
Library 2000

ANNUAL REPORT July 1, 1993-June 30, 1994

Academic Staff

J. H. Saltzer

Research Staff

M. N. Charity

Graduate Students

A. Alavi J. A. Hylton
Y. Ho A. J. Kass

Undergraduate Students

E. Adar G. M. Lee Seyon
C. P. Brown K. Tukovinit
G. D. Elcock Y. O. Yerushalmi

Reading Room Staff

N. M. Loui
P. Mickevich
M. T. Sensale

MIT Library Staff

T. G. Anderson L. J. Eisan, Jr.
M. W. Cook K. Glavash

Support Staff

L. M. Eastman
C. M. Molnati

Library 2000

1.1 Introduction

Library 2000 is a computer systems research project that is exploring the implications of large-scale online storage using the future electronic library as an example. The project is pragmatic, developing a prototype using the technology and system configurations expected to be economically feasible in the year 2000. Support for Library 2000 has come from grants from the Digital Equipment Corporation; the IBM Corporation; ARPA, via the Corporation for National Research Initiatives; and uncommitted funds of the Laboratory for Computer Science.

The basic hypothesis of the project is that the technology of online storage, display, and communications will, by the year 2000, make it economically possible to place the entire contents of a library online and accessible from computer workstations located anywhere. The goal of Library 2000 is to understand the system engineering required to fabricate such a future library system. The project's vision is that one will be able to browse any book, journal, paper, thesis, or report using a standard office desktop computer, and follow citations by pointing—the thing selected should pop up immediately in an adjacent window.

Our research problem is not to invent or develop any of the three evolving technologies, but rather to work through the system engineering required to harness them in an effective form. The engineering and deployment of large-scale systems is always accompanied by traps and surprises beyond those apparent in the component technologies. Discovering relevant documents, linking archival materials together (especially across the network), and maintaining integrity of an archive for a period of decades all require new ideas.

This is the third reporting year of the project. The first two annual reports laid out the vision and described the overall architecture of a testbed system. This year we report the first completed academic theses, the start of production scanning, and significant progress in designing and implementing the architecture. We have also met several times with the other members of an informal consortium, known as the CSTR group, which by agreeing to exchange online materials is developing both a useful library of computer science technical reports and interoperability experience. The other members of the CSTR group are Stanford University, the University of California at Berkeley, Carnegie-Mellon University, Cornell University, and the Corporation for National Research Initiatives.

1.2 Production Scanning

One of our cornerstone hypotheses is that scanned images of documents will be an important component of many future electronic libraries. To this end, an important milestone was reached during the year: We designed, acquired, implemented, and started running a production scanning system. In accordance with our view of its importance, this work was done as a cooperative activity with the Document Services group of the MIT Library System, which provides the professional staff that operates the scanning station.

The scanning station currently consists of a network-attached Macintosh Quadra 840AV equipped with 64 Mbytes of RAM, 4 Gbytes of disk, one DAT, two SCSI channels, and two SCSI-attached scanners: an Hewlett-Packard IICx and a Fujitsu 3096G. Software was initially a package named Optix, from Blue Ridge Technologies; we recently switched to a package named Cirrus, from Canto Technologies, together with a workflow manager written in AppleScript. As we gain production experience and conquer work flow bottlenecks, we expect this configuration to increase in disk storage and DAT capacity.

Our overall scanning strategy is to locate the earliest-generation originals that are available, and extract the maximum amount of information possible with production-capable hardware. The two scanners are capable of a resolution of 400 pixels per inch, with eight bits of grey-scale per pixel. (The HP scanner can also provide 24-bit color, but we are using that ability only for individual pages that intentionally make use of color.) This resolution produces image files that are quite large, about 16Mbytes per scanned page, thus requiring careful organization of storage and workflow.

At the current rate of scanning, we are acquiring about 1 Gigabyte of raw data per day. Our approach is immediately to compute a checksum signature of the raw data, archive that raw data to tape, and then to (reversibly) compress and transfer it out of the scanning station to a server site for additional processing before storing it for distribution. Currently, this additional processing consists of reducing the data to an agreed-upon standard interchange form: 600 pixels per inch, one bit per pixel, using CCITT Group IV FAX compression in TIFF format.

In interchange form, a typical page occupies about 100 Kilobytes of storage, about 0.5 percent of the size of the raw data. Only about 7 percent of the originally scanned information is retained in this reduction; the remaining factor of ten comes from compression. The archived raw data remains available for experiments and for future uses when management techniques for larger volumes of data are more refined.

Because the volumes of data are so large, bottlenecks of various consequence turn up at every step that precedes reduction to interchange form (scanning time, disk writing time, DAT writing time, checksum calculation time, compression time, network file transfer time, and reduction itself). The development of an effective workflow has involved quite a bit of trial, backtracking, and searching for alternatives. Initial production work therefore used second-generation originals, and because of complications in matching software and hardware capabilities, was carried out at 300 dpi, 8-bit grey-scale. About 40 technical reports, totalling 3900 pages and about 40 Gbytes of raw data, have been scanned as of June 30, 1994. Software is now in place to allow scanning at 400 dpi, and enough experience has been gained that a cutover to first-generation originals is expected sometime later in the summer of 1994.

It is apparent that when scanning a document, one should capture more than just the bits of the images, but it is not yet clear exactly how much more. Information about the resolution of the original scanner is needed to display the images correctly. Information

Library 2000

that image 59 is a color scan and image 60 is a monochrome scan of the same page is needed to interpret the set of images properly. Information about pagination and position of blank pages is needed for a browser to display the images in the proper order when a user says *next page* or *go to page 43*. Scans of standard test targets may also be added to the set of images of a document to allow calibration of output devices. We have begun to design a standard scanned document description record, whose purpose is to capture these various categories of information in a uniform, machine-readable form.

We are also exploring various issues that arise when scanning in gray-scale, such as scanner calibration and quality control in the absence of output devices that have information density as high as that of the scanners. We are experimenting with test targets from the Association for Information and Image Management and the Institute of Electrical and Electronic Engineers to see whether or not scanner calibration can be checked by computer analysis of the scanned images of those test targets, rather than visually. With this approach, we hope that visual quality control can be simplified to looking for obvious errors such as missing, askew, or upside-down pages.

In a closely-related task, Hylton and Yerushalmi wrote utilities to convert 300 dpi and 400 dpi gray-scale images to 600 dpi monochrome form for printing. These utilities are two of several needed at various processing points for images. The basic issue being explored here is that different image transformations are needed for printing, display on various output devices, and long-term storage. To speed up any particular display situation, one can do precomputation of the required form, but holding the results until they are needed requires storage; precomputation takes time that is wasted if the result is not actually used. The tradeoffs are complex and constantly changing with technology.

Detailed design, configuration, and debugging of the scanning station and its internal workflow have been done largely by Eisan, Hylton, and Kass, with software contributions by Adar. Overall image processing workflow has been addressed by the entire group, with additional advice from Ladd and Richter, who manage the publications of the Laboratory for Computer Science and the Artificial Intelligence Laboratory, respectively.

1.3 Linking Documents

One of the more interesting problems in an electronic library is implementing cross-references among documents in a way that will persist across hardware and software technology changes, library reorganizations, and decades of time. We have explored several different aspects of this problem.

The CSTR group has agreed to use RFC-1357 as a standard bibliographic exchange record. That standard contains an open-ended (which really means not-yet-defined) *retrieval* field, intended to provide the information needed to allow automated location of the rest of the document, such as the name of an FTP site and the name of the file at that site to transfer. It also defines an *ID field*, a standard two-component document identifier, such as STAN//CSL-TR-91-004. The first component (before the double slash)

identifies the distributor of the document; the second component (after the double slash) is the identifier that this distributor uses for the document.

In discussions about use of RFC-1357, it became apparent that the concept of embedding retrieval information in a retrieval field is flawed. All of the other information in a bibliographic record is relatively permanent, because it describes the document itself, which does not change very often. But the to-be-defined retrieval field presumably describes the network site, the method of obtaining the document, and the available formats, all three of which we can expect to change (perhaps frequently) with technology and with development of new ideas. In addition, this kind of retrieval information is properly the province of technologists, not of librarians; it is the only field in the bibliographic record that is not determined by bibliographic considerations. One can thus argue that including this kind of retrieval information explicitly in the bibliographic record violates some kind of packaging or abstraction boundary. The value in maintaining a set of information items about a document that is purely bibliographic is that the resulting bibliographic record can be freely circulated and there need be no inhibition about storing it for an indefinite time and forwarding copies to others.

Alavi undertook a M.Eng. thesis to explore the following idea: rather than adding a retrieval field to the bibliographic record, consider the ID field to be a universal name for the desired document, and present that name to a network name service, which would return the temporally fragile information needed for retrieval. With help from Charity, Alavi designed and implemented this system using the Internet Domain Name Service (DNS) to associate each ID as defined in RFC-1357 with an Internet Uniform Resource Locator (URL), a standard form for retrieval information that was developed elsewhere for the World-Wide Web. Alavi primed a DNS server with a set of ID-URL pairs for several thousand currently online technical reports of the five universities in the CSTR group.

To demonstrate the usefulness of this approach, Alavi's thesis went one step further, modifying the user interface of the Library 2000 catalog system so that whenever a search result includes a technical report, the system submits the ID found in the catalog to the DNS ID-URL server. If the server returns a URL, the modified user interface pops up an extra active button on the user's screen, giving the user the ability to retrieve the online version of the report.

This demonstration was successful, and it is now the basis for discussions with the CSTR group on the relation between IDs and *handles*, a document registration system proposed by CNRI. We are also monitoring the work of the Internet Engineering Task Force on proposals for *uniform resource names* and *uniform resource characteristics*, two closely related ideas.

At a higher level of linking, Adar developed a program to take technical report citations as found in typical published documents and identify the corresponding documents by performing a catalog search. This problem, which seems easy at first, is severely

Library 2000

complicated by the existence of varying citation styles, multiple forms for author and publisher names, incorrect and incomplete citations, and other similar real-world variations. As a result, heuristics seem to be an essential component of any automated lookup process. Adar's program selects several independently random sets of words from the original citation, performs a catalog lookup for each such set, and compares the results. It presents to the user the document or documents that matched several of the independent sets. The system has proven to be surprisingly effective for such a simple heuristic.

This citation lookup system is especially interesting when coupled with the text-image map developed by Hylton. This map relates scanned images to the corresponding ASCII text. The goal here is to allow two symmetric features. In one direction, the user selects a region of interest in a scanned image that contains, for example, a citation to another document or a page number. The image-text map allows the system to identify the ASCII text that corresponds to the selected region. In the case of a citation, the system can then invoke Adar's citation lookup program to locate the cited document. Or, if the selected region contains a page number, the system can, using the scanned document description record mentioned earlier, identify which page image corresponds to that page number and display it.

In the other direction, when a user initiates a catalog search using particular words or phrases and asks to display a document that contains them, the text-image map allows the system to locate the appropriate page of that document and highlight the region of the image that contains the words or phrases of the original search.

Hylton worked out methods of representing text-to-bounding-box mapping information and did initial demonstrations of both page number lookup and citation search. He also looked into exploiting a feature of some OCR software, such as Xerox ScanWorX, that returns not just ASCII text but also the coordinates of the bounding boxes in which it found the text.

As another linking and interoperability demonstration, group members created a tool to print technical reports from images. This package, when given the name and page numbers of a technical report at a different CSTR member site, will locate that technical report at the other site's storage service, grab the requested page images, and send them to a nearby printer.

1.4 Integrity through Replication and Continual Testing

A similarly interesting problem for an electronic library is maintaining integrity of a large collection of scanned images, some of which are used less often than the technology changes. Tukovinit undertook an Advanced Undergraduate Project (under a recent revision of the MIT EECS department curriculum, the Advanced Undergraduate Project replaces the undergraduate thesis) to begin developing techniques for integrity checking to help ensure long-term survival of massive quantities of online data. The basic survival concept we are exploring is to create geographically separated complete replicas of the

data. Most research on replication concentrates on atomic, reliable update rather than on long-term survival of rarely updated data. The most common operations on library data are to read and to add new data; the rate of intentional modification is probably less than the rate of accidental bit decay, so a different emphasis in design seems to be required.

Our current thinking on the design of this survival system is that it should be organized in independent components. A first component is a data reader, associated with each replica, that continually compares that replica's current data with a record of yesterday's data. A second component is a set of cross-checkers that compare the replicas. A third component performs any needed repairs to the data, based on the discoveries of the first two. The objective of the project is to create a design for these components that is so straightforwardly simple that one can make a credible claim that it will work correctly. Tukovinit documented this multicomponent model; Yerushalmi and Charity are now refining and prototyping the model. A major point of departure (and simplification) of the model they are currently working with is that a storage server should be immutable except for addition of new files of data.

1.5 Minimal Perfect Hashing

Ho undertook a M.Eng. thesis exploring the idea of using minimal perfect hashing to speed up indexing in large RAM indexes. With standard hash indexing, several different keys may hash to the same index position, and some mechanism (buckets, open-addressing, etc.) that involves multiple probes of the index is needed to separate them. Perfect hashing is a technique for precomputing a hash function that, for a given set of keys, produces a unique index for each different key. With minimal perfect hashing, the index set is, in addition, compact. The advantage of using minimal perfect hashing, then, is that one need probe the hash table only once for any lookup.

This idea has been explored theoretically, and some experimental implementations have been done, but as far as we can tell, it has not been tried in practice. Ho located one of these experimental implementations, upgraded it to production quality, installed it in the search engine of the Library 2000 testbed system, and did a series of performance measurements. His conclusion was disappointing: The time required to invoke the pre-computed hash function was comparable to or greater than the time required to perform multiple probes in an ordinary hash table, so there was essentially no performance advantage to perfect hashing, at least within the range of hardware configurations that we anticipate will be useful in a library. A full report on this result appears in Ho's M.Eng. thesis.

1.6 Library-Related Activities

Library 2000 is working cooperatively with the MIT Libraries to allow that organization to incorporate Library 2000 research ideas on large-scale system architectures and

Library 2000

data management techniques as an operational component of its developing electronic information services.

To this end, Library staff are working on the organizational issues of processing data: the bibliographic record, the database, and scanning and image management. Service issues will also be explored: the delivery and presentation of this information to users, and the integration of this service into the Libraries' array of services. Finally, the Libraries will serve as an evaluation testbed for the service as it is made available widely to the MIT community.

Within the five-member CSTR group, library representatives have also engaged these discussions. The CSTR group is planning for the day when the research project will end, at which time the CSTR libraries must be prepared to continue the services developed and to support an operational system that is sustainable, maintainable, and scalable to include other academic materials beyond computer science technical reports. Greg Anderson is coordinating an effort by the CSTR library representatives to identify, understand, resolve, and build the structures necessary for an operational system. At the CSTR research group meeting in Palo Alto in February 1994, he presented an outline of those issues identified to date. There are three areas of focus: processing, services, and evaluation.

Specific issues related to processing of technical reports include agreement on continued use of the bibliographic record format (RFC-1357) and that format's relationship with library catalogs and MARC (Machine Readable Cataloging) records. The library group is participating in the discussion of handles and the linking conventions between index records and the information content object, and is trying to understand how to create and manage handles in a scalable, production workflow. Each campus will have ongoing responsibilities for the database of its own CSTR images, and the library must be prepared to support, grow, and maintain the database. Libraries must be able to continue to convert existing paper-based technical reports for the foreseeable future and therefore must understand the issues related to operational scanning operations and image file management. Over time, however, some technical reports may exist only in electronic form, and libraries must be prepared to accommodate them in their normal workflow.

The CSTR group libraries have many issues related to the delivery of this information to their campus communities. The libraries have begun to address the presentation of technical reports online, their indexing and retrieval, and their integration into the library's existing array of networked electronic services. The library group will try to list and analyze the various clients being used on the consortium campuses and will work to integrate the CSTR services coherently into the libraries' electronic offerings.

The library group believes that the CSTR project should begin planning for the evaluation of the project as it nears completion—and should be prepared to feed those evaluation results into the ongoing development of the operational system.

The CSTR library group will continue to delineate this set of issues for the consortium and will intensify its planning effort in the next year.

1.7 Facilities

Under a hardware grant from the IBM Corporation, we installed processor and memory upgrades to three IBM RS/6000 servers used for indexing and storage services and for program development, along with four development workstations. Our server complement now consists of two model 530Es and one model 530, each with 0.5 Gbytes of RAM. The large RAM complement permits us to demonstrate software ideas using hardware configurations that we expect to be routine within a few years. The full-text index for the entire MIT library catalog is now resident in one of these servers, and another was used in Ho's thesis on alternative hashing techniques.

Charity explored the current market in magnetic disks and disk packaging. That exploration led to two conclusions.

1. The discounts available for purchasing magnetic disks in large quantities (on the order of 10 percent) are small relative to the rate at which market prices for disks are dropping with time (5 percent per month) so there is no point in buying ahead of need. A just-in-time ordering system (perhaps having the computer system note that it is filling up the last disk and automatically printing a purchase order for the next one) seems to be the right strategy.
2. Packaging at the level of placing 3.5-inch magnetic disk drives in arrays has not kept pace with the disk drives themselves. The best available disk packaging systems are very space-inefficient, wasting 75 percent of the box volume. We did find one very promising example; Digital Equipment Corporation has developed a nicely space-efficient system of disk packaging, but DEC has not yet figured out that it should market this packaging system as a separate product.

On this basis, we acquired an interim packaging system to allow us to store the images coming from the scanning system. We are hopeful that future orders of packaging systems may be able to take advantage of the developments at DEC. As of the end of June 1994, the storage servers were equipped with 22Gbytes of storage, mostly in the form of 2.0Gbyte 3.5-inch disks.

To obtain an additional storage server for replication experiments, we worked out an arrangement with MIT Information Services in which they provide an otherwise disused RS/6000 520, we equip it with magnetic disks, and we share the resulting server between the CSTR project and the TULIP project, another image-delivery experiment in which we are jointly participating. In addition to the opportunity to share hardware resources, we are getting some additional benefit by mixing and matching software packages, some of which we are developing and some of which they are developing.

Library 2000

J. H. Saltzer. Storage, the unnoticed revolution. Digital Equipment Corporation Systems Research Center, Palo Alto, CA, August 25, 1993.

