HOW EFFECTIVELY CAN COMPUTERS BE USED FOR THE SKILLED TASKS OF PROFESSIONAL LIBRARIANSHIP?[1]

William Y. Arms

Key-words
Librarians; Librarianship; Digital Libraries; Computers and Man - Relationship

Palabras-llave:
Bibliotecarios; Biliotecologia; Biblioteca Digitais; Computadores y Hombres - Relacionamiento

[1] Este artigo foi apresentado originalmente por William Y. Arms na D-Lib Magazine, v.6, n. 7/8, July/August 2000 - Copyright © 2000.

THE COST OF ACCESS TO RESEARCH INFORMATION

Libraries are expensive and research libraries are particularly expensive. Even in the United States, few people can afford good access to primary scientific, medical, legal and scholarly information. Members of major universities have excellent library services. So do people who work in teaching hospitals, or for drug companies or rich law firms. Others have access to information only through the tedious, inefficient system of interlibrary lending. In less affluent countries the situation is worse; even the best universities cannot afford good libraries. Must access to scientific and professional information always be expensive, or is it possible that digital libraries might change this sad situation?

The costs of a conventional research library fall into three main categories: facilities (which include buildings), library materials and staff. In digital libraries, the facilities costs are small, since digital libraries avoid the need for expensive buildings. Digital libraries require computers and networks, but these are relatively inexpensive, and the costs to users are shared with other services, such as electronic mail and word processing.

To build digital libraries that are inexpensive for users requires dramatic reductions in the costs of materials and of staff. Progress is being made in reducing material costs. Open access materials on the Internet are making many primary materials available at no cost. Some open access materials are identical to those that are available commercially. Others provide an acceptable substitute, such as Amazon.com as an alternative to Books in Print. For several disciplines, the open access materials are already good enough to support research. In an earlier paper, I discussed the economic forms that are supporting these open access publications and the strong economic reasons to believe that the volume of open access materials will increase (ARMS, 2000).

Hence, the key to inexpensive access to information lies in lower staff costs, which is the subject of this article. Big libraries are labor-intensive. Although salaries are low, staff costs are the largest item in most budgets, typically about half. The libraries at Harvard employ a thousand people and the Library of Congress more than four thousand. If professional and research information is to be available more widely, either users must bypass libraries, or libraries will have to employ fewer people. Over the past thirty-five years, libraries have automated routine clerical tasks, such as circulation or filing catalog cards. Is it possible that, at some future date, computers might assume the skilled tasks that now require professional librarians?

The term "automated digital library" can be used to describe a digital library where all tasks are carried out automatically. Computer programs substitute for the intellectually demanding tasks that are traditionally carried out by skilled professionals. These tasks include selection, cataloguing and indexing, seeking for information, reference services, and so on. The common theme is that these activities require considerable mental activity, the type of activity that people are skilled at and computers find difficult. Automated digital libraries should not be confused with library automation, which uses computing to reduce routine tasks in conventional libraries.

EQUIVALENT LIBRARY SERVICES

The remainder of this paper discusses the question of whether automated digital libraries can give good service to users. The short answer is that many aspects of automated libraries are a mirage -- always just over the horizon -- but some are surprisingly close or actually with us today. The underlying question is not whether automated digital libraries can rival conventional digital libraries today. They clearly cannot. The question is whether we can conceive of a time (perhaps twenty years from now) when they will provide an acceptable substitute.

Quality of service in automated digital libraries will not come from replicating the procedures of classical librarianship. More likely, automated libraries will provide users with equivalent services that are fundamentally different in the way that they are delivered. For example, within the foreseeable future, computer programs are unlikely to be much good at applying the Anglo American Cataloguing Rules to monographs. But cataloguing rules are a means to an end, not the end itself. They exist to provide services to users, notably information discovery. Automatic methods for information discovery may not need traditional cataloguing rules. The criterion for evaluating the new methods is whether the users find what the information that they require.

Consider the contrast between web search engines and conventional abstracting and indexing services or library catalogs. Almost everything that is best about a library catalog is done badly by a web search service. The selection of which materials to index by a web search engine relies on arbitrary considerations, the indexing records are crude at best, authority control is non-existent, and the elimination of duplicates leaves much to be desired. On the other hand, web search services are strong in ways that catalogs are weak. While cataloguing is expensive, indexing the web is cheap. The leading web search engines index several hundred million web pages every month, more than the total number of MARC records that have ever been created. It is wrong to claim that conventional catalogs or indexes are superior because of their quality control, and it is equally wrong to claim that the web search services are superior because of their coverage and currency. The value to users depends on what the user wants to achieve.

For medical research, no web search engine can approach the National Library of Medicine's Medline service. Medline has over 11 million references and abstracts. It is built by a team of indexers who have knowledge of bio-medical research, using indexing rules and MeSH subject headings that have been developed laboriously over decades. In contrast, web search services such as GoogleSM are entirely automated. The indexes are built by a team of computers with no knowledge of what they are indexing. Google has the advantage over Medline of indexing hundreds of millions of web pages, and doing so repeatedly every month. It is quite useful for finding general information on medical topics, but it does not index the major scientific journals, its indexing records are crude, it has no understanding of medical terminology, and makes no attempt to separate sound medicine from quackery.

It is a long way from being a substitute for Medline.

On the other hand, consider the trade-off between Google and Inspec, which is the leading abstracting and indexing service for computing. I used to be a regular user of Inspec, but have largely abandoned it in favor of Google. In many areas of computing, Google's restriction to open access web materials is relatively unimportant, since almost every significant result first appears on a web site and only later reaches the printed journals, if ever. Google is more up to date than Inspec, its coverage is broader and its indexing records are good enough for me to find what I am looking for. But its greatest strength is that everything in its indexes is available online with open access. In computing, substantially the same information is often available from several sources. Google provides a direct link to an open access version. Inspec references a formally published version, which is usually printed or online with restricted access. For my purposes, Google's broad coverage and convenient links more than compensate for its weaknesses.

This discussion highlights important differences between disciplines. For example, a scientist judges a library catalog simply as a tool for information discovery, while, for a humanist, the catalog may be a bibliographic source in its own right. Any predictions of the future value of web search engines compared with conventional abstracting and indexing services depend heavily on the publication and reading habits of people in different disciplines. However, as we look ahead twenty years, the most difficult part of the comparison is to guess how the automated tools will develop. We know that they will improve dramatically and we can anticipate that research habits will accommodate themselves to the new tools, but only a fool would attempt to forecast the precise changes that will occur.

BRUTE FORCE COMPUTING

The first serious study of what is here called automated digital libraries was at MIT in the 1960s, under the leadership of J. C. R. Licklider, and discussed in his 1965 *book "Libraries of the Future*" (LICKLIDER, 1965). This remarkable book described the design of what he called "procognitive systems" for the year 2000. It envisaged digital libraries based around "information-processing schemata" that would be free from the physical constraints of books and library shelves. He listed twenty-five desiderata listed for procognitive systems. Many of these are definitely in the realm of artificial intelligence. For example, one of his desiderata was, "*Converse or negotiate with the user while he formulates his requests and while responding to them*."

At the time that Licklider was writing, early experiments in artificial intelligence showed great promise in imitating human processes with simple algorithms. (For a contemporary view of this work, see Feigenbaum and Feldman, "Computers and Thought" (FEIGENBAUM, FELDMAN, 1963) Therefore, Licklider was optimistic that, within thirty years, advanced algorithms in fields such as natural language understanding would enable intellectual processes to be carried out automatically.

Thirty-five years later, we can see that many of the results that he predicted have come to fruition, but not all and not in the manner that he expected. The development of sophisticated natural language processing has been slower than hoped, with general-purpose software still on the distant horizon. However, while Licklider and his contemporaries were over-optimistic about the development of sophisticated methods of artificial intelligence, they underestimated how much could be achieved by brute force computing, in which vast amounts of computer power are used with simple algorithms.

The rate of progress in computing power is described by Moore's Law, that the number of transistors on a semiconductor doubles every eighteen months. This is roughly equivalent to saying that computing power increases 100-fold in 10 years or 10,000-fold in 20 years. Few people can appreciate the implications of such dramatic change, but the future of automated digital libraries is likely to depend more on brute force computing than on sophisticated algorithms.

Many of the most successful methods of artificial intelligence apply simple methods to huge volumes of data. An interesting example comes from the computer programs that play chess. The IBM system that is now equal to the world's greatest grandmasters is descended from Deep Thought, a student project at Carnegie Mellon University. The members of the Deep Thought team were not chess experts. Their breakthrough came from expertise in developing exceptionally fast hardware, which could analyze immense numbers of chess variations.

The potential for automated digital libraries lies in the simple observation that:

- Simple algorithms plus immense computing power often outperform human intelligence.

- Moore's Law tells us that the computing power will be available.

THE STATE-OF-THE-ART IN AUTOMATED DIGITAL LIBRARIES

The introduction of automated digital libraries is a continual process, much of it happening outside conventional libraries. Here are some current examples.

Information Discovery

Information discovery illustrates the complementary skills of computers and people. Humans are skilled at reading a few thousand words and extracting complex concepts. Faced with a billion pages (roughly the size of the web), they are helpless. Computers can index every word in a billion pages and search the indexes for simple patterns almost instantaneously.

The web search services represent the state-of-the-art in automated information discovery. Within each service lie a number of separate processes, each of which is carried out automatically, and each of which is constrained by the current state of computing. To build the indexes, a web crawler must decide which pages to index, eliminate duplicates, create a short index record for each page and add the terms found on the page to its inverted files. To search the index, the search engine must convert the user's query to a search command, match it against the inverted files, rank the results and return them to the user.

To anticipate the potential of automatic systems to rival the functionality of Medline in all disciplines, we need to examine the components of a system like Google and see how they could be improved. Moore's law predicts that computers will be 10,000 times more powerful in twenty years. With such computer power available, we know that the automatic search systems will be extremely good, even if no new algorithms are invented.

For example, to decide how closely a document matches a query would seem to require human judgment, yet standard methods of information retrieval do remarkably well. They use the power of computers to match simple patterns as a surrogate for the human ability to relate concepts. As humans, we use our understanding of language to observe that two texts are on similar topics, or to rank how closely documents match a query. Computers can estimate closeness of match by comparing word frequencies. One basic concept, developed by Gerald Salton at Cornell University about 1970, represents each document as a multi-dimensional vector and uses the angle between their vectors as a measure of the similarity of twodocuments (SALTON, McGILL, 1983).

Evaluating the importance of documents would appear to be another task that requires human understanding, but Google's ranking algorithm does remarkably well entirely automatically (PAGE, BRIN, 1998). The idea behind this algorithm is simple. Google ranks web pages by how many other pages link to them. It gives greater weight to links from higher-ranking pages. Calculating the ranks requires the algorithm to iterate through a matrix that has as many rows and columns as there are pages on the web, yet with modern computing and considerable ingenuity, Google performs this calculation routinely. As a result, Google is remarkably successful in presenting a user with the most important page on a topic or a well-respected overview. This algorithm was developed as part of the NSF-funded Digital Library Initiative.

Archiving and Preservation

The Internet Archive, directed by Brewster Kahle, provides a topical example of the economic advantages of automated digital libraries. The Internet is an extremely important part of modern culture and contains many materials that should be preserved for future generations. Each month, a web crawler gathers every open access web page with associated images. The Internet Archive preserves these files for the future and mounts them on computers available for scholarly research today.

The Internet Archive is not perfect. Only HTML pages and images are collected, no Java applets or style sheets; the materials are dumped into a computer system with no organization or indexing; broken links are left broken; and access for scholars is rudimentary. Yet the simple fact is that without the automated approach of the Internet Archive, these materials would already have been lost. Attempts to catalog and collect web materials using skilled librarians and archivists have floundered on the scale of effort needed to do even a rudimentary job.

Citations, Hyperlinks and Reference Linking

Citation analysis is a long-standing success story of applying computers to library information. Inspired by the efforts of Eugene Garfield, the founder of Science Citation Index, there is a long tradition of using citations as bibliographic measures (GARFIELD, 1979). Hyperlinks are the web's equivalent to citations. Since they are already in machine-readable form, they are amenable to algorithmic analysis. Google's ranking algorithm can be seen as applying the concepts of citation analysis to the web.

Hyperlinks refer to items or copies of a work, but citations normally refer to the work itself or a specific manifestation. Automatic systems are becoming capable of extracting a reference from a document and linking it to the digital object that it references. Currently, the most fully automated system for reference linking is the SFX system, created by Hebert Van de Sompel and colleagues at the University of Ghent (SOMPEL, HOCHSTENBACH, 1999).

Reference linking is one of the building blocks that are being used to build large-scale automated digital libraries. ResearchIndex is a digital library of computer science materials, created entirely automatically by Steve Lawrence and colleagues at NEC. It makes extensive use of citation analysis and reference linking (LAWRENCE, GILES, 1999).

It downloads papers from the web. If they are in PostScript or PDF, it converts them to text. It parses the papers to extract citations and the context for the citation. It provides users with services such as searching the entire text or the citations, listing the references within a paper, following the citation links, or displaying the context in which references appear. An interesting aspect of ResearchIndex is that it provides a way for users to submit corrections, automatically of course.

Beyond Text

Metadata is one of the foundations of librarianship. There are a number of projects that extract metadata from digital objects automatically. Perhaps the most remarkable is the Informedia project, led by Howard Wactlar at Carnegie Mellon University (WACTLAR et al., 1999). Informedia has the extremely ambitious goal of providing access to segments of video, such as television news, entirely automatically. Thus it includes algorithms for dividing raw video into discrete items, for generating short summaries (called "skims"), for indexing the sound track using speech recognition, for recognizing faces and for searching using methods of natural language processing. Each of these methods is a tough research topic and, not surprisingly, Informedia provides only a rough-and-ready service, but overall it is surprisingly effective. Moreover, many of the weaknesses of Informedia could be overcome by applying huge amounts of computing power. Informedia was another project of the Digital Library Initiative.

REFERENCE LIBRARIANSHIP

The job of a reference librarian ranges from helping users with the mechanics of using a library to tasks that require deep intellectual understanding. The ugly term "disintermediation" is used when users perform for themselves tasks that used to be carried out with the help of a librarian. Could we conceive of an automated digital library that disintermediates all the services that reference librarians now provide?

Information retrieval provides a good test case. The mechanics of searching have been almost completely assumed by computing. The current generation of scholars never experienced the tedium of reading through long lists of abstracts, searching huge card catalogs and following citations laboriously from journal to journal. Searching a card catalog required skill. Because of the labor in creating and filing cards, only a small number of entries were provided for each work. Since few users ever mastered the complex rules for the main and supplementary headings, or the intricate filing conventions, serious users turned to reference librarians for help.

Automated digital libraries can clearly help with the mechanics of searching, but information seeking is more complex. Some years ago, I wanted to find data to support Moore's Law. Specifically I wanted to compare the rate of progress in semiconductors, magnetic media and telecommunications. After searching for half an hour using the standard online tools, I gave up and asked a reference librarian. Half an hour later she provided me with the data that I wanted. There was nothing magical about the methods that she used.

She simply had more expertise in the idiosyncrasies of the information available and how to navigate through it. Automated libraries are a very long way from providing such insights.

In disciplines with complex organization of information, searching for information remains a skilled task. Often, knowledge of the subject matter is paramount and the experts develop their own skills. For example, carrying out legal research online is a basic skill that every law student learns. Historically, most doctors needed the help of a medical librarian to carry out an in-depth search. Now, even in medicine, the tools available to the user are sufficiently good that most searches can now be carried out directly by the user. It seems that automatic tools are steadily reducing the need for reference librarians in these fields.

However, consider a problem once set to a student by Marvin Minsky of MIT. How would we create a computer system to answer questions such as, "Why was the space station a bad idea?" (MINSKY, 1991). For many years, we have had computer systems that can search enormous collections of text for the phrase "space station", or simple variations. For this purpose, computers clearly out-perform human searching for both speed and accuracy. But consider the concept "a bad idea". What is the possibility of computers being able to examine any arbitrary text and look for such abstract concepts within it? The recognition that no existing computer could address such questions stimulated the student (Danny Hillis) to design new computer architectures and to found the company Thinking Machines, but even with the most advanced parallel computers, nothing on the horizon approaches human judgment in understanding such subtleties.

COST

This article ends as it began, with cost. Undoubtedly the greatest advantage of automated digital libraries is cost. Computing power is much cheaper than human expertise, more so every year. When money and time are available in abundance, skilled professionals have no equal. But they are always in short supply. In the United States, the National Library of Medicine is funded by the government and provides open access to Medline, but only rich lawyers can afford to use the legal services provided by Westlaw and Lexis.

Because of the cost, traditional library systems are selective. Indexing and abstracting services restrict their coverage to a carefully selected set of publications. Catalogs do not include potentially useful information from monographs, such as individual items in anthologies, subheading, captions, and so on. There are essentially no attempts to catalog the content of web sites, below the level of the whole site. Even so, these services are very expensive.

Automatic systems have no trouble with being inclusive; they have problems when they attempt to be selective. Their weakness is lack of precision. They exhibit what, in a human, would be called very poor judgment. Both ISI's Web of Science® and ResearchIndex provide users with interlinked scientific documents. Web of Science is a splendid service, but its production is not fully automatic, since ISI relies on skilled personnel for election and for key parts of the input process. Because the ResearchIndex methods are automatic and are applied to raw data, ResearchIndex inevitably has errors, while the skilled staff at ISI eliminate almost all errors, but at a cost. A library subscription to the Web of Science costs $100,000 per year. ResearchIndex is free.

To conclude, automated digital libraries combined with open access information on the Internet offer to provide the Model T Ford of information. Nobody would claim that the Model T Ford was a peer to the handcrafted cars of its generation, and automated digital libraries cannot approach the personal service available to the faculty of a well-endowed university. But few people could afford a hand-built car, and few people have easy access to a major research library. The low cost of automated digital libraries is already bringing scientific, scholarly, medical and legal information to new audiences.

FOOTNOTE

The examples in this paper have emphasized the sciences (notably computer science) and professions, such as law and medicine. These fields have characteristics that make them very different from the humanities and social sciences. A convenient argument would be that these differences are so fundamental that automated digital libraries will never extend beyond a few specialized fields. But this argument lacks depth; the distinction may be simply one of timing. Superficially, there appear to be no fundamental reasons why automated libraries cannot be effective in any field where a substantial proportion of the source materials are available in digital formats. There are tough technical and organizational problems, but nothing that cannot be solved in twenty years of natural evolution.

Finally, there is the organizational question, what is the role of libraries in developing automated digital libraries? It is no coincidence that, of the examples of automated digital library services listed in this paper, only SFX was developed within a library. Informedia is being developed in a center of computer science and robotics research, Brewster Kahle the leader of the Internet Archive has a background in supercomputers, ResearchIndex comes from an industrial research laboratory and Google from the computer science department at Stanford. The teams that build automated digital libraries are small, but they are highly skilled. (In March 2000, the Internet Archive had a staff of 7 and Google had 85 of whom half were technical and 14 had Ph.D. degrees in computing.) Research libraries, as organizations, have great difficulty in developing the technical skills and implementing the revolutionary changes that are needed for automated digital libraries.

BIBLIOGRAPHICAL REFERENCES

ARMS, William Y. ***Economic models for open-access publishing. iMP***, March 2000.

FEIGENBAUM, Edward A. , FELDMAN, Julian (Ed.). ***Computers and Thought***. Massachusts : MIT Press,1963.

GARFIELD, Eugene. **Citation *indexing*** *: Its Theory and Application in Science, Technology, and Humanities*. New York : Wiley, 1979.

LAWRENCE, Steve, GILES, C. Lee, BOLLACKER, Kurt. Digital Libraries and Autonomous Citation Indexing. ***IEEE Computer***, v.32, n.6, p.67-71, 1999. <http://www.neci.nj.nec.com/homepages/lawrence/papers/aci-computer98/aci-computer99.html>

LICKLIDER, J. C. R. ***Libraries of the Future***. Massachusts : MIT Press, 1965.

MINSKY, Marvin. ***Reported***. [S.l.] : Woods Hole, 1991. <http://www.cisp.org/imp/march_2000/0300arms.htm>

PAGE, Larry, BRIN, Sergey. The anatomy of a large-scale hypertextual Web sSarch Engine. In: ***Proceedings of WWW7***, Australia, 1998.

SALTON, Gerald, McGILL, Michael J. ***Introduction to modern information retrieval***. New York : McGraw-Hill,1983.

SOMPEL, Herbert Van, HOCHSTENBACH, Patrick. Reference linking in a hybrid library environment, Part 1: frameworks for linking; Part 2: SFX, a generic linking solution. ***D-Lib Magazine***, April1999. <http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html> <http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt2.html>

WACTLAR, H. et al. Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library. ***IEEE Computer***, v.32, n.2, p.66-73, 1999.

William Y. Arms
Cornell University
e-mail: wya@cs.cornell.edu