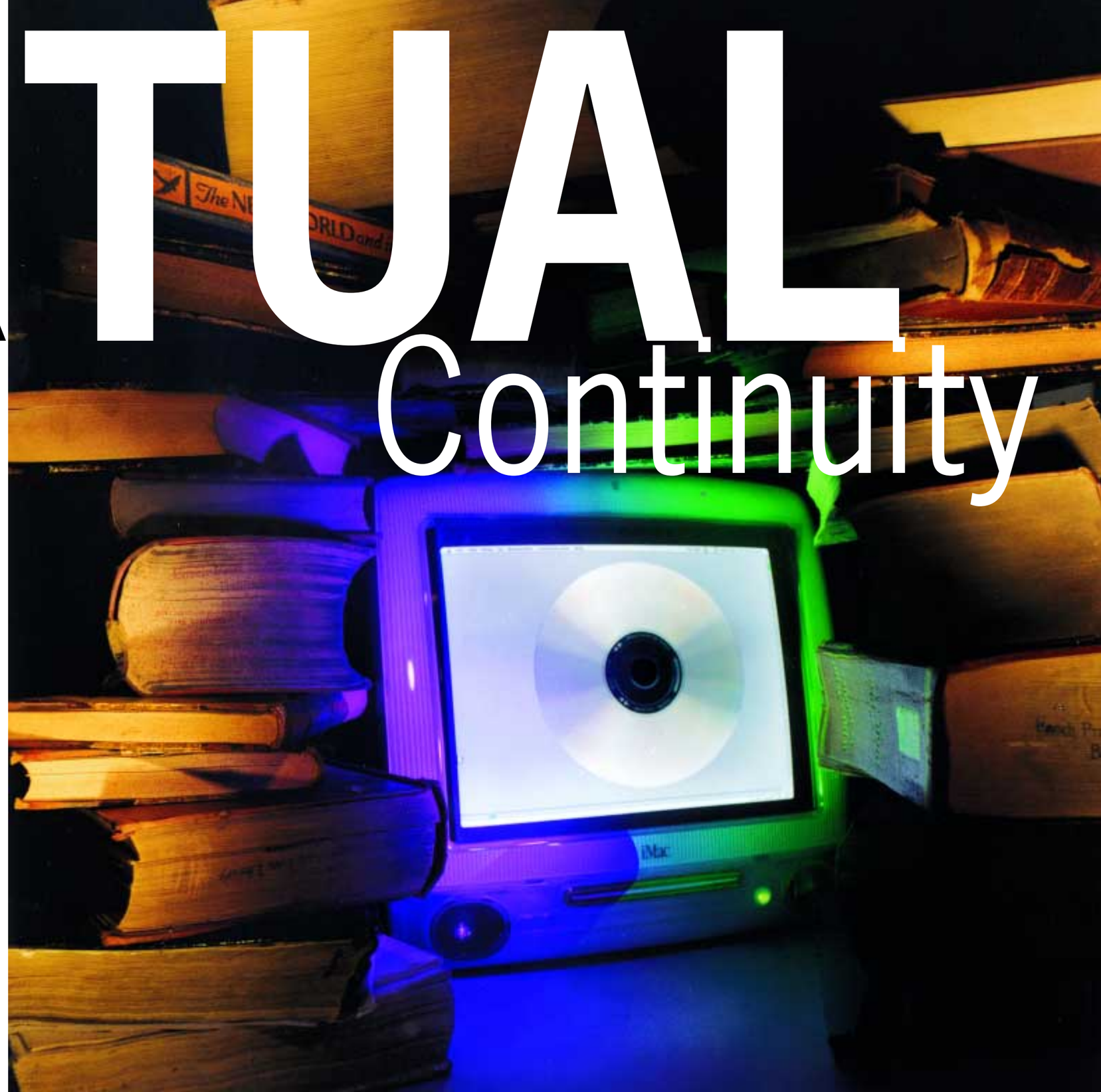
VIRTUAL Continuity

The Challenge for Research Libraries Today

By Nancy M. Cline

I have long been an advocate of the use of information technologies to improve access to information, to achieve efficiencies in library processing, to facilitate communication, to improve reference services and document delivery, and to support independent learning. In 1988, in remarks to an advisory committee at the Library of Congress I noted, "The challenge is how to correlate traditional *methodologies* with new *technologies* so that we ensure a *system of access* to information and knowledge." Today, it is certainly an understatement to say that the Internet is revolutionizing information access. But amid the proliferation of information, are we creating sustainable *systems* of access? Are we building reliable databases and durable objects? In our enthusiasm for access, are we overlooking important issues of reliability, redundancy, the ability to replicate results—important elements of continuity for scholars? While we work to incorporate vast amounts of digital information into our libraries, schools, universities, and colleges, how much should we concern ourselves with "virtual continuity"?

Nancy M. Cline is Roy E. Larsen Librarian of Harvard College at Harvard University.



Photography by Strobophoto ©

Let me be clear: I am not referring to *all* information. Not all information is the same. I am talking about the information used in research. Research depends on the ability to scrutinize earlier work and build upon it. Earlier publications are in many ways an audit trail for the verification of so much work—in science, business, industry, architecture, transportation, medicine, and other fields.

In the past, research libraries have worked with diverse collections, spanning hundreds of years and representing different methods and different physical forms of recording knowledge. For much of this time, they have operated on the assumption that these collections were of lasting permanence. We know that there are faults with this assumption: acidic paper becomes embrittled, and collections of print, microfilm, videos, sound recordings, and other formats fail to hold up to repeated use or suffer from neglect as well as intentionally destructive behaviors. This erosion in the content and reliability of research collections can often be detected quite readily. We can see the trail of crumbling, yellowed shards of paper found at the base of shelves or the pages that fall apart in our hands when a book is opened, or we notice the visibly missing pages, the mold, the water damage, the acidic ink on acidic paper, the unstable acetate tapes, the work of insects.

But when such problems are noted, there are often redundant sources to rely on for substitution. If the copy at one library is damaged, in many instances there are copies at other institutions, or a copy may exist in microfilm. There also have been catastrophes—fires, floods, thefts, willful and malicious damage—yet overall, redundancy and reformatting, along with a system of safeguards protecting against fire, earthquakes, and other factors in physical facilities, have enabled scholars to work with certain assumptions about the continuity and durability of the materials on which they depend.

The Proliferation of Digital Information

The massive physical spaces required for library collections are evident in our colleges and universities. But once, not very long ago, these collections were much smaller. I worry when people tell me that I should *not* worry about the storage requirements and the costs for digital information, since there was a time when book collections were much smaller too, when faculty knew the contents as well as the places in which the books were located, and when browsing and selecting were much simpler than today. With the rapid growth in digital content, we need to be concerned with its reliability and durability over time and how researchers will continue to access this content.

At Harvard, I am responsible for nearly ten million books plus millions of visual images, microforms, manuscripts, documents, maps, photographs, slides, sound recordings, films, scrolls, and the like. They are in dozens of locations. Nearly two million items are in an off-campus depository. Yet, only one hundred years ago it was possible for one of my predecessors to recall *all* library materials from circulation and have them properly placed on the shelves for an annual inventory and dusting! I use this as an example to make the point that we, today, have no idea how vast—and how different—the information universe may become. The techniques of managing today's information "collections" may not carry us smoothly into the future.

Collections—in all the earlier, "predigital" formats—have been costly to acquire, catalog, store/house, and manage. For many valid reasons, including the convenience of distributed access, institutions now consider digital resources to be important investments. Some are hoping that the digital titles will completely replace printed or microform collections, as well as eliminate the related costs of managing them. Some feel that digital resources will so greatly expand access that users will no longer have to go to a physical facility, the library. Some institutions envision convenience and outreach to new markets for distance learning programs. Others are simply excited by the vast array of information that can be found on the Web, much of it at "no cost."

In the past, research libraries have worked with diverse collections, spanning hundreds of years and representing different methods and different physical forms of recording knowledge.

As we move about

rapidly in the stateless environment of the Web, we sometimes struggle for the right words to describe what we saw and used. How often in conversation do we hear people asking others *when* they used a version of a Web site or whether they were on the *development* version or the *public* version when trying to pin down a problem? The significance of words like "citation" and "publication" is fading. Replicating one's research activities takes on new dimensions.

For example, a library's online catalog may represent both the bound holdings and the link to the electronic issues of a journal, but who maintains the commitment to provide access to the full run of the publication? With online journals, the library often leases or licenses access and no longer acquires and builds a lasting resource. Where does that leave the researcher who expects continuity? What happens if/when publishers avoid any responsibility for maintaining e-archives of back issues? What happens if/when they maintain only those back issues that are profitable to archive? How do we shield our users from the risks of depending on proprietary systems that can be withdrawn with little warning?

During the fall semester, the following e-mail was sent to a member of my staff: "Access to [Publisher X's] electronic journals is down. The publisher has reconstructed their site and has implemented changes to their authentication mechanisms. Unfortunately, news of this change failed to reach Harvard in time to test and implement changes on our end. As a result the following eight journals are inaccessible...." Another example is this e-mail sent by Vendor Z: "The system crash at the main U.S. server is more extensive than initially realized, and we are still trying to resolve the problem. It appears that the server will be down the remainder of the day." A full day later, the vendor had still not restored access. Who is responsible for which assurances in a "virtual" library?

On the Web, anyone can be an author, can maintain a Web site, and can claim names. The ease with which we depend on the sites of publishers, vendors, and other universities and colleges—without knowing where the sites are, if they are backed up, how they are archived, when they were last bought or sold—is amazing. The seductive lure of "free information," which can be found at all hours without having to go to library buildings or to deal with library catalogs, *is* intriguing, but will that information still be there next fall? Will the user ever be able to find it again? Will it be in the same version or in revised form?

Developing Reliable Digital Information

Many exciting developments are taking place in libraries today. There are libraries that are relying heavily on digital content, libraries partnering in distance education, libraries developing new databases, libraries studying users' behavior with online resources, libraries working on policy issues affecting access, and libraries addressing technical challenges. There are numerous experiments that will demonstrate libraries' involvement with digital content. Projects abound.

However, most of these projects have short life lines. Many are funded with "soft" or one-time money. There is little assurance that researchers will be able to reliably use the majority of these resources for a long period of time into the future or that the resources will continue to exist as a means for replicating today's research environment. This uncertainty is quite different from the ways in which researchers could expect continuity and reliability from libraries in the past. Yet we seem to be carrying forward our assumption that digital information will be there in the future, just as the printed medium has "been there" for centuries.

This issue of the permanence of digital records (I hesitate to use the word *permanent*) is far from new. In 1990 the federal government's report "Taking a Byte Out of History: The Archival Preservation of Federal Computer Records" stimulated discussion in various sectors. The report predicted that within ten years, by the year 2000, 75 percent of all federal transactions would be handled electronically. The same report pointed to the well-known problems experienced with the 1960 census data, citing it as one example of the rapid pace at which computer information can become obsolete. The report asserted the need for improvements: "The Federal Government must take steps to identify, preserve, and provide for the practical use of information of historical interest created and stored on computers.... Managing electronic records to ensure long-term availability is the most significant challenge facing the archival community."

In 1990, at the time of that report, the National Archives and Records Administration (NARA) already had *twenty years* of experience with accessioning electronic records. But as alarmed as many were, their earlier experience could barely predict the chaos and creativity that would arrive with the Internet and the powerful personal computing devices that we now use.

How do we deal with the rapid obsolescence in hardware and software produced by industry? When a leading officer of a major technology corporation bluntly states that he creates products for only a three-year market presence, this is not reassuring to libraries—institutions whose mission is to ensure a reliable trail of information. Likewise, it is not reassuring to know that there are data-refreshing schemes that will lose *only* a small percentage of the data—or to hear publishers and vendors propose that "what should be saved for the future is only that which is used today," leaving popularity to be the measure of significance.

However much we intend to keep, there are still no standards for the permanence of digital items. We do know that certain kinds of paper can last for many centuries. Microfilm has been stress-tested to show that under good storage conditions, it too can last hundreds of years. The jury is out on digital materials: some formats are rated to last fifty years, but others have a much shorter life span. And if indeed we do maintain the bits, will they be readable if the operating software changes? Obsolescence of *both* software and hardware must be considered.

Traditional Roles and Future Issues

In expressing concerns about the permanence of digital resources, I am *not* saying that the future should look like the past or that the Web should work like a library. Each has value and purpose. But I do propose that we look at what *libraries have done* and determine whether some of the same attributes and strengths are needed in the emerging electronic environment.

Let me briefly highlight some of the traditional roles played by libraries in the support of scholarship and research:

- Libraries bring people and information together. Libraries purposefully select materials. The selection of materials is active, intentional, and consonant with the institution's mission. In a research institution, a significant number of items are acquired with the expectation that they will be needed not now but rather at some time in the future. It is the nature of research materials that it often takes years for information to be needed, years to "find" its audience. Universities have a responsibility not only to provide information for their faculties and students but also to preserve the intellectual and cultural heritage of society.
- Libraries add value by organizing information into "collections," by placing items into classifications. Libraries construct systems of access, both intellectual and physical.
- Libraries manage the collections and allow for the retrieval of items so that they may be used by individuals in-house, locally, or through interlibrary lending. While managing the use of materials, libraries also protect the privacy of users. In subscribing to the American Library Association Code of Ethics, libraries agree "to protect each user's right to privacy and confidentiality with respect to information sought or received and sources consulted, borrowed, acquired, or transmitted."
- Libraries engage in the preservation of collections, work that encompasses a substantial range of materials. In music collections alone, there are wax cylinders, acetate disks, vinyl disks, magnetic tapes, performance videos, CDs, and DVDs, in addition to materials in sheet and book form.

The creation of new knowledge is founded on the ability to rigorously pursue diverse routes of inquiry. New work builds on previously recorded knowledge, in all formats. Putting aside libraries for the moment, we must also deal with the researcher's inquiries in the years ahead:

- How can I know that this e-text is the latest edition? Is it the British version? Is it the one with the author's emendations? Do I have the first conference paper or the pre-print?
- How can I find what I know I saw yesterday on this terminal?
- How can I download this to be sure that I can use it again tomorrow?
- How do I know this is authentic?

It is certainly fair to say that not all information is intended to last. But how do we determine—or who determines—what *should* last? What if none of the "temporary" or "less meaningful" information had survived from Tiananmen Square, or Desert Storm, or World War II, or the Civil War, or expeditions to discover the American West? What if diaries and personal correspondence had not survived?

With all the enthusiasm for *using* digital content, there is relatively little understanding of what we expect to build and how well it will expand to the future. With little question, digital content excels in terms of providing access. There are exciting opportunities to bring together in a digital context items that in the current physical context would require an individual to use in many different facilities. The increase in digital texts such as journals, books, manuscripts, photographs, and other research material provides an array of resources available around the clock. There are exciting developments that will support new ways for learning, for teaching, and for delivering library services.

For example, imagine holding in your hand a music manuscript from South India and at the same time being able to work online with the field notes of the anthropologist and to hear the recording *as it was captured in the field years ago* and to click on photos that were taken at that time to see buildings, the marketplace, train stations, clothing, a culture that no longer exists. Imagine what it would be like to know the context in which the music was produced and enjoyed, to know what the composer or artist may have experienced, to know that particular "slice" of time and place. This is why we build digital libraries, to stimulate learning and research, even though we know there are risks and challenges.

With all the enthusiasm for *using* digital content, there is relatively little understanding of what we expect to build and how well it will expand to the future.

Solutions

Digital content is springing up everywhere. Everyone has a simple solution for creating digital libraries. And it *is* easy to scan thousands of pages or to create digital resources, but sooner or later the questions of scale, sustainability, reliability, support for other users, and assurance of authentic

copy rise to the surface. The convenience of plug-and-run devices gives way to serious discussions of the need to learn about such things as metadata, digital object identifiers, Z39.50, GIS, the licensing landscape, "freedom, privacy and the network," legal issues in the digital environment, markup languages, intelligent systems for indexing and retrieval, knowledge access structures, cross-catalog searching, linked catalogs, keyword, online thesauri, and the like.

So, in a field of doubts, how can we make progress? In the United States, in contrast to some European countries, there is no national library charged with responsibility for solving these issues for the nation. Instead we have a flourishing mix of talents at work in academic institutions, not-for-profit enterprises, commercial businesses, and consortia: for example, the Coalition for Networked Information (CNI); the Online Computer Library Center (OCLC); the Association of Research Libraries (ARL); the Research Libraries Group (RLG); the Digital Library Federation (DLF); and numerous individual libraries.

Association of Research Libraries (ARL)
The ARL is host to SPARC (Scholarly Publishing and Academic Resources Coalition), which intends to stimulate digital publishing in the sciences, reduce the cost of information access and use, expand the dissemination of research, and support practice and teaching. It is especially looking to stimulate and accelerate the creation of new nonprofit information communities in key fields in science, technology, and medicine. SPARC recently announced awards for three new projects:

- Columbia University's Columbia Earthscape, an online resource in earth sciences to be managed by the Electronic Publishing Initiative at the university and involving the university press and the library, will include reports of research projects and conference proceedings as well as curricular materials and will link to data sets, computer models, and an online journal.
- The University of California's Digital Library's eScholarship will support innovations in scholarly communication by providing an infrastructure for experimentation. It will include an e-print database system, support new and linked digital journals, and integrate digital publishing and access.
- MIT's CogNET, an Internet Gateway to the Cognitive and Brain Sciences, will be managed by the MIT Press with ties to the Institute's Digital Project Lab and the Libraries. CogNET will integrate a range of online utilities in a customized workspace, delivering access to the very best, most accurate, and most timely technical information in contemporary cognitive and brain research.

Research Libraries Group (RLG)
With an increasing number of museums and special collections libraries in its membership, there is a growing need for RLG to deal with digital collections of cultural objects. There is a sense among members that the time is now to define ways of developing richer, more robust repositories of digital content to support the way that people want to access many of the outstanding, yet often small, collections or to bring together what are now disparate pieces into virtual collections of significance. This will require considerable attention not only to search structures, metadata, linkages, and user interfaces but also to the aspects of developing sites responsible for the long-term storage and dissemination of digital material.

Digital Library Federation (DLF)
The DLF, within the Council on Library and Information Resources, began as a consortium of fifteen research institutions. It has grown to twenty-three and has alliances with CNI, NARA, OCLC, and RLG. The participants share a common goal: to create a system of independent, distributed repositories of digital works. The priorities of the DLF are to focus on materials that are "born digital," to integrate digital materials into the fabric of academic life, to stimulate the development of core digital library infrastructure, and to develop the organizational support needed for the management of digital libraries. The following are among the focus areas within the DLF:

- Social sciences data: finding strategies to address the dual challenge of preserving digital data—the challenge to maintain the tools to read the data files while also preserving the codebooks needed to interpret the data output
- Visual resource imaging: documenting the science of imaging as an assurance of quality and reliability
- Authorization systems: developing a prototype system among several institutions, with an emphasis on a shared protocol and the expectation of producing a statement on an authentication and authorization architecture
- Reference linking; distributed finding aids; metadata; technical architecture of digital libraries; standardized means of maintaining persistent links between citations and the digital objects they refer to; and researchers' tools for migrating digital materials

Individual Libraries
At many institutions, a "digital library" is based within the library but maintains close ties to the academic computing organization. Some libraries, like that at the University of California (UC), have taken a bold, defining step. UC has made its digital library the tenth library in the UC system. At Cornell, computer scientists and librarians are working together to develop a working prototype of a digital library system with built-in mechanisms to preserve documents, protect intellectual property rights, and permit interconnections with other digital library systems worldwide. The challenges the group faces are summed up in the acronym coined for the project: PRISM, which stands for preservations, reliability, interoperability, security, and metadata.

Harvard has chosen not to create a separate entity but instead intends to integrate digital resources as one more evolutionary stage in its libraries. It has established the Library Digital Initiative, a five-year specially funded program to support digital acquisitions, cataloging, collection management, reference, and preservation among its nearly one hundred libraries. It intends to develop the university's capacity to manage digital information by

- creating the technical infrastructure to acquire, organize, deliver, and archive digital library materials,
- establishing a team of specialists to advise librarians and others at the university on key issues,
- providing librarians and staff with experience in a wide range of technologies and digital materials, and
- enriching the Harvard University Library collections with a significant set of digital resources.

Among the initiative's recent developments are several Web-based union catalogs:

- VIA (Visual Information Access), a catalog that can describe prints, photographs, drawings, and paintings held by libraries, museums, and archives
- OASIS (Online Archives Search Information System), a catalog of finding aids that provide access to archival collections
- Geodesy, a catalog that provides access to geospatial data

Major areas of development include the following: a digital repository for long-term management and access to digital objects of all types (text, image, sound, multimedia); naming services to provide digital objects with persistent, location-independent identifiers; metadata; and reformatting services.

Harvard's Library Digital Initiative includes an internal challenge-grant program to stimulate projects and to serve as a catalyst for infrastructure development. One project just now getting under way involves the Harvard Business School, which holds a large collection of advertising trade cards. From 1876 to the end of the century, these three-by-five-inch cards, printed on both sides, were the chief medium of advertising for merchants and manufacturers. Harvard Business School currently plans to digitize a sizable number of these cards and to create a database. A valuable reference for those who need to know about products and marketing in the late nineteenth century, these cards can be approached as research materials in a variety of ways—as industrial products, cultural artifacts, or works of commercial art. Though housed in a business library, they have been used by scholars from the fields of cultural anthropology, ethnic and gender studies, industrial archaeology, sociology, fine and decorative arts, and engineering.

At Cornell, computer scientists and librarians are working together to develop a working prototype of a digital library system with built-in mechanisms to preserve documents, protect intellectual property rights, and permit interconnections with other digital library systems worldwide.

Another Harvard project involves using digital information as a preservation strategy by creating surrogates for user access when the physical form of the original makes it difficult to use. In one case, the libraries created a virtual collection of daguerreotypes. Hundreds of daguerreotypes were housed in fourteen different repositories; content encompassed the history of medicine, science, and anthropology as well as institutional history. The Harvard class portrait of 1852 includes eighty-five daguerreotypes fitted into a wooden cabinet, a rather unwieldy item for research! In digital form, these are much easier to work with and will help protect the originals as well.

Conclusion

In recent centuries, research libraries have held significant roles in research and education: selecting and organizing materials for collections; developing systems of intellectual access; organizing items for physical access and retrieval; and preserving items for long-term use. These attributes signified a durability that is now challenged in today's fast-paced digital environment of networks, Web interfaces, and proliferating search engines. We cannot ignore the rapid acceleration of digital dependence in all aspects of education and research, nor can we overlook the researcher's need for permanence, reliability, and continuity in this digital environment.

Thus as we look to the new century, we must shape an information environment that has sustainable systems of access to enduring information resources so that users, now and in the future, can rely on them with confidence. Defining this future calls for new combinations of talent and expertise, for short- and long-term collaborations, and for experimentation and risk-taking in order to develop the best strategies for managing the rapidly expanding amounts of digital information. Our challenge is to ensure the viability, the continuity, of information for the scholars of 2020, 2050, and beyond.

This article is based on a speech delivered at the EDUCAUSE national conference, Long Beach, California, October 27, 1999.