The Impact of Electronic Publishing on the Academic Community

Session 5: Digital libraries and archiving of electronic information

Choices in digital archiving: the American experience

Donald J. Waters

Director, Digital Library Federation, 205 Church Street, Third Floor, New Haven, CT, U.S.A. dwaters@clir.org

---

The rearranging effect

The use of technologies, including the use of digital technologies such as those we have been discussing in this volume, are subject to what historian Edward Tenner calls a rearranging effect. Malcolm Gladwell called attention to the phenomenon in a recent commentary in the *New Yorker* [1]. He observed that "anyone standing on a New York city subway platform on a hot summer day" experiences a rearranging effect.

> *Subway platforms seem as if they ought to be cool places, since they are underground and are shielded from the sun. Actually, they're anything but. Come summer, they can be as much as ten degrees hotter than the street above, in part because the air-conditioners inside subway cars pump out so much hot air that they turn the rest of the subway system into an oven. In other words, we need air-conditioners on subway cars because air-conditioners on subway cars have made stations so hot that subway cars need to be air-conditioned. It's a bit like the definition the Viennese writer Karl Kraus famously gave of psychoanalysis: "the disease of which it purports to be the cure".*
>
> *Not all technological advances result in this kind of problem, of course. But it happens often enough so that when someone comes along making spectacular claims in behalf of a new technology...it's worth asking whether that technology really solves the problem or simply rearranges the hot air from the car to the platform [1].*

And so it is with digital technologies. In this chapter, I ask you to think about the digital environments we are creating and question to what extent we are revolutionizing scholarly communication and to what extent we are experiencing merely a rearranging effect --- a movement of hot air from car to platform --- in which the balance of scholarly communication is shifting from enduring to immediate access, and in which future members of the academy will have to pay dearly for the loss today of the enduring nature of the scholarly record. What choices do we have to influence the balance between immediate and enduring access?

To identify these choices, I suggest that we organize this discussion around the following topics: the limits of digital technology, the nature of information integrity and how to preserve it, the organizational aspects of digital archives and the steps needed to create an infrastructure that facilitates digital archiving.

The limits of digital technology

Rapid changes in the means of recording information, in formats for storage, in operating systems and in application technologies threaten to make the life of information in the digital age much like life in Hobbes' state of nature: "nasty, brutish, and short". At the end of 1994, the Commission on Preservation and Access and the Research Libraries Group (RLG) created the Task Force on Archiving of Digital Information to help relieve building anxiety about the fragility of culturally significant digital information. The Commission and RLG asked the Task Force to frame digital archiving as a set of problems and tasks and to suggest an orderly, perhaps even manageable, approach to their resolution.

The Commission and RLG selected members with a breadth of experience from a broad range of disciplines and backgrounds, including many from the research library community. I am sure that it was an accident, but as if to emphasize the strangeness of the new land they were asking the group to chart, the Commission and RLG selected two co-chairs --- John Garrett and I --- who are both anthropologists by training. In addition to research librarians and anthropologists, the Task Force included archivists, publishers, technologists, bibliographic service vendors, and legal and copyright specialists. The Task Force sponsors then asked the group to seek input from a still wider array of specialists and interested parties by issuing a draft report, distributing it widely and inviting comment before composing a final report.

The Task Force incorporated what we learned from the comment period into our final report. We corrected the most flagrant errors and infelicities contained in the draft report and, in revisions and an extensive set of annotations, addressed most of the questions and additional issues that arose during the comment period. We completed our work and submitted our final report on 1 May 1996.

One of the key distinctions that we introduced in the Task Force report was the distinction between digital libraries and digital archives. I notice in its discourse that the Academia Europaea workshop has not clearly articulated the difference between these concepts. In the Task Force, we introduced a simple functional distinction for the purposes of argument: digital libraries collect, store and provide access to culturally significant knowledge in digital form, while digital archives ensure the integrity and long-term accessibility of culturally significant

knowledge in digital form. Many functions --- selection, storage and access --- overlap in these two forms of organization, but digital libraries may or may be digital archives. Moreover, as the Task Force observed, libraries, museums and archives as they are presently structured may or may not be able to perform as either digital libraries or digital archives.

The problem before the Task Force was how to stimulate the creation of digital archives in the face of the limits of digital technologies, which make digital information so fragile. The line of reasoning that the Task Force pursued included the following arguments. First, new knowledge comes from confronting the old; preserved or archived work from past generations of scholars is a necessary foundation for present and future work (see also [5]). It follows, second, that the emerging knowledge economy thus cannot survive without provision for the archiving function. However, third, there intrudes what Donald Norman [4] calls the psychopathology of everyday things: despite its many advantages, digital information remains difficult and costly to use because its present design makes it difficult and costly to maintain use, especially over the long term. So, fourth, we need to consider a different approach in which we design and create a technical, legal and organizational infrastructure to afford the long-term preservation of digital information.

Of what does such an infrastructure consist? The Task Force [7] identified a wide range of factors the interaction of which provides fertile ground for the development of an archiving infrastructure. The factors include the various kinds of digital information objects --- text, images, numeric data, sound, video, simulations, geographic information systems, hypermedia and so on --- and the various claims of stakeholders with interests in the creation, management, dissemination, use and retention of digital information. Perhaps the most significant factors are those affecting the integrity of information objects in whatever form they may appear, and those required specifically for the organization of archives.

Preserving the integrity of digital information

The central goal of preservation must be to preserve the integrity of the object. Knowing how to preserve a digital information object depends on being able to define and preserve the features that give it a distinct identity and define it as a whole and singular work. In the digital environment, the features that determine information integrity and which deserve special attention for archival purposes include the following: content, fixity, reference, provenance and context. Choices about each of these features significantly affect the economy of archiving.

Choices about preserving the content of digital information objects range over a continuum of abstraction. At the lowest level of abstraction, preserving content simply means preserving a collection of bits. An archival choice at this level often means preserving the hardware and software that may be uniquely capable of interpreting the bits associated with a particular information object. Preserving content may also refer to preserving the composition of ideas in a particular structure and form. Encoding characters in ASCII or UNICODE provides varying ability to represent multiple languages, and formulae and equations. Markup languages, such as TEX, SGML (standard generalized markup language) and HTML (hypertext markup language), offer both advantages and disadvantages in representing lad and document structure compared to the use of proprietary word-processing systems and interchange formats. In the realm of digital

images, consideration of resolution, colour and compression often pits the quality of content representation against storage efficiency and loss of content. Finally, preserving content may refer, at the highest level of abstraction, to preserving the knowledge and ideas embodied in an object in a way that transcends the limits of the hardware and software needed to read bits or to render the information for use in a specific format or structural representation.

Preserving the fixity of information objects is especially troublesome in the digital world, where objects are frequently subject to change or withdrawal. Outside the digital arena, there are various methods of fixing information in objects: business records contain evidence of transactions, the acts of production and broadcast record specific radio and television programs, and publishers generate specific versions or editions of works. In the digital environment, however, the use of cryptography and other techniques is still maturing to support digital archives in establishing trusted channels of distribution, and to help them discriminate among multiple versions and to identify canonical versions. Moreover, some digital information objects are better modelled as continuously updated databases for which the preservation choice is whether to compile a complete record of changes or to capture snapshots of the database as the means of preserving information integrity.

Systems of citation, description and classification provide the necessary means of reference for consistent discovery, identification and retrieval of information objects over time. Preserving reference is thus an essential means of preserving the integrity of digital information, but it is problematic for several reasons. Self-referential information in digital objects seldom meets conventional citation quality. Moreover, consistently resolving names and locations of digital objects is, given the current state of the art, either difficult or unreliable. Finally, conventional reference mechanisms, such as online catalogues, do not easily accommodate certain kinds of reference data, such as information about the terms and conditions of licenses for intellectual property, which increasingly govern the use and cost of culturally significant information objects in the digital world.

Provenance is another essential feature of information integrity, and refers to the origin and chain of custody through individuals, organizations and instrumentation, including within the archive itself. By documenting provenance, archives create the presumption that an information object is authentic. Compared to conventionally published objects, which employ well-known techniques for establishing their origin that are usually shown on a title page or its verso, the means of establishing the provenance of information published digitally are not yet well established. In addition, there are special problems in the digital world, as in other arenas, for establishing the provenance and authenticity of individual records, such as mail, diaries and personal databases, and of corporate records, the understanding of which depends fundamentally on an appreciation of their origins in policies, procedures, and organizational roles and responsibilities. Of special note are the integrity problems associated with digital information objects produced by digital instrumentation in scientific experiments, clinical services and remote sensing. Establishing provenance of these objects --- and thus their integrity --- requires a detailed understanding of the calibration, units of measure, sampling rate, recording conditions and other features of the instrumentation that generated the information (see [2,3]).

The fifth attribute of information integrity that bears on the preservation of digital information objects is their context, the ways in which they interact with elements in the wider digital world. Among the various dimensions of interaction, there is a technical dimension, in which digital objects depend for their existence on specific hardware and software. There is also a dimension of linkages to other objects. In the World Wide Web, the integrity of many objects resides in the network of linkages. To preserve both the objects and the linkages is a daunting challenge for which there exists no good solution today other than to take periodic snapshots of the network objects. A communications dimension of information context defines the effects of the medium of transmission, such as CD-ROM or networks of varying bandwidth, on the types and characteristics of digital information objects. Finally, a social dimension, in which government policies, role relationships, and other political and organizational factors shape the creation and use of digital objects, also affects information integrity and the ability of archives economically to preserve it.

The organization of digital archives

Another set of factors that the Task Force on Archiving of Digital Information identified as grounds upon which to develop an affordable infrastructure for digital archiving are the features required specifically for the organization of archives. The digital environment today is so fragile that those who disseminate, use, re-use, re-create and re-disseminate various kinds of digital information can easily, even inadvertently, destroy valuable information, corrupt the cultural record and ultimately thwart the common pursuit of knowledge. Digital archives build and maintain reliable collections of well-defined digital information objects and they preserve the features --- content, fixity, reference, provenance and context --- that give those objects their integrity and enduring value. They do so by managing costs and finances within an operating environment that has a core set of features, including the means of migrating digital information to maintain its vitality as hardware and software environments change.

Among the core set of features in the operating environment of digital archives is a selection and appraisal process. Archives cannot save everything. To identify the most valuable objects for preservation, archives must appraise the content of the object --- its subject and discipline --- in relation to the collection goals of the digital archives, the quality and uniqueness of the object, its accessibility in terms of available hardware, software and legal status, its present value and its likely future value. Once an object is selected for inclusion, it needs to be accessioned --- that is, prepared for the archives. Accessioning involves describing and cataloguing selected objects, including their provenance to authenticate them, and securing them for storage and access. Storage, depending on expected use and the kind of performance needed in retrieval, may be online in magnetic media, near-line in optical or tape media in a jukebox retrieval system, or off-line in media that require manual intervention to retrieve. Access systems must facilitate discovery, retrieval and use, including the management of intellectual property rights as appropriate, in a distributed, presumably networked, environment. Finally, digital archives need a high level of systems engineering skill to manage the interlocking requirements of media, data formats, hardware and software, and to help determine when objects should migrate to new systems or system components.

Migration is the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. As the Task Force defined it, "the purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display and otherwise use them in the face of constantly changing technology" [7]. Digital archives have various migration strategies available to them. Internally, they can build hardware or software emulators to preserve the technical operating environments of the information objects, they can change, or 'refresh', the media on which the objects are stored as storage technology evolves, or they can reformat the objects to accommodate changing technology. In addition, they can work externally with creators so that digital information incorporates standards that simplify the migration issues. They can work with systems designers to engineer cost-effective migration paths into the hardware and software on which information objects depend. Finally, they can use processing centres that develop best practices and achieve economies of scale in certain kinds of migration techniques.

In this complex mix of factors by which digital archives operate to preserve the integrity of digital information objects, there is much room for the play of specialization, division of labour and competition that will not raise archiving costs, but drive the knowledge economy vigorously to lower them and thereby to encourage the expansion of the archiving infrastructure. Division of labour and specialization are already evident. For example, some key services, such as rights management and network charging facilities, are emerging generally in the commercial marketplace and will undoubtedly serve the interests of archiving as well as other segments of the knowledge economy. The development of other services, such as durable naming conventions and expanded metadata facilities, are well under way. Still other kinds of specialized archival services --- those, for example, that require the complex weaving of information holdings in particular disciplines from among a variety of providers and custodians --- will require time and a commitment to a complex iteration and re-iteration of exploration, development and solution as the relevant issues emerge and become clearer and more tractable.

The mechanics of digital archiving

There is an apocryphal story about the government service agency that formulated its record-retention rules as follows: (i) discard all records when they become 30 years old; and (ii) retain all records over 50 years old for their historical value [3]. The Task Force designed most of its recommendations explicitly to avoid the paralysis of this kind of thinking about the emerging knowledge economy.

The primary objective of the Task Force recommendations was to build the infrastructure necessary to make the long-term preservation of culturally significant digital objects easy and natural. The effort involves multiple, interrelated factors, only some of which are technical. In particular, the Task Force argued that the first line of defence rests with creators, providers and owners of digital information who recognize their personal interest in preservation. The Task Force also argued for the development of trusted organizations devoted to digital archiving. A climate of trust implies the existence of a fail-safe mechanism; that is, the legal right and duty of digital archives to rescue information objects that are in danger of destruction, neglect or abandonment. Trust also depends on the actual emergence of working, digital archives. Recent evidence from the United States is that archives capable of storing, providing access to and

migrating digital collections are emerging from key pilot projects, from work on essential support structures and from the development of best practices.

Among the pilot project initiatives, I want to highlight, first, Brewster Kahle's Internet Archive (see http://www.archive.org). The Archive is currently demonstrating the costs and feasibility of taking periodic snapshots of World Wide Web sites and providing long-term archival storage of the snapshots. Access, intellectual property and privacy issues, however, are highly problematic and need significant attention in subsequent phases of Kahle's venture. Second, I bring to your attention several archival projects that OCLC (Online Computer Library Center) has undertaken. It has taken responsibility as the fail-safe backup of the JSTOR database of digitized journals (see http://www.oclc.org/oclc/press/970214c.htm). OCLC has also launched the Electronic Archiving Pilot Project, a testbed for "the study of usage patterns and issues related to electronic archiving" (see http://www.oclc.org/oclc/press/970127a.htm).

In addition to these pilot projects, work on various support structures is helping to create an archiving infrastructure. For example, government funding agencies in the United States are promising to build appropriate incentives for archiving by including it among funded applications in the development of the Next Generation Internet (NGI). In March, there was a significant advance in the development of support structures organized around a discipline-based interest in digital information. Parties with interests in agricultural information met to develop a national preservation plan for digital publications of the United States Department of Agriculture. I also note, following the lead of my colleague Ann Okerson, that legal support structures for preservation are emerging not only in the national copyright arena, but also in the contractual arena as university libraries and other parties raise the issue of long-term rights and duties with respect to intellectual property in their negotiations of content licenses (see Okerson's chapter in this volume; Session 4, Chapter 2).

Finally, I draw your attention to the ways that an archiving infrastructure is taking shape in the development of best practices. The Social Science Library at Yale University has a project underway to link the migration of online data files with the digitization of related code books (see http://statlab.stat.yale.edu/SSDA/cpa.html). The RLG has created several working groups to advance standards for digitization, for recording metadata for digital objects and for preserving digital media (see http://www.rlg.org/preserv/index.html). Finally, I want to highlight an extraordinary conference of technical, archival and legal experts held in San Francisco in February 1997 on "Documenting the Digital Age" (see http://dtda/mci/com). A particularly notable feature of this conference is that it was sponsored not only by the National Science Foundation and the History Associates, a firm of consulting historians, but also by the telecommunications giant, MCI, and Microsoft. Indeed, Nathan Myhrvold, Microsoft's chief technology officer, attended the conference and provided its keynote theme in a widely circulated memorandum in which he asked: "who will save the Net?"

Conclusion

I conclude this discussion by observing that the notions of archives and archiving today have much currency and import, even outside the context in which we have been discussing them here. Such currency is evidence, perhaps, of the democratizing effects of the knowledge

economy. In the *New York Times Magazine* at the end of 1995, William Safire devoted one of his "On Language" columns to the topic of kids' slang. He advised that "if you want to stay on the generational offensive, when your offspring use the clichéd 'gimme a break', you can top that expression of sympathetic disbelief with 'jump back' and the ever-popular riposte 'whatever'." However, he noted that some expressions, such as 'I'm outta here' or 'I'm history', are now very much dated. 'I'm history', Safire quoted a forthcoming study of slang, is "a parting phrase modeled on an underworld expression referring to death, and the phrase has both inspired and been replaced by the more trendy expression, 'I'm archives'" [6].

With regard to the future of digital information in the pursuit of knowledge, I have no doubt that the expression 'I'm archives' will apply truthfully to those of us trafficking in electronic information. The choice before us, both individually and collectively, is to decide in what sense it will apply.

References

1. Gladstone, M. (1997) Chip Thrills. *New Yorker*, 20 January 1997, 7--8

2. National Research Council (1995). *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources*. National Academy Press, Washington DC

3. National Research Council (1995). *Study on the Long-Term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers*. National Academy Press, Washington, DC

4. Norman, D. (1988) *The Psychology of Everyday Things*. Basic Books Inc., New York

5. Pelikan, J. (1992) *The Idea of the University: A Reexamination*. Yale University Press, New Haven

6. Safire, W. (1995) Kiduage. In *The New York Times Magazine*. 8 October 1995, 28--30

7. Task Force on Archiving of Digital Information (1996) *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.* Commission on Preservation and Access, Washington, DC, and The Research Libraries Group, 1 May 1996, Mountain View, CA. Also available at: WWW:http://www.rlg.org/ArchTF
