depends critically on progress in our ability to describe successfully objects and repositories (the second area).

1. Interoperability

The difficulty in defining the objectives for interoperability have already been discussed; clarifying these objectives, mapping the spectrum of interoperability, and establishing the key challenges at points along this specturm are key research issues in their own right.

The more technical interoperability research involve protocol design that supports a broad range of interaction types, inter-repository protocols, distributed search protocols and technologies (including the ability to search across heterogeneous databases with some level of semantic consistency), and object interchange protocols. Interoperability is not simply a matter of providing coherence among passive object repositories. Digital library systems offer a range of services, and these services must be projected in an interoperable fashion as well. One particular issue that emerged was that existing Internet protocols (such as HTTP, the basis of the World Wide Web) are clearly inadequate. Research must move beyond the current base of deployed protocols and systems. This raises complex questions about how to deploy prototype systems and the tradeoffs between advanced capabilities and ubiquity of access.

The practical question of the nature of the installed technology base and the need to support this installed base will increasingly frame and influence interoperability research. Access to digital libraries is not an end in itself for most users, but rather a support service; many will be willing to sacrifice advanced functionality for consistency, stability, and ability to use familiar, common access tools. Just as the installed base has become the greatest barrier to meaningful large-scale trials of new approaches that improve existing services (as opposed to providing entirely new services which do not compete with an installed base) in the overall Internet environment, user expectations and the installed base will ultimately impede progress in fundamental technology research within the large-scale experiments necessary to gain insights into interoperability among digital libraries. Managing this tension will be a critical element in the continued development of the community's research agenda.

It should be noted that, at this relatively early stage in the evolution of digital library technology, it is of vital importance that projects strive for approaches that incorporate high functionality and extensibility. A high level of functionality in the standards and protocols used, even if not fully exploited initially, will postpone the time when the inertia of the installed base begins to confine research opportunities. Careful design of extensibility in digital library systems will facilitate continued research progress and understanding of the impact of new approaches on the user community without the need to attempt to displace an installed base.

2. Description of Objects and Repositories

In order to provide a coherent view of collections of digital objects, they must be described in a consistent fashion which can facilitate the use of mechanisms such as protocols that support distributed search and retrieval from disparate sources. Research in description of objects and collections of objects provides the foundation for effective interoperability. Interoperability at the level of deep semantics will require breakthroughs in description as well as retrieval, object interchange, and object retrieval protocols.

Issues here include the definition and use of metadata and its capture or computation from objects, the use of computed descriptions of objects, federation and integration of heterogeneous repositories with disparate semantics, clustering and automatic hierarchical organization of information, and algorithms for automatic rating, ranking, and evaluation of information quality, genre, and other properties. Other key issues involved knowledge representation and interchange, and the definition and interchange of ontologies for information context. The idea of active "information matchmaking" emerged in several group reports.

Research is also needed to understand the strengths and limitations of purely computer-based technologies for describing objects and repositories, and the appropriate roles for the efforts of human librarians and subject experts in the digital library context as a complement to these technology-based approaches.

3. Collection Management and Organization

Collection management and organization research is the area where traditional library missions and practices are reinterpreted for the digital library environment. Progress in this area is essential if digital library collections are to meet successfully the needs of their user communities.

Policies and methods for incorporating information resources on the network into managed collections, rights management, payment, and control issues were all identified as central problems in the management of digital collections. Approaches to replication and caching of information and their relationship to collection management in a distributed environment need careful examination. The authority and quality of content in digital libraries is of central concern to the user community; ensuring and identifying these attributes of content calls for research that spans both technical and organizational issues. Research is also needed to clarify the roles of librarians and institutions in defining and managing collections in the networked environment.

With the enhanced potential to support nontextual content effectively in the digital library environment, issues in nontextual and multimedia information capture, organization, and storage, indexing and retrieval are clearly key research areas.

However, textual digital documents remain a vitally important research area in their own right, and are far from fully understood. The role of knowledge bases in digital libraries remains a poorly explored but potentially important question.

The preservation of digital content for long periods of time, across multiple generations of hardware and software technologies and standards is essential in the creation of effective digital libraries. This is an extraordinarily difficult research problem which has not received sufficient attention.

4. User Interfaces and Human-Computer Interaction

While user interfaces and human-computer interaction issues are an extensive field of research in their own right, there are some specific problems that are central to progress in digital libraries.

Display of information, visualization and navigation of large information collections, and linkages to information manipulation/analysis tools were identified as key areas for research. The use of more sophisticated models of user behavior and needs in long-term interactions with digital library systems is a potentially fruitful area for research. The necessity for a more comprehensive understanding of user needs, objectives, and behavior in employing digital library systems was stressed repeatedly as a basis for designing effective systems. Finally, it was observed that digital library systems must become far more effective in adapting to variations in the capabilities of user workstations and network connections (bandwidth) in presenting appropriate user interfaces; new technologies such as personal digital assistants and nomadic computing models will emphasize this need.

5. Economic, Social, and Legal Issues

Digital libraries are not simply technological constructs; they exist within a rich legal, social, and economic context, and will succeed only to the extent that they meet these broader needs. Rights management, economic models for the use of electronic information, and billing systems to support these economic models will be needed. User privacy needs to be carefully considered. There are complex policy issues related to collection development and management, and preservation and archiving. Existing library practice may shed some light on these questions. The social context of digital documents, including authorship, ownership, the act of publication, versions, authenticity, and integrity require a better understanding. Research in all of these areas will also be needed if digital libraries are to be successful.

Conclusions

This workshop has made substantial progress in refining and focusing a research agenda for digital libraries, as well as in developing insights into questions about

interoperability among digital libraries and the infrastructure necessary to support such interoperability. Interoperability is likely to continue to be a useful organizing theme in refining this agenda in the coming years. The outcomes of the workshop also suggest that a focus on broad architectural issues in digital libraries will be fruitful. Several working groups commented on the need to develop component software strategies that would facilitate the transfer of technology among the current digital library pilot projects and from these projects to other new digital library research efforts. The Internet working group went further in suggesting that the development of a broadly available software base for the digital library community would contribute to rapid progress, and we believe that this suggestion deserves careful consideration.

Scaling was identified as a major area of concern. The common vision is one of tens of thousands of repositories of digital information that are autonomously managed yet integrated into what users view as a coherent digital library system. Accommodating this very large number of repositories -- a very different environment than that in which today's handful of pilot projects operate -- will clearly have major implications for infrastructure definition and design. We must move rapidly towards an infrastructure that can support and facilitate research towards this common vision. The full range of issues here are unclear. Some immediate needs are evident; these are reflected in the emphasis on establishing naming systems for digital objects as a high priority, for example.

We don't know how to approach scaling as a research question other than to build upon experience with the Internet. However, attention to scaling as a research theme is essential and may help in further clarifying infrastructure needs and priorities, as well as informing work in all areas of the research agenda outlined above. For example, reliability questions are poorly understood; in a sufficiently large system, some components will inevitably be out of service during the processing of any given query. The need to support large-scale deployment projects (in terms of size of user community, number of objects, and number of repositories) and to study subsequently the effectiveness and use of such systems was emphasized repeatedly. It is clear that limited deployment of prototype systems will not suffice if we are to understand understand the research questions involved in digital libraries.

Research in scale-up is very difficult to perform except by building and deploying a large-scale digital library system. Establishing infrastructure and tools to facilitate experimentation with large-scale systems is essential, as is funding to study use and behavior of large-scale systems once deployed through this infrastructure. The Internet as a context for deploying digital library systems offers an unprecedented opportunity -- not only technically by providing connectivity to an enormous potential user base but also culturally, given the Internet community's models and traditions of technology diffusion through the distribution of publicly available prototype software -- to move ahead large-scale experiments. Research efforts should exploit these opportunities.

Finally, it seems clear that the inevitable presence of large amounts of commercially valuable, proprietary information in the future -- which can be viewed as another form of scale-up in digital libraries -- will also shape the research agenda in new ways. The near-term focus is on overcoming the infrastructural barriers to supporting proprietary information (such as authentication, billing, and rights management). There are research issues in the design of such an infrastructure, but also operational and policy problems impeding deployment. While some of the research issues are complex and will require ongoing exploration, putting at least the first steps towards the necessary infrastructure in place to accommodate such commercially valuable information is a high priority in advancing the research agenda and addressing scale-up issues. It will also stimulate commercial developments that will complement existing research initiatives. The development of an increasingly rich marketplace of information resources under a wide range of economic and legal constraints will create new opportunities in all areas of the research agenda presented above, and will allow us to explore vital new research questions in the development of description, navigation, access, and resource discovery technologies and systems that can function in this broader environment.

Report of the Multimedia Perspective Working Group

IITA Digital Libraries Workshop

Terry Smith

---

1. Introduction

Identifying important research issues concerning the development of digital libraries requires a focused discussion. A useful focus is provided by defining the essential nature of a digital library and by restricting discussion to special classes of collections. A particularly useful focus emerges from a consideration of multimedia collections, since

digital technology offers powerful techniques for handling queries involving heterogeneous collections of such materials.

1.1. An Approach to Defining a Digital Library

It is helpful to recast the question "What is a Digital Library?" into a set of simpler questions. In particular, one may ask: "What is a library?"

"What is the role of a librarian in a library"? "In what ways does digital technology extend and enhance the nature of a library and the role of a librarian"?

At its heart, a library is a collection of items containing representations of information with some intended meaning. The single most important property characteristerizing whether a collection of informational items belongs to a library is that the collection is organized and managed in a manner that optimizes access to the information for a given class of users. In particular, the organization and management of a library's collections should facilitate the processing of the information in the items and the extraction of useful knowledge represented either explicitly of implicitly in the items.

The major role of a librarian is in organizing and managing the library's collections, and in facilitating the communication of the information between the library and its users. In a "traditional" library, such management and organization involves the creation of catalog information facilitating access to appropriate items in the collections. Cataloging information essentially provides a mapping between the items and abbreviated representations of the items and their content. The management and organization also involves a physical organization of items that accords with the cataloging procedures. It should be noted that important metadata associated with the items in a library's collections concerns the "authenticity" and "validity" of items. Such information may be implicit in the fact that a librarian has decided to add an item to a library's collections.

A digital library may be viewed as a library that has been extended and enhanced by the application of digital technology. Important aspects of a library that may be extended and enhanced include:

1. the collections of the library;
2. the organization and management of the collections;
3. access to library items and the processing of the information contained in the items; and
4. the communication of information about the items.

2. Digital Technology and Extensions of Libraries and the Roles of Librarians.

In discussing the extensions and enhancements for the four aspects of libraries that may be supported by digital technology, we first provide examples of key extensions, a brief overview of important issues associated with such extensions, and a list of research problems germane to these issues.

2.1. Extensions and Enhancements to Library Collections

Providing digital representations of library items, with all the attendant advantages of such representations, is clearly a major enhancement Digital technology, however, also permits the extension of library collections into new domains. We briefly discuss five examples of such domains and a few associated issues.

"User-centered" collections involve the construction of personalized collections by users and may involve, for example, the reorganization of parts of existing items into new items, as well as their extension with various annotations. A critical issue raised by this possibility relates to the procedures by which certain classes of items are "authenticated" as being part of a "core" library. In general, it appears important that such a core collection be identified in terms of certain admissibility criteria.

Multimedia collections may involve digital representations of

- textual items;
- graphical and spatially-indexed items;
- acoustical items; and
- video items.

An important issue relating to the items of such collections is that they may require significant levels of intermediate processing or interpretation, such as image or acoustical signal processing, that are not required for collections of traditional textual materials. A second important issue relates to integration of such materials. A relatively simple example demonstrating the need both for intermediate processing and for integration is provided by the following query:

Find all quadsheets containing towns with over half a million inhabitants in the Mississippi Valley that are within 50 miles of Indian burial sites for which the library has digitized photographic records dating from the last century.

A third important issue arising in multimedia collections concerns the need to construct, store, access, and process multiple concrete representations of items. For example, different representations of digitized map information currently exist, with some favoring given information processing operations more than others. A fourth issue concerns metadata about the "lineage" of the information contained in some digital object, since it may embody a complex history of information processing to material from a variety of different sources. Lineage issues are frequently important in determining the value of information in certain applications.

"Procedural" collections involve collections of information-processing operations that may be applied by users or librarians in order to extract information from other library items. Multimedia materials represented in digital form may require the application of a large variety of procedures in order to extract the information required by users. An

important issue is how libraries should support, organize, and manage procedural information. A related issue is about "dynamic" collections, by which we mean collections that grow as information is extracted from other items already residing in a library's collections. Such information may be extracted by various procedures stored as library items and applied in some "automated" manner. Finally, we mention "knowledge bases," which we may interpret to be representations of large domains of knowledge, such as those contained in online encyclopedias. Such knowledge bases, when viewed as "ontologies," may be used both as a basis for metadata in a library's catalog and as an information source enhancing a users access to the information in other library items. An important issue is the construction and maintenance of such knowledge bases.

2.1.1. Specific Research Problems Relating to Extended Collections

A few of the many important research questions that relate to the issues identified above include:

- What procedures should librarians adopt in deciding which collections and items should constitute the "core" of libraries containing multimedia and other non-traditional items?
- What procedures should librarians adopt in deciding which items to discard from the "core" collections of a library?
- How should libraries support the need to apply information processing procedures to library items?
- What are the user requirements concerning the integration of multimedia materials and how should they be supported in libraries?
- What procedures and protocols should be followed in authenticating such collections and items?
- What standards and protocols need to be adopted to enable interoperability of libraries with respect to transfers of multimedia and non-conventional items?
- What issues arise when multimedia collections cohabit the same storage device/file structure?
- How should libraries approach the issues relating to collections of procedural, as opposed to declarative, information?
- In what manner should libraries support the concept of knowledge bases?

2.2. Extensions and Enhancements to Library Organization and Management

If there is one single criterion that defines a library, it concerns the generally accepted procedures and protocols for organizing and managing the collections. An important enhancement that is provided by digital technology is the possibility of presenting to the user many alternative ways of viewing any subset of the library's collection. Such dynamic "reorganizations" may be based on the different ways in which library items may be indexed in the catalog according to various criteria that include the medium and

format of the item as well as many aspects its content. In the context of digital libraries, such dynamic reorganizations of the catalog may be viewed as providing the user with a variety of different browsing contexts, as if the items had been reorganized on the "stacks." Important issues relate to the set of organizing principles for a library's collection that are supported by a library and the ways of making such multi-organizations interoperable between libraries.

A second and related enhancement involves the great variety of metadata that it is possible to extract, store, and retrieve about items and the various organizations that may be imposed on a library's collections using such metadata. It is useful to categorize metadata according to whether it is domain-dependent or domain-independent and whether, in the latter case, it is related to "low-level" content or to aspects of the origin and representational aspects of the item. Such distinctions are very important in the case of multimedia libraries. Items such as images and maps, for example, have huge numbers of interpretations concerning their content, and require significant extensions of traditional cataloging practices in order to characterize them in a manner that is useful for user access and browsing. An enhancement to traditional libraries that arises in relation to metadata involves the variety of "annotations" that may be stored in association with items. Critical issues that arise concern the classes of metadata to be extracted from library items of different types and the procedures for extracting such metadata.

At the highest level, metadata may be based on "ontologies," or organized sets of concepts concerning both the representational aspects of library items and their content. Ontologies provide the basis for the many indexing schemes that are possible. For example, ontologies that relate to the origin, lineage, format, and representational aspects of library items may be viewed as extensions of the "author catalog," and are important for representing catalog information about the many classes of non-standard items that may occur in multi-media collections and about the multiple representations of such items. Ontologies that relate to the content of library items may be viewed as extensions of the "subject catalog" of traditional libraries. As an example, an important class of ontologies concerning geographical objects provide bases for indexing schemes for library items in terms of the "spatial projection" of the objects to which the items refer. Ontologies may be multiple, overlapping, hierarchically organized, and amenable to object-oriented representations. In particular, they may be used to define "sublibraries." Important issues concern the construction and use of various ontologies and the interoperability of libraries with respect to such ontologies.

Important general issues concern the role of the librarian in constructing the extended metadata and catalogs that are required, as well as issues relating to standards and protocols for metadata extraction, organization, and access.

2.2.1. Specific Research Problems Relating to Organization and Management

A few of the research problems relating to the issues identified above include

- Standards and protocols concerning metadata in its widest sense.
- Metadata and catalog support for multimedia items, and their special attributes, such as lineage and multiple concrete representations.
- The librarian's role in constructing authenticated metadata, catalogs, and organizations of library items.
- Tools and resources for building and communicating "ontologies."
- he cross-referencing of items using metadata.
- Automated metadata extraction.
- The organization of "declarative" versus "procedural" information and the role of the librarian in defining metadata for procedural information.

2.3. Extensions and Enhancements to Information Access

The extensions and enhancements that digital technology offers in the area of library access relate the universality and ubiquity of the access, the nature of the information accessed, the large variety of means for accessing information, and the extraction of knowledge with the use of procedures that convert information in implicit form to information in explicit form.

Accessible information includes the extended metadata discussed previously, as well as information about other libraries, their catalogs, their collections, and their items. Digital technology makes it possible to access information by any of the extensions to metadata that were discussed above, including access by medium, structure, and content.

Important issues in relation to finding digital objects in integrated, multimedia digital libraries include query languages and support for complex query construction, particularly in the case of complex queries that require synthesis and the application of transformations. Perhaps the single most important criterion is the ability to search content. Other issues include the translation of queries into domain-independent and domain-dependent metadata, the use of similarity matching in answering queries, multidimensional indexing, and the correlation of multimedia items with the use of information concerning the content of the items.

2.3.1. Specific Research Problems Relating to Information Access

A few of the many important research problems that relate to the issues identified above include

- The design of, and support for, query languages and particularly query by example.
- Support for answering queries that involve the application of transforming procedures to accessed information.

- Support for answering queries based on the content of library items.
- The development of similarity matching procedures.
- The development of multi-dimensional indexing techniques.
- The development of search tools for the Web and the Internet.
- Interoperability that involves users being able to search on, and perform on, the information stored in the collections of different libraries.
- The evaluation of usage patterns, user performance, and user satisfaction.

2.4. Extensions and Enhancements to Communications:

Communications between users and libraries and among libraries themselves are major aspects that assume critical importance in digital extensions of libraries.

Important issues that arise from the preceding discussion involve protocols and standards for data and metadata, high-level languages for users and librarian, and low-level protocols to support library interoperability

2.4.1. Specific Research Problems Relating to Information Access

- Protocols and standards for the communication of data and metadata.
- High-level languages permitting easy communications for users and librarians.
- Low-level protocols to support library interoperability.