Converting and Preserving the Scholarly Record

An Overview

By Jeffrey L. Horrell

The author provides an overview of the issues related to preservation in the digital environment and describes initiatives that promise to address these issues. He considers the mutability of electronic content, the mission of libraries to preserve, the long-term ownership of digital content, the nature of preservation in a digital age, and promising digital preservation initiatives. The paper, drawing on the work of a collaborative Duke/Dartmouth Mellon-sponsored project, concludes with recommended elements for a campuswide digital repository.

To set the stage for this article, let me begin by sharing my first experience of realizing what it meant to preserve or potentially lose scholarly content. Picture a library school (when they were called library schools) student in his first semester, working at the reference desk of a major research institution on a Sunday evening, when a user appears at the desk with a handful of author catalog cards clearly ripped from the public card catalog (again, when there were such things), presents them, and quite casually asks, "Where do I find these books?" The shock of what I was witnessing was overwhelming to a soon-to-be librarian! I quickly composed myself, politely offered a stacks chart, and offered to write down the call numbers as I extended my hand to retrieve and secure the cards. I can remember the almost reluctant tug of the cards from the individual. The transaction ended smoothly, with the cards in my hand and the person off to find the materials. At that moment, in a very small, but what seemed dramatic way, I felt what the responsibility for preserving access to content really meant. No doubt, scholarship as we know it would not have come to a standstill without that handful of cards, but the potential loss might, indeed, have had an effect.

Forward several decades later, and ask yourself if you have had the experience of searching on the Internet, locating a Web site using Google or Yahoo, and returning to search for it again to find that the site has disappeared. Have you expected to find particular information on a Web site, but cannot because the site has been updated and there is no easily accessible archive? Have you encountered a difference between the print and electronic versions of a document and wondered which one is correct? There are countless examples of this "now you see it, now you don't" phenomenon.

Consider Britannica Online. For decades, print editions of the *Encyclopaedia Britannica* have been considered authoritative and reliable, but were out of date soon after publication. By comparing one edition with subsequent editions, one could see when a topic was introduced or an entry changed. With the elec-

Jeffrey L. Horrell (Jeffrey.L.Horrell@Dartmouth.edu) is Dean of Libraries and Librarian of the College, Dartmouth College, Hanover, New Hampshire.

This article is based on a presentation given at "Converting and Preserving the Scholarly Record," Eighth Annual Symposium on Scholarly Communication, State University of New York, Albany, October 24, 2006.

Submitted February 12, 2007; tentatively accepted pending revision April 2, 2007; revised and resubmitted June 30 , 2007, and accepted for publication.

tronic version, however, changes are often transparent, and information moves in and out of the work without notice. Another example is the online version of the journal *Nature*, which several years ago embargoed or delayed publishing parts of its content in the electronic version. In some ways, this practice could be an effective, strategic marketing initiative designed to preserve the print subscription base, but nonetheless, it was not clear if and when the content of the online version was complete. Finally, I think we can all point to redesigning or writing over Web sites in our own institutions, if not our libraries, and in the process losing content that traditionally had been preserved in print, but lost in subsequent electronic versions.

In the predigital era, an understanding of what actually constituted a record was less complicated, and libraries, together with records management units, provided preservation and archival services for their institutions that, in large part, preserved our history and ensured regulatory compliance. With multiple libraries acquiring the same titles, redundancy was a safeguard for the materials. Volumes and archival records rested neatly on shelves in our libraries or in climate-controlled storage facilities. Guthrie, former head of JSTOR and now Ithaka, has pointed out that costs certainly were associated with preserving the print copies beyond simply storage, including maintaining the physical environment, shelving, repair, microfilming, and replacement in some instances.[1] But with the pace at which new forms of digital formats are replacing print, photographic film, video, and audio recordings, combined with extraordinary amounts of digital data that can be readily acquired or licensed, stored, and disseminated, institutions and society are challenged to consider ways of maintaining archives of digital objects of all descriptions.

As we think about this challenge, I believe it is important to remind ourselves why preserving information is important in the first place—and this takes us to our mission. The basic elements of the mission of an educational institution include researching, teaching, publishing research outcomes, and maintaining these results in a record of some sort. Data replication is essential and at science's core. Building on the scholarly record is central to the humanities and social science traditions. The mission statements of our libraries reflect the same elements. Dartmouth's is as follows:

> The Dartmouth College Library fosters intellectual growth and advances the teaching and research missions of Dartmouth College by supporting excellence and innovation in education and research, managing and delivering scholarly content, and partnering in the development and dissemination of new scholarship.[2]

Managing the Scholarly Content Means Preserving It

Waters, program officer of scholarly communications at The Andrew W. Mellon Foundation, in a 2006 paper titled, "Managing Digital Assets in Higher Education: An Overview of Strategic Issues," indicated that dissemination, preservation, and access refers to the life cycle of scholarly resources that are used and produced in teaching and research, and are the objects of scholarly communications.[3] He went on to outline a serious and deeply troubling scenario that centers on the transition from print to electronic publishing and from owning to licensing information. When a library purchased materials outright, it could do with it as it liked within the guidelines of copyright. There were instances of libraries giving content to microfilm publishers and then buying back the products, and of libraries allowing publishers to convert their microfilm content to digital format and then licensing it. Now, in some cases, this content is maintained on remote systems controlled by publishers. In effect, libraries and their institutions have an ongoing mortgage for the content that they owned in the first place, in print. No doubt access is improved, but something is seriously wrong with this model. Waters argued that one could see a business model for publishers that offers data-mining services for the large aggregation of content that could enable greater opportunities for scholarship. In an institution of higher education, one could easily imagine preservation at the center of such an endeavor, but is there a compelling business interest for it in a profit-driven company? The results could be that libraries will not own or have the rights to the scholarly products, nor will they have a true archive of them, and publishers could apply whatever pricing model they choose for such data-mining services. This presents a scenario where the control of the scholarly record moves from the academy to the publishing, and mostly for-profit, sector.

In addition to considering our mission in light of the business model just described, we also are faced with new questions about the nature of document preservation. What, exactly, is preservation? For example, can we say that a document has been preserved if we save the text, but our digital systems cannot reproduce its original typeface or style?[4] Related issues surrounding the context and thinking behind manuscripts or policies that were once captured in letters, memoranda, drafts, and other ancillary documents need to be considered in a world driven by e-mail and instant message. As a society and as educational institutions, we have a collective responsibility to preserve and make available, along a continuum of a life cycle, our digital heritage, but an understanding of what preservation means in the digital world is complicated.

Promising Initiatives

How do we proceed? Several examples of promising preservation initiative are worth noting. One that began several years ago and now has more than forty libraries as partners is Lots of Copies Keep Stuff Safe (LOCKSS).[5] It also had the engagement and support of more than thirty publishers. The goal of LOCKSS is to provide a low-cost, low-tech system of ensuring continued access to journal literature. It collects newly published content using a Web crawler similar in nature to those used by commercial or other search engines. It compares the content it has collected with the same content on other distributed computers and repairs or reconciles any differences. Earlier this year, a project called Controlled LOCKSS (CLOCKSS), which uses the LOCKSS methodology, was developed as a dark archive intended to serve as a fail-safe repository for this content.[6] The content from CLOCKSS would only be used in the event that it was no longer available from the publisher. A group representing publishers, learned societies, and libraries would be responsible for deciding the trigger conditions by which the content could, or should, be made available.

Another important preservation initiative is Portico.[7] Sponsored by The Andrew W. Mellon Foundation, Ithaka, the Library of Congress (LC), and JSTOR, Portico provides limited access for audit purposes and institutionwide access in the event content is no longer available from a participating publisher. Portico intends to provide a reliable methodology for ongoing access to an institution's scholarly collections. A growing number of large and important publishers are partnering with Portico, including Elsevier, Oxford University Press, the University of Chicago Press, John Wiley and Sons, the UK Serials Group, the American Anthropological Association, and the Berkeley Electronic Press, with more than five thousand journals already slated to be archived. A list of nearly two hundred libraries of varying sizes and missions have either become partners with Portico or are seriously considering becoming involved.

The developing work of LC and its National Digital Information Infrastructure and Preservation Program (NDIPP) also is significant.[8] NDIPP's mission is to work closely with a number of federal agencies and private partners to provide a national focus on important policies, standards, and technical components required to preserve digital content. Various options and technical solutions are being explored and tested. With an appropriation of nearly $100 million, LC's role will be an important part in helping address these issues. A recently announced Web site related to this work is the WEB Capture site.[9] Since 2000, LC has been selectively capturing and preserving sites in such areas as Hurricane Katrina, recent Supreme Court nominations, and the transition subsequent to the death of Pope John Paul II. Current projects include the crisis in Darfur, Sudan, the Iraq War, and the 2006 elections. Also, related to the earlier description of electronic journal archiving initiatives, LC and the British Library, under NDIPP's auspices, have agreed to support a common archiving standard for electronic content migration.[10] Finally, the MetaArchive Project, a collaborative venture of eight institutions, including LC, is a three-year project to develop an infrastructure to capture at-risk digital content related to the culture and history of the American South.[11]

Also on the federal level, the National Archives and Records Administration and the San Diego Supercomputer Center have recently agreed to work together to preserve valuable digital collections.[12] This unprecedented partnership between the National Archives and an academic institution marks an opportunity for securing critical data created for, or by, agencies of the United States federal government's executive branch.

Brewster Kahle's Internet archive, Wayback Machine, is another effort to archive digital information, specifically Web pages.[13] Begun a decade ago, it provides the ability to browse, not search, more than 55 billion Web pages; by design, it is not comprehensive and efficient in mining its content. However, the Internet Archive has collaborated with the Smithsonian and LC, and has developed a number of important collections, including the United Kingdom Central Government Web Archive, a collection devoted to sites instrumental in the early development of the Internet, and a number of election sites.[14] It now offers a Web Archive on Demand Service, which is a subscription-based archiving service targeted to a range of institutions at costs lower than some other archiving platforms. Called Archive It, subscribers can capture, organize, and theoretically preserve material from the Internet as well as their own institutions and collections. Users can then search these collections fairly easily.[15]

Campus-wide Asset Management: The Duke/Dartmouth Project

Work is nearing completion as part of a shared planning grant undertaken by Duke University and Dartmouth College and sponsored by The Andrew W. Mellon Foundation. Many projects are underway at a variety of research institutions, but they have not necessarily been undertaken in the context of a broad, campuswide asset management plan. Rather than looking for technological solutions, the focus of this grant has been on designing institutional strategies and policies for managing scholarly and administrative assets in digital form. Taking an institutional approach potentially brings advantages: a common understanding of the value of asset management, a shared commitment to building a campuswide repository, economies of scale, and the possibility of a consistent set of policies that will apply to a wide range of

materials. The decentralized nature of most institutions and the resulting silos of content and policies and proceedings associated with them is a considerable barrier to a common view of the overall challenge. There are cultural obstacles that come into play. Faculty members are not accustomed or always comfortable in placing their work in an institutional repository and have traditionally managed it themselves. Funding also is a challenge. Without a comprehensive plan to support the development of the systems, it is difficult to contemplate a successful outcome.

An institutional program, as identified in the collaborative planning project at Duke and Dartmouth, has several elements:[16]

- *Structure*. An overall program to manage a university's digital assets requires a formal organizational structure that is part of the institution's overall administration. There should be a steering committee with broad oversight. The committee's charge should include responsibility to research, develop, and implement an enterprisewide program for the institution. Policies will need to be established in addition to hiring staff, charging subgroups, developing implementation teams, and assigning accountabilities to individuals and groups across the institution. Three areas are key: priorities, policies, and implementation.
- *Priorities.* Setting priorities for attention and resources is essential. A census of asset areas should be developed to serve as the basis for priority setting. Information considered vital to the operation of the institution, or of irreplaceable value, will be of highest priority. There may be stopgap measures necessary to prevent loss. Timing may well be critical in these instances, and the committee will need to be flexible and act quickly.
- *Policies*. A set of principles and policies for digital asset management for the university will provide continuity and clarity for its users and stakeholders:
  - Think globally. Because of the scope and complexity of the effort, it is important initially to think globally and act locally. Decisions made in designing and implementing specific solutions should take into account issues of scalability, institutional capability, future data migration, available support, and overall institution efficiency.
  - Incentives for participation. While the digital records of university administration are clearly owned by the university (and staff can be required to deposit them in an asset management system), the issues are not so clear for faculty, as intellectual property ownership and workflow processes are much less straightforward. Therefore, planners should carefully think through how to make it clear that participation in a university digital asset management program is in the best interest of faculty, not just of the university. It should be self-evidently useful to all whose participation is required for its success, and should meet their needs and save them time.
  - Confidentiality and openness. Planning for a university digital asset management system should include development of policies that differentiate among a variety of use cases and provide for different levels of access and security, depending on the submitter's and university's needs for openness or confidentiality in those cases.
  - Terms of use, stewardship, and governance. The policies that govern the rights and responsibilities of the various program stakeholders must find a way to balance potentially conflicting requirements between depositors and the institution, and account for changing needs as digital assets mature through different phases of their life cycle.
- *Implementation*. While each college or university may have different dynamics, there are some points that will improve the chances of success in most academic environments:
  - Sponsorship and the Digital Asset Management Steering Committee. The first step in establishing an enduring institutional culture of information stewardship is to secure explicit, high-level endorsement and support. A sponsor or sponsors at the level of the provost or executive vice president can facilitate a good start and ensure ongoing support. The second step is to create a broadly based steering committee to manage the program's establishment and foster collaboration across the university.
  - Steering committee activities. the steering committee will need to assign tasks to short-term teams. While the committee should retain the priority-setting and oversight activities, the development of day-to-day practices and support should be tasked to functional groups, such as administrative departments, the library, and information technology organizations.
  - Establishing a permanent organizational structure. Informed by the work of the program coordinator and start-up teams, the steering committee needs to identify a permanent organizational home for the program within the university.
  - Commonalities and differences among different parts of the organization. As the system is developed, it must be responsive to the diverse requirements of distinctive units of the institution. The differences in managing material for

teaching, research, and administration can be dramatic. Intellectual property and access issues arise frequently on the academic side, while administrative data may need to be summarized and analyzed on an ad hoc basis by many organizations on campus.

- System attributes. Attributes of an underlying technology system can influence the program's success or failure. The steering committee should identify the technology characteristics needed for success, including:
  - The technology that supports the digital asset management effort will evolve over time.
  - The institution must retain the ability to export and migrate materials to new technology as systems evolve.
  - It must be possible to remove materials from the system.
  - The time it takes to access materials must be perceived as reasonable.
  - There must be appropriate tools to access, analyze, or transform the materials stored.
  - There must be selective degrees of access provided to materials.
  - Policies must provide guidance for administrators, faculty, and other users of the system on what can be added and what cannot be added, as well as how long they should be retained.
  - Measuring success. The steering committee should require an evaluation/assessment model to be developed to measure the success of the endeavor.
  - A final key role of the steering committee and the sponsors is to assign responsibilities and identify institutional custodian(s) of the materials.

Concluding Thoughts

Ultimately, we must provide a secure infrastructure to ensure the enduring viability of digital content for the business aspects of our institutions and for the intellectual assets produced and acquired by and for the scholarly community.[17] Future generations of students, faculty, administrators, and scholars are depending on us. The Duke/Dartmouth report, submitted to the senior leadership at Dartmouth, is a call to action. Wess Jolley, Dartmouth's records manager, describes it as follows:

> It is a call for leadership within our institution in response to a critical need. Inaction based on concern about the potential difficulties is not an option. Indeed, the accelerating transition from paper to digital records has already caused the irreversible loss of vital historical information. Unless we act decisively and immediately, the first years of the twenty-first century will be forever known as the era of lost history. Besides the intellectual impact, from a legal standpoint the transition to digital record keeping is a ticking bomb for our institutions. As digital systems replace paper, our carefully formulated records retention programs are becoming null and void. Without a digital equivalent to lifecycle controls traditionally established for paper records, each new digital record is a potential legal liability, and our ability to conduct the business of our institution becomes increasingly difficult. Simply put, we are losing control of our records with every passing day.[18]

A centralized approach to digital asset preservation can reduce legal liability and begin to ensure digital records are captured, maintained, disposed, and preserved over time. We should not underestimate the scale, scope, and ongoing nature of this task, and we cannot disregard or fail to meet this challenge. Too much is at risk and at stake. Now is the time for leadership and action.

References

1. Kevin Guthrie, "Archiving in the Digital Age," *EduCause Review* (Nov./Dec. 2001): 57–65, www.educause.edu/ir/library/pdf/erm0164.pdf (accessed Jan. 16, 2007).
2. Dartmouth College Library, "Dartmouth College Library Mission and Goals Fiscal Year 2005–2006," www.dartmouth.edu/~library/col/DCL_Mission_Goals.pdf (accessed Jan. 16, 2007)
3. Donald J. Waters, "Managing Digital Assets in Higher Education: An Overview of Strategic Issues," *Association of Research Libraries Bimonthly Report* no. 244 (Feb. 2006): 1–10, www.arl.org/newsltr/244/assets (accessed Jan. 16, 2007).
4. Jeffrey Horrell and Martin Wybourne, "Archiving the Digital Age: How Do We Preserve Our Present for the Future?" *Vox of Dartmouth,* July 25, 2005.
5. LOCKSS, www.lockss.org/lockss/Home (accessed Jan. 16, 2007).
6. CLOCKSS, www.lockss.org/clockss/Home (accessed Jan. 16, 2007).
7. PORTICO, www.portico.org/ (accessed Jan. 16, 2007).
8. Library of Congress, National Digital Information Infrastructure and Preservation Program, www.digitalpreservation.gov/index.html. (accessed Jan. 16, 2007).
9. Library of Congress, "WEB Capture," www.loc.gov/webcapture/index.html (accessed January 16, 2007).
10. Sustainability of Digital Formats Planning for Library of Congress Collections, "NCBIArch_1, NCBI/NLM Journal

Archiving and Interchange DTD, version 1" (Apr. 19, 2006), www.digitalpreservation.gov/formats/fdd/fdd000174.shtml (accessed Oct. 27, 2007).
11. "MetaArchive," www.metaarchive.org (accessed Jan. 16, 2007).
12. San Diego Supercomputer Center, News Center, "The National Archives and the San Diego Supercomputer Center Sign Landmark Agreement to Preserve Critical Data" (June 27, 2006), www.sdsc.edu/News%20Items/PR062706.html (accessed Jan. 16, 2007)
13. Internet Archive, Wayback Machine, www.archive.org/web/web.php (accessed March 31, 2007).
14. UK Government Web Archive, www.nationalarchives.gov.uk/preservation/webarchive (accessed Jan. 16, 2007).
15. Internet Archive, "Archive-It: Archiving the Internet for Future Generations," www.archive-it.org (accessed Jan. 16, 2007).
16. *Digital Asset Management: Elements of an Institutional Program—Final Report on the Duke/Dartmouth Project* (Draft of 30 November 2006), www.dartmouth.edu/~library/col/docs/0607/DukeDartmouth.pdf (accessed Dec. 9, 2007).
17. Horrell and Wybourne, "Archiving the Digital Age," 7.
18. Wess Jolley, College Records Manager, Dartmouth College, personal conversation with the author, March 17, 2006.