Toward a Digital Library Strategy for a National Information Infrastructure

Robert A. Coyne
Harry Hulen

IBM Federal Systems Company
3700 Bay Area Blvd., Houston TX 77058

Abstract

Bills currently before the House and Senate would give support to the development of a National Information Infrastructure, in which digital libraries and storage systems would be an important part. A simple model is offered to show the relationship of storage systems, software, and standards to the overall information infrastructure. Some elements of a national strategy for digital libraries are proposed, based on the mission of the nonprofit National Storage System Foundation.

1. National Information Infrastructure Background

Two bills before the current session of Congress call for the creation of a National Information Infrastructure. The bill before the House of Representatives is called the "National Information Infrastructure Act of 1993" [1], and a somewhat similar bill before the Senate is called the "National Competitiveness Act of 1993" [2]. Whether or not either of these bills are passed, the fact that these bills have reached the level of serious committee discussion has a far reaching impact on the storage industry. The Senate bill states that "While the private sector must take the lead in the development, application, and manufacture of new technologies, the Federal Government should assist industry in the development of high-risk, long-term precommercial technologies which promise large economic benefits for the nation, ... and cooperate with industry and academia to help create an advanced information infrastructure for the United States. The term "information infrastructure" is defined in the Senate bill as "a network of communications systems and computer systems designed to exchange information among all citizens and residents of the United States."

Both bills propose to support the development of digital libraries as part of the information infrastructure. Some of the key provisions which relate to the underlying storage systems are,

- "Development of advanced data storage systems capable of storing hundreds of trillions of bits of data and giving thousands of users simultaneous and nearly instantaneous access to that information;"
- "Development of means for simplifying the utilization of networked databases distributed around the nation and around the world;"
- "Encourage the development and adoption of common standards and, where appropriate, common formats for electronic data."

The references to information infrastructure, storage systems, and digital libraries in the proposed legislation demonstrate an important shift in the perception of what constitutes our national information assets. In 1987, the Executive Office of the President, Office of Science and Technology Policy, published "A Research and Development Strategy for High Performance Computing" [3], which found its way into the High Performance Computing Act of 1991. The four areas of research supported by the program which came to be called the High Performance

Computing and Communications Initiative were high performance computers, software technology and algorithms, networking, and basic research and human resources. Storage systems were supported only indirectly, to the extent that they were needed by the computing and communications elements. The proposed legislation, which is heir to and which references the HPCC legislation, still emphasizes networking and various aspects of computation, but the acknowledgement is there that storage systems are an integral part of the information technology that forms our national information infrastructure.

2. A Model for National Information Infrastructure

A simple model for the components of a national information infrastructure is diagrammed in Figure 1. At the top of the figure is a layer representing users. As defined by Congress, the users are the American people - children in school, individuals in their homes, and entrepreneurs creating new opportunities and new jobs. To say this another way, the users are not just academic and government researchers.

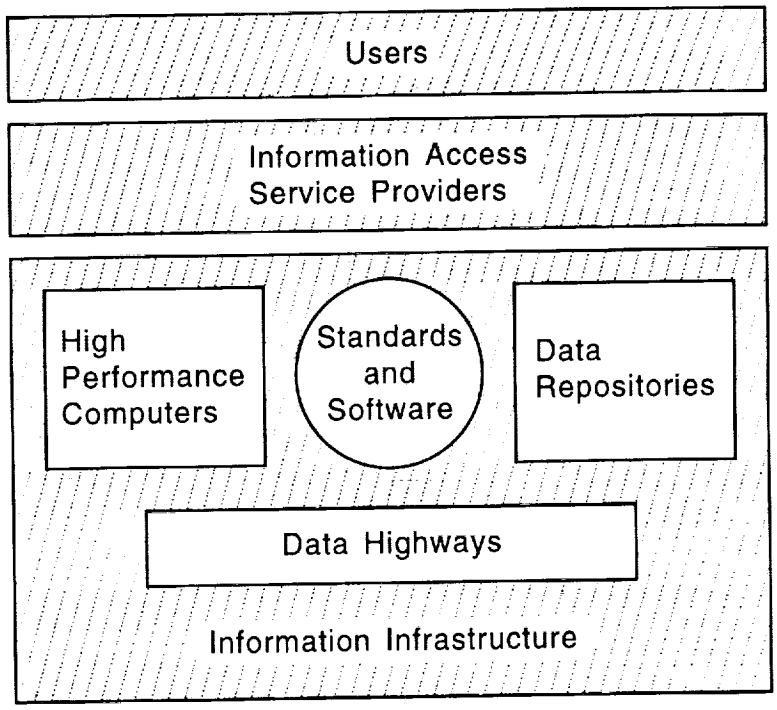

Figure 1. A Model for Information Infrastructure

To access the information infrastructure, there must be a layer of service providers. In our model the information access service providers are a layer of entrepreneurial service offerings which make the nuts and bolts of the infrastructure available to users. In a recent Business Week article entitled "The Cleavers Enter Cyberspace" [4], the lead sentence refers to this layer by asking, "Is Middle America ready for Internet?" The article describes how services such as Prodigy, CompuServe, GEnie, and America Online are offering Internet access to hundreds of thousands of users. Other examples of services available today are access to stock market information, airline reservations, banking, and thousands of forums and bulletin boards. Will industry find it profitable to deliver access to the medical research libraries of NIH, weather and environmental data bases of NOAA, earth observation data bases of NASA Goddard, and the astronomical catalog from JPL? The expectation is that by developing the infrastructure and educating the public, opportunities will abound to provide value-added access to the nation's information assets.

The information infrastructure layer itself is the focus of our interest in this paper. In our model, the infrastructure consists of four components: high performance computers, data repositories, the networks which it is now fashionable to call data highways, and the standards and software which enable the other components to work together.

High performance computers and networks have had center stage for the last several years. Our model takes a cue from the proposed NII legislation and elevates the data repositories to an equal billing. Repositories are the massive storage systems comprised of disk arrays, high density tape, optical media, and robotic media libraries. The fourth item, Standards and software, has been acknowledged by HPCC and NII sponsors to be a key enabling component. In our model, standards for information infrastructure would encompass standards for storage systems, such as the ones being developed by the IEEE Storage System Standards Working Group. Software would encompass the file systems, database management systems, storage servers, intelligent data movers, and physical volume managers.

3. Toward the creation of a United States digital library strategy

We must all seize the moment to formulate a strategy in plain, simple English to support the National Information Infrastructure with technologies for digital libraries, storage systems, and software. One organization which is working toward formulating such a strategy is the National Storage System Foundation (NSSF), a newly formed division of the National Storage Industry Consortium. NSIC is a not-for-profit business league chartered in California. The mission of the NSIC's NSSF division is to promote and support joint academic, industrial, and governmental research in information storage systems and software. The following outline for a digital library strategy is based on the mission statement of NSSF.

Develop core technology for high capacity, high performance digital libraries and storage systems

Digital libraries and storage systems can be very large, they can be geographically distributed using high speed networks, and they can be complex systems containing many kinds of data and many varieties of hardware and software components. We must, of course, be concerned with developing the core hardware technologies which provide the physical storage for digital libraries. We must also be concerned with developing the overall software architecture of the digital library systems. This includes research and development of key software components that are not yet available and integrating the hardware and software to create digital library systems. Other architectural issues are the ability to scale storage system both in size and in performance, to distribute them geographically, to make them secure, and to allow nondisruptive insertion of new technologies.

Develop technology for simplifying access to digital libraries

The successful deployment and utilization of the National Information Infrastructure is dependent upon massive amounts of data, stored in digital library systems, to be readily and easily available to consumers of this information. We must work to promote the development of software and systems which will make this possible. This includes technology to categorize and organize data, methods for optimizing data organization for rapid retrieval, and technology for extracting metadata. It includes technology to search, filter, and summarize large volumes of data. It includes technology for handling text, images, sound, and numerical data. It includes user interfaces using graphical and expert system technologies as well as automated access from other computers.

Define a coherent digital library infrastructure

The components which define the effectiveness of a digital library system, whether on a local or national scale, include computers, software, storage hardware and networks. Subcomponents include the security environment, the systems management environment and many other facets of

information access and retrieval. For a large number of these component and subcomponent areas, standards exist or are being developed. We must work to identify usable, coherent environments from the available choices and to identify areas where additional work needs to be done.

Establish requirements for buildable components

Not every company has the interest or resources to build an entire storage system. Our goal should be to define storage systems in such a way that specialists can build a software component with reasonable certainty that it can operate in a system with software components from other sources. As with hardware, the definition of components is a combination of historical precedent, feasibility, and standards. By defining building blocks and interfaces that conform to standards, we can enable different organizations to develop components in their areas of expertise with the confidence that these components will work with components from other developers.

Encourage the development and adoption of standards

We must encourage the establishment of standards for the storage industry through the IEEE Storage System Standards Working Group, ANSI X3, and other standards organizations. The IEEE Mass Storage System Reference Model has already taken significant steps to define the broad outlines of a future scalable standard for open storage system interconnection.

Promote interoperability among components from different vendors

A critical factor in the rapid market acceptance and deployment of digital library systems is the ability of hardware and software products from many vendors to seamlessly operate together. We must promote the idea of interchangeable and/or interoperable hardware and software components. To this end, we should support the development and use of standard test cases and reporting procedures to validate compliance with defined standards and interface definitions. Compliance enforcement is not a function of the IEEE and other standards organizations. This makes it possible to establish clearinghouses in which interoperability of digital library components and systems can be tested and demonstrated. Clearinghouses could be commercial operations, nonprofit organizations such as NSIC/NSSF, or an informal network of researchers and users.

Promote collaborative research projects for next generation digital libraries and storage systems

We need to initiate more joint research projects among industrial, university, and governmental research units to focus on the core technologies of the next generation of information management systems, including scalable, high performance, high capacity digital libraries and storage systems. Collaborative research proposals from the storage system community will help to focus governmental and industrial research funds on the development of digital library and storage system technology. Collaborative research will go a long way toward ensuring openness and interoperability among components. Collaborative research will lower the cost for everyone of building the national information infrastructure.

References

1. U.S. Congress, *National Information Infrastructure Act of 1993 - H.R. 1757*, Washington, DC, July 13, 1993.
2. U.S. Congress, *National Competitiveness Act of 1993 - S. 4*, Washington, DC, January 21, 1993.
3. Office of Science and Technology Policy, Executive Office of the President, *A Research and Development Strategy for High Performance Computing*, Washington, DC, November, 1987.
4. Schwartz, Evan I., "The Cleavers Enter Cyberspace," *Business Week*, October 11, 1993.