Report on the May 18-19 1995 IITA Digital Libraries Workshop

FINAL DRAFT FOR PARTICIPANT REVIEW
August 4, 1995

Clifford Lynch (clifford.lynch@ucop.edu)
Hector Garcia-Molina (hector@db.stanford.edu)

Introduction

This report summarizes the outcomes of a workshop on Digital Libraries held under the auspices of the US 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 scaling and interoperability and the infrastructure needed to enable this research agenda.

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 has 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 by ARPA, NASA and NSF. While Digital Libraries are now a vibrant research area, and also an 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. This workshop offered an opportunity to consider questions such as interoperability objectives that might be defined among the projects now underway in light of insights already gained from current research.

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? In particular, how does it differ from an information repository or from today's Worldwide 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? In particular, 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? In particular, 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; this is reproduced in Appendix 2.

Participants spent the majority of the workshop in one of five groups; unlike many workshops where 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 the 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 out common themes and also key points of disagreement from the work of the five groups and places them in broader context. It does not attempt to completely reflect any of the five group reports.

This report addresses responses to the first two of the questions posed to the attendees (the definition of a digital library and infrastructure needs to support digital libraries) as well as discussing 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 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, inteoperability 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 proposed that this organization of information was characterized by the absence of prior 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 useable in traditional library collections, such as multimedia, geospatial data or numerical datasets. There is thus in reality a very strong continuity between traditional library roles and missions and the objectives of digital library systems.

While there would be many digital repositories, from the user's perspective 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 be a single digital library system. Users also 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.

Infrastructure Issues, Enabling Technologies and Interoperability

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 understanding interoperability issues called for operational experience that could only be gained by large scale deployment of digital library systems; speculating about interoperability in the abstract was of very limited value.

Views of interoperability ranged from 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 one extreme, through primarily syntactic interoperability (metadata and digital object transmission protocols and formats), to deep semantic interoperability at the other end of the spectrum. The precise definition of deep semantic interoperability was the subject of some debate but deals with the ability of a user to consistently and coherently access similar (though autonomously defined and managed) classes of digital objects, distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations. 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.

A key theme was the need 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 one of the most important infrastructural component requiring immediate effort. Deploying such systems would enable progress in a wide range of key areas.

Public key cryptosystems and their associated infrastructure of key servers and standards was also identified as essential to progress in digital libraries; this is closely linked to questions of security, privacy, rights management, and payments for the use of intellectual property. While the need for public key cryptosystems is hardly unique to digital libraries, it is particularly acute for digital library applications since the lack of such an infrastructure is impeding progress in a number of crucial areas. Until these impediments are removed it seems unlikely that we will see large amounts of high-value copyrighted information broadly available to digital library users.

Research Issues and Priorities

The working groups outlined a wide range of important research issues; most groups were less successful at prioritizing them, beyond immediate infrastructure needs already discussed. Broadly, the key research areas fall into the following (unprioritized) categories:

1. Interoperability

The difficulty in defining the objectives for interoperability have already been discussed; clarifying these objectives and measuring their feasibility is itself a key research issue.

The more technical research questions in supporting interoperability 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. One particular issue that emerged was that existing protocols (such as HTTP, the basis of the Worldwide Web) are clearly inadequate and that research must move beyond the current base of deployed protocols and systems; this raises some 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.

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.

Understanding 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 is also clearly a central problem.

3. Collection Management and Organization.

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 help define 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 as 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 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.

Display of information, visualization 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 need for a more comprehensive understanding of user needs, objectives and behavior in employing digital library systems was stressed repeatedly as a necessary 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 other nomadic computing technologies 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 availability 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 today's handful of pilot projects -- 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 infrastrcture 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 subsequently study 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 fully 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.