Technology, Networks, and the Library of the Year 2000

Jerome H. Saltzer

Massachusetts Institute of Technology
Cambridge, Massachusetts

Abstract: An under–appreciated revolution in the technology of on–line storage, display, and communications will, by the year 2000, make it economically possible to place the entire contents of a library on–line, in image form, accessible from computer workstations located anywhere, with a hardware storage cost comparable to one year's operational budget of that library. In this paper we describe a vision in which one can look at any book, journal, paper, thesis, or report in the library without leaving the office, and can follow citations by pointing; the item selected pops up immediately in an adjacent window. To bring this vision to reality, research with special attention to issues of modularity and scale will be needed, on applying the client/server model, on linking data, and on the implications of storage that must persist for decades.

1 Overview

The idea that computer technology could somehow be useful in libraries has been around for decades, and inspired and visionary proposals have never been in short supply.[1] Unfortunately, until very recently these ideas were interesting but academic, because the computer technology that was available was simply not capable enough to do the job.

The theme of this paper is that accumulated computer technology changes of the last decade, together with those expected in the next, finally make possible some of the ideas proposed at the dawn of the computer age. In fact, the rate of technology change has been so great that it will soon be commonplace to store page images, a possibility that seems hard to envision even today.

This paper has three main purposes:

- To describe a new vision of the high–technology electronic library.
- To examine the driving technologies.
- To identify the system engineering challenges.

Our overall approach is from a computer systems design perspective. That is, we consider a real application that would stress current systems technology, and then look at the stressed points for guidance on what systems problems need research.

1. For examples, see Vannevar Bush, 1945; John Kemeny et al., 1962; J. C. R. Licklider, 1965; J. Francis Reintjes 1984; Edward Feigenbaum, 1986.

2 The Vision

2.1 What, exactly, is a library?

The library is one of several closely related—even overlapping—information–rich applications that are being created or revolutionized by advancing computer and communication technology. These applications include:

- the library;
- news gathering and dissemination;
- on–line bulletin boards and discussion groups;
- electronic mail;
- personal files and databases;
- scientific, engineering, and business data banks;
- electronically assisted publishing;
- government & business reports;
- collaborative work.

The boundaries separating these several areas in some cases are appropriate, while in others they are artificial and only loosely related to technological realities. Many of these boundaries will be the subject of battles over the next decade, as people stake out revenue streams and novel ideas enter the arena. However, even though we can be certain that some boundaries will move, it isn't plausible to try to innovate across the whole area and at the same time within each area. Instead, we suggest that these various information–intensive fields will proceed by a process of successive approximation, with individual areas first working under the assumption that each will maintain roughly its traditional interfaces with the others. Then, as adjacent areas become comfortable with new paradigms, they will explore pair–wise negotiation of the boundaries that separate them. (There is no reason to believe that this approach is the *best* way to proceed, just that this is the way things will probably work out in practice.)

2.2 The Library's defining properties

Even if we aren't certain where the edges of the future library may lie, we need to locate its center. We therefore take as the defining properties of the future electronic library the following more or less traditional characteristics. The materials of a library are:

- *Selective.* A publisher or editor selects things to make available and a librarian or curator chooses ("collects") from among these published items. This selectivity characteristic distinguishes the library from, say, a public bulletin board, to which anyone can contribute without review, and the ultimate reader must perform all selection.
- *Archival.* The contents of the library are expected to persist for time periods measured in decades, and a user can depend on again finding things that were found there once before.

- *Shared.* The collection is used by many people. The activity of collecting is thus a shared and centralized one, and there are generally specialists (reference librarians) who stand ready to help users find things in the collection.

2.3 Technology and collections

The traditional concept of a library collection involves both the physical books and the catalog that lists those books. As our first observation about the impact of technology, we may note that in an electronic library these two parts can, and probably will, become much more independent. In an electronic library, the physical collection comprises a set of bits in computer storage that represent the words or page images of books, reports, and journals. The catalog is a set of references to those bits, organized in ways to make it easy to find things. The interesting opportunity is that, thanks to communications networks, the catalog (which we should now call the "logical collection") can refer not only to things in the local physical collection but also to things in the physical collections of other libraries. That opportunity carries significant implications.

In an electronic world, an item can be collected simply by including it in the catalog; if any other library anywhere in the network already has the item in its physical collection, it is not necessary for this library to acquire another physical copy of the file of bits that represent the item. Instead, it can simply place in its catalog a cross–reference to the physical copy in the other library. Communications thus make it possible to share physical collections, and one can even imagine future electronic libraries that consist exclusively of logical collections, a kind of space–age inter–library loan system.

Several interesting consequences flow from this single observation. One might expect to see new kinds of specialization in which some libraries concentrate on building up very large physical collections, while others instead focus on creating catalogs for specialized audiences. Publishers will be very interested in understanding how such sharing of physical copies will affect their revenue streams, and they may conclude that they should use copyright to restrict placement of their own publications to physical collections over which they have some control.

2.4 How Computers might help

Noticing the potential for separation of physical and logical collections has caused us to digress slightly from the first question we should have asked: how can computer technology help in a library? Traditional views of how computers might be useful in libraries concentrate on one of two quite different concepts, and advances in computing technology prompt us to propose that a third is now feasible:

1. *Discovery of relevant documents* ("Search" or "Information Retrieval"). For over thirty years, computer scientists have strongly focused on tools to help people discover things because of the potential both for finding things that would otherwise be missed and for saving time. Study has ranged from simple database queries ("find papers by Einstein") to knowledge–based measures of document "relatedness" ("show me documents like this one") and concepts with a distinct flavor of artificial intelligence ("find Broadway plays

that use plots from Shakespeare.") Progress has been slow, for several reasons. Probably the prime one is that "relevance" has proven to be an elusive concept. A second reason is that information retrieval ideas are hard to test—it takes a lot of effort to acquire a large enough body of on–line material with which to practice. A review of current research in information retrieval was recently published in *Science* magazine.[2]

Despite this limited progress, over–optimistic computer people have occasionally announced that they have just developed exactly the retrieval technology needed for the library of the future. The natural result of this series of premature announcements has been that librarians have learned to be very wary of the claims of computer people.

The second traditional use of computing in the library is

2. *Back office automation.* Behind the scenes in a library are several record–keeping and organizing activities that are quite amenable to computer support: acquisition (ordering books), preparing catalogue records, circulation (keeping track of checked–out books), overdue notices, serials control, and inter–library loans. Librarians have usually embraced this form of automation with enthusiasm, because it reduces drudgery and releases time for the intellectual aspects of librarianship. One side–effect of use of computers to input, edit, and review catalogue records has been the creation of on–line catalogs and, more recently, making those on–line catalogs available to library patrons. There are now more than 100 research library catalogs available on the internet.

The relevance of changing technology to libraries is that there has quietly emerged a third, new way in which computer technology can be useful in a library:

3. *Storage, browsing, and identification.* The computer system can be used as a bulk storage and browsing device, enhancing the speed and ease of access to a very large body of material. One way that access can be eased is by navigation: moving from one work to another by following citations. We can think of storage, browsing, and navigation as extensions of traditional cataloguing activities along two dimensions—to include the documents themselves, and to catalog their entire contents, for example in a full–text index.

Browsing and navigation involve *identification* of documents from their citations, a concept distinct from discovery (the traditional computer science interest mentioned earlier). Identification and discovery can be viewed as being at the opposite ends of a spectrum; on the one hand we have a more or less complete description of a desired document, as found in a citation, while at the other extreme we have only a vague inquiry in mind, not knowing whether or not anything in the library satisfies the inquiry. Once this identification/discovery spectrum is in mind, it is apparent that tools such as full–text search fall somewhere in the middle and are likely to be among the first available in practical systems.

In these terms, we can now more specifically identify the opportunity: if modern technology can support storage, browsing, and identification at an attractive price, a very

---

2. Gerald Salton, 1991.

useful system can be constructed, even without the potential enhancement of advanced discovery tools. The market for such a system could be vast—every communication–capable desktop workstation and personal computer in the world is a potential client for some form of this service. And the existence of such a system (particularly the large resulting collection of on–line books, journals, and reports) would speed up the rate of research on better discovery tools. So a bootstrapping opportunity is apparent; all we have to do is convince ourselves that the technology is capable of storage, browsing, and identification.

2.5 The Vision

Pulling these observations together leads to the following two–component vision of the future electronic library:

1. Anyone with a communication–capable desktop workstation or personal computer can browse through any book, paper, newspaper, technical report, or manuscript without actually visiting the library.

The primary implication of this first component is that the full text of all documents is on–line, in image form.

2. While reading a document, if one notices an interesting–looking reference or citation, one should be able to point to that citation, press a button, and expect the cited document to appear immediately in an adjacent window.

The primary implication of the second component is that there be a robust mechanism of connecting references with physical documents, thereby giving the library the feel of a hypertext system.

Note that this vision does not propose to replace books, but rather to augment them. We assume that there will still be a way to obtain a paper copy of the book for detailed study. The primary goal of the envisioned system is to allow the library user to browse the book to ensure that it is of interest, before going to the trouble of calling it from the stacks. Anyone who has found it necessary to go beyond the reference collection in a large research library and call books from compact stacks, closed stacks, or repository storage will immediately recognize the potential for saving huge quantities of time, both for library staff and for themselves.

3 The Four Advancing Technologies

Four technologies are driving the opportunity to create an electronic library:

- High–resolution desktop displays.
- Megabyte/second data communication rates.
- Client/Server architecture.
- Large capacity storage.

We explore each in turn.

3.1 High–resolution desktop displays

Displays commonly seen today are not very comfortable to use in reading scanned images. However, it turns out that they are just below a critical psycho-optical threshold, above which they become quite acceptable for browsing and perhaps even for extended reading. The change required to cross that acceptance threshold is the addition of shades of grey—at least eight levels. Since this feature can usually be implemented by simply adding random–access memory to the display controller, it is already standard on many high–end desktop workstations, and it is likely to become a standard feature of virtually all computer displays. Thus we can expect that usable display technology for the electronic library will be widely available well before the library itself will be on–line.

3.2 Higher data communication speeds.

Megabyte per second data communication speeds are gradually becoming available over community–sized distances such as from the office to the nearest library, and Megabit per second data communications from there to more distant major libraries. Thus data communications, both campus–sized and nationwide, now or soon will permit moving a page image from library storage to a display workstation in about a human reaction time, again at reasonable cost. There is both bad news and good news associated with this change in data communication technology. The bad news is that, because of the very large installed base of older, slower, communication equipment, it will probably take quite some time for these higher speeds to become widely available, for example, to residential locations. As a result, the range of locations from which an electronic library will initially be usable may be limited. The good news is that the technology and economic improvements that have become available are so dramatic that entrepreneurs are looking to devise ways to bypass the traditional telephone-based data communication installations, using techniques such as radio and cable. One would expect the good news eventually to overcome the bad news; the only question is when. Since we are looking toward a time that is nearly ten years away, there is hope that the available technology will be in place.

3.3 Client/Server architecture.

The client/server model, an organizing method in which a network links multiple computers each separately dedicated to distinct functions, has proven very effective in large–scale systems such as Project Athena[3]. The client/server model has matured to the point where it is directly applicable, and it looks like it may provide exactly the right modularity tool for dealing with several problems that traditionally inhibit technological progress in the library. Of the four advancing technologies, applying this one appears, perhaps, to be the easiest: it is an off–the–shelf technology that provides natural solutions to several problems: ubiquity, competition, cooperation, stability, modularity, performance, and so on. A later section of this paper describes these problems and explores the research aspects of verifying this claim.

---

3. George A. Champine, 1991.

3.4 Large capacity storage.

If there is a single technological advance that is most strongly driving the opportunity to build an electronic library, it must be the rapid decrease in cost of magnetic disk storage. Today's (1992) two-Gigabyte disk costs about $2500. It is instructive to examine the rate at which that technology has been changing. Figure one shows the size, in bytes, of the high–end, $2500, 5.25–inch "Winchester" hard disk drive over the last ten years. The slope of the line in that figure is astonishing—note the logarithmic scale on the left—the number of bits available for this price has been doubling every year for the entire decade.[4]

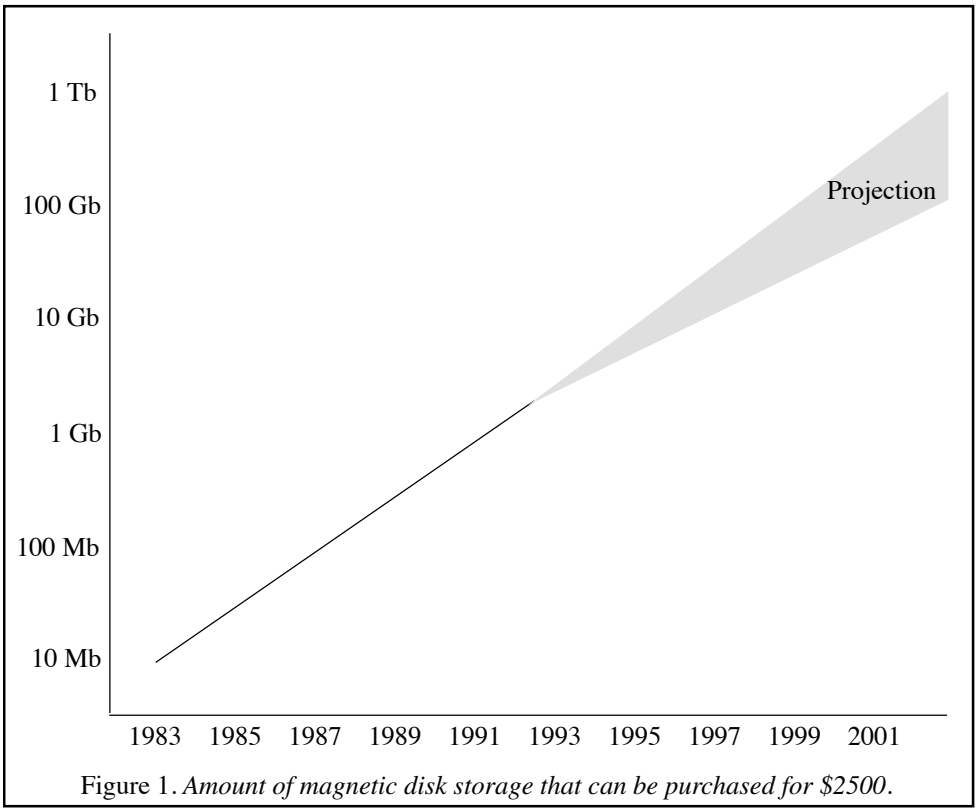

Figure 1. *Amount of magnetic disk storage that can be purchased for $2500.*

Going into the laboratory to find out whether or not this rate of cost improvement can be sustained, one learns that a primary constituent of the slope has been improvement in magnetic areal density (MAD), or number of bits per unit area. MAD has been climbing 40% per year for as long as anyone can recall, and current progress suggests that it will continue to rise at that rate for at least another decade. The greater slope (100% per year) for products in the field apparently came about because early disks did not attempt to approach the MAD limit. More recent designs have begun to narrow the gap, but at the

4. Based on street prices for 5.25–inch hard disks as appearing in advertisements in popular magazines over the period.

same time, arm positioning technology borrowed from the world of optical storage is beginning to show up in magnetic disks; this and other innovations may allow the slope to continue at the higher rate for a little longer. Based on these observations, we can with some confidence predict that disk cost over the next decade will track somewhere within the shaded area of the figure, and the bottom edge of the shaded area represents the minimum improvement that one might expect to see.

The bottom line is that there have been at least two factors of ten improvement since 1983, and two more factors of ten will probably accumulate by 2002, perhaps as early as 1999. The cumulative impact of these four orders of magnitude is that all prior assumptions about what is feasible to do with disk storage must be revisited; 100 Gigabytes of magnetic disk storage will cost about $2500 at the turn of the century.[5]

Probably the biggest impact of this cost improvement is that it will soon be economically attractive to store scanned page images in on–line disk storage.

If one scans a typical book in monochrome at 120 pixels/cm, one obtains about 300 Megabytes of raw data, or after compression, about 30 Megabytes.[6] Other resolutions should scale, after compression, with the square of the log, so doubling the resolution might add 20% to the bit count. For comparison, the character content of the same book would be about 1.2 Megabytes, about one twenty–fifth as much[7]. This ratio will be of some importance in choosing appropriate technology for indexing.

At 30 Megabytes per book, one of today's two-Gigabyte disk drives would hold the page images of about 66 books, at a cost of $40 per book and occupying a storage volume of 400 cm^3 per book, about one third of the space occupied by the book itself. By the end of the decade the equivalent disk drive should hold 3300 books, at a cost of 80 cents per book, and the space required will be less than 1% of the space occupied by the book.

For a library the size of M. I. T., which currently has about two million books, an array of six hundred such disk drives, costing about $1.5 million and occupying the space of a single small office, would suffice—about 60 Terabytes of storage. The U. S. Library of

---

5. Although much attention has been focused on the potential of optical storage, both read–only and write–once, the mechanical engineering of optical media happens to be driven by the requirements of the entertainment industry rather than the computer business. As a result, even though writable optical storage costs about one tenth that of magnetic disk, its performance in one critical area—its access time—is three orders of magnitude worse. This devastating performance difference makes it difficult to apply optical storage outside of a few very special applications.

6. This calculation assumes a 23 x 30 cm book of 400 pages. Such a book would have an average of about 450 cm^2/page of printed area and a total of 180,000 cm^2 of printed area to be scanned. Using the suggested scanning resolution would produce about 1800 data bytes/cm^2, or 300 Megabytes for a book. The group IV FAX compression algorithm, when applied to text, normally achieves about a ten to one compression ratio. Assuming that most books are dominated by text (images such as photographs do not compress so well) the average book would yield about 30 Megabytes.

7. Assuming 70 characters per line and 40 lines per page gives 2800 characters per page or about 1.2 Megabytes per book.

Congress holds about 90 million works of various types and collects another 30,000 per day[8]. Those works probably are on average somewhat smaller than the book on which we based our estimate, but if we assume they are all that size the Library would require about 2.7 Petabytes to start and an additional 225 Terabytes each year.[9] Putting scanned images on magnetic disk, in ten years, storage for the entire Library of Congress will fit on one floor of a small office building and the storage equipment will cost $60 million.

We can interpolate between these two cases to conclude that for most libraries, in a space much smaller than the present library and at a cost in the ballpark of one year's budget, the purchase of magnetic disk storage to contain scanned images of every document is feasible. Put another way, by the year 2000, storage of scanned images will seem so reasonable that it will be difficult to avoid.[10] Note, however, that this observation applies exclusively to the cost of the storage media; it does not offer any help in figuring out how to go about getting all those books, journals, and reports scanned; the cost of scanning, which is potentially quite labor–intensive, is another matter. Images of future publications may be materializable from the machine representations that were used in their preparation, but scanning of existing materials will probably not make much progress until it is forced by conservation requirements or storage space costs.

4 The Research Challenges

So much for the driving technologies. Availability of those technologies only enables the solution; creating a workable system involves tackling many interesting engineering problems. At the highest level, workable engineering of a system is a grand challenge: finding the proper modularity, finding techniques that simplify operations and maintenance, finding algorithms that allow working at very large scales. At the next level down, there seem to be three major problems, plus a list of more modest ones. The three major ones are:

- applying client/server design
- how to represent links
- persistent storage

---

8. From Michael R. Lawrence, *Memory and Imagination: New Pathways to the Library of Congress,* [Los Angeles: KQED: 1990]

9. One of the secrets of keeping up with the computer business is that one needs to learn a new metric prefix each decade. *Tera–* is beginning to show up as a common prefix just now, and *Peta–* will be with us by the end of the decade. If you want to keep one step ahead, *Exabytes* will be the unit of discourse in 2010.

10. Again, to check on the corresponding situation with optical storage, off–line CD ROMS can put 20 scanned books in a jewel box and thus already with today's technology apparently take up much less space than the books themselves. But jewel box storage is off–line, so we are discussing a very different kind of system. On–line storage would require a "jukebox," which brings the volume requirement to about one–tenth that of the book. Unfortunately the severe performance penalty that comes with jukeboxes makes the idea feasible only for rarely used materials, or with a complex and relatively expensive system of staging to magnetic disk.

4.1 Applying the client/server model

Although the client/server model appears to be well adapted to solving several obvious problems of applying technology to the library, each of these obvious examples needs to be verified in the field:

- *Hardware ownership.* Traditionally, the presentation device (for example, the display of a public access catalog of a library) has been owned by the library, so library budgets limit the range of locations and ubiquity of those displays. But with a client/server model, the presentation device can equally well be a workstation or personal computer that is owned by the customer. Thus ubiquity of access points can be achieved without the need for extravagant capital budgets.
- *Third–party sources.* With the client/server model, presentation management, customizing, and inquiry state all become the responsibility of the program that runs on the customer's workstation. Improvements in these areas can thus go on somewhat independently of the library itself, and can be the subject of third party competition. Network protocol standards can assure stable interfaces in the face of evolution of both user facilities and searching systems.
- *Function separation.* With a client/server model, one can easily and naturally separate indexing and search systems (the logical collection mentioned above) from storage devices (the physical collection.) This separation brings both the administrative benefit of decoupling physical from logical collections, thereby permitting sharing of physical collections, and also the performance benefit of allowing bulk storage to take place on a large, slow, cheap system without impeding search speed. Similarly, the circulation management system, which needs fast response, has traditionally operated in the same computer as the bibliographic search program, which soaks up lots of computing capability and degrades response to other activities. With the client/server model, one can separate these functions and place each on an appropriately configured computer.
- *Modular evolution.* Modularity also simplifies change. In traditional, monolithic library systems, any change is a big deal; the effort involved in changing everything at once inhibits needed change. But client/server components plug together like a hi–fi system, allowing modular replacement of any obsolescent component without replacing whole system. One would expect modular replacement to be the key way of achieving the system longevity requirement of a library.
- *Parallel inquiry.* One can make multiple, parallel inquiries to several search services, for example, to the local library, to the Library of Congress, and to a Books–in–Print server, so that the system can respond, "We have identified the thing you asked for, but we don't have it in the local library." Traditional catalog systems cannot distinguish between the two outcomes "Cannot identify the thing you asked for" and "This library doesn't have that item." They return the single response "Search failed," even though the appropriate thing to do next may be completely different in the two cases.
- *Reconciling alternative methods.* Libraries do bibliographic cataloguing of books, while for–fee services do the same service for journal articles, and the two worlds use independent, different, search methods. Similarly, different collections may call for different search techniques. These alternative views can be reconciled with a

client/server model that encompasses multiple collections with distinct search engines.

- *Cooperation*. Traditionally, inter–library cooperation is done at arms–length, yet much of any collection duplicates other collections. The client/server model makes inter–library cooperation technically straightforward.
- *Unregistered collections.* If one wants to extend a search to unregistered private collections, the client/server model again provides a natural mechanism.

Each of these concerns appears superficially to be well addressed with a client/server model, but field experience with real designs is needed to see if it actually works. As with all modularity proposals, the challenge is not just a matter of cutting the system into modules, it is finding the *right* cuts.

4.2 Links

The second area of research interest is links. A link is the cross–reference that allows one data object to mention another one, perhaps stored elsewhere in the network. In a library system, links potentially appear everywhere:

- Many papers, reports, and books contain explicit references to other papers, reports, and books.
- Some books are bibliographic reference works that consist of nothing but links, together with descriptions of their content and perhaps opinions as to their value.
- A user's request for "other things by this author" is actually an inquiry about an implicit work consisting of a list of works by the author; the user has asked about the links in that implicit work.
- Similarly, a request for "other things in this journal" follows links from its table of contents.
- A request for "other things catalogued as being on the same subject" invokes links provided by the librarian in preparing a traditional card catalogue.

Most research on distributed systems has been on a program–oriented model of cooperation (remote procedure call), in which one machine asks another to run a program. Links call for a different model of cooperation in which one machine needs to maintain over a long time references to data stored by the other. They require a carefully engineered blend of direct reference (for performance) and stand–offishness (for insulation against failures, change, and lack of cooperation). Links appear to involve, but are not limited to, the rendezvous provided by naming services.

The mechanics of links seem superficially to be straightforward. Given a citation, one needs to be able to locate the object cited by identifying it, what library holds it, and some method of obtaining it. The systems challenge is how to represent links, considering that the target of a link may be on the same machine, on a machine elsewhere running the same program, on a machine elsewhere running the same program but administered by someone else, on a machine elsewhere running a different program that is alleged to meet the same specification, or on a machine elsewhere running a different model of the universe.

A more extensive discussion of the research problems surrounding links appeared on the agenda of a recent SIGOPS European workshop.[11]

4.3 Persistence

A third research challenge is persistence—managing archival storage with a time horizon measured in decades, rather than years. Most experience in computer file system design is with data that is expected to persist with a lifetime of at most a few years. The occasional data set that must last longer may be handled as a special case. But the storage system for an electronic library must be designed for data that will virtually all be kept around for fifty years or more. Current systems aren't designed to handle data that is meant to be retained for times that are one or two orders of magnitude greater than the lifetime of storage media, data compression techniques, forward error correction techniques, and representation standards. Several observations come to mind.

At about the time the system reaches its storage capacity, one should expect that the disks that the system started with will be on the verge of becoming obsolete. But the proper goal is to preserve the information, not the disks, so part of the system design must be a component, similar in spirit to backup in time–sharing systems, that automatically moves data from obsolescent storage devices to newer technology without getting in anyone's way. This technology refresh component may well be running in the background much of the time, and its correct operation is critical to the success of the system. A carefully designed ceremony is required to copy the data, to ensure that it all gets copied and the new copy hasn't been corrupted by the copying process itself. Because copying will undoubtedly be a long–running job, it must be coordinated with updates and additions that are going on at the same time.

A similar concern arises surrounding using data compression to reduce disk space. Is it safe to compress data? How does one read data that was compressed 75 years ago, using techniques, algorithms, and programming languages that have long been superseded and then forgotten? One possibility is to try to store the compression algorithm with the data. But this possibility leaves one wondering how to devise a timeless description of the algorithm. Another, perhaps more plausible, answer is to decompress and recompress the data, using the latest compression technology, whenever the disks are being replaced, as part of the data copying procedure. If this kind of technology refresh procedure is used, it would seem advisable to avoid using the non–reversible (lossy) compression algorithms that have been suggested for moving video. Over the course of 50 years, one may have to decompress and recompress with newer standards five or ten times; losses from incompletely reversible algorithms would be expected to accumulate in the form of increasing degradation.

An almost identical argument about occasionally refreshing the technology applies to the use of forward error correction, or coding, to insure that data will be readable despite occasional media errors, but with an extra edge. In order to replace forward error correction coding it is necessary to remove the old coding, so while the data is being copied, it is

---

11. J. H. Saltzer, 1992.

unprotected and vulnerable to undetected errors. Thus a very carefully designed copying ceremony is required.

Finally, a threat that looms larger when decades are involved is media loss through disaster—fire, hurricane, earthquake, civil disturbance, flood, war, or whatever. For reliability, there must be more than one copy of the data, but traditional backup methods involving full and incremental copies made to tape do not appear to scale up well in size and they are notoriously complex and error–prone. In addition, the kinds of disaster listed suggest that the copies should not be in the same room, building, or city. One hypothesis is that one can approach the necessary reliability with geographically separated multiple copies, plus a change log that allows recovery from mistakes in updating. Of course if the data is replicated at multiple sites, then the previous discussion of forward error coding should be revisited. Perhaps error detection will suffice, and after a new copy, with new error–correcting codes, compression algorithm, and disk technology, is made at one site it can be compared with the older copy at another site to insure that nothing went wrong while the data was unprotected. Yet another aspect is that for information that isn't regularly used (e.g., the least–used 50% of a library's collection,) trade–offs among the number of copies, reliability, and geographical dispersion need to be explored; the best parameters may be quite different from those applicable to frequently–used materials.

Persistence, replication, backup, technology refresh, and the interactions among them in the electronic library application were explored briefly in two recent workshop papers.[12]

4.4 Other Research problems

There are quite a number of other interesting problems raised by the prospect of the electronic library. Some of them are specific to the library application, while others probably apply to any application that is enabled by the same four technology advances.

1. *Caching and replication.* The opportunity to share part or all of a physical collection among several libraries opens a question of when to share and when to collect a copy of a document. One might expect that more than one library should collect a physical copy of any particular document for reliability. Those libraries that find that their own users frequently use a document might collect a copy, either temporarily or permanently, to improve availability or to reduce communications costs. Finally, a library might collect a copy because it is not satisfied with the administrative arrangements and assurances that surround the copies already collected under other administrations. The trade–offs and balances among these three pressures are quite interesting and some field experience will be needed.

2. *Administration.* The administrative aspect of deciding when to collect a physical copy calls attention to a cluster of other political and administrative problems involved in negotiating the transition from the traditional paper–based library to the electronic version. First, one must maintain production continuity. Second, the cost of hardware is continually changing, generally in the downward direction. Although lower costs generally improve the

---

12. J. H. Saltzer, 1991 and J. H. Saltzer, January, 1991.

situation, they also present a dilemma of when to buy in. Third, the prospect of sharing physical collections, and the prospect of very effective access from the office or home both threaten to disrupt traditional revenue flows that have been negotiated among authors, publishers, booksellers, libraries, and users. (This concern usually shows up in discussions labelled "copyright" or "intellectual property protection.") More deeply, the concern for revenue flows may affect the fundamental structure of a library as a resource shared by a community. In an academic community, for example, there is a tradition that once a scholar has been admitted to the community, he or she can carry out library research without limit, which means without a fee proportional to usage. But for–fee information search services are already changing this tradition, in ways that may act to inhibit scholarly research. It seems likely that a scholar will behave differently when a meter is ticking, as compared to when one is not. Finally, new modes of organizational cooperation need to be worked out. Different organizations will be the natural providers of different physical collections, logical collections, and indexing services. New relations of inter–dependence among players must be worked out.

3. *Reference support.* One important function of a traditional library is that of the reference librarian, who helps users find their way through the collection. The corresponding concept in an electronic library is probably that a reference librarian works with a user remotely, using "collaborative work" techniques. The exact techniques, as well as the effectiveness, of remote reference help, remain to be discovered.

4. *Representation.* The question here is how to represent documents in storage. There appear to be many possibilities, for example, bit–maps representing scanned images, ASCII, PostScript, SGML, FAX Group IV, etc., but the requirement of storing the data for decades seems to reduce the field drastically, to forms that are simple and self–describing. It is possible that the right thing to do is to collect scanned images and minimally–tagged ASCII for every document, on the basis that those are the only representations likely to survive for a long time. A sub–problem is cross–representation coordination: how to identify a scanned image with the corresponding ASCII representation of the text. For example, how does one relate a mouse click on a displayed image to the corresponding words in the ASCII form?

5. *Variant copies.* When large numbers of documents go on–line, the need for coordination of variant copies will become pressing. It may be common that the local system has an old copy of something, while some remote system has an up–to–date copy but is currently out of touch or not available. The user interface, as well as the underlying storage and search systems, need to provide semantics to deal with this situation gracefully. To the extent that the information is textual and will be examined by a person, it may be reasonable to go interactive and offer the user an opportunity to choose, especially if well–thought–out defaults are part of the design. An interesting related question is how one discovers that two things from different collections are actually the same object, or a minor variation of one another.

6. *Large RAM.* As mentioned earlier, the space occupied by the scanned images of the pages of a book is about twenty–five times as great as the space occupied by the corresponding ASCII text. We can draw another interesting observation from that ratio by

comparing the cost of magnetic disk with that of random access memory (RAM). Specifically, we notice that the cost of RAM has followed a similar trend line to that of magnetic disk, and will probably continue to do so for the rest of the decade. Taking current street prices of $25 per Megabyte for RAM and $1 per Gigabyte for magnetic disk, we note that the cost of RAM is about twenty–five times as great as that of magnetic disk.

The somewhat startling implication is that if we can afford to place scanned images on magnetic disk, we can also afford to place full–text indexes of the contents of those images in RAM. Evidence that such an implication is reasonable abounds; today's personal computers are being delivered with eight Megabytes of RAM, and desktop workstations can be configured with as much as 0.5 Gbyte of RAM already.

Large RAM indexes provide another interesting subject for research; most research on full–text indexing is based on strategies intended to minimize the number of disk arm movements; completely different algorithms may be appropriate when the index for a large collection can reside permanently in random access memory.

7. *User interface.* A major challenge in a library system is to provide the user with a simple, intuitive model of what is going on, especially if multiple collections are being searched.

8. Resale architecture. The client/server model opens another opportunity, that a value–added reseller can repackage and offer to the public alternative access paths to library collections. The terms and conditions, as well as the technical aspects, under which such value–added services might be offered will require considerable thought.

Even the most casual reader will quickly think of several things to add to this laundry list of research problems.

5 Conclusion

In summary, we have claimed that advances in computer technology (especially in magnetic disk storage) will make a library of scanned page images with full–text search feasible within the decade. Networking will make it possible to share physical collections, and introduce the option of creating logical collections for purpose of searching and indexing. Networking will also make every desktop workstation and personal computer a potential access point for the electronic library. However, the availability of the technology is only one part; bringing it together involves many challenging tasks of system engineering, including getting the modularity right, arranging for an orderly transition from traditional methods, and identifying solutions that scale up in size in a satisfactory way. There are quite a number of research problems that need to be explored before an electronic library will actually be feasible.

5.1 Acknowledgement

Work on this subject began during a sabbatical at the University of Cambridge, and the author is grateful for extensive discussions there with Roger Needham, Karen Sparck–

Jones, Sam Motherwell, and many graduate students. Work continued during temporary assignments at the Digital Equipment Corporation Systems Research Center, where discussions with Paul McJones, John Detreville, Michael Burrows, John Ellis, Chuck Thacker, and Andrew Birrell were very helpful. At M. I. T., discussions with Gregory Anderson, Tom Owens, Mitchell Charity, David Clark, and David Gifford provided many ideas. Finally, over the course of the last two years, legions of librarians made extensive contributions by patiently explaining to me how their collections are organized in ways that are different from all other collections. Research support was kindly provided through grants from Digital Equipment Corporation and the IBM Corporation. Finally, because they had a significant influence on the thinking behind this paper, several otherwise uncited sources appear in the bibliography that follows.

Bibliography

Arms, Caroline R., ed. *Campus Strategies for Libraries and Electronic Information*. [Bedford, Massachusetts: Digital Press: 1990] ISBN 1-55558-036-X.

The Bibliothèque de France: a Library for the XXIst Century. [Paris: Etablissement Public De La Bibliothèque de France: October, 1990].

Brindley, Lynne J. "Libraries and the wired-up campus: the future role of the library in academic information handling," *British Library Research and Development Report 5980* (August, 1988).

Bush, Vannevar. "As we may think," *Atlantic Monthly 176,*1 (July, 1945) pp 101–108.

Champine, George A. *M. I. T. Project Athena: A Model for Distributed Campus Computing.* [Bedford, Mass.: Digital Press: 1991].

Dertouzos, Michael L. "Building the information marketplace," *Technology Review 94,* 1 (January, 1991) pp 29-40.

Evans, Nancy H., Troll, Denise A., Kibbey, Mark H., Michalak, Thomas J., and Arms, William Y. "The vision of the electronic library," *Carnegie Mellon University Mercury Technical Report Series 1* (1989).

Feigenbaum, E. A. "The library of the future," lecture given to mark the opening of Aston University's new Computing Suite, Manchester, England, November 11, 1986.

Kemeny, John G., Fano, Robert M., and King, Gilbert W. "A library for 2000 A. D.," *Management and the Computer of the Future.* Martin Greenberger, ed. [New York: The M. I. T. Press and John Wiley & Sons, Inc.: 1962] pp 134–178.

Licklider, J. C. R. *Libraries of the future*. [Cambridge, Mass.: M. I. T. Press: 1965].

Lynch, Clifford A. "Image retrieval, display, and reproduction," *Proceedings of the 9th National Online Meeting,* (May, 1988), pp 227-232.

Lynch, Clifford A. “Information retrieval as a network application,” *Library Hi Tech 32,* 4 (1990) pp 57-72.

Reintjes, J. Francis. “Application of Modern Technologies to Interlibrary Resource-Sharing Networks,” *Journal of the American Society for Information Science 35,* 1 (January 1984), pp 45-52.

Salton, Gerard. “Developments in automatic text retrieval,” *Science 253*, 5023 (August 30, 1991) pp 974-980.

Saltzer, J. H. “Fault–tolerance in very large archive systems,” O*perating Systems Review 25*, 1 (January, 1991), pp 81–82.

Saltzer, J. H. “File system indexing, and backup,” in *Operating Systems for the 90's and Beyond, Lecture Notes in Computer Science 563*. Arthur Karshmer and Juergen Nehmer, eds., [New York: Springer–Verlag,: 1991] pp 13–19.

Saltzer, J. H. “Needed: A systematic structuring paradigm for distributed data,” to appear in ACM SIGOPS 5th European Workshop, September 21–23, 1992, Le Mont Saint–Michel, France.

Tilburg University. *The New Library and the Development of Innovative Information Services at Tilburg University*. [The Netherlands: Tilburg University Press: 1989] ISBN {90-361-9662-0.