

Archiving the World Wide Web

Peter Lyman
School of Information Management and Systems
University of California, Berkeley

PROBLEM STATEMENT: WHY ARCHIVE THE WEB?

The Web is the largest document ever written, with four billion public pages and an additional 550 billion connected documents on call in the “deep Web” (Lyman and Varian 2000). The Web is written in 220 languages (although 78% is in English) by authors from every nation. Ninety-five percent of Web pages are publicly accessible, a collection fifty times larger than the texts collected in the Library of Congress, making the Web the information source of first resort for millions of readers. Yet the Web is less than ten years old and the process of economic, social, and intellectual innovation it is causing is just beginning.

The Web is growing quickly, adding more than seven million pages daily, yet at the same time it is continuously disappearing. The average life span of a Web page is only 44 days, and 44% of the Web sites found in 1998 could not be found in 1999.¹ Web pages disappear every day as their authors revise them or servers are taken out of service, but users notice only when they enter a Web address, a Universal Resource Locator (URL), and receive a “404–Site Not Found” message. As ubiquitous as the Web seems to be, it is ephemeral, and today’s Web will have disappeared by tomorrow. The implication is clear: if we do not act to preserve today’s Web, it will disappear.

In the past, important parts of our cultural heritage were lost because they were not archived, in part because past generations did not—or could not—recognize their historic value. This is a *cultural problem*. They did not address the *technical problem* of preserving storage media—nitrate film, videotape, vinyl—or the equipment to access new media. They did not solve the *economic problem* of finding a business model to support new media archives, for in times of innovation the focus is on building new markets and better technologies. And they did not solve the *legal problem* of creating laws and agreements that protected copyrighted material and at the same time allowed for its archival preservation. Each of these problems faces us again today in the case of the Web.

The cultural problem. The very pace of technical change makes it difficult to preserve digital media. How many people can retrieve documents from old word processing diskettes, or find yesterday’s e-mail? All documents have a life cycle: from valuable to outdated, but then, perhaps, to historically important. Archivists often rescue boxes of printed documents as they leave the attic on their way to the dump. But the Web isn’t stored in attics; it just disappears, so preservation is urgent. The hard questions are how much to save, what to save, and how to save it.

¹ The lifespan estimate is from “The Size of the Web” cited in Lyman and Varian 2000. The Web site estimate is from OCLC’s Web Statistics (June 1999) cited in Lyman and Varian 2000. The original source document is available on the Internet Archive “Wayback Machine” at web.archive.org/web/.

The technical problem. Every new technology takes a few generations to become stable, so we do not think to preserve the hardware and software necessary to read old documents. Digital documents are particularly vulnerable, since the very pace of technical progress continuously makes the hardware and software that contain them outmoded. A Web archive must solve the technical problems facing all digital documents, plus its own unique problems. First, information must be continuously collected, since it is so ephemeral. Second, information on the Web is not discrete; it is *linked*, so the boundaries of the object to be preserved are ambiguous.

The economic problem. Who has the responsibility and the resources to collect and preserve the Web? The economic problem is acute for archives. Since their mission is to preserve primary documents for centuries, the return on investment is very slow to emerge—and may be intangible and hard to measure. Archives serve the public interest in the very long run, with immediate benefits only for a few scholars. For this reason, they tend to be small and specialized. However, a Web archive will require a large initial investment for technology, research and development, and training—and be built to a fairly large scale—if it is continuously to save the entire Web.

The legal problem. Many believe that current intellectual property laws concerning digital documents are optimized to develop a digital economy; thus, the rights of intellectual property holders are emphasized. Copyright holders have reason for caution since the technology is so new, and the long-term implications of new laws are unknown. Although the Web is popularly regarded as a public domain resource, it is copyrighted; thus, archivists have no legal right to copy the Web. And yet it is not *preservation* that poses an economic threat to new markets, it is *access* to archives that might damage new markets, and this is the most urgent problem to be solved.

Access is a political as well as a legal problem. And like all political problems, the answer lies in establishing a process of negotiation among interested parties. Who are the stakeholders and what are the stakes in building a Web archive?

- For librarians and archivists, the key issue is to ensure that the historically important parts of the documentary record are preserved for future generations.
- For owners of intellectual property rights, the problem is digital asset management, the flexibility to manage and experiment with the creation of new information products in order to create sustainable markets, and to protect this investment.
- The Constitutional interest is twofold: the innovation policy derived from Article I section 8 (“progress in the useful arts and sciences”) and the First Amendment. Copyright law has operationally defined this interest to include the right to quote and criticize published works, and providing for personal educational uses of information for learning and the creation of new ideas.
- The citizen’s interest is in *access* to high quality authentic documents, through markets, libraries, and archives.
- Schools and libraries have an interest in educating the next generation of creators of information and knowledge by providing them with access to the documentary record; this means access based upon the need to learn rather than the ability to pay.

In sum, the policy problem is to find a process for balancing these interests in the long run, including finding a way for significant experiments to be conducted and evaluated by each of the parties, and to reach negotiated solutions that strike a balance among legitimate, contending interests.

TECHNICAL DESCRIPTION OF THE OBJECT

The literature on digital preservation is focused on solving five key technical problems (Besser 2000).

- *The viewing problem* is the maintenance of an infrastructure and the technical expertise necessary to make digital documents readable.
- *The scrambling problem* is decoding any compression or technical protection service software protecting the Web page.
- *The inter-relation problem* is preserving the contexts that give information meaning, such as links to other Web pages.
- *The custodial problem* is defining the standards, best practices, and collection policies that define the boundary of the work and its provenance and authenticity.
- *The translation problem* concerns the way the experience and meaning of the Web page are changed by migrating it into new delivery devices.

In building a Web archive these problems translate into three questions: What should be collected? How do we preserve its authenticity? And how do we preserve or build the technology needed to access and preserve it?

What is the digital object to be collected?

Ultimately, the scope and scale of a Web archive will be determined by the definition of the digital object to be collected, the “Web page.” This is not a simple matter. From a user’s point of view, a Web page is the image called forth by placing a URL address into a Web reader. This operational definition is necessary but not sufficient, for an archive must be sure that the document is *translated* in an authentic manner. In this case, authenticity means that the document must include both the context and evoke the experience of the original.

The average Web page contains fifteen links to other pages or objects and five sourced objects such as sounds or images, so the boundaries of the digital object are ambiguous. If a Web page is the answer to a user’s query, what must be preserved is a set of linked Web pages sufficient to provide an answer. From this perspective, the Web is like a reference library; that is, it is the totality of the reference materials in which a user might search for an answer. If so, the object to be preserved might include everything on the Web on a given subject at a given point in time, e.g., the 2000 election or the World Trade Center terrorist attack. Thus, there is a temporal dimension: must we preserve the context of the Web page at every point in time, or when it was created, or when it was at its best? And there is the issue of quality: are we to preserve all pages relevant to a query, or just the best ones? And who is to judge?

None of these options is easy to accomplish, for the Web is not a fixed collection of artifacts. Today, the *surface Web* contains all of the static HTML pages that can be accessed by Web URL addresses. Some of the surface Web, especially in the commercial sector, requires passwords or encryption keys; this might be called the *private Web*. To archive these Web pages would require permission of the owners, and the private Web is often encased in security protection services that make copying and preservation doubly difficult. But today, surface Web pages are often generated on the fly, customized on demand from databases in the *deep (or dark) Web*. The deep Web is estimated to be 500 times larger than the surface Web. It includes huge data sources (such as the National Climatic Data Center and NASA databases) and software code that provides information services for surface Web pages on the fly (such as the

Amazon.com software that creates customized pages for each customer by name). The deep Web is the architecture that produces what we read on the surface; the surface itself only exists as long as a reader is using it, then it disappears. This deep Web cannot easily be archived, since the data are guarded by technical protection services. It is also potentially protected by privacy concerns, since if Amazon.com owns a profile of my use of information, it is not necessarily available for archiving without my consent. Here there are not only tensions between markets and archives, but conflicts between privacy concerns and the interest of history.

The ambiguous boundaries of Web objects are also problematic because they are compounds of design elements, including texts, pictures, graphics, digital sound, movies, and code—the list expands as innovation continues. Each of these elements has intellectual property rights attached to it, although they are rarely marked and sometimes impossible to trace. Yet, at least in principle, a digital archive would have to have permission from each of these rights holders. In the words of the National Research Council's report, *The Digital Dilemma: Intellectual Property in the Information Age*, "for the digital world, one must sort out and clear rights, even of ephemera" (National Research Council 2000, 12).

Even if the Web page could be copied technically and we knew what we wanted to preserve, Web pages are protected by copyright law. Even now there are sophisticated debates about how a Web archive should collect data: should the default be that copyrighted information is collected and the owner has to opt out; or may it not be collected or disclosed unless the owner actively gives permission, or opts-in? This is one of those questions that legislation or the courts may resolve. It is important to remember, however, that the Web is a global document, so there are likely to be many different jurisdictions making laws and rules, and enforcement across national borders will be very difficult without treaty agreements.

The authenticity and provenance of the object collected

Defining the boundaries of the object to be collected also requires a decision about authenticity and provenance. These decisions, whatever they may be, must be recorded as part of the archive; the preservation community calls this kind of information "metadata," which is information about information, and often builds records of what is in the collection using this metadata. A standard way of recording the metadata must be created to record the historical and technical context in which the document(s) were found. Among many other facts, metadata might record answers to the following questions (Besser 2000):

- What is the name of the work? When was it created, and when has it been changed? Who created it, changed it, or reformatted it?
- Are there unique identifiers and links to organizations or files or databases that have more extensive descriptive metadata about this record?
- What technical environment is needed to view the work, including applications and version numbers needed, decompression schemes, and other files? If the Web page is generated on the fly, what database generated it, and what is known about its provenance?
- What technical protection devices and services surround it, if any?
- If the Web page contains more than text, what applications generated the sound, video, or graphics?

What copyright information is there about each of the elements of the Web page, and what is the contact information for them?

Work to define standard answers to these and other questions is ongoing through the Dublin Core metadata project.

What technologies are needed to preserve the Web collection?

Technologies to reproduce the Web object—however defined—must be preserved, including the hardware and software necessary to access the information in an authentic context, or to recreate it. This is difficult in the best of cases. Have we authentically preserved a computer game if we just preserve the graphics, or must we preserve the look and feel of the game in use? Every solution changes the context of information in ways that affect its authenticity: one strategy tries to preserve the original equipment; another uses contemporary technology to *emulate* the original “look and feel” of the information in use; while another migrates the digital signal to new storage media.²

Migration is not just a technical problem. Storage media for digital documents are not yet stable for long-term preservation. Magnetic storage media such as tape and disks eventually dissolve. Moreover, hardware and software eventually become obsolete, hence very expensive to preserve and operate. A Web archive must *migrate* from one technical environment to another over time as generations of technology succeed one another. Yet, under today’s law such migration could be a violation of copyright law because it involves copying the signal from one medium to another.

These problems are typical of an early stage of innovation in which getting to market quickly is more important than perfecting the product. Digital information products are not designed for longevity, and even if they were, it is likely they would become obsolete quickly. As a consequence, the technologies of digital preservation are complex and expensive. The problems are understood far better than the solutions at this point, but it is already clear that a Web archive will require substantial investment in technological infrastructure and technical research and development, and commercial entities are unlikely to lead this effort.

ORGANIZATIONAL ISSUES

Both archives and libraries collect, organize, preserve, and provide access to the documentary record. But the distinguishing function of archives is *to preserve the integrity of documents for the long run*.³ Preservation for centuries invariably requires new technologies, hence the Council on Library and Information Resources and other organizations are investigating long-term storage and migration of data.⁴ While the technical problem of preservation is difficult, it is well understood, but the problem of access involves legal and economic issues that have not yet been adequately explored. While print archives provide a useful model, the economic and legal environments surrounding print are quite different from those surrounding digital documents (National Research Council 2000, 113–116).

² A comprehensive description of the technical issues in digital preservation is provided in Rothenberg 1999. Migration is discussed on page 13, emulation on pages 17–30.

³ For functional descriptions of the terms “digital library” and “digital archive,” see Task Force on Archiving of Digital Information 1996, page 7.

⁴ The Council on Library and Information Resources has published numerous papers on digital preservation. See <http://www.clir.org>.

Economic and legal issues cannot be separated. In 1998, the Digital Millennium Copyright Act (DMCA) gave copyright owners rights to protect their works in digital formats. The DMCA implements the 1996 WIPO Copyright Treaty and WIPO Performances and Phonograms Treaty. Among the purposes of these treaties was harmonizing copyright policy around the world to encourage global commerce in digital information.

As a public policy, the DMCA was focused upon making the Internet safe for intellectual property. If digital information is easily moved from place to place on a network, such movement is copying, protected by copyright. If Internet information is easily accessed, making it difficult for a rights holder to control distribution, the DMCA encourages the development of technical protection services (such as encryption) by making it illegal to develop technologies to break them.

Historically, copyright policy has balanced information markets with public goods, such as education, the First Amendment, and libraries to provide access to information.

- The *First Sale doctrine* allows libraries to circulate copyrighted works to library patrons. In the digital realm, however, information is more often licensed than sold under copyright. With licenses, the provisions of the contract determine the uses that are allowed, which may or may not include library circulation or Fair Use. While printed works may also be sold with “shrink wrap” licenses, the print market has not accepted them as readily as have markets for digital information.
- *Fair Use* allows for copying for personal educational purposes, within limits that are designed to protect information markets from damage. Here again, if licenses govern commerce in digital information, these copyright provisions do not govern the contractual agreement reached between buyer and seller.

The Digital Dilemma makes a constructive case for extending the Fair Use doctrine in the future (National Research Council 2000, 137–139).

The rationale for the market approach, embodied in the DMCA, was twofold. New information markets are expensive to develop, and from the industry perspective public interest doctrines like First Sale and Fair Use are taxes on this investment. Second, the global scale of the Internet means that millions of copies can be made and distributed in seconds, causing economic damage that cannot be repaired. Thus, while copyright laws governing print placed emphasis upon *ex post facto* remedies such as litigation, the DMCA places emphasis upon prevention. Thus, every digital copy requires the permission of the copyright holder, perhaps even digital copies made temporarily for system management purposes. The DMCA explicitly allows archives to make digital copies of print works for the purpose of preservation.

To prevent illegal copying, the DMCA encourages the use of technical protection services (such as encryption) by making it illegal to use software to break them, and also making it illegal to develop and distribute such software. Software developers feel that this provision raises free speech issues, and perhaps property issues if it makes it illegal for the owner of a legal copy to make a backup. Congress recognized the complexity of some of these issues, empowering the Library of Congress to advise Congress whether

this provision in Section 104 prevents non-infringing uses of certain classes of copyrighted works.⁵

What is the impact of these new legal regimes upon archives? Print archives are permitted to collect copyrighted materials and copy them for preservation purposes. For example, it is legal to copy print materials from one medium to another as part of a migration strategy over time, but it may not be legal to do so with digital collections, or to reformat them (e.g., from CD-ROM to a hard disk).

Moreover, differences between the production and distribution of printed and digital works raise additional legal issues for Web archives. When something is published in the print world it is registered for copyright; thereafter the laws governing it are largely unambiguous. On the Internet it is not always clear when something has been “published.” At this point, it is not clear to most users whether placing information on the Web places it in the public domain or under copyright protection. *The Digital Dilemma* concludes that the Web is copyrighted in principle, but notes public confusion on the issue, and explores ambiguities that make it unclear whether archives have the right to make preservation copies and preserve them using migration strategies.⁶

In the print world, it has been possible to develop a copyright regime that balances the needs of markets and archives. Clearly, the Internet makes it difficult simply to transfer copyright doctrine from print to the digital environment. Yet many of the problems for the Web archive outlined above seem to be unanticipated consequences of laws intended to support the digital marketplace, thus which might in principle be resolved by negotiation. This process might begin by discussing the possible damage to the marketplace caused by long-term archives, and seeking solutions.

IMPLICATIONS FOR LONG-TERM PRESERVATION

The most urgent problem at this point is to create an organization capable of managing the process of building a Web archive, including negotiating to solve these problems. Inevitably, a Web archive will be a new kind of organization, one that responds to the problems and interests surrounding the Web. It may not be a place at all—it may be a function distributed among institutions over many locations on a global network.

The starting point for building a Web archive is to envision organizational strategies to manage this process. Two different technical and organizational strategies are emerging, one from the archival and library professions, another from the discipline of computer scientists. These strategies are not opposites, and are not mutually exclusive, but contrasting them helps to frame the strategic choices.

One library and archival strategy for digital archives is presented in *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* (1996), published by the Commission on Preservation and Access and the Research Libraries Group. In contrast, Brewster Kahle’s for-profit Alexa Internet and nonprofit Internet Archive might be used to illustrate the computer science vision of the Web archive.

⁵ In August 2001, the Copyright Office at the Library of Congress released the DMCA Section 104 Report, available at <http://www.loc.gov>.

⁶ See the more detailed discussion in National Research Council 2000, 113–119.

Two Technical Strategies

Which profession should develop digital archives—librarians or computer scientists? In other words, who owns this problem?

- One technical strategy is offered by the library community, which has developed sophisticated cataloging strategies. The MARC record is used to build print library catalogs that may be searched by users to identify the best information resources. MARC records include fields to describe every aspect of printed documents; the Dublin Core metadata project is defining a standard for cataloging digital documents.
- Computer scientists funded by the National Science Foundation's Digital Library program are developing a second model. While the Dublin Core is designed to enable searches of library catalogs of digital collections, the NSF digital library projects are developing search engines that directly parse the digital documents themselves.

Records identify the best information source described in a catalog, while search engines and data mining technologies go directly to the source itself; each has its advantages. But the point is that these technologies are optimized for two different kinds of archive. The computer science paradigm allows for archiving the entire Web as it changes over time, then uses powerful search engines to retrieve the necessary information. An archival catalog supports high-quality collections built around select themes, saving only the Web sites judged to have potential historical significance or special value, and describing these special qualities in collection records and catalogs which could be searched.⁷

This is a fundamental debate about the nature of the Web as a technical object as well. The librarian tends to look at the content of the Web page as the object to be described and preserved. The computer scientist tends to look at the Web as a technology for linking information—a system of relationships (hence the name “Web”). This implies not only a difference in scale; it is a difference in philosophy. Should Web archives include everything, or should they include only carefully selected samples? Should the end user make decisions about the quality of the Web page, or should the selector who chooses which Web pages to save?

Preservation Powers

Copyright requires that copies of a published work be deposited in the Library of Congress, and the National Archives have the legal responsibility for archiving Federal documents; in each case responsibility is clearly located in a funded institution. How do the librarian/archivist and computer science models solve this organizational problem?

Preserving Digital Information proposes that the digital archive begin with the following principles (among others):

- The copyright holder has initial responsibility for archiving digital information objects to ensure their long-term preservation.
- This responsibility can be subcontracted or otherwise voluntarily transferred to others, such as certified digital archives.
- But, if important digital objects are endangered because the owner does not accept responsibility for preservation, “certified digital archives have the right and duty to

⁷ On the issue of the quality of information, see, for example, Conway 1996.

exercise an aggressive rescue function as a fail-safe mechanism” (Task Force on Archiving of Digital Information 1996, 20). Clearly, this “rescue function” would require a revision of the Copyright Act to create such a right and duty.

Alternatively, the task force suggests the creation of a system of legal deposit, on the model put forth by a European Union proposal to require publishers to place a copy of their published digital works in a certified digital archive. The word “certified” in each proposal is important, for it refers to a professional and legal code of conduct so that access to the archive would not be misused.

The strength of this proposal is that it creates clear institutional responsibility for the Web archive (“certified”), and describes necessary legislation to extend proven print models (such as deposit) to the digital realm. However the “rescue” proposal has not gathered political support, and the model relies upon already scarce library subsidies for economic support.

Alternatively, consider the model of Alexa Internet and the Internet Archive. Alexa Internet is a for-profit corporation that measures the quality of Web pages by tracing consumers’ use of the Web. These measurements are made using an enormous Web archive, built by Alexa Internet using Web spiders (robots or agents) that roam the Web copying everything they find, unless forbidden entry. In this model, commercial use provides a viable economic base for the creation of the Web archive; note that Yahoo! and Google and other search engine companies have also built large Web archives for commercial purposes. Alexa Internet then turns over the Web archive to the nonprofit Internet Archive, which is to provide for long-term preservation of the digital archive.

This linkage between corporate archives and nonprofit philanthropic archives is not unprecedented; many print archives have been built through philanthropic gifts from corporations or their owners after the economic value of the collection has faded. It relies upon the philanthropic vision of individuals, which may seem unreliable, but may be more realistic than the legal establishment of a last-resort rescue power. However, it is problematic in that its funding depends upon the sustainability of a dot.com business model. And, it is not clear that it is legal for a Web crawler to copy the Web without permission; Alexa Internet proactively copies, but removes Web pages from the archive upon request of the creator or copyright holder—an “opt out” strategy.

The models developed by librarians and computer scientists are not opposites, in fact they overlap in significant ways. Each relies upon a partnership between the for-profit and non-profit realms, for in practice the digital archive is much more likely to rely upon the voluntary transfer of preservation responsibility from the copyright holder to certified archives than a controversial rescue power. Alexa Internet is an example of a philanthropic transfer from a commercial entity to an archive. Each model ultimately relies upon the resolution of legal ambiguities concerning the right to copy the Web. And to some extent, each uses an element of eminent domain over copyright, the digital archive in its rescue power and Alexa Internet in its “opt out” philosophy.

Access and Market Failure

Preservation does not threaten markets, access does. How can the Web archive protect markets from the potential damage of competition from illegal copies preserved by the nonprofit sector? Four current practices might help to provide a solution to this problem.

- *Delay.* The archive can delay making the archive available to the public until the economic value of the copy has been extracted. For this reason, Alexa Internet holds the tapes of the Web archive for six months before releasing them to Internet Archive. The length of the delay is an important subject for negotiation, since different kinds of content have different economic value cycles.
- *Opt-out.* The copyright holder can opt-out of the archive. First, the Web crawler or robot making the copy can be automatically excluded from the Web site. Second, even if copied by the crawler, the owner could ask that the copy be removed. This would allow the default to be that the Web is preserved, accomplishing the goal of the *Preserving Digital Information* Task Force, yet provide space for the owner and the archive to negotiate an agreement about the terms of access, if any.
- *Restricted access.* The archive can restrict access to the collection to those judged by the copyright holder to pose no threat, a category that might include scholars.
- *Motive.* Finally, like the Fair Use policy, it could be required that the archive user have an educational motive, and sign an agreement that the use of the archive would be restricted to certain purposes.

These ideas are not comprehensive, but are described only to suggest that current practice offers fertile ground for stakeholders to discuss.

UNRESOLVED ISSUES

Every law ultimately relies upon the perception of citizens that it is fair. Within this general cultural approval of the legitimacy, a political consensus must be built among those with significant stakes in the issues. Often this kind of consensus begins with an agreement about a fair procedure for resolving differences, such as the Conference on Fair Use (CONFU) process that attempted to build a consensus which defined the fair use policy.

The building of each kind of public consensus depends in turn upon developing a shared understanding of digital information. It is clear that Web pages have intellectual and economic value, but thus far the new kinds of value created by Web pages, and digital information generally, have not been well described.

- How do the creators of intellectual property use information? Specifically, what is the role of Fair Use in creating new information? Is copyright law the best way to govern the role of digital information in the creative process, or is the public interest best served by an emphasis upon innovation, that is, the output of the creative process?
- What value comes from distributors or publishers in a networked environment? This is clear in print, but digital commerce is still in a highly experimental state of development, making the market value of digital commodities difficult for consumers to understand.
- Consumers give value to any commodity, in a sense, by sustaining markets that ultimately justify investment in innovations, but this relationship is unexpectedly novel in the case of Web pages. For example, Web pages collect information on users and often place cookies on the Web browsers of the readers. This information has commercial value, both enabling more customized services to be provided to the consumer, and, it is hoped, building brand loyalty and justifying advertising rates on Web pages. In this sense, we might now try to understand the consumer's role in the value chain, and define how the consumer adds value to information.

Old intellectual and organizational paradigms are not easily adapted to new digital markets because they do not describe them well; thus they constrain innovation in markets that are still experimenting and evolving. Ultimately, legal and policy frameworks for the digital economy must be consistent with the citizen-consumer's own experiences if they are to be perceived as legitimate.

If the social and political framework for the Web archive is still evolving, so, too, are other key elements.

1. Evolving Technology. The Web has grown to global scale very rapidly; it may represent the fastest diffusion of a new technology in human history. But, at the same time, the technology of the Web has not stopped evolving. Even now, significant evolution is occurring as, for example, new architectures replace static Web pages with customized Web pages generated on the fly. Because innovation is not linear, the development of the Web is unpredictable. For stakeholders, the best option is to participate in the new organizations which, if they do not govern the future of the Web, at least attempt to analyze and influence its direction. To participate in discussions about the technical future of the Web it is worthwhile to follow the discussion of the World Wide Web Consortium.

2. Evolving law. Copyright law protects all of the Web. Yet the Web is global; hence, a practice that is legal in one jurisdiction may violate the law in another. For this reason, Web law needs to become harmonized, which suggests that international treaty making (like the WIPO treaty) may be as important as national legislation.

3. Evolving economic issues. The Web began as software for the exchange of documents among scientists and researchers, using an Internet that was subsidized for education and research purposes. Today the Internet is increasingly commercial, and the Web has been the subject of vigorous investment as a technology for the digital economy. The search for sustainable business models for Web business has undergone a rapid evolution, ranging from Web advertising models to banner ads, sponsorship ads, subscription models, and B2C enterprises. Investment in these enterprises and technologies has stopped for the moment because there is little sense that viable economic models have been identified.

4. Public policy. In recent years, information policy leadership has been moved from the Department of Education to the Department of Commerce, because the Internet was seen as a medium for commerce and international economic competition. At the same time, the public sector policy goal governing the Web was focused on e-government, requiring government agencies to develop Web resources and to move from print to Web publishing. Thus, at one pole the market was treated as the best way to deliver content onto the Web, and, at the other pole, the public good was defined solely in terms of online government information. There is a vacuum between these two poles, where the public interest ought to be. In between is a territory that might be called innovation policy, which is the ground upon which a Web archive, among other innovations, might be created.

REFERENCES

Besser, Howard. 2000a. Digital Longevity. In *Handbook for Digital Projects: A Management Tool for Preservation and Access*, edited by Maxine Sitts. Andover, Mass.: Northeast Document Conservation Center, pages 155-166.

Conway, Paul. 1996. *Preservation in the Digital World*. Washington, D.C.: Commission on Preservation and Access.

Lyman, Peter, and Hal Varian. 2000. How Much Information? Available at <http://www.sims.berkeley.edu/research/projects/how-much-info/>.

Lyman, Peter, and Howard Besser. 1998. Defining the Problem of Our Vanishing Memory: Background, Current Status, Models for Resolution. In *Time and Bits: Managing Digital Continuity*, edited by Margaret MacLean and Ben H. Davis. Los Angeles: Getty Information Institute and Getty Conservation Institute, pages 11-20.

National Research Council. 2000. *The Digital Dilemma: Intellectual Property in the Information Age*. Washington D.C.: National Academy Press.

Rothenberg, Jeff. 1999. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*, Washington, D.C.: Council on Library and Information Resources. Available from <http://www.clir.org/pubs/abstract/pub77.html>.

Sanders, Terry. 1997. Into the Future: Preservation of Information in the Electronic Age. Santa Monica: American Film Foundation (16 mm film, 60 minutes).

Task Force on Archiving of Digital Information. 1996. *Preserving Digital Information*. Washington, D.C.: Commission on Preservation and Access and Research Libraries Group. Available from <http://www.rlg.org/ArchTF/tfadi.index.htm>.

Web sites:

Alexa Internet: <http://www.alexa.com>

Dublin Core: <http://dublincore.org>

The Internet Archive: <http://www.archive.org>

World Wide Web Consortium: <http://www.w3c.org>
