Expanding Digital Library Research: Media, Genre, Place and Subjects

Michael Lesk
National Science Foundation

Abstract

Is the most important word in *digital library* the word *digital* or the word *library*? Two major efforts, writing software to search text, and scanning existing books for online access, represent these two threads. But now digital libraries are expanding well beyond scanning books and text searching. We are in the middle of a great increase in the scope of digital libraries, without enough knowledge of what users either want or can use effectively. Research is making major progress in areas such as image searching, video data bases, and sound libraries, while major US libraries are expanding their digital collections in all areas of knowledge. However, we don't yet know what kind of information will make major changes in the life of the users. We can see fields such as molecular biology that have been transformed by the availability of digital knowledge, but we don't know yet how to produce that kind of transformation in other areas. We are still bridging the gap between traditional libraries and digital information retrieval, but need to take heart: an almost-completed bridge has incurred much of the cost, but the benefits are yet to come.

Introduction

The phrase "digital library", to some people, conveys the idea of a scholar, sitting at a wooden table and covered with dust, but facing a computer terminal instead of a book. Certainly the digitization of existing content is of vital importance. Yet to discuss only this activity would omit many of the most exciting projects now going on.

In particular, the new round of DLI research funding is supporting work in (among other things):

new media: sound, data, and software, along with more ambitious work with text, image and video collections;

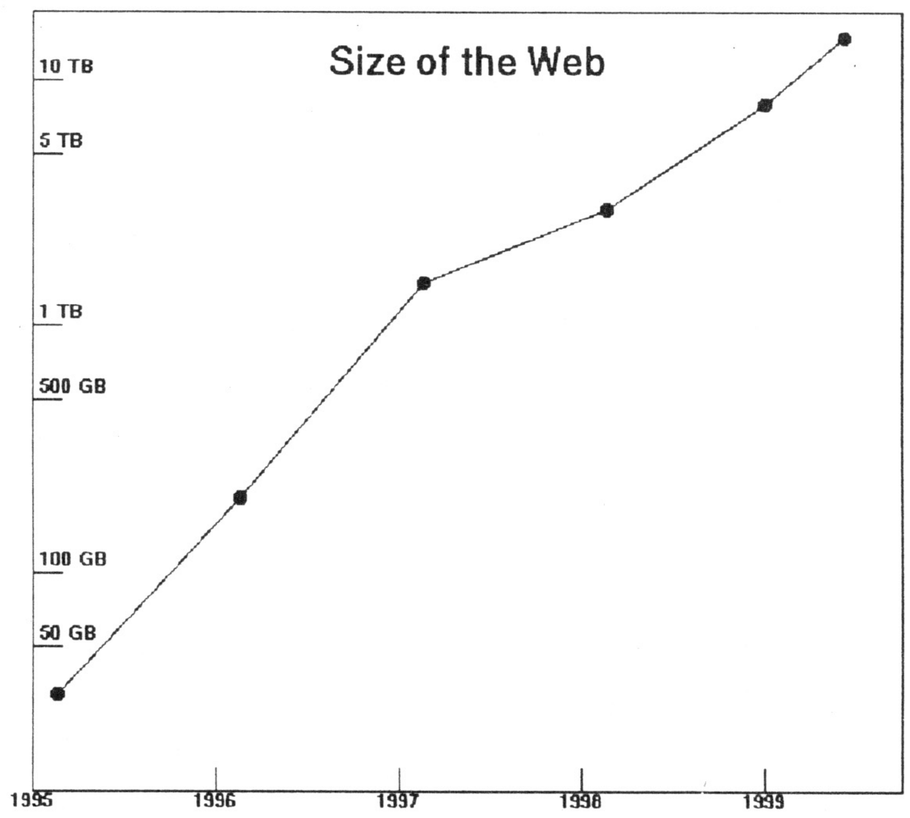
new genres: folk literature, popular songs, and literary manuscripts, adding to the explosion of material on the Web;

new places: up from six universities in four states to twenty universities in fourteen states, plus international projects linked to six different UK universities with more countries to come

new subject matter: anthropology, classics, social science, music, and many other academic disciplines.

Even the existing digital libraries have drastically changed some areas of research. Molecular biology research, for example, has been transformed by the online availability of the Protein Data Bank and the various genome data banks. Even more dramatic has been the change in the behavior of undergraduates. In universities and in secondary schools, professors and teachers are forcing students to use traditional libraries, since left to their own choice, many students do all of their research on the Web.

There is certainly enough material on the Web. The chart below shows the growth of the Web from 1995 to 1999. The change in the slope is



probably caused by the change from exhaustive webcrawling to sampling as the basis for the estimate; the Web is now too big for any crawler to easily gather everything. The popular search engines, for example, have not attempted to do complete crawling for some years, as they find it

more important to invest their effort in keeping up-to-date by revisiting important sites, than to gather every single page within some very large sites.

What we do not know is where the greatest support should be provided. Should we put our effort into conversion of the old, extensions into online publishing, or what? The US DLI-1 initiative was criticized as "90% digital and 10% library", and although the new DLI-2 initiative is broader and more balanced, we still don't know the right balance of different efforts. Although it would be convenient to pass the question to economists and let the market answer it, much of the funding for libraries has not been market-driven in the past. Governmental and university groups that have supported information collection and distribution in the past now face an embarrassment of new choices, and need more information to understand what should be done now.

Conversion of traditional materials.

Certainly, we wish not to lose touch with the past; the importance of conversion was shown with online catalogs. Libraries which attempted to convert their card catalogs on the principle of "old material remains accessible through paper cards, new books are cataloged only in the on-line system" found that after perhaps 1/3 of the catalog was on-line, students stopped taking the time to look in the card file. Retrospective conversion of the old catalog cards is the answer for many libraries, including the very biggest (Library of Congress, the British Library, and Harvard).

Among the most exciting projects in this area is the Universal Library at CMU, led by Raj Reddy. This has the ambitious goal of converting everything to machine-readable form; as a start, they would like to convert one million volumes. As of mid-1999, they estimate that about 60,000 books are in machine readable form. This is a complicated number; for example, the Bibliothèque Nationale de France has scanned 80,000 books, but since only about 5,000 are online, CMU does not count the others.

There are many other important retrospective conversion projects. Both the University of Michigan and the University of Virginia are speculating about converting all their 19th century books. The American Memory project at the Library of Congress is converting 5 million items to digital form, and the associated Ameritech Foundation projects administered through the Library of Congress support a variety of other conversion projects at universities throughout the United States. Project Gutenberg is an unfunded, volunteer effort in which people scan things and donate them in a cooperative effort reminiscent of "open source" software; it has been imitated in Italy, for example.

Perhaps the most significant conversion project is JSTOR, started by the Andrew W. Mellon Foundation but now independent, which converts major journals back to their first issue (and typically stopping five years before the present day). JSTOR is very broad, with a list of over 100 journals and growing, but focusses in areas such as history, politics, economics, and mathematics. Most important, JSTOR is aiming at self-sufficiency; it sells subscriptions to major libraries at a price set with the goal of permitting JSTOR to support itself without further operational funding from foundations.

Sometimes one is skeptical about the value of conversion of old material, thinking that it is rarely used. In a large research library, the average book is not touched in the average year. Yet routinely, when old material is converted to machine-readable form, users appear. Far more people routinely use the catalogs of the major libraries today than when you had to travel to a library in order to consult them. The American Memory Project is used by students of all levels, including primary schools. Amateur genealogists appear willing to read any 19th century document with a proper name in it.

The truth is that we do not know how many people wish to look at old documents of any sort. For comparison, no one looking at the number of movie theatres showing other than first-run movies in the 1970s would have predicted the interest in cable channels such as American Movie Classics or Turner Classic Movies.

Expanded sponsorship.

The US DLI-2 program, compared with the first set of projects which began in 1994, is a larger and broader effort. It received around three times as many proposals (215 in 1998), and they went to twice as many government agencies. The 18 funded projects cover a substantially wider range of subjects and media, and the program involves about twice as much money in total as the DLI-1 round of projects five years ago. The increase in activity, sponsorship, and breadth reflects the success of the field and in particular the success of the DLI-1 projects and the public attention and interest they achieved with their results. We can only regret that funding limits prevent still larger and more ambitious projects.

Most important administratively is the expansion of the group of government agencies sponsoring the program. DLI-2 is an effort of

1. the National Science Foundation (NSF),
2. the Defense Advanced Research Projects Agency (DARPA),

3. the National Endowment for the Humanities (NEH),
4. the National Library of Medicine (NLM),
5. the Library of Congress (LOC),
6. the National Aeronautics & Space Administration (NASA), and
7. the Federal Bureau of Investigation (FBI), in partnership with
8. the Institute of Museum and Library Services (IMLS),
9. the Smithsonian Institution (SI), and
10. the National Archives and Records Administration (NARA).

The new agencies joined the program as a result of seeing the DLI-1 results, and their participation has permitted widening the efforts in digital libraries, particularly into the medical and humanities disciplines. This is a clear instance of positive feedback operating: good research results attracted more supporting agencies and more financing.

In addition, major efforts in this area are funded by organizations such as the Andrew W. Mellon Foundation, the Ameritech Foundation, the Kellogg Foundation, and the Packard Humanities Institute. And a great many universities have programs that they fund themselves. The University of Virginia's Institute for Advanced Technology in the Humanities, for example, has been advancing this field for over six years. Harvard and Stanford both have substantial self-funded efforts at digital libraries. as do many other universities. Perhaps the most dramatic effort in this area is the California Digital Library, a joint effort of the entire University of California System and Stanford plus various State agencies and other California educational institutions. This will attempt to create the equivalent of another major research library, but in digital form.

And some private corporations are also creating digital libraries. IBM is well known for its involvement with the Vatican Library and the Seville Archives, as well as many other projects. University Microfilms is creating its "digital vault initiative," and planning to scan 5.5 billion pages. And Microsoft's *Terraserver* is an example of a giant database, in this case aiming for a complete set of photographs covering the Earth.

We still do not understand, however, the balance between government, private charitable, and free-market support of digital information. Much material that used to be sold for cash is now on the Web for free; for example, the US is discontinuing the National Technical Information Service since so many of the reports it sells are now available free on the Web. It seems unlikely that all the information now paid for can be provided free; there is simply not enough money floating around either the university sector or the research funding sector to replace the $20B or so now spent on non-fiction books in the United States. But the boundaries between the free publication and market-publishing sectors are certainly shifting.

Expanded media.

The new digital library research projects extend the media being studied. For example, Michigan State's new digital library will contain sound recordings of voices, while Johns Hopkins will contain both sheet music and software to render it into audio you can hear. Harvard will expand its center for providing political and economic data, including large archives of opinion polling data. At the University of South Carolina, the digital library research will look at software for the social sciences as well as data and publications.

Meanwhile, other advanced media studies are continuing. Carnegie-Mellon has important results in video libraries, including methods for searching based on speech recognition, closed-captioning, and image search. Other image search and classification projects will be at the University of California Santa Barbara and Stanford University. Some very broad projects, such as those at the University of California Berkeley and Tufts University, are looking at several kinds of media with the aim of looking at the most efficient interactions between them.

Production methods for new media are coming quickly. New software will help make the creation of video or audio material much less tedious than it has been in the past. And even newer media, such as interactive simulations, will become commonplace in digital libraries. DARPA is funding some exciting work called 'interactive drama.' The idea of this research is to allow a user to have a 'conversation' with some expert whose time is not actually being used; a system with a large body of the expert's writing to draw upon chooses the sentences to use to reply to the questioner. Even without this, there are already major resources in video as more and more universities (and other institutions) videotape courses, seminars, and other activities.

What media should digital libraries concentrate upon? Perhaps people prefer to look at images, or videos, or listen to voices rather than read the words being spoken. Does the greater comfort in hearing a voice rather than reading words outweigh the difference between a 125 word-per-minute narration and perhaps a 200-300 word-per-minute reading speed? Would it make a difference if software make skimming and searching

as easy for sound as it is for text? We don't know. We do know that it is important to give people a choice; some users, with limited vision or hearing, have strong needs to hear or see a particular form. But in terms of information delivery, the balance between formats is unclear and requires more research.

Expanded genres.

Traditional libraries focussed on serious scholarly publication. Most university libraries do not have the resources to collect popular fiction, amateur writing, and the like. Sometimes they have regretted this. For example, as literature departments started to look at the fiction being written in African countries, libraries often wished they had bought popular fiction from these countries earlier in this century, perhaps even before they were independent, now that these books are hard to find. Digital libraries can be much broader in their grasp.

For example, looking at the new DLI funded efforts, the University of California (Davis) is making a digital archive of folk literature. The University of Kentucky is looking at historical literary manuscripts. There are other projects looking at children's books, TV news, and many other kinds of efforts that expand the traditional content of libraries.

About two years ago I looked at 200 random queries from Excite and tried to classify them. The categories were

1. traditional reference queries, of the sort given to school librarians or public librarians. Examples are: *alexander graham bell*, *australian business commerce*, *apartments and virginia and manassas*, *solar eclypse*, or *plant+light+growth absorption.* A few queries are a suitable topic for traditional reference and I included them even though one might hesitate to actually ask them. An example was *making explosives pyrotechnics explosive propellants improvised detonation.*

2. Queries about popular concerns, that might be asked of traditional librarians, but which are so focussed on popular culture as to make them more frequently subjects of lunchtime conversation. They might be answered better in a library than by random friends, however. Examples of this category were: *kiss music concerts*, *honda accord*, and *celebrities addresses.*

3. Queries about the net and computers, such as *http://www.alc.co.jp./vle*, *corel*, *http://www.jpl.nasa.gov*, and *office97.* These queries make no sense in the absence of computers.

4. Sex-oriented queries, usually requests for pornography, such as *nudes sex pictures.*

5. Queries I could not easily classify, such as *aunt peg*, *apac*, *sol*, or *lh 4314.* Typically these queries are too short, or use abbreviations I didn't recognize, or where I didn't feel comfortable guessing (e.g. LH 4314 might be a Lufthansa flight number, but who knows?).

Obviously, this is sketchy; I only looked at a few more than 200 queries, and I might have made mistakes putting them in piles. For example, I don't play golf but I do know that *Big Bertha* is a brand name for golf clubs; otherwise I might have chucked that question in the sex pile. It was also the name of a German gun in World War I, a rocket, and a few other things, but the typical query involving Big Bertha seems to be part of an attempt to buy or sell second-hand golf equipment.

The results were:

Kind of Query	**Frequency**
Traditional reference	43%
Pop culture	13%
Sex	19%
Net and computers	12%
Unclear	12%

The implication to me is heartening: more than half the questions asked are basically on the turf of traditional librarians, even if many of them would not have been asked of librarians for reasons of convenience. They represent, however, something of a shift in the kinds of information that ought to be available online. A simpler example of the same thing is that a major online service is weather forecasting. Traditional libraries simply were too slow to serve such a function; not only do the patrons have to travel to the library, but except for daily newspapers most libraries did not shelve material fast enough to be of interest for weather reporting.

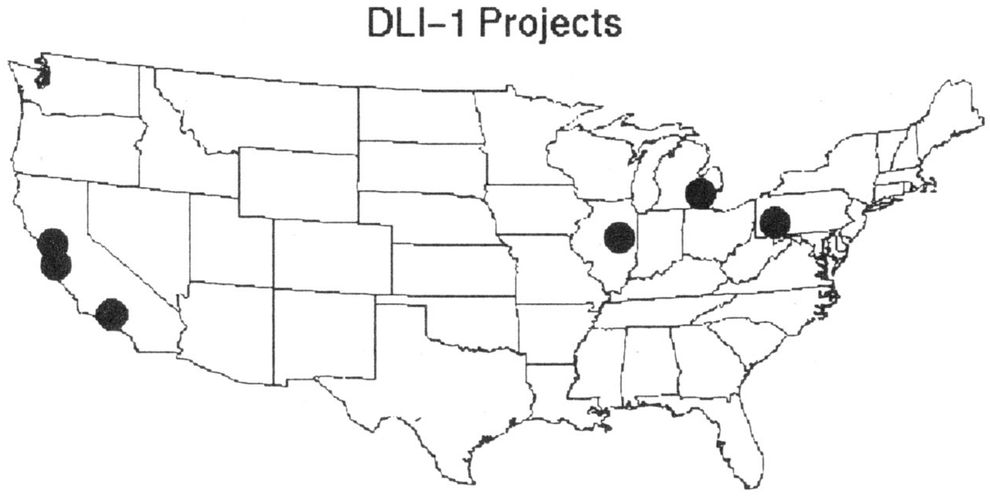
If libraries are going to expand the genres of material they hold, so that vastly more kinds of stuff are represented on their virtual shelves, we need to understand how, if at all, they should choose among this wide variety of material, what kinds of it are most valuable to provide, and what the users need and in what form.

Expanded locations.

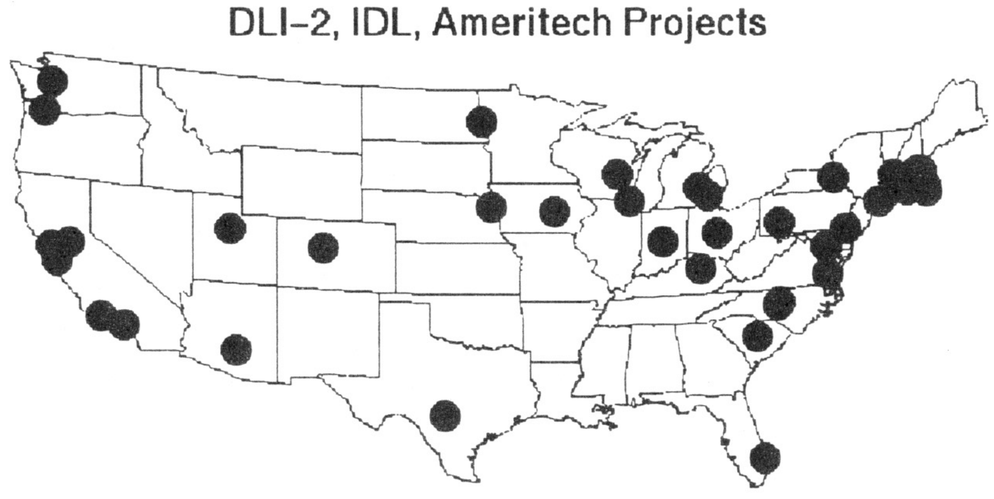
Digital libraries in the United States are no longer restricted to the biggest and most important universities. The two images below show the places in the United States where the DLI-1 projects were located and the places which have either DLI-2 awards, awards from the NSF international digital library program, or Ameritech

Foundation awards. Looking at all of the university self-funded digital library projects would cover even more of the map.

DLI–1 Projects



DLI–2, IDL, Ameritech Projects



International coverage is also expanding. The National Science Foundation announced, about a year ago, a program to support the US portion of digital library research by international teams. Almost simultaneously the United Kingdom announced its international program, and the awards from the first round of these applications are already made. The second round of reviews for international US applicants was held in August, and we look forward to additional international awards as more countries apply.

Conversion of past material is particularly important for expanding the geographic spread of digital library work. In terms of access to material, digitization tends to centralization; there is little reason to care whether the copy of the item you are reading resides nearby or far away, aside from the issue of having enough copies to be secure against loss. The spread of projects relates rather to the desire to publish, particularly to provide access to specialized materials in a particular library.

Expanded subjects.

Although the subject areas of the collections represented by the funded DLI projects have increased as we went from six projects in DLI-1 to triple that number in DLI-2, we are still dealing mostly with serious scholarly content. Some are quite interesting and innovative. The University of Texas, for example, will have a collection of anthropological models and images, while Columbia University works with medical informatics related to patient care. Earlier paragraphs have mentioned the Tufts work on classics, the UC Davis work on folk literature, and the Kentucky research on literary manuscripts in cooperation with the British Library. But these are still all subjects in which universities offer courses and degrees. Any random probing of the Web will turn up lots more - subway maps, used car listings, movie reviews, and conspiracy theories.

Pure technology is also common in the DLI-2 projects. Cornell and Stanford will look at as interoperability and security, the University of Arizona will study automatic classification. and research at Indiana will explore information filtering. The new topic of "data provenance" will be explored at the University of Pennsylvania. Summarization will be key to the work at Columbia.

Expansion of research groups.

When asked to explain the benefits of digital library research, the spin-offs come most quickly to mind. For example, one of the earliest search engines - Lycos - and one of the most effective new search engine ideas - Google - can both be traced to people that participated in the DLI-1 program. Yet the most important influence of the research funding may be on people, not on projects. Senior professors with successful careers in other disciplines, easily able to keep doing exactly what they have been doing, have changed their research to the digital library area. For DLI-1 this group included Hector Garcia-Molina and Robert Wilensky, and the DLI-2 group of investigators adds people such as Sidney Verba and Gio Wiederhold. Attracting the top researchers into a field matters more than the exact subjects they start pursuing. With the kind of people now doing this research, something good is bound to result.

Future needs.

Despite exhortations at conferences, the funded DLI-2 projects still leave major gaps in knowledge we need to move towards a goal of "everything on-line." We lack knowledge of

1. economic models for self-sustaining digital libraries,
2. user and societal needs, and
3. the necessary balance between access and preservation.

Economic models are probably the most directly important problem. Steve Harnad argues eloquently that publication will be so cheap in the

future that we don't really have to worry about subscription prices, and it is certainly attractive to think of everything one wishes to read being available free. Yet it seems unreasonable to think that the $1B of academic library purchases, or the $20B of non-fiction book purchases, can be replaced with an entirely free system. Librarians argue, quite correctly, that even if in the future the cost of running an electronic system is less than the cost of running the paper system today, for a while they will have to provide users with both paper and electronic information, and during that period things will be more expensive than either. Since every library, and in fact every university, is under financial pressure right now, there is no easy way to deal with this.

There are, fortunately, a few experiments being run. JSTOR has been mentioned above; the JSTOR pricing model involves a substantial up-front membership fee, and then relatively low costs for continuing access to the material. This strategy discourages universities from dropping JSTOR after the initial purchase. The overall cost is very cheap; it is the equivalent of buying journals at $1/year. JSTOR now has over 500 subscribers, and is approaching the break-even point at which it can continue to digitize additional journals based on its current revenues.

We may also get valuable data from High Wire Press, an offshoot of the Stanford University Library which serves as an electronic publisher for scientific societies. High Wire Press gives the societies it serves the choice of how to price their journals, and so it provides a mix for free and charged journals.

The most successful online subscription service is the *Wall Street Journal Interactive Edition,* which claimed over 300,000 subscribers by mid-1999 in their announcement of July 7, 1999. It is not clear if it is breaking even yet (the answer undoubtedly depends on cost allocation between the paper and electronic services). Dow Jones charges $59 per year, including some rights to back issue searching. *The New York Times*, by contrast, offers the current day's paper free, but access to back issues costs some $3.50 per article.

The ACM (Association for Computing Machinery) Digital Library is another successful example of electronic publishing. All of the ACM journals are available in electronic form, and some 30,000 ACM members have signed up for them (this includes some 10,000 students getting a discounted rate, but even eliminating them some 1/3 of the ACM membership has bought into the digital library.

Monograph publishing also has some experiments. Everyone is now familiar with the electronic ordering but paper fulfillment systems of Amazon.com, Barnes and Noble, and others. As of now these systems are losing money; Amazon, for example, is losing about $7 per book sold. Many university librarians must wish that they had the option of being as far from break-even as the free market seems to accept. Perhaps more relevant to digital libraries are the electronic delivery systems such as NetLibrary, which offers books from publishers such as university presses in an entirely on-line environment.

To the surprise of many, online delivery may not compete as directly as we think with paper delivery. Both the National Academy Press and Columbia University Press (international books) have tried putting full-text online for free, and in both cases have seen their paper sales increase. If such a model were to succeed in general, we would not have to worry about online charging.

Another major issue is that of user and societal needs. We do not know what users actually need in libraries. Page hit data, the new version of circulation records, tell us something about what is being used, but very little about who is using it, or for what purpose. A digital librarian trying to decide on what to spend scarce resources has little help figuring out whether the user community wants journals, monographs, videos, or other modalities.

A complicating difficulty is that the traditional scientific publication process is driven more by authors, seeking credentials for promotion and tenure, than by readers. Thus, acceptance of digital libraries involves not only readers but also authors, and in this case the author community seems more conservative than the reader community. Scholars fear they will not get adequate credit for material that appears online, and this has prevented the exclusively online journals from gaining wide acceptance. The strategy of simultaneous electronic and paper publication, at least so far, seems more effective.

Yet usage is very high online. George Gilder estimated that Web traffic on the Internet was 4 petabytes per month in 1998 (see the *New York Times* for July 5, 1998), which was 1000 times the size of the Web. Certainly the typical book in a library is not used 1000 times a month. People are certainly fetching, if not reading things. A check of Web logs kindly provided by Alexa Internet showed time spent per URL at a median of 23 seconds per page, and a modal time of 5 seconds per page (compared with time per page on chemistry journals, for example, of about 90 seconds).

We also do not understand the societal implications of moving from a world of paper libraries to a world of electronic libraries. Much has been written about the possibility of a 'digital divide'

in which poorer people, racial minorities, or rural inhabitants would have less access to information because they are less likely to own computers or have Internet access. Yet rural inhabitants clearly benefit from online services, since they are also more likely to have a long trip to a traditional library. And for reasons not yet thoroughly understood, rural inhabitants of the West are more likely to have Internet access than rural inhabitants of the South (see TechWeb for July 8, 1999). If the goal is to have good information everywhere, what do we have to do to achieve that?

Perhaps most important to me while I am at NSF is the implications for scientific research of having information online. Molecular biology and climatology, for example, have been transformed by online information and computational capabilities. Can we produce similar effects in other disciplines? This is not to say that the reason for progress in molecular biology is the existence of the Protein Data Bank, but it certainly helps accelerate work in an already thriving field. This may be as much a social question as a technical one: how can we arrange a reward system in which placing research results online is the encouraged as well as desirable behavior?

The tension between saving the old and providing the new is certainly not a surprise. Traditional libraries have dealt with this conflict for a long time. The British Library strategic plan commented that access and preservation were both important, but that access was something done for today's users and preservation was done for tomorrow's users. In the digital library context, there are two important kinds of preservation work: one is carrying forward digital formats into the future, and the other is the conversion of printed material to digital form.

Carrying digital material forward is the subject of considerable discussion (and some of our funded projects) at the moment. The problem is not storage cost, but the human effort required to do format conversion or even format understanding and recognition. Bill Arms has made an interesting triage-like suggestion. He proposes that a computer archivist will divide material into three categories: that which is valuable enough to justify the effort of converting to a new format, that which is so clearly worthless that it can be abandoned, and that which will be saved without conversion in the hope that some future archivist will have better automatic tools to do the conversion in an affordable way. He suggests that the vast majority of computer records may wind up in that third category, since storage is so cheap that almost everything can be saved if we wish it.

Conversion of printed material is more directly expensive, and amounts to spending our current effort on decisions made in the past. How much of that should we do? If all of our effort were spent keeping up with the past, we'd never get anything new done; and yet presumably our predecessors knew what they were doing when they chose material for publication, ordered it in libraries, and saved it from periodic shelf-weeding. Fortunately, conversion is getting steadily cheaper, and decisions that may be hard today will be easy tomorrow. Once upon a time (in the 1960s) the conversion of a single work of literature to machine readable form was a notable effort; today it seems trivial. Raj Reddy's million-book goal will probably seem as trivial in ten years.

We are clearly closer to merging the words *digital* and *library* than we have been in the past. But we still need more knowledge to do this effectively. We need to know what information is most beneficial for users; what methods will be most effective at getting that information online; and how to balance new and old, text and video, popular and scientific information. Sustainable funding for digital libraries is still elusive. But the quality of the people doing research today gives confidence that we will find the answers to these questions.