Preserving Digital Information to Support Scholarship

Clifford A. Lynch

In a number of arenas, the development of strategies to preserve digital information on a large scale is of great importance and urgency to the higher education community, and indeed for society as a whole. Higher education and the research libraries within our universities, in alliance with other cultural memory organizations (including museums, archives, historical societies, and other libraries), have traditionally taken primary responsibility for preserving our cultural, scholarly, and intellectual record on behalf of society broadly. The tacit assumption—indeed faith—on the part of the public is that higher education and its allies will continue to do so. But it has now become clear that digital information is different, both in the kinds of preservation approaches it demands and the kinds of problems it presents: the customary legal, organizational, technical, and economic models no longer apply. We face a major new challenge. The consequences of failing to rise to this challenge are serious, both for higher education and for cultural memory institutions, and for the society they serve. Yet the scope and magnitude of this challenge, while increasingly clear to the organizations facing it, have not yet been communicated well to the broader society.

This chapter builds on chapter 8 by examining obstacles to successful development of digital preservation strategies and briefly summarizing some promising developments under way in various arenas of interest. It is not intended for specialists in digital preservation; rather, it is aimed at academic faculty and higher education executive leadership, with the goal of explaining the nature and importance of the problems involved in digital preservation.

Preserving Digital Materials

Digital technology (particularly word processing) has long been used to create digital versions of printed pages, but we are increasingly moving away from digital works that can be reduced to printed form without loss of information. As more authors incorporate multimedia, interactive programs, models and simulations, databases, hypermedia, and similar extensions to their works, these "born digital" objects grow more and more distant from the printed pages that we have historically preserved through mechanisms such as publishers and libraries. In this regard, it is important to note that a growing number of scholarly journals, particularly in the sciences, have declared their digital rather than print versions to be the version of record, even though the publishers continue to produce both online and printed versions of their journals. The publishers have decided to no longer be restricted by the capabilities of print. While authors today make only modest use of the digital medium's capabilities, particularly in the sciences, the threshold has been crossed.

To clarify, note that digital technology is now often applied to help preserve and improve access to fragile physical materials such as rare books, manuscripts, photographic collections, ephemera, and the like. These works are being digitized and the digital images placed online for much broader access than the delicate physical versions could survive. Having invested in these digital surrogates (particularly if they are high quality), we must consider their preservation as well. These materials, however, are not the primary cause of the crises surrounding the preservation of digital content.

To keep digital information alive into the indefinite future, several things need to happen. First, the bits that constitute the information must be kept alive—they need to be copied periodically from storage technologies that are becoming obsolete to newer ones. Further, these bits need to be replicated, to protect against hardware failures or disasters, and safeguards against accidental or deliberate corruption or erasure should be put in place.

Second, periodic curatorial work is needed

- to ensure that the bits continue to be interpretable, meaningful, and usable, and thus can be rendered and manipulated in current software; or

- so that older software can be brought forward into new computing environments.

If only the bits themselves are preserved, without appropriate attention to the necessary software technology, then we are at best establishing fertile fields for future generations of digital archeologists who would have to mount heroic and speculative efforts to re-derive meaningful interpretations from collections of bits left behind in the evolution of software technologies.

The technical details are complex and controversial, particularly with regard to the curation of bits (as opposed to simply their literal migration). We remain far from a widely accepted general theory or practice of digital preservation. Experimentation is hard; we do not know how to accelerate time in the context of software evolution in order to test approaches synthetically. We can never prove that we have succeeded in preserving digital information—only that we have not yet failed! But the key point for our purposes here is to recognize that, to survive, digital information needs active, consistent, and competent stewardship, with all the steady economic support and organizational focus that implies. Contrast this to the situation with most traditional printed works, which have survived rather well for decades or sometimes centuries in an environment of benign neglect.

The technical problems are indeed difficult, but the details are largely irrelevant here. The even more challenging intellectual problems involve rethinking fundamental curatorial principles and scholarly needs in the context of digital works—understanding the characteristics of works one is trying to preserve, developing technical approaches that capture these characteristics, or, for example, dealing with questions of authenticity and provenance. Let us leave these issues aside as well. The hardest problems—the ones I will focus on here—are organizational, economic, legal, political, and social.

The broad organizational questions include

- Which entities should preserve which materials?
- How do we ensure enough redundancy to give confidence that the materials will survive changes in organizational mission, funding, governance, regulation, and other factors, as well as simple incompetence or bad luck?

The economic questions involve paying for preservation, dealing with "free riders," ensuring steady funding for fragile digital materials that will vanish if neglected too long, and estimating costs. Clearly, organizational and economic issues interact in complex ways.

Recognize that in the world of published printed materials, multiple, redundant distributed copies under autonomous control have been the key to preservation. Many different institutions independently decide to acquire the same work, whether and for how long to keep it, and what needs to be done to preserve their copy. It seems clear that this widespread approach is an essential characteristic of any effective preservation system operating as a complement to a system for the dissemination of scholarship or cultural materials. The notion of a publisher, for example, serving as the sole archive for its publications is nonsense. It's clearly nonsense in a world of physical markets for published materials, and it is no less nonsense in a world of digital distribution—even though some publishers might suggest the contrary. (I will note, without going into further detail, that for unique unpublished materials—the treasures that make up the holdings of archives, museums, and library special collections, for example—the most common model has been to put our trust largely in the responsible stewardship of single institutions, since the costs and practicalities of replicating and distributing high-quality copies of unpublished materials typically has been prohibitive. In the digital world, as we couple access with at least limited preservation, we can and should do much better, although this will require major and disconcerting shifts in thinking among our cultural memory institutions.)

The legal, social, and political challenges involved in preserving digital materials are particularly fascinating. They include the creation and maintenance of a public and political consensus that it is necessary and important to preserve our cultural, intellectual, and scholarly record—despite opposition from well-funded and vocal commercial interests with massive political influence. This is not just a question of arguing for sufficient funding for cultural memory institutions. Legal rights of way are needed as well. In the world of printed materials one purchases a book, and then, under a legal doctrine called “first sale,” one has control of that copy of the book: you may keep the book, discard it, or loan it at will—but not make copies (except for some very specific and limited preservation purposes). This is the key both to allowing libraries to operate and to our ability to preserve a great deal of our cultural heritage.

In the world of digital information, consumers face license agreements, pay-per-view business models, subscription music and e-book services, digital rights management technologies, and other obstacles to keeping and maintaining private copies of digital information long-term—particularly when one realizes that the curatorial process for preserving digital information involves making lots of copies and periodically applying various transformations to the work. In essence, under the current legal regime it is very hard for any institution—other than perhaps the Library of Congress, which operates under a special and unique legal status—to preserve commercial digital material without the explicit and active consent of the publisher. There is no obligation for the publisher to extend any special privileges to cultural memory organizations. Indeed, these cultural memory institutions increasingly face an arsenal of digital rights management technologies intended to enforce pay-per-view or use-only-while-subscribed business models. They also face draconian legal sanctions under laws such as the notorious Digital Millennium Copyright Act for any attempt to circumvent such technologies, even for preservation purposes.

Arenas of Crisis and Emerging Responses

There are at least six relatively discrete areas where we face some form of digital preservation crisis. All are different, and some are more obvious than others: published scholarly materials, published non-scholarly materials (consumer marketplace information), digital ephemera, new genres of scholarly work, new teaching and learning materials, and digital records.

Published Scholarly Materials

Traditional journals and monographs are migrating to digital form. At present they are most commonly published both in print and digital formats (with the digital form, as mentioned, the authoritative version). This is expensive and inefficient. The lack of a widely accepted, credible digital preservation strategy is one of the main reasons that print publication continues. Printed editions might gradually become more imperfect representations of the contents of these journals, but we have a track record of successfully preserving print formats.

In this arena, the key problems are economic and organizational—who will do the preservation, and who will pay for it? Legal barriers are minimal, as the scholarly publishers—be they commercial or noncommercial—share common values with higher education institutions, libraries, authors, and readers about the need to preserve the scholarly record. Further, since higher education and libraries represent the vast majority of the market for these publications, they have sufficient negotiating leverage to overcome any divergence of values resulting from market pressures.

Interestingly, the biggest legal problems in this context may not be with commercial publishers, but rather with some of the scholarly societies that wish to retain exclusive control over their publications rather than permit others to archive them, based on the belief that the societies should serve as exclusive stewards for their publications. Because these societies are ultimately governed by the scholarly community, it seems likely that these problems will be resolved within a reasonable period of time. There are technical problems as well, such as establishing standards that allow large numbers of publishers to submit journal issues to archiving organizations efficiently in volume, then getting all the scholarly publishers—particularly the smaller ones—to implement these standards.

The Andrew W. Mellon Foundation has funded a series of university-publisher collaborative projects to explore these issues. The reports of this work are now available through the Council on Library and Information Resources.[1] JSTOR, a not-for-profit organization that serves the library and higher education community, is also actively exploring a role in archiving published scholarly journals, as are organizations such as OCLC. Major scholarly publishers, such as Elsevier Science and the American Physical Society, have established deposit agreements with various national and university libraries to create archives for their journals. The key problems lie in financial models and in scaling up these efforts, and in particular making the models work not only for large and wellresourced publishers, but for smaller and less technically sophisticated ones.

Published Non-Scholarly Materials

Today's mass-market novels, music, film, newspapers, magazines, television and radio broadcasts, and other cultural products represent essential raw materials for tomorrow's scholarship, as well as an integral part of our social and cultural record. More and more such material is appearing in digital form, though not by direct substitution of digital versions for physical artifacts but in a rather complex way: there are a few, unique, born-digital genres appearing, but in the main digital content is partially replicating and partially complementing the content now available through traditional channels, such as printed newspapers and magazines or broadcasts.

In the world of scholarly publishing, authors, readers, and the marketplace largely share common values. Higher education libraries and related cultural memory institutions represent an insignificant part of the marketplace for consumer-oriented cultural materials, however, and hence have no real negotiating leverage with their publishers. Indeed, in many cases, the producers and distributors of these goods would be happy to eliminate libraries, along with resale marketplaces for used products, as drains on the profitability of their products. Further, as these products move to digital form, interest is growing in newly possible business models that are fundamentally incompatible with long-term preservation, such as pay-per-view or monthly music subscription services. On the other hand, there are confusing countervailing trends in some areas. Film preservation, for example, got an enormous boost from the VCR and more recently the DVD, which have for the first time put copies of films into many hands, transforming them from experiential goods to properties. We are seeing a growing market emerge in DVDs of collections of television broadcasts, which helps ensure these will be preserved as well. (Note, however, that preserving digital material on DVDs is problematic due to reformatting and copying considerations—more for legal than technical reasons.)

Perhaps one of the key questions for mass-market materials is whether a more diverse set of business models—embracing both traditional "purchase" models structured by the first-sale doctrine and "rental" models enabled by the new technology—will emerge as these works shift to digital formats. Or it might be that the purchase models, which are friendly to cultural memory organizations and preservation objectives, will simply be eclipsed by the newer rental models in the hopes that they will be more profitable.

The difficulties of preserving digital consumer-market cultural products are numerous. It should be emphasized that in this regard we actually are facing many different markets as a result of the nature of the content and the physical media from which it emerges—the traditions, biases, and economics for music or film are very different from those in the book or magazine market. While cultural memory organizations can work with the publishers and distributors of digital mass-market materials both to help make the case for why preservation is important and to provide guidance and assistance in accomplishing it, the legal and economic problems remain massive.

National libraries, such as the U.S. Library of Congress, have special roles to play here. They can convene producers to discuss archiving, and they also may be able to use legislative provisions to compel rights holders to deposit material for archiving. For example, the Library of Congress is now in the second phase (estimated to cost $20 million) of a large-scale program for preserving our digital cultural heritage for the nation.[2] This phase will involve further convening of discussions with content producers, as well as the launch of a series of pilot projects in various areas.

Digital Ephemera

The World Wide Web has many functions. It serves as a portal to many commercial (both subscriber and advertiser supported) and nonprofit information services. It also serves much like a printing press, housing vast amounts of ephemera—political broadsides, leaflets, menus, transportation schedules, announcements, posters, and similar materials that historically have been collected by cultural memory institutions and that represent essential raw material for future scholarship. In the multimedia world of the Web, these materials are supplemented with collections of photographs, movies, musical performances, and other materials.

Historically, if you could obtain a copy of something in a physical format, you could add it to a collection and preserve it. Under current copyright law, it's not at all clear you can simply copy a document from a Web site without obtaining permission of the author (although maybe the Library of Congress can do so because of its special role with regard to copyright—or maybe not; there's little case law here). In most cases creators don't care if their material is copied from the Web or are happy to have their work preserved. If they don't want their material archived, easy opt-out mechanisms could be devised. The logistics of obtaining literally millions of permissions per year from people who are not set up to give such permissions are utterly intractable, however.

Several organizations have taken initial steps to capture and preserve the materials on the Web, most notably Brewster Kahle's Internet Archive in San Francisco[3] and, more recently, several European national libraries. In the United States, the Library of Congress has also done some pilot projects with the Internet Archive.

There are technical problems about how much of the Web can be captured and what might be missed. There are economic problems as well. While the total cost of what the Internet Archive does is tiny compared to what libraries in the United States spend per year (and also tiny compared to the huge amount of material that it captures and protects!), secure long-term funding needs to be put in place. Approaches are also needed to manage the legal risk involved in large-scale capture and preservation of the Web. This could be handled with some fairly simple—and, if properly crafted, perhaps not even overly controversial—legislation.

New Genres of Scholarly Work

While traditional scholarly publishing venues such as journals and monographs are slowly moving beyond the conceptual limitations of works designed for the printed page, some of our most creative and dedicated faculty are aggressively exploring the frontiers of what the digital medium can offer in communicating and documenting research, and in enhancing teaching and learning. These scholars are beginning to develop the various genres that represent the intellectual descendents of the monograph for the digital age, for example.

This work, which is centrally important to the future of scholarship and the academy, faces many burdens. Particularly for junior faculty, there are questions of how to evaluate such work within the context of promotion and tenure and how it should be weighted in comparison to more traditional products such as journal articles and books. Further, because it does not fit within the framework of traditional scholarly publishing, there is no automatic mechanism for ensuring that novel digital work gets replicated and preserved for the long term—which in turn further jeopardizes the credibility of the work in tenure and promotion settings.

It's not just new genres of authorship. The practices of scholarship are changing in other ways as well; for example, by becoming much more data-intensive. In addition, it's increasingly commonplace for authored works to be linked to data sets in various ways. Sometimes the data sets are deposited in disciplinary repositories (as is the case in some areas of molecular biology, for example). In other cases the author retains responsibility for stewardship of the data. Here the legal and political problems are relatively minimal. Responsibility is increasingly clear: the institutions that employ the scholars producing the work bear primary responsibility, although some disciplinary communities are organizing data archives along disciplinary lines (with funding from government agencies that support scientific research, for example).

The key approach developing in this area is the concept of an institutional repository, which is a set of services offered by a university or other organization that provides a professionally managed system of dissemination, stewardship, and preservation for works created by the organizational community. Perhaps the best known of these projects is the Massachusetts Institute of Technology's DSpace system,[4] jointly developed with Hewlett-Packard. DSpace has been made available as free open-source software and is now being replicated at a number of institutions both in the United States and abroad with funding from the Mellon Foundation and other sources. For more on institutional repositories, see "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age."[5]

New Teaching and Learning Materials

This is to some degree an extension of the previous category. Universities increasingly use advanced information technology platforms such as learning management systems to deliver teaching and learning materials and also to support interactive processes both for distance education and face-to-face classroom experiences. These represent new genres of scholarship.

Again, this is an area where MIT has taken considerable leadership with its OpenCourseWare Program,[6] which seeks to place faculty materials for the vast majority of its courses online for public access over the next few years.

I have treated teaching and learning materials as a category separate from research for the important reason that, unlike new genres of research publications, they are not simply faculty work. They typically involve student participation and represent not new ways of communicating research but new ways of facilitating and documenting the processes of teaching and learning. As a result, enormous policy and legal questions surround the desirability and practicality of archiving these works. To the best of my knowledge, little systematic work is being done in this area other than the initiation of policy discussions, particularly as relates to student work. In this area, I would argue that the first step is to decide what we are trying to accomplish; then we will have to address all the questions of legal, technical, and economic feasibility and strategy.[7]

Digital Records

In the discussion of digital materials thus far, I began with traditional published scholarly materials that are moving to digital form and progressed steadily farther away from well-established publishing traditions into new-media scholarship and teaching and learning materials. Another important class of materials includes those that were never published in the physical world, nor will they ever be published in the digital world; these comprise the large and important class of organizational and personal records.

Many important materials fall under organizational stewardship. The ability of organizations to preserve these materials even as their creation and management shifts to digital formats has not been impressive to date.

Key materials stem from government sources. The National Archives and Records Administration (NARA), for example, is charged with maintaining the federal government's records; increasingly, this record is constituted in digital materials. Similar issues arise at the state and local levels. All of these records are critical parts of the intellectual and social record and form vital components of the raw materials for future scholarship. Most of the non-federal governmental efforts are terribly under funded—if they are funded at all—and lack the technical expertise to deal with the complex problems they face. Yet these organizations are custodians of vital, essential evidence that is part of our society's record and that scholars will continue to need to access. The records of corporations and nonprofit organizations are also invaluable resources for future scholarship.

Universities themselves generate records that are important parts of society's record, and these too are in jeopardy as they move to digital form. Consider materials such as transcripts, which need to be preserved for very long periods of time (relative to the lifespan of administrative systems). Note also that, in order to interpret a transcript, one needs the related records of the institution's annual course catalogs, yet these are rapidly transforming from publications that can be preserved to up-to-date, dynamically maintained databases of courses that make no provision for preservation. Universities have a particular responsibility to display responsible leadership in this area. For more information specifically on higher education records, see the valuable series of ECURE conferences.[8]

Not all records are institutional. If one looks at what makes some of our great research libraries and archives such magnificent and unique resources for scholars, it is their special collections. In substantial part, these are built from personal papers of important authors, artists, scholars, politicians, or other public figures. As masses of personal papers morph into masses of files on a few hard drives on personal computers, our libraries and archives struggle to develop the appropriate professional practices and supporting technologies to deal with the shifting technology and behavior patterns. This too constitutes a real threat to the future of our research collections and our cultural memory institutions.

Conclusions

Substantial and critically important problems involving the preservation of digital content by our cultural memory institutions—universities, libraries, museums, and archives—abound, each with its own specific frame and implications. The good news is that a great deal of effort is mobilizing to address these problems.

What does higher education leadership need to do today? Most importantly, recognize these problems, explore them, and help others understand them. Advocate for support to address them through government policies and legislation on intellectual property, and for government funding of cultural memory organizations. Help the public understand how the world is changing and what is now at risk. Focus action and investment on areas where higher education has direct responsibilities and can make a distinct and rapid impact on the problems. These areas, in my view, include planning for collective action to archive traditional scholarly publishing materials as they move to digital form; building institutional repositories to support new-media scholarship and teaching and learning materials; exploring policies for digitally enabled teaching and learning materials and records; and seriously engaging the issue of university records in a digital world. Finally, higher education leaders should seek opportunities for the campus community to intellectually engage the complex and fascinating questions of records, memory, persistence, and accountability in the digital world; these are essential questions for our society today and will continue to be so in the coming decades. Insights and intellectual leadership from the academy will be sorely needed to address them.

NOTES

[1] See the Council on Library and Information Resources Web site, http://www.clir.org.

[2] See the Library of Congress project description for the National Digital Information Infrastructure and Preservation Program on the Web at http://www.digitalpreservation.gov/.

[3] See the Internet Archive on the Web at http://www.archive.org.

[4] See the DSpace section of the MIT Web site at http://www.dspace.org.

[5] Clifford A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," *ARL Bimonthly Report,* Vol. 226, February 2003, pp. 1–7; available on the Web at http://www.arl.org/newsltr/226/ir.html.

[6] For more on MIT's OpenCourseWare Program see http://ocw.mit.edu.

[7] For more information see Clifford A. Lynch, "The Afterlives of Courses on the Network: Information Management Issues for Learning Management Systems," EDUCAUSE Center for Applied Research Bulletin, Vol. 2002, Issue 23, November 26, 2002; available on the Web at http://www.cni.org/staff/cliffpubs/ECARpaper20024.0.pdf.

[8] The ECURE conferences focus on preservation and access for electronic college and university records. See http://www.asu.edu/it/events/ecure/.

Clifford A. Lynch is Director of the Coalition for Networked Information (CNI).