8

Today and Tomorrow: What the Digital Library Really Means for Collections and Services

Clifford Lynch

This chapter will take a hard and realistic look at some of the ramifications of digital information for library services, collections, and strategies. I used the phrase "digital library" in the title because I want to begin with this idea as a way of setting a context for the broader discussion that follows.

Digital Library

The phrase "digital library" is very fashionable these days. Everyone seems to be building a digital library. Not just libraries, but an amazing number of other institutions, organizations, and random individuals are building what they are calling digital libraries. It would be useful to explore what this term "digital library" really means to most people, because it is a problematic term to me. On one hand, it has an oxymoronic feel; on the other, it resonates with me, and I think with many other people, as a way of capturing how technology is really transforming not just libraries, but many profound social and cultural structures that are linked to institutions like libraries—for example, the acts of authorship and publication.

This paper is an edited transcription of the presentation Clifford Lynch gave at the Virtually Yours Institute.

One view of digital libraries that may be disconcerting to the "traditional" library community is that they are *information manipulation and use system environments.* (The adjective "traditional" is used only to distinguish between existing libraries that are trying to come to grips with the digital transformation, as opposed to all the other groups that are now positioning themselves as builders of digital libraries but that stand outside the historic institutional library traditions of selection, organization, access, and preservation.)

Many of the NASA/ARPA/NSF-funded digital library projects convey this perspective. These projects are designed for information users working in specific data-intensive environments. They are designed for people working with, for example, geospatial data or environmental data and related materials, who want to operate in a data-immersive environment.

When one looks at the computer science research world's view of digital libraries—as expressed by some of their systems—one sees no clear line between the library that is a storage place and active systems that facilitate communication and collaboration among researchers (e.g., control of experimental apparatus and sensor systems, data analysis environments, and authoring or annotation environments). The separation between readers and authors has become murky in some of these systems. One certainly does not get the sense they are thinking in terms of information that will persist for hundreds of years and of authors who will create works that will reach, and be reexamined by, many generations of future readers. There is a very strong focus on supporting current research activities. There is an important, although uncomfortable, message here for libraries as they think about how to draw the boundaries around their services in the digital environment. I will return to this issue several times in this chapter.

Before the computer scientists hijacked the term "digital library," many of us were really thinking more about "digital collections." We were thinking about libraries as we understand them today—extending their collections to encompass substantial amounts of digital materials, making use of information systems to provide access to them, and providing a coherence of access between their digital materials and their printed materials. People from the traditional library community look at the "digital library" and use that term as shorthand for any library that includes in its collection large amounts of digital material. They continue to see maintaining this coherence of access and coverage across collections, be they print or electronic, as an important goal.

None of us look forward to continuing and expanding the dichotomy we have historically constructed between books and journals to one in which you look one place for digital material, another place for printed monograph material, and yet another for journals. The problem is compounded by not being sure where to go depending on whether the journals are electronic or print. The trend and the objective over the last decade have been to provide greater

coherence among collections previously treated disparately. If one looks at the evolution of the online catalog into complex systems that encompass large amounts of abstract and indexing information (and the University of California's Melvyl system is a good example here), one can see a road map that leads to convergence and greater coherence. There are some very powerful forces—market forces, in particular—that profoundly threaten our movement toward that coherence.

Redefining Library Service

Having provided some context about the term "digital library," I will now feel free to abuse it as shorthand for "digital collections in the libraries." Let me suggest some issues to help us redefine library services and shape the evolution of an environment increasingly characterized by digital information and the systems required to access and coordinate that digital information. Thinking about these boundaries is difficult and makes many uneasy, because we really do not want to have any limits. Librarians want to do everything, even as the resources get tighter and tighter. We get this queasy feeling that we should be doing an infinite number of things, even though our expertise and our resources are so limited in the face of an infinite agenda of service demands. Somehow we have to set priorities.

Data Sets

One issue we need to consider is what I call data sets. This is digital content that you do not read. You may render it or manipulate it with the aid of computer programs, but it is clearly not just some pile of paper that someone has transferred to digital form. There are a few classic examples. Many libraries have struggled with what to do with census data. There is no question that this information is a fundamental part of the primary research material for a wide range of scholarship. Geospatial data, digital maps, remote sensing information, and molecular biology databases are important and will be on any short list of working scholars' resource requirements in the relevant disciplines. No one would suggest this is fringe or optional material.

These data have some key characteristics that make a computer essential for its use. Most of what people want to do with it is computation-intensive. Contrast this kind of resource to an electronic journal. While you need a video monitor to read the electronic journal, reading is not an computation-intensive act. Most libraries easily can afford to buy all the computer power needed to enable large numbers of people to read this electronic information on a screen, but when you talk to people about geospatial and census data, you hear terms like "super computer." Most libraries do not have budgets for super computers.

Given the other pressures on their budgets, most probably are not eager to set up a budget line and have demand for super computer cycles compete against the latest round of journal price increases, but using these data requires computation. Where is the computation going to happen? Who's going to provide it? How does the library as perhaps manager or repository of these data relate to those computational resources?

This is a difficult boundary question right now. How much computation is acceptable inside the libraries' core systems and how much is the library purely an information warehouse or repository that feeds computation going on elsewhere? As we see more and more data becoming essential to more and more scholarly pursuits, this will become an increasingly critical issue. It is easy to think that this problem is limited purely to the sciences and social sciences. Don't believe it. A whole new discipline of computational humanism, usually lumped under the phrase "computing in the humanities," is developing. These folks are generating a lot of data. Think about all of the marked-up manuscripts and other texts being created now that the Text Encoding Initiative standards are in place. Consider computational linguistics, computation-intensive studies, or word use patterns and statistics and how they change across manuscripts. This is serious computation, too. Most libraries are fairly comfortable with the notion that text-encoded, marked-up manuscripts are a reasonable part of their collection. Having said that, what service framework do they need? What do you do with them?

Let me raise another point about this computation-intensive information. Suppose that computing becomes so cheap that it is not an issue. (I don't believe it is going to get that cheap, because every time computational capacity doubles, people think of more things to do with it, resulting in a perpetual pushing of the limits of feasible computation.) Who is going to teach people about what data sets there are, how to use them, and how they relate to these tools? I don't think this will be purely a library function. Certainly, I do not anticipate large numbers of librarians having second careers as instructors in statistical software or image manipulation software.

Yet it is clear that libraries are going to have a piece of the action. This is a new collaboration that will grow in the next decade between computer support people and faculty and librarians in a much more profound way. In some universities, I already am seeing computer lab support within the individual departments or schools, often in partnership with the library. It is increasingly integrated in the curriculum, particularly at the graduate level, rather than confined to a centralized university-wide computer lab system. The whole complex of issues around data sets is one situation where the boundaries are becoming very blurry and the stakes are very, very high.

Electronic Journals

Let me turn to another set of issues and another set of environmental factors that may affect service priorities—the transition of the printed journals to the electronic environment. Something very significant has happened in the last twelve to eighteen months. Much of the publishing industry has moved beyond experimental mode. We have seen a decade of interesting experiments and collaborations, looking at issues of translating journals into electronic form, how they'll be used, and what problems will arise. This year we have seen many of the major publishers, particularly in the scientific, technical, and medical areas, make commitments to bring up all or significant parts of their product line as standard commercial ventures that can be licensed on reasonably short time frames. What is fascinating is that about the time the publishers'mindset changed, the business model also changed.

Consider, for instance, most of the experiments in the early nineties: the Elsevier/Tulip Project, work with the American Chemical Society at Cornell, and the core project for the collaboration between the IEEE (Institute of Electrical and Electronics Engineers), the IEE (Institute of Electrical Engineers), and the University of California. What characterized all those projects is that libraries got the data and the material was mounted locally. This presented serious problems. Everybody thinks it is easy until they try it. Talk to any of the participants in the Tulip Project who figured casually at the beginning of the project, "Oh, we'll have this up in a few months; it's just bitmap images; this can't be very complicated, can it?" A year later they were still trying to fight their way through some of the implementation problems. People considered the experiments and then thought about scaling up and realized that it was not viable. Almost every one of the implementations for those experiments was custom software. Most commercial local library systems do not have facilities to import half a terabyte of bitmap images as a production input stream, process it, link it to the appropriate abstracting and indexing databases, and make it available for patron use on an ongoing basis. Doing this for several hundred publishers in parallel is almost unthinkable, so the publishers moved to a model where they are network-based information service providers, either directly or by subcontracting with a private service bureau or an aggregator. Elsevier, Springer, and Academic Press have put up Websites with their material. At some level, this is a relief for libraries. They can move forward into the world of largely electronic journals and bypass their local systems'inability to mount the material and their budgetary inability to make the necessary investments to develop or create local systems that can allow them to mount the material.

With all the publishers putting up individual Websites, journal literature is dissolving into incoherence. My sense is that most users do not structure their work strongly along publisher lines. Last time I looked, we did not have academic disciplines devoted to studying the works of individual publishers. Most people work in particular disciplines and most scholars, let alone most students, have not got a clue who publishes what.

One of our challenges, as we look at this evolving world of journals in electronic form, is how to recapture the progress we have made with the generations of online catalogs and their extensions. How do we retain coherence in our collections? I believe this is one of the most fundamental service and collection issues today. We are going to see abstracting and indexing databases form a spine that gives coherence to these distributed collections.

Saying this is easy. It sounds really good as broad theory, but some very complex issues arise in the practice. Linkage is one. We need linkage mechanisms between abstracting and indexing databases and these distributed repositories of journals that publishers are managing. This is not a simple problem. Check with your favorite publisher about whether direct addressability at the article level has even been considered. I have had some interesting conversations with people who build these Websites. I have said, "People want to make pointers directly into your journal articles so that authorized users might follow a citation from the bibliography of one article to the next article or make such a citation," and I've been met with blank looks. "They want to do that? We never thought of that. Gee, we'll have to redesign our whole system." Whoops! The primary question is do the publishers have the technical design to permit such access? The second issue is what the linkage mechanism is going to be. URLs? URNs? What kind of URNs? What kind of linkage code? A serial item contribution identifier (SICI)?

The next question is another good one. Where are these links going to come from? Some institutions have been building their own links. When I was at the University of California, we were tying up a lot of resources building links between A&I databases and publisher Websites. This is not going to scale. There is something profoundly crazy about having thousands of libraries trying to accomplish this individually. Links will have to be provided by the publishers, the A&I vendors, or from some third party. We are already seeing signs of all kinds of players trying to move into that area.

Another nasty issue is time synchronization. Abstracting and indexing databases trail the published literature by anything from a couple of weeks to a couple of years. The average working scientist is not going to be happy with a world of electronic journals in which access to information comes weeks after it would have come out in print because it is coming through the abstracting and indexing databases. The result will be a dual mode of literature access,

in which most searching is done through the abstracting and indexing databases, but tables of contents for more recent publications will be found on the publishers' Websites. This will not be greeted with enthusiasm by most of our user base. They want a coherent approach to their research literature.

Solving this problem will require greater cooperation between primary and secondary database suppliers—between publishers and A&I vendors. It also has a serious operational implication for libraries. I think libraries have been amazingly casual about the time constraints on projects like database updates. Here is the scenario. You have an abstract and indexing database mounted locally. You get the update tape, load it overnight, maybe tomorrow, and nobody gets excited. Maybe you are loading weekly or monthly. The tapes come in weekly, but if they get held up in the mail for a day, nobody notices. We did not have that problem with primary publication in print. It came in and you got it out there. As soon as it was out there, it was accessible. If we need to update some kind of abstracting and indexing apparatus to make these new resources visible, users are going to be sensitive about how quickly it is done. We will have to get very serious about update schedules and time constraints in a way that we have not had to be. This wonderful world of weekly and even monthly tape updates in the mail is going to be replaced by daily or maybe even hourly FTPs that trigger processing programs when they come in across the Internet. It is going to create a new and different management world. There are many implications as we look at what the next generation of local systems look like. What do we want from our abstracting and indexing databases vendors? Now they are no more prepared to provide hourly updates to their database than libraries are to receive them, but I believe users will drive us toward this new world.

The final issue I want to raise about coherence in journal databases is coverage. We have several kinds of coverage problems. Abstracting and indexing databases do not cover everything in most journals. Some of them practice "selective" indexing. Some have selection criteria you cannot deduce. Some change editorial policies at least biweekly, and generally do not bother to tell you unless you pound on them because you noticed that something has changed in the latest database updates. We need to understand what we want and what we expect in the correspondence between the description of the literature and the literature itself in the electronic environment. The first step is getting informed about what is going on. We also need to think about what we really want. There is an additional factor. Contrary to what a growing proportion of higher education institutions (and particularly their undergraduate populations) in the United States are coming to believe, the published journal

literature did not start in the early seventies. It only looks that way because that is the electronic coverage we have on it.

Incidentally, if somebody is looking for a fun study to do sometime, I would like to see an analysis of how citation in faculty publications or student papers at a given institution relates to how much of a backfile of A&I databases they can access in the relevant discipline. I suspect that if they only go back to 1985, or whatever the significant year is, one will see a sudden decline in the citations to literature published before that year.

The point is that we need to deal with retrospective conversion for the journal literature. The first step is getting some kind of bibliographic apparatus around it. Some efforts to digitize the retrospective journal literature—for example, JSTOR—are generating the abstracting and indexing apparatus for the material they cover as a by-product of their work. We will need to spend more time and money on the apparatus and less time digitizing, early in the process. JSTOR, wonderful resource though it is, has the same fragmentation problem. The access apparatus does not integrate reasonably with other A&I databases and the primary content is not easily linked to any access apparatus other than the one that JSTOR provides. It is an insular resource. Having a closed bibliographic apparatus that comes with particular runs of digitized resources merely contributes to the fragmentation of the research literature as it becomes digital.

Interaction

Another service boundary that we need to struggle with is interaction. We talk about the network as a publishing medium. In fact, it is two intertwined media, which are part of its richness and complexity. It is a medium for information dissemination and access and it is a medium for communication. Communication as scholarly discourse is moving into the network environment in more creative ways than just cranking up the scanner and transforming print pages into electronic pages. Interaction enters the picture very quickly. It is unclear how much we want interaction to come under the libraries' service umbrella. Many libraries struggle with the issue of whether their workstations should be public service points for general electronic mail transactions. Do we really want to host people chatting on the Internet all day long? What is going to happen when packet-type video technology becomes commonplace? Are we going to put little video cameras on top of all our public terminals and invite people to watch each other as they continue to chat on the Internet all day? This dilemma is not unrelated to that of people doing data

analysis, for example, at public workstations. The answers are going to be different for different libraries.

One will see constant pressure to permit interaction because interaction is going to be more and more part of the new mode of scholarly communication and discourse. Content will be difficult to separate from interaction. Systems are now coming online that make this definition more and more confusing. Consider the experimental collaborative filtering or community filtering systems. Basically, either as a deliberate act or as a by-product of doing something else, you express likes and dislikes about some class of objects. This may be as simple as signing onto a system and typing in your ten favorite movies. Perhaps every time you read a Usenet news group post, you are invited to give it a rating from one to five, from truly memorable to truly useless, or anything in between. Many of these systems on the Internet take opinion polls for books or sound recordings. Some online stores are using software that catalogs this information. Firefly is probably the best-known company. People might visit amazon.com, for example, repeatedly making purchases and generating an information trail. These companies'databases compare the books you bought to the books others bought. If similar tastes show up, users might be solicited by the software to buy some book "that a number of other people with interests similar to yours are buying." These systems are still somewhat experimental. They are similar to some of the high-end, full-text information retrieval technology that respond to a few typed words. Sometimes magic happens, you get a really good, useful result, and you think it is wonderful. Then the next time you use it, it fails abysmally. You get a result so stupid and bizarre that you think the technology will never be ready for general production use.

This imperfect technology is something very powerful because it begins to emulate how human beings in communities find information. You ask people whom you respect, word passes around inside scholarly circles or other communities of discourse, and work gets a little buzz around it that causes people to start looking at it. This technology attempts to emulate such communication. How do these interactions fit in with library service offerings? Do we want to offer them inside library services? Keep in mind that whether or not libraries accept them, other organizations out there—in some sense, competitors to libraries—will offer this service. I don't have an answer.

User Privacy

Another related issue is how well we are really willing to get to know our user communities. In many ways, the less we know about our users, the happier we

are. There are some good and legal reasons for this attitude and some tradition of defending user privacy: "Let's not collect and keep certain data, because if we don't have it the government (or other organizations) can't get it." The library community has kept this policy for all the best reasons, particularly to protect patron privacy. Yet in this new era, people are building personalized systems. The fact that these systems really know the users and know what they have been doing improves their quality of service. Current awareness, collaborative filtering, remembering what you've read, correlating it with other people, remembering preferences constitute a complex of systems and services that fundamentally rely on maintaining increasingly extensive and comprehensive dossiers on what users are doing, where they have been, where they are going, and what they are interested in. I think we need to grapple with that issue.

Clearly, you are going to build such systems with the informed consent of the user. Protecting anonymous access for people who want it is terribly important. As I look at the activity on the Web—for example, the changing relationship between publishers and readers—I am beginning to believe that within a decade, libraries may be the only place where you can look at anything anonymously. Perhaps bookstores and newsstands also will continue to fill this role to some extent. It depends on how rapidly and how much material migrates to electronic formats. There will be a class of information that you either will have to buy over the Internet with very little privacy from the rights' holders or read at the library. Anonymity may become a significant role of libraries much more so than today because so many other options are being squeezed out of existence. We need to address the service implications of truly knowing our users, not just their names and visiting days, but their research and their preferences, and getting that knowledge into our information systems.

Distance Education

As far as I can tell, many who are promoting distance education strategies either do not believe this or it has not occurred to them, but libraries are going to be an important part of delivering distance education effectively, particularly as it moves beyond very narrow vocational courses. Certainly, if you are teaching a class on photocopier repair, you can send out the manual and then give a three-hour video course without requiring library support. When the concept expands to projecting substantial parts of the higher education experience outside the walls of the university, you are going to need substantial participation from the library. We do not understand some aspects of this con-

cept at all and we have made considerable progress on others. Libraries have been aggressive in getting a handle on the issue of electronic reserves, which can be used to support distance education programs. Different libraries have developed different policies, of course, and taken different strategies. This is an area that people seem to be comfortable addressing.

However, people are not so comfortable addressing one area. A lot of distance education is broadcast, and these broadcasts are recorded. This kind of information resource is going to be commonplace as digital video gets cheap over the next few years. Eventually, it's likely that every seminar and many classes at most major educational institutions will be videoed routinely. All this video will have to be cataloged and stored and someone will have to make decisions about what to keep and for how long. Is that a library function or can someone else take responsibility? Look at the EDUCOM National Learning Infrastructure Initiative (NLII). Information technologists and distance learning people are building an instructional management system (IMS), which they wisely did not call a digital library. Once people start looking at what it does, they are going to be very comfortable calling it a digital library. The relationship between libraries in institutions and this new kind of system is an absolutely open issue at this point and one that we really need to consider.

As video technology in education becomes ubiquitous, how will we organize, index, and provide access to it? Today, we are cataloging this information at a fairly gross level. There is a lot of technology in experimental development. One of the six NASA/ARPA/NSF digital library projects is the Intermedia project at Carnegie Mellon, which specifically focuses on what we will do with this flood of video. People are building systems that summarize videos, trying to provide key frames, trying to offer some framework other than the fast forward button for navigating through an hour of video that may contain something pertinent. Are we going to run these indexes in the libraries or will they be archived somewhere else? How well will automated systems, as opposed to intellectual indexing by human beings, perform? Instructional video material is not mass produced, as are books and journals. Copy cataloging strategies are not going to work because most of this will be unique material generated at individual organizations. This is another area that requires our attention as we try to draw our boundaries and set our priorities.

Special Collections

We know that a wonderful revolution is taking place in the realm of special collections. We are blowing open the doors through digitization and making a

great deal of this fragile, hard-to-find, esoteric, unique material available in digital form to people through the Internet. No longer will researchers have to trek to individual institutions to mine the treasures of their special collections. Undergraduates can enjoy the thrill of working directly with primary source materials. We are going to be digitizing away at these massive collections for the foreseeable future. Setting priorities and working through the process of digitizing all or even most of this material will take decades. We have taken some very important initial steps with projects like the EAD (encoded archival description) work and the electronic finding aids. These can provide a coherent context for ongoing digitizing projects.

Digitization of special collections will make an immense amount of previously remote material accessible and relevant. Perhaps a library has a special collection of a famous author. Perhaps a museum's special collection of an obscure artist contains important materials that relate closely to the first collection. These might be parts of a long correspondence between the two. Geography will no longer be an organizing principle for collections unless we want it to be. Somebody can build virtual special collections that bring together material from multiple sites. Who is going to do that? Is that a curatorial function? Is that a library function? Is that a function that is best left to the scholars? Who will maintain these virtual collections? Links die, links change, and electronic content needs to be maintained to be vital and useful.

A closely related issue is what to do about exhibits. Many large libraries organize modest exhibits and some support impressive exhibits. Museums, of course, offer exhibits all the time. Exhibits come and go. We can keep all of them, however, in the electronic environment, because we can re-purpose, rearrange, re-present the same material multiple ways simultaneously. Such is the wonderful flexibility of the electronic environment. How much of this responsibility falls under the libraries'purview?

Preservation

That question leads me directly to the final issue—preservation of digital content. We feel philosophically and morally responsible for the preservation of everything, even if we are not exactly sure what it all is. Yet we are building systems, for example, the publisher-based Websites, that do not correspond to that philosophy. We are also missing something else. Our conversations about digital preservation are profoundly schizophrenic. Two separate conversations are occurring. One discusses information that is akin to the published literature but is more digital. We are worried about preserving it as much as we have always been concerned with preserving the printed literature. When we

look at the scholarly journals going electronic, we continue to worry about how we are going to preserve them. While we have been worrying, a vibrant new communications medium has emerged on the Internet containing many new genres. As with virtually every new communication medium, almost no one has quite noticed it yet. We really do not understand the latest new communication medium any more than anyone had a good grasp on radio or television in the beginning. How much of the information disseminated from those media was archived in the early days? As we are discussing what is really important to preserve, lots of information is coming and going and going and coming and falling off the face of the earth.

In the 1996 presidential elections, both Clinton and Dole had Websites. Some scholar working fifteen years from now is going to be interested in charting how the world of the Internet and the communications that it enabled affected the political process during the 1990s. Such a scholar might have more than a passing interest in what was on those sites and how the site contents changed and evolved on a day-to-day or even hour-to-hour basis. Did anyone save them? I have not looked, but this scenario came to me when I was trying to think of good examples of event-driven material that is just coming and going on the Internet, and that we may one day regret not saving. We need a serious dialog about what our preservation priorities should be for these new genres. We need to look back to the last few new communication media to learn from our experiences.

Conclusion

I have made quite a survey here through various areas where difficult and controversial decisions are to be made. There are no absolute right answers here—at least, that is my view. Choices are going to be made by individual institutions based on resources, on expertise, and on the needs of the specific user communities that the institutions must support. They also will be made in response to visions of what libraries should become in the digital culture. There will be choices, not globally right or wrong decisions. Many of these choices are going to be about boundaries, rather than actions—about what organizations will not do, what they will define as outside their scope, rather than within their mission. I hope that this chapter has been helpful in beginning to define those choices, particularly for collections and services, and has offered some case studies that organizations can use to test their views of where their own boundaries should be drawn.