CHAPTER 6 *Summary and Recommendations*

Sketch of a Digital Library Initiative

E.A. Fox

Editorial Note: This subsection lays out my attempt to sketch key aspects of a national initiaative. It must be remembered that NSF normallywill fund research, development, and prototyping efforts such as those which are essential to the success of a broader National Initiative in this area.

A. Purpose
To advance U.S. science and engineering efforts, particularly research, education, and technology transfer, by improving the availability and supporting technology for access to useful information.

Note: We launch this in 1993, the 250th anniversary of the birth of Thomas Jefferson, who insisted that the free and vigorous pursuit of knowledge was essential to a democracy.

In modern terms: To develop vehicles and intelligent navigation methods for our nations' information highway and knowledge network.

B. Goals

- To develop a prototype national digital library that will be a proof of concept, showing it can be constructed, and that it yields important benefits.
- To develop digital information bases in key areas, that will be usable and useful, and should assist those working in the selected areas to be more productive.
- To develop improved methods and technologies for accessing (searching, browsing, routing, etc.) digitally encoded information.
- To strengthen U.S. research and, thereby, industry in this area, making it more competitive in the global community.
- To set a clear direction for coordinated development of larger national digital libraries, so that efforts in this area are compatible and systems are interoperable.

C. Objectives

- To leverage the network and computer technology from HPCC by adding content to be transmitted over the network, and by providing methods for more efficient and effective access to that information.
- To learn what methods work best in building a digital library, by observing the efforts and monitoring the progress of several top-rated groups as they work on the construction.
- To pick the best areas for building information bases, maximizing what can be learned in the process and maximizing the benefits to users that result from having the information.
- To learn what methods work best in accessing a digital library, by observing the efforts and monitoring the progress of several top-rated groups as they research this problem. To ascertain what works best as a function of the content (area, level), users (knowledge, experience, skill), and tasks. Key user communities should be included: K-12, undergraduate, graduate, practicioner, unemployed being retrained in new technology.
- To demonstrate the interdependence of information bases, methods for accessing them, and the quality of scientific and engineering research and education.

- To set in motion a self-sustaining government/industry cooperation that will ensure all publicly available information and knowledge becomes easily and universally accessible.

D. When Can We Tell This is Done?

- NSF's role in developing information bases will diminish once there are enough data/information/knowledge bases to support most of the types of important research that is needed, or when other agencies or corporations have taken over this effort. The first will be measured by asking proposers to justify why further development is needed for their research. The second will be observed when bases needed for research are existent or being developed by other groups.

- Development of information bases will be undertaken by government agencies like Library of Congress, National Library of Medicine, National Agriculture Library, or by other groups or corporations charged with that responsibility, or embarking upon this work as part of the free enterprise system.

- Research in these areas should be ongoing, though not necessarily as well supported if good efficiency, effectiveness, and usability has been scientifically verified based on known approaches.

E. Activity Context

- This responds to NSF's long range goal for "investment in human resources."
- The research undertaken deals with central concerns in computer and information science, especially those dealing with data, information, and knowledge — including their accumulation, processing, transformation, access, and utilization — as well as large scale systems to support those processes.
- This relates to other NSF goals: discovery of knowledge, building upon knowledge, ...
- This continues funding that goes back to the 1960's in the area of information access, but which has not been supported at this level in recent years. Thus, the aim is to build upon and discover new knowledge.
- This is an active area of research, supported by branches of NSF like CISE. That support, however
    - is at a relatively low level
    - is not coordinated or applied to help with the development of Digital Libraries

F. Impact Monitoring Plan / Evaluation Plan

- Annual progress reports

- Annual meeting of all funded by the initiative to report on progress and to ensure coordination and interoperability. Part of this will be "testing" of several access systems at a time, controlling the information base, user, and tasks -- to compare efficiency, effectiveness, and usability. Part of this can be "canned" comparisons, and part should be live.

G. Management

- NSFshould be the lead agency. IRISshould be the lead division at NSF. Other agencies, foundations, state governments, etc. should be invited to assist so that maximal leveraging of NSF funds takes place.
- A coordination/standard setting function is needed so that all research projects can share results, especially digital library data and software. Once the scope of this function is agreed upon, all funded groups would be obligated to participate in discussions and to abide by decisions falling in that scope and agreed to by a majority.
- If possible, the prototype information bases should be launched on a rolling basis, so that lessons learned from earlier efforts can be incorporated in work on others.
- If possible, the key areas should be dispersed around NSF, so that all of the important areas are included: hard science, social/behavioral science, K-12 education (preferably, say, building a 10K volume equivalent from out-of-copyright works that are at the core of the K-12 education scene). Another area might be patents, to demonstrate technology transfer, with funding from the Department of Commerce. DARPA might pick some key area to fund, such as materials, based on strategic defense needs.
- A monthly caucus, possibly co-directed by NSF and DARPA,should be held, with all government agencies invited to the entire event, and at least part of the event (e.g., presentations, demonstrations, talks) open. These would help ensure coordination, and lead to spin-off technical or administrative task forces (e.g., to agree on suites of standards).

Included should be: Departments (Agriculture, Commmerce, Defense, Education, Energy, Justice), Libraries (LoC, NAL, NLM), and Agencies (CIA, DARPA, FBI, NASA, NIST), Key contacts include: Larry Hunter, Barry Jacobs, Alexa McCray, Gio Wiederhold.

CHAPTER 7 *Bibliography*

References

From the June issue of BYTE 1991 in the State of the Art section and in the Solutions Focus section:

State of the Art Section - Managing Infoglut:

Making Knowledge Pay by Christopher Locke, pp. 244-246, 248, 250, 252.

A good overview of the way the new information managing technologies can be put to efficient use in organizations, particularly in support of collaborative projects. The potential for expert systems is down played, while the expanding potential of hyptertext and structured forms of information is emphasized. There is mention of

the article "The Dark Side of DIP" (BYTE, April 1991, p. 193) on the increasing need for indexing to help manage document images, which is an article that every information professional should read.

The New Age of Documents by John E. Warnock (CEO of Adobe Systems), pp. 257-260.

Discusses the critical role of universal document interchange, including the tranfer of images and standard fonts, in efficient communication. The author describes the technologies of Matching Fonts, a techniques for transferring fonts between computers with a minimum of information overhead, and Carousel, the Adobe designed format to enable universal document exchange and sharing.

Electronic Books by Louis R. Reynolds and Steven J. DeRose, pp. 263-268.

The practical prospects for the introduction of electronic books are described, as are some of the likely immediate applications of such a medium. A number of large scale electronic book projects are discussed, along with the Motif, Intermedia and DynaText technologies which are presented as examples of what is now available. They also talk about the importance of SGML and other publishing standards - as well as those standards still under development such as the ISO HyTime and the DSSSL (Document Style, Semantics and Specification Language) standards - in implementing this technology.

Search and Retrieval by Earlene Busch, pp. 271-274, 276.

A nice short review of some of the new text retrieval methodologies such as document clustering, probabilistic pattern matching, term weighting and other techniques which replace or enhance Boolean searches in large textual databases.

SGML Frees Information by Haviland Wright, pp. 279-282, 284, 286.

For those of you that have not yet seen an SGML encoded document, this article is a must. It describes the impact SGML will have on efficient text management and describes how they are authored and tagged.

Resource Guide: Search and Retrieval Systems, pp. 287.

Prices and sources for 32 products that allow you to create and search text databases.

Referenced and Related Documents

[Editor's Note: Throughout this Source Book, many entries make references to their origins. This section has supplementary references to other influential publications.]

After the Electronic Revolution, Will You Be the First to Go?: Proceedings of the 1992 Association for Library Collections & Technical Services President's Program, 29 June 1992, American Library Association, 1993

Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, 1989 by Gerard Salton

Capitalizing on the Nation's Knowledge: Making Use of What We Know, A Statement by John Seely Brown before the Subcommittee on Science, Technology, and Space of the Committee on Commerce, Science, and Transportation, United States Senate, 15 September 1989

Excerpt from Introduction:

In its potential to amplify productive and innovative practices in the crucial years of international competition that lie ahead, the National Research and Education Network offers an extremely exciting prospect. Along with its generally recognized possibilities for research and education, we find latent with this high-performance network means to overcome some of the present and future challenges to the nation's entire manufacturing and technological base.

Communications in Support of Science and Engineering: A Report to the National Science Foundation from the Council on Library Resources, pp 5, 7, 9, 10, 18, 24, 25, 31, August 1990

Excerpt from Abstract:

During the summer of 1989, the National Science Foundation asked the Council on Library Resources to explore selected aspects of scientific and engineering communication with the objective of learning more about the relationships between information resources and scientific productivity. Such information is of use in the preparation of the Science and Engineering Indicators series of the National Science Board. The enclosed papers, which constitute the report of the project, are the result of CLR's effort. The topic is a complex one, but it is the Council's hope that what has been done will point the way toward productive areas for investigation.

Draft Preliminary Findings of the Rights for Electronic Access to and Delivery of Information (READI) Project, Robert Ubell Associates for Coalition for Networked Information, September 29, 1992

Excerpt from Objective:

Under contract with the Coalition for Networked Information, Robert Ubell Associates convened a series of three expert panels to explore whether and, if so, how contract law, in the form of licenses and related agreements between creators and users of published works, could be applied within the context of copyright law to ease the flow of networked information. The particular objective was to determine whether it is possible to develop a common set of terms and conditions for managing relationships in the emerging market for networked information.

EDUCOM Update, March/April 1993, Volume 2, Number 2, "EDUCOM Calls for A National Network Not a Federal Network"

Editor's Note: This article can be obtained by sending e-mail to LISTSERV@BITNIC containing the command:
GET HETERICK UPMAR93 EDUNEWS

RG Fiddles and DR Winterbottom, *The Feasibility of Electronic Accession of Archive Material at the Printers: British Library R&D Report 6061*, British Library Research and Development Department, 1991

Excerpt from Background Summary:

This report is the result of a study into the feasibility of acquiring material for storage in an electronic archive during the production process when it is first typeset and is held by the typesetter in electronic form. The premise to be investigated was that it may be feasible to take a copy of the electronic form of the material held by the printer or typesetter rather than converting the printed copy back to an electronic form at a later date. This would result in a copy of the material which contained text in ASCII code, which is compact and which would allow the recipient to use electronic search techniques to enhance the use of the material. A survey of printers was carried out to establish the current state of technology and of practives in the printing industry in the UK to determine the likely technical problems that would be encountered in copying electronic data.

Frakes, William B., and Baeza-Yates, Ricardo, Eds.; *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, 1992

Garrett, John R. and Alen, Joseph S., *Toward a Copyright Management System for Digital Libraries*, Copyright Clearance Center, Inc. March 1992

Excerpt from the Executive Summary:

"What is it about digital information that makes designing a workable copyright management system so important, and so complicated?"

Prepared by the Copyright Clearance Center under an agreement with the Corporation for National Research Initiatives, this Report seeks to define the key issues in copyright management in digital environments, and to describe the legal and institutional contexts in which any workable system must operate.

The Report discusses several assumptions about copyright management of electronic information in detail, outlining and responding to counter-arguments. There is also an extensive analysis of the U.S. and international legal framework for copyright management of digitized information, with a detailed discussion of major issues of concern to both rightsholders and users.

A Hybrid Systems Approach to Preservation of Printed Materials: A Report to the Commission on Preservation and Access, by Don Willis, Commission on Preservation and Access, November 1992

Internet Society, *The Internet Society News*, Winter 1992, vol 1, no. 1

Joint Study in Digital Preservation: Report Phase I: Digital Capture, Paper Facsimiles, and Network Access, The Cornell/Xerox/Commission on Preservation and Access, Project Managers: Anne R. Kenney and Lynne K. Personius, Cornell University, September 1992

Excerpt from Executive Summary:

Cornell University and the Xerox Corporation, with the support of the Commission on Preservation and Access, have collaborated for the past two years in a Joint Study to investigate the use of digital technology to preserve library materials. The primary emphasis of this study has been on the capture of brittle books as digital images and the production of printed paper facsimiles. Of equal interest, however, has been the role of digital technology in providing networked access to library resources, and preliminary work in this area has also been accomplished.

The Joint Study has led to a number of conclusions regarding preservation, access, electronic technology, and the role of the library. In particular, participation in this study has convinced Cornell of the value of digital technology to preserve and make available research library materials. Such digital preservation presents a cost effective alternate to photocopying, and -- subject to the resolution of certain remaining problems -- a potential adjunct or alternative to microfilm preservation. The greatest promise of digital technology as a preservation option is to improve access to materials. Cornell expects to work with others to find ways to resolve the remaining issues surrounding the use of digital technology.

Krol, Ed; *The Whole Internet: User's Guide and Catalog*, O'Reilly & Associates, Inc., September 1992

Methodologies for Intelligent Text Processing and Information Resources Management: Achievements and Opportunities: Report of the NSF Invitational Workshop on Future Directions in Text Analysis, Retrieval and Understanding, Jitender S. Deogun, Chair, NSF & University of Nebraska, 1993

The Organizational Phase of Project Open Book: A Report to the Commission on Preservation and Access, by Donald Waters and Shari Weaver, Library and Administrative Systems, Yale University, Commission on Preservation and Access, September 1992

> Excerpt from Introduction:
>
> The Yale University Library is now organized to move ahead with Project Open Book, the conversion of 10,000 books from microfilm to digital imagery. In the final phase of the Project -- the organizational phase -- Yale established a Steering Committee, including several faculty members, and created a project team. In addition, Yale conducted a formal bid process and selected the Xerox Corporation to serve as its principal partner in the project.
>
> In this report we review the purpose and scope of the project, outline the steps taken in this first, organizational phase, and finally, present a summary of the results to date.

Preservation of New Technology: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access, by Michael Lesk, The Commission on Preservation and Access, October 1992

Proceedings of the Workshop on Future Directions in Text Analysis, Retrieval and Understanding, NSF & University of Nebraska, February 1993

Talbutt, Palmer, *Rationalization and Focused Retrieval*, 1992

Technology, Scholarship, and the Humanities: The Implications of Electronic Information, Summary of Proceedings for September 30 - October 2, 1992, The American Council of Learned Societies and The J. Paul Getty Trust, 1993

> Excerpt from Foreward:
>
> This volume constitutes a summary record of an important conference. Its brevity is intended to offer a rapid grasp of the main discussion points and conclusions produced by the conference, rather than to convey the spirit and diversity of the event and its participants.
>
> The findings of the working groups are offered as guides to those whose decisions affect the creation and use of electronic resources in institutional, technical, scholarly, and philanthropic settings. Reflecting not just the voice of the scholar nor only that of the librarian or systems specialist, this volume, like the conference, has captured the conjoint thinking of diverse, highly placed experts from the many areas that bear on scholarly computing in the humanities.

Time Magazine, “Cyberpunk!”, pp 59-65, February 8, 1993

Time Magazine, “The Electronic Superhighway”, pp 50-55, April 12, 1993

A Workshop on Scientific Collaboration and Interaction with Remote, Unique Facilities: Planning a Sondre Stromfjord Upper Atmospheric Research Facility Testbed; A Final Report to the National Science Foundation, University of Michigan, 1992