The Stanford Integrated Digital Library Project

Summary

The goal of the Stanford Integrated Digital Library Project (SIDLP) is to develop the enabling technologies for a single, integrated and "universal" library, providing uniform access to the large number of emerging networked information sources and collections. These include both on-line versions of pre-existing works and new works and media of all kinds that will be available on the globally interlinked computer networks of the future. The Integrated Digital Library is broadly defined to include everything from personal information collections, to the collections that one finds today in conventional libraries, to the large data collections shared by scientists. Our technology will provide the "glue" that will make this worldwide collection usable as a unified entity, in a scalable and economically viable fashion.

A consortium of information vendors, publishers, end users, commercializers, and researchers has been formed to carry out the proposed research. They will also implement a SIDLP Testbed system that will provide access to significant collections of computing and related literature. The testbed will be used to develop a variety of novel library services, and to experiment with protocols for the integration of heterogeneous information sources.

I. Executive Summary

We are currently witnessing an explosion of available digital information. Some of it is in unstructured repositories, some in primitive digital libraries. This trend is certain to accelerate as the National Information Infrastructure becomes a reality.

The goal of the Stanford Integrated Digital Library Project (SIDLP) is to develop the enabling technologies for a single, integrated and "universal" library, composed from the large numbers of emerging individual heterogeneous repositories. Our definition of a constituent repository is broad enough to include everything from personal information collections (e.g., a file of electronic messages) to the collections that one finds today in conventional libraries to the large data collections shared by scientists. Our technology will provide the "glue" that will make this worldwide collection usable as a unified entity, in a scalable and economically feasible fashion.

An analogy with the development of the Internet may help in clarifying our objectives. Over the past twenty years, as computer communications networks became a reality, separate and disconnected networks arose, such as the ARPANET, BITNET, and private company networks. Later, gateways were established between networks, but end users still needed to be aware of the different networks. For instance, an electronic mail message to a user on a different network had to be explicitly addressed via a gateway. Today, network protocols have become much more integrated, at least for primitive services such as electronic mail. That is, one can send messages to users on most networks simply by using a "logical" address such as "hector@cs.stanford.edu" with no knowledge of what gateways or networks the message will go through. It is important to note, however, that each network is still relatively autonomous, using its own local protocols and conventions.

Intuitively, we aim to provide a similar type of integration at a much higher level over the emerging collection of autonomous and heterogeneous digital libraries and information repositories. While each repository and component digital library retains its identity and independence, we will provide a collection of protocols and services that will integrate them into the *Inter-Library*, to use a term analogous to the Internet.

Our analogy with the Internet is limited because we want to go beyond the primitive exchange protocols that comprise the current state of integration on the Internet. That is, today one can exchange electronic messages, files, and packets on the Internet, but there are few universal protocols for higher level information and functions. In the case of the Inter-Library, our goal is to provide the high level concepts and protocols that can allow powerful library operations across multiple sites. To make this vision a reality, a number of challenging problems must be addressed:

- *What is the proper architecture for the Digital Library?* We need one that is scalable and allows for a variety of information sources and services. A very first cut at such an architecture is presented in Section II.A.

- *What are appropriate models for representing documents (and more generally information) and library services?* Such models must also encompass "meta-information," that is, information about information. For example, meta-information may indicate where can a document be found, who is authorized to see it, how does one pay to see it, and so on.

- *What protocols for interoperation between repositories, libraries, and services are required?* Mechanisms are required, for instance, to express interests in particular types of information, to pay for services received, and to announce new available services.

- *What are the best human interfaces to such an integrated library?* We need to enable end users to visualize the available resources and to express their interests.

- *How can one make the Digital Library economically viable?* We must be able to protect intellectual property by preventing or detecting violations, and a framework is needed to charge for use of the Library and its resources.

- *How does one locate information of interest, in a very large, distributed, and possibly disconnected collection of repositories?* There must be ways to identify sources that are "likely" to have the information of interest, without having full global knowledge.

- *How does one represent and manipulate the information processing "activities" that occur or will occur in a Digital Library?* For example, the process of publishing an article can be captured by an automated agent that can find a good publisher, handle requested revisions, and perhaps negotiate a price. Similarly, a "personal assistant" agent could be tailored to manage the personal information needs of a particular person, even as that person travels and becomes disconnected.

The research problems we have outlined will be the main focus of our project. In Section II.B we discuss them in more detail, and explore some promising initial directions to be pursued.

To maintain focus, we have consciously omitted certain topics from our project. This is *not* because we do not consider the other topics important or relevant, but simply because we have limited resources. One example of an omitted topic is multimedia information. Obviously, multimedia information will play a central role in any Digital Library, and the protocols and systems we develop as part of this project will be able to handle multimedia information. However, we do not propose to do research in this area (e.g., compression of images, searching an audio library). We believe that multimedia and the other omitted problems will be addressed by other projects under the NSF/ARPA/NASA initiative, and we plan to incorporate their results.

The central component of our project will be an experimental testbed facility for demonstrating and evaluating the technologies developed. This SIDLP Testbed will be implemented at Stanford University and will provide access to a significant and varied collection of information. At the same time, it will be accessed by a substantial community of users. Their operations will be traced for further study, and their feedback will be solicited. On the testbed, we will implement library services that arise out of our basic research.

The collection of materials we plan to provide is detailed in Section II.C.2. Its theme is *computing literature* of interest to computer professionals, business people, and scientists. Given the focus of our project, this theme has two important advantages for us: (1) there is already a substantial body of literature available digitally, with much more on the way; and (2) all of the researchers on our team are both generators and clients for this type of information. This creates a tight research-use-feedback loop that has historically been very beneficial in developing new concepts into working technologies.

Our proposed collection includes:

- Full documents, such as ACM journal articles provided by ACM (one of our partners), Computer Science Technical Reports provided by the CNRI Consortium (of which Stanford is a member), and books provided by O'Reilly Associates.

- Abstracts of technical documents, such as those provided through the Stanford University Folio System. An example of such collections is INSPEC, which includes abstracts for computer science, electrical engineering, and physics papers.

- Information on commercial products, such as catalogues produced by computer and electronics vendors that will become available though the Smart Valley CommerceNet Project (EIT, the prime contractor, is one of our partners), and DataPro reports and analysis (full text) available through Dialog.

- Legal information, such as a patent database, and DMS/FI Contract Awards, which is a comprehensive file of all nonclassified U.S. government prime contract actions of $25K or more (both available through Dialog).
- The public information available through the Internet, on a variety of services such as the World-Wide Web, and the Wide Area Information System (WAIS), including the relevant news groups in the vast NetNews collection, specialized bibliographies, such as the HCI bibliography, and the extremely rapidly growing collection of reports, abstracts and other materials being put on-line by computer professionals around the globe.
- Private information, such as electronic mail files, file archives, and papers in progress.

As one can see, this is a very diverse set of information sources, and will give us an opportunity to truly experiment with heterogeneity.

The testbed will not only provide mechanisms for accessing the collections, but at the same time will demonstrate a variety of novel *services* for finding, manipulating, and safeguarding the information. We illustrate here by listing three of these services; others are given in Section II.B.

- *A Finding Service.* This service will collect "statistics" from a variety of potential sources (e.g., frequency of occurring terms, concepts covered) and, based on these, will route queries to promising sources. It will also have facilities for conducting multi-site searches even when all sites are not available at once. In addition, the service will provide for selective dissemination of information (e.g., alerts). In this case, users subscribe to information of particular types, and the information is routed to the users as it becomes available. We note that this service, like all the others we illustrate here, may be distributed across several computers.
- *A Copy Detection Service.* Publishers are currently reluctant to make their materials available in a public digital library because once there, it is easy to duplicate them. The Copy Detection Service helps in identifying holders of illegal copies. As new documents are created, they are registered at the service. Questionable documents can be compared against the body of registered documents. By having appropriate index structures, the service can detect not just exact replicas, but also documents that have significant overlap with registered documents. The service can be used in a variety of ways, for example, a bulletin board service could automatically check new postings for originality. The Copy Detection Service is just one example of the new types of services we believe need to be developed to safeguard intellectual property in a Digital Library.
- *An Agent Creation Service.* The services we have illustrated can be generalized to the concept of an *agent*, a knowledgeable program that can help in library-related functions. Hand programming agents is time consuming, so we plan to develop tools for automatic generation of agents. These tools will take as input "scripts" that describe the functionality of the agent and the knowledge it manages, and generate the appropriate agent. This service will be used to construct, among others, a personal information assistant (PIA). This PIA agent can be thought, as a first approximation, as a personal "secretary" that knows one's information needs and tastes, and assists in locating information and filtering it once received.

All of these services will need to be accessed by people through interfaces that hide the unimportant details of diversity of materials and provide ways to navigate and manage the "information space" in a consistent and unified way. We will apply new techniques of information visualization and view structuring to provide users with a window into the Inter-library that can help locate and identify relevant materials in a vast collection, without swamping them with too much information.

The services and collections we have described must of course be integrated into a seamless Digital Library. Thus, a key component of the SIDLP Testbed will be its interoperation protocols for communication and information exchange among heterogeneous components. Our initial design is based on the concept of a common "information bus" or "highway" that links all components. Like a physical bus or highway, our Information Bus allows any constituent, in our case digital library repositories, interfaces, or services, to connect with others in a standard way. However, unlike physical computer busses, ours is a *logical* concept; each constituent can actually be connected to different types of networks and run on different brands of computers. Another important difference is that our Information Bus is high level and library-specific. For example, a typical bus operation may be to send a document from one site to another, where the bus protocols would take care of translating the document into the desired one at the destination. As part of this request, appropriate charges may be made to the account of the recipient user, a summary provided for browsing through sets of documents, a check be made for related documents of interest, etc. Additional discussion of our proposed Information Bus concept can be found in Sections II.A and II.B.

Since the Information Bus protocols impact all aspects of the Digital Library, they cannot be developed in isolation. Fortunately, for this project we have assembled a very strong and diverse set of participating institutions. All the relevant groups—users, archival establishments, information generators, commercializers, and researchers—are well represented. The list of participating organizations is:

Association for Computing Machinery (ACM). ACM is the premier publisher of computer science journals. They will provide in digital form most of the materials they publish, perform research on economic issues, and participate in the design and evaluation of the testbed.

Bell Communications Research (Bellcore). The research arm of the Bell local telephone companies will represent the telecommunications and information delivery perspective. They will participate in the development of the Information Bus and information finding services.

Dialog. Dialog is one of the major information vendors, with over 400 repositories and over 150,000 customers. They will provide data and participate in the evaluation of the testbed, especially commercial aspects such as safeguarding intellectual property and charging for information.

Enterprise Integration Technology (EIT). EIT will contribute their integration and network access technologies, including those of the Smart Valley CommerceNet Project. They will participate in the implementation and evaluation of the testbed and its services, especially the interface clients.

Hewlett Packard Laboratories. Hewlett Packard, a major equipment manufacturer, will work with us on the development of library protocols, searching services, and personal information assistants. They will also contribute equipment for our effort.

Interconnect Technologies Corporation (ITC). ITC is a start-up recently formed by NASA scientists to commercialize information access technologies. They will facilitate access to NASA information sources and participate in the design and evaluation of the testbed.

Interval Research Corporation. Interval is a new research laboratory which is charged with developing the computer and communication systems of ten years from now. They will participate in the design of the Information Bus, and in the research on agents and interface clients.

O'Reilly and Associates. O'Reilly is one of the first publishers to look towards the Internet as a major venue for publication. They will provide access to their published materials and participate in the design of the Information Bus and interface clients.

NASA Ames Artificial Intelligence Group. The NASA Ames AI group is involved in many research areas that impact Digital Libraries. They will provide a user community, and participate in the evaluation of the testbed and in the research on agents.

NASA Ames Library. The NASA Ames Research Center employs over 2,000 government and contract scientists and engineers. Their library will work with us to make our testbed system available to their scientists, and will participate in evaluating the testbed.

Stanford Computer Science Department (CSD). Their researchers will coordinate the overall project, implement the testbed, and carry out research in all areas described here.

Stanford Libraries. The Stanford University Libraries will contribute their digital collections available through the Folio System, will participate in testbed evaluation, and will provide a user community of 12,000 students and 1,400 faculty.

WAIS Inc. WAIS Inc. sells a supported, commercial version of the highly successful WAIS software. They will provide the project with their commercial software, will help in accessing WAIS collections on the Internet, and will participate in the evaluation of the testbed.

Xerox Palo Alto Research Center (PARC). Xerox PARC is a world-renowned research laboratory that pioneered many of today's technologies, such as the personal computer and the Ethernet. They will help implement the SIDLP Testbed, and will actively participate in much of the proposed research.

The research team we have described proposes to investigate and demonstrate the technology for an integrated, scalable, economically feasible, and usable Digital Library. The expected accomplishments of this endeavor are:

- An architecture and protocols that make it possible to combine large numbers of heterogeneous and autonomous information repositories.
- Novel Digital Library services for accessing, manipulating, and safeguarding information.
- Innovative interfaces that provide understandable access to diverse collections of information and make it easy for users to become information providers themselves in adding to the on-line collection.
- A functional testbed that demonstrates the developed technology on a substantial body of information and a varied number of information sources.
- Fundamental insights into the principles of Digital Libraries and their implementation.

In summary, current trends indicate a tremendous growth in the number of digital information sources. Without technology like the one we plan to develop, these sources will remain independent "atoms," closed off from others. Our technology will "crystallize" these atoms into a solid and coherent Digital Library, offering seamless access to the constituent, autonomous information sources.