Practical Limits to the Scope of Digital Preservation

Mike Kastellec

ABSTRACT

This paper examines factors that limit the ability of institutions to digitally preserve the cultural heritage of the modern era. The author takes a wide-ranging approach to shed light on limitations to the scope of digital preservation. The author finds that technological limitations to digital preservation have been addressed but still exist, and that non-technical aspects—access, selection, law, and finances—move into the foreground as technological limitations recede. The author proposes a nested model of constraints to the scope of digital preservation and concludes that costs are digital preservation's most pervasive limitation.

INTRODUCTION

Imagine for a moment what perfect digital preservation would entail: A perfect archive would capture all the content generated by humanity instantly and continuously. It would catalog that information and make it available to users, yet it would not stifle creativity by undermining creators' right to control their creations. Most of all, it would perfectly safeguard all the information it ingested eternally, at a cost society is willing and able to sustain.

Now return to reality: digital preservation is decidedly imperfect. Today's archives fall far short of the possibilities outlined above. Much previous scholarship debates the *quality* of different digital preservation strategies; this paper looks past these arguments to shed light on limitations to the *scope* of digital preservation. What are the factors that limit the ability of libraries, archives, and museums (henceforth collectively referred to as archival institutions) to digitally preserve the cultural heritage of the modern era?[1] I first examine the degree to which technological limitations to digital preservation have been addressed. Next, I identify the non-technical factors that limit the archival of digital objects. Finally, I propose a conceptual model of limitations to digital preservation.

TECHNOLOGY

Any discussion of digital preservation naturally begins with consideration of the limits of digital preservation technology. While all aspects of digital preservation are by definition related to technology, there are two purely technical issues at the core of digital preservation: data loss and technological obsolescence.[2] Many things can cause data loss. The constant risk is physical deterioration. A digital file consists at its most basic level as binary code written to some form of

Mike Kastellec (makastel@ncsu.edu) is Libraries Fellow, North Carolina State University Libraries, Raleigh, NC.

physical media. Just like analog media (paper, vinyl recordings), digital media (optical discs, hard drives) are subject to degradation at a rate determined by the inherent properties of the medium and environment in which it is stored.[3] When the physical medium of a digital file decays to the point where one or more bits lose their definition, the file becomes partially or wholly unreadable. Other causes of data loss include software bugs, human action (e.g., accidental deletion or purposeful alteration), and environmental dangers (e.g., fire, flood, war).

Assuming a digital archive can overcome the problem of physical deterioration, it then faces the issue of technological obsolescence. Binary code is simply a string of zeroes and ones (sometimes called a *bitstream*)—like any encoded information, this code is only useful if it can be decoded into an intelligible format. This process depends on hardware, used to access a bitstream from a piece of physical media, and software, which decodes the bitstream into an intelligible object, such as a document or video displayed on a screen, a printout, or an audio output. Technological obsolescence occurs when either the hardware or software needed to render a bitstream usable is no longer available. Given the rapid pace of change in computer hardware and software, technological obsolescence is a constant concern.[4]

Most digital preservation strategies involve staying ahead of deterioration and obsolescence by copying data from older to current generations of file formats and storage media (migration) or by keeping many copies that are tested against one another to find and correct errors (data redundancy).[5] Other strategies to overcome obsolescence include pre-emptively converting data to standardized formats (normalization) or avoiding conversion and instead using virtualized hardware and software to simulate the original digital environment needed to access obsolete formats (emulation). As may be expected of a young field,[6] there is a great deal of debate over the merits of each of these strategies. To date, the arguments mostly concern the quality of preservation, which is beyond the scope of this work. What should not be contentious is that each strategy also imposes limitations on the potential scale of digital preservation. Migration and normalization are intensive processes, in the sense that they normally require some level of human interaction. Any human-mediated process limits the scale of an archival institution's preservation activities, as trained staffs are a limited and expensive resource. Emulation postpones the processing of data until it is later accessed, potentially allowing greater ingest of information. As a strategy, however, it remains at least partly theoretical and untested, increasing the possibility that future access will be limited. Data redundancy deserves closer examination, as it has emerged as the gold standard in recent years.

The limitations data redundancy imposes on digital preservation are two-fold. The first is that simple maintenance of multiple copies necessarily increases expenses, therefore—given equal levels of funding—less information can be preserved redundantly than can be preserved without such measures. (Cost considerations are inextricably linked to every other limitation on digital preservation and are examined in greater detail in "Finances," below.) There are practical, technical limitations on the bandwidth, disk access, and processing speeds needed to perform

parity checks (tests of each bit's validity) of large datasets to guard against data loss. Pushing against these limitations incurs dramatic costs, limiting the scale of digital preservation. Current technology and funding are many orders of magnitude short of what is required to archive the amount of information desired by society over the long term.[7]

The second way technology limits digital preservation is more complex—it concerns error rates of archived data. Non-redundant storage strategies are also subject to errors, of course. Only redundant systems have been proposed as a theoretical solution to the technological problem of digital preservation,[8] though, so it is necessary to examine their error rate in particular. On a theoretical level, given sufficient copies, redundant backup is all but infallible. In practice, technological limitations emerge.[9] The number of copies required to ensure perfect bit preservation is a function of the reliability of the hardware storing each copy. Multiple studies have found that hardware failure rates greatly exceed manufacturers' claims.[10] Rosenthal argues that, given the extreme time spans under consideration, storage reliability is not just unknown but untestable.[11] He therefore concludes that it cannot be known with certainty how many copies are needed to sustain acceptably low error rates. Even today's best digital preservation technologies are subject to some degree of loss and error.

Analog materials are also inevitably subject to deterioration, of course, but the promise of digital media leads many to unrealistic expectations of perfection. Nevertheless, modern digital preservation technology addresses the fundamental needs of archival institutions to a workable degree. Technological limitations to digital preservation still exist but the aspects of digital preservation beyond purely technical considerations—access, selection, law, and finances—should gain greater relative importance than they have in the past.

ACCESS

With regard to digital preservation, there are two different dimensions of access that are important. At one end of a digital preservation operation, authorized users must be able to access an archival institution's holdings and unauthorized users restricted from doing so. This is largely a question of technology and rights management—users must be *able* to access preserved information and *permitted* to do so. This dimension of access is addressed in the Technology and Law sections of this paper. The other dimension of access occurs at the other end of a digital preservation operation: An archival institution must be able to access a digital object to preserve it. This simple fact leads to serious restrictions on the scope of digital preservation because much of the world's digital information is inaccessible for the purposes of archiving by libraries and archives.

There are a number of reasons why a given digital object may be inaccessible. Large-scale harvesting of webpages requires automated programs that "crawl" the Web, discovering and capturing pages as they go. Web crawlers cannot access password-protected sites (e.g., Facebook) and database-backed sites (all manner of sites, including many blogs, news sites, e-commerce sites,

and countless collections of data). This inaccessible portion of the Web is estimated to dwarf the readily accessible portion by orders of magnitude. There is also an enormous amount of inaccessible digital information that is not part of the Web at all, such as emails, company intranets, and digital objects created and stored by individuals.[12]

Additionally, there is a temporal limit to access. Some digital objects only are accessible (or even exist) for a short window of time, and all require some measure of active preservation to avoid permanent loss.[13] The lifespans of many webpages are vanishingly short. Other pages, like some news items, are publicly accessible for a short window before they are hidden behind paywalls. Even long-lasting digital objects are often dynamic: the ads accompanying a webpage may change with each visit; news articles and other documents are revised; blog posts and comments are deleted. If an archival institution cannot access a digital object quickly or frequently enough, the object cannot be archived, at least not completely. Large-scale digital preservation, which in practice necessarily relies on periodic automated harvesting of content, is therefore limited to capturing snapshots of the changes digital objects undergo over their lifespans.

LAW

Existing copyright law does not translate well to the digital realm. Leaving aside the complexities of international copyright law, in the United States it is not clear, for example, whether an archival institution like the Library of Congress is bound by licensing restrictions and if it can require deposit of digital objects, nor whether content on the Web or in databases should be treated as published or unpublished.[14] "Many of the uncertainties come from applying laws to technologies and methods of distribution they were not designed to address."[15] A lack of revised laws or even relevant court decisions significantly impacts the potential scale of digital preservation, as few archival institutions will venture to preserve digital objects without legal protection for doing so.

Given this unclear legal environment, efforts at large-scale digital preservation are hampered by the need to secure permission to archive from the rights holder of each piece of content.[16] This obviously has enormous impact on preserving the Web, but even scholarly databases and periodical archives may not hold full rights to all of their published content. Additionally, a single digital object can include content owned by any number of authors, each of whose permission is needed for legal archival.

Without stronger legal protection for archival institutions, the scope of digital preservation is severely limited by copyright restrictions. Digital preservation is further limited by licensing agreements, which can be even more restrictive than general copyright law. Frequently, purchase of a digital object does not transfer ownership to the end-user, but rather grants limited licensed access to the object. In this case, libraries do not enjoy the customary right of first sale that, among other things, allows for actions related to preservation that would otherwise breach copyright.[17] Preservation of licensed works requires that libraries either cede archival responsibility to rights

holders, negotiate the right to archive licensed copies, or create dark archives that preserve objects in an inaccessible state until their copyright expires.

SELECTION

The limitation selection imposes on digital preservation hinges on the act of intellectual appraisal. The total digital content created each year already outstrips the total current storage capacity of the world by a wide margin.[18] It is clear libraries and archives cannot preserve everything so, more than ever, deciding what to preserve is critical.[19]

Models of selection for digital objects can be plotted on a scale according to the degree of human mediation they entail. At one end, the *selective* model is closest to selection in the analog world, with librarians individually identifying digital objects worthy of digital preservation. At the other end of the scale, the *whole domain* model involves minimal human-mediation, with automated harvesting of digital objects. The *collaborative* model, in which archival institutions negotiate agreements with publishers to deposit content, falls somewhere between these two extremes, as does the *thematic* model, which can apply either selective- or whole-domain-type approaches to relatively narrow sets of digital objects defined by event, topic, or community.

Each of these approaches results in limits to the scope of digital preservation. The human mediation of the selective model limits the scale of what can be preserved, as objects can only be acquired as quickly as staff can appraise them. The collaborative and thematic models offer the potential for thorough coverage of their target but by definition are limited in scope. The whole domain model avoids the bottleneck of human appraisal but, more than any other model, is subject to the access limitations discussed above. Whole domain harvesting is also essentially wasteful, as it is an anti-selection approach—everything found is kept, irrespective of potential value. This wastefulness makes the whole domain model extremely expensive because of the technological resources required to manage information at such a scale.

FINANCES

The ultimate limiting factor is financial reality. Considerations of funding and cost have both broad and narrow effects. The narrow effects are on each of the other limitations previously identified—financial constraints are intertwined with the constraints imposed by technology, access, law, and selection. The technological model of digital preservation that offers the highest quality and lowest risk, redundant offsite copies, also carries hard-to-sustain costs. While the cost of storage continues to drop, hardware costs actually make up only a small percentage of the total cost of digital preservation. Power, cooling, and—for offsite copy strategies—bandwidth costs are significant and do not decrease as scale increases to the same degree that storage costs do. Cost considerations similarly fuel non-technical limitations: Increased funding can increase the rate at which digital objects are accessed for preservation and can enable development of systems to mine deep Web resources. Selection is limited by the number of staff who can evaluate objects or

the need to develop systems to automate appraisal. Negotiating perpetual access to objects or arranging to purchase archival copies creates additional costs.

The broad financial effect is that any digital preservation requires dedicated funding over an indefinite timespan. Lavoie outlines the problem:

> Much of the discussion in the digital preservation community focuses on the problem of ensuring that digital materials survive for future generations. In comparison, however, there has been relatively little discussion of how we can ensure that digital preservation activities survive beyond the current availability of soft-money funding; or the transition from a project's first-generation management to the second; or even how they might be supplied with sufficient resources to get underway at all.[20]

There are many possible funding models for digital preservation,[21] each with their own limitations. Creators and rights holders can preserve their own content but normally have little incentive to do so over the long-term, as demand for access slackens. Publicly funded agencies can preserve content, but they may lack a clear mandate for doing so, and they are chronically underfunded. Preservation may be voluntarily funded, as is the case for Wikipedia, although it is not clear if there is enough potential volunteer funding for more than a few preservation efforts. Fees may support preservation, either through charging users for access or by third-party organizations charging content owners for archival services; in such cases, however, fees may also discourage access or provision of content, respectively.

A Nested Model of Limitations

These aspects can be seen as a series of nested constraints (see figure 1).

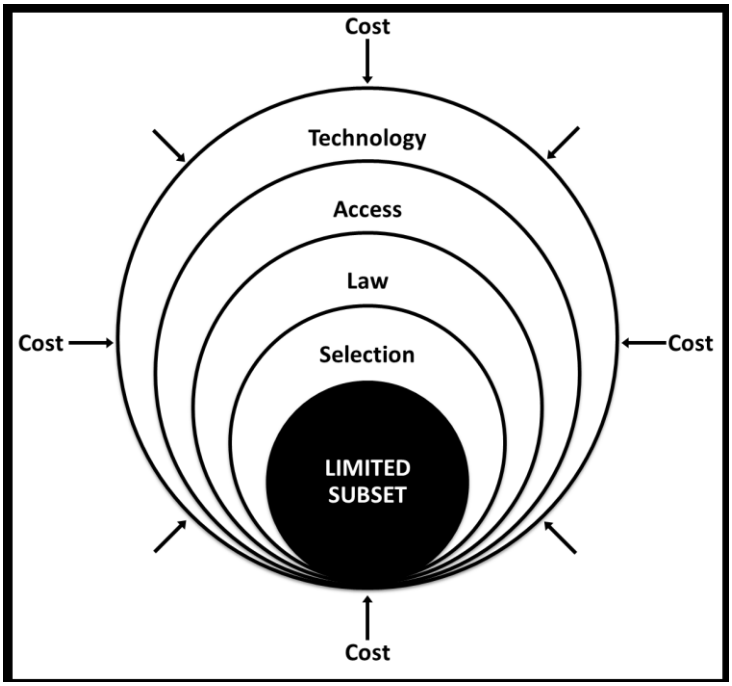

Figure 1. Nested Model of Limitations

At the highest level, there are technical limitations on how much digital information can be preserved at an acceptable quality. Within that constraint, only a limited portion of what could possibly be preserved can be accessed by archival institutions for digital preservation. Next, within that which is accessible, there are legal limitations on what may be archived. The subset defined by technological, access, and legal limitations still holds far more information than archival institutions are capable of archiving, therefore selection is required, entailing either the limited quality of automated gathering or the limited quantity of human-mediated appraisal. Finally, each of these constraints is in turn limited by financial considerations, so finances exert pressure at each level.

CONCLUSION

It is possible to envision alternative ways to model these series of constraints—the order could be different, or they could all be centered on a single point but not nested within each other. Thus, undue attention should not be given to the specific sequence outlined above. One important conclusion that may be drawn, however, is that the identified limitations are related but distinct. The preponderance of digital preservation research to date has understandably focused on overcoming technological limitations. With the establishment of the redundant backup model, which addresses technological limitations to a workable degree, the field would be well served by greater efforts to push back the non-technical limitations of access, law, and selection. The other conclusion is that costs are digital preservation's most pervasive limitation. As Rosenthal plainly states it, "Society's ever-increasing demands for vast amounts of data to be kept for the future are

not matched by suitably lavish funds." [22] If funding cannot be increased, expectations must be tempered.

Perhaps it has always been the case, but the scale of the digital landscape makes it clear that preservation is a process of triage. For the foreseeable future, the amount of digital information that could possibly be preserved far outstrips the amount that feasibly can be preserved. It is useful to put the advances in digital preservation technology in perspective and to recognize that non-technical factors also play a large role in determining how much of our cultural heritage may be preserved for the benefit of future generations.

REFERENCES AND NOTES

1. Issues specific to digitized objects (i.e., digital versions of analog originals) are not specifically addressed herein. Technological limitations apply equally to digitized and born-digital objects, however, and the remaining limitations overlap greatly in either case.

2. Francine Berman et al., *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information* (Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010), http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf (accessed Apr. 23, 2011).

3. Marilyn Deegan and Simon Tanner, "Some Key Issues in Digital Preservation," in *Digital Convergence—Libraries of the Future*, ed. Rae Earnshaw and John Vince, 219–37 (London: Springer London, 2007), www.springerlink.com.proxy-remote.galib.uga.edu/content/h12631/#section=339742&page=1 (accessed Nov. 18, 2010).

4. Berman et al., *Sustainable Economics for a Digital Planet*; Deegan and Tanner, "Digital Convergence."

5. Data redundancy normally will also entail hardware migration; it may or may not also incorporate file format migration.

6. The Library of Congress, for instance, only began digital preservation in 2000 (www.digitalpreservation.gov/partners/pioneers/index.html [accessed Apr. 24, 2011]).

7. David S. H. Rosenthal, "Bit Preservation: A Solved Problem?" *International Journal of Digital Curation* 5, no. 1 (July 21, 2010), www.ijdc.net/index.php/ijdc/article/view/151 (accessed Mar. 14, 2011).

8. H. M. Gladney, "Durable Digital Objects Rather Than Digital Preservation," January 1, 2008, http://eprints.erpanet.org/149 (accessed Mar. 14, 2011).

9. Rosenthal, "Bit Preservation."

10. Ibid. Rosenthal cites studies by Schroeder and Gibson (2007) and Pinheiro (2007).

11. Ibid.

12. Peter Lyman, "Archiving the World Wide Web," in *Building a National Strategy for Digital Preservation: Issues in Digital Media Archivin*g (Washington, DC: Council on Library and Information Resources and Library of Congress, 2002), 38–51, www.clir.org/pubs/reports/pub106/pub106.pdf (accessed Dec. 1, 2010); F. McCown, C. C Marshall, and M. L Nelson, "Why Web Sites are Lost (and how they're sometimes found)," *Communications of the AC*M 52, no. 11 (2009): 141–45; Margaret E. Phillips, "What Should We Preserve? The Question for Heritage Libraries in a Digital World," *Library Trend*s 54, no. 1 (Summer 2005): 57–71.

13. Deegan and Tanner, "Digital Convergence"; McCown, Marshall, and Nelson, "Why Web Sites are Lost (and how they're sometimes found)."

14. June Besek, *Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessmen*t (The Council on Library and Information Resources and the Library of Congress, 2003), www.clir.org/pubs/reports/pub112/contents.html (accessed Mar. 15, 2011).

15. Ibid., 17.

16. Archival institutions that do not pay heed to this restriction, such as the Internet Archive (www.archive.org), claim their actions constitute fair use. The legality of this claim is as yet untested.

17. Berman et al., *Sustainable Economics for a Digital Planet*.

18. Francine Berman, "Got Data?" *Communications of the ACM* 51, no. 12 (December 2008): 50, http://portal.acm.org/citation.cfm?id=1409360.1409376&coll=portal&dl=ACM&idx=J79&part=magazine&WantType=Magazines&title=Communications (accessed Nov. 20, 2010).

19. Phillips, "What Should We Preserve?"

20. Brian F. Lavoie, "The Fifth Blackbird," *D-Lib Magazin*e 14, no. 3/4 (March 2008): I, www.dlib.org/dlib/march08/lavoie/03lavoie.html (accessed Mar. 14, 2011).

21. Berman et al., *Sustainable Economics for a Digital Planet*.

22. Rosenthal, "Bit Preservation."