Information in Digital Libraries: Document Preservation and Access

Rebecca L. Fitzsimmons

University of South Florida

Abstract

Borgman (2000) raises questions about the longevity of digital documents noting that, "digital preservation looms as one of the greatest challenges of information technology management and policy" (p. 66). Digital libraries have the capacity to transcend geographical borders and deliver highly accessible content to users, but they face daunting challenges from copyright, authenticity, and format issues related to digital documents. In order to ensure access to the wealth of information contained within their collections, information professionals must address serious questions related to the long-term preservation of digital resources. The alternative is to lose large quantities of information that form the base of digital libraries. The following paper will examine issues of access related to digital preservation.

Information in Digital Libraries: Document Preservation and Access

Digital libraries (DL) that contain full-text digital documents have the unique ability to transcend physical borders in ways that traditional libraries cannot. Users can both search collections and immediately access materials that are available for on-screen use or download without regard to geographic location; in many cases this eliminates the need for waiting on interlibrary loan services or traveling to a particular library to view a physical document. Obviously, the convenience afforded by this service has the potential to markedly increase access to collections of information, a core goal of every library. As the construction of DLs has grown in recent years, many users have also come to expect this type of access to a variety of materials and may take for granted that this model is the new norm, putting pressure on libraries to digitize existing collections and subsequently manage huge volumes of digital information. This raises a host of concerns related not only to providing immediate access to content, but also to the long-term management and sustainability of digital documents.

Borgman (2000) raises questions about the longevity of digital documents noting that, "digital preservation looms as one of the greatest challenges of information technology management and policy" (p. 66). This is perhaps one of the most central issues regarding access to information, especially when information is increasingly being created in a digital format (known as born-digital). Loss or corruption of files that have no analog originals threatens the record of human knowledge and therefore becomes a pressing concern for DLs. The remainder of this paper will address issues DLs are facing with regards to dealing with timely access to and preservation of digital materials, including concerns with copyright, file formats, sustainability, digitizing, authenticity, and provenance. The long-term success of DLs is dependent on continued and efficient user access to all possible sources of digital information and to the preservation of digital materials, all of which hinges on resolving many of the concerns listed above.

Preservation

Preservation is a difficult concept when applied to digital materials, yet for DLs it is always looming due mainly to the fact that foregoing preservation will eventually equate to the inability of users to access information. Complicating this issue is that preservation goals can be difficult to define and execute; a Council on Library and Information Resources report (2000), titled *Authenticity in a Digital Environment,* demonstrates this point through the disagreement among the authors about what constitutes authenticity and subsequently how preservation should be handled in relation this concept. Cullen (2000), for instance, raises the issue of authenticity by suggesting that digital documents migrated to other formats may lose the very essence of their value; Smith (2000) questions the concept of a digital original and notes that preservation techniques relate largely to the specific definition of authenticity that is applied. Rothenberg (2000), on the other hand, points out that, "an informational entity that is 'preserved' without being usable in a meaningful and valid way has not been meaningfully preserved, i.e., has not been preserved at all" (p. 54). While serious disagreement exists over how digital materials should be treated, this latter concept of meaningful preservation is especially significant to DLs, whose primary purpose is ensuring that content is usable. Taking the stance that original digital formats are the only authentic documents has a certain academic merit, however it is also arguable that storehouses of information that appear in now-obsolete formats are not fulfilling the information needs of anyone. Furthermore, if the main concern of libraries was ensuring *only* access to originals, the widespread practices of photocopying, emailing, or faxing documents would be strictly taboo. Instead, access to the *content* of material is a primary concern with DL preservation efforts.

Digitization and Digitally-born—Issues with Document Care

Borgman (2000) suggests that while digitization and the collection of digitally-born documents allows DLs to provide better models of access, collecting materials and maintaining such access raises a number of problems. First, she notes the most obvious pitfall, which is that "digital information is not eye legible," (p. 66); this means that hardware and software for viewing this information must be kept current somehow, in addition to maintaining the integrity of the actual file. If not, serious problems occur, often on a massive scale. Dougherty (2009) relates a sense of the urgency of dealing with digital file obsolescence, noting that, "Britain's National Archive holds the equivalent of 580,000 encyclopedias of information in file formats that are no longer commercially available" (p. 600). Such information is at best highly difficult to access, requiring machinery and programs that few institutions or individuals retain and at worst completely inaccessible due to the lack of original equipment and software. Complicating this issue is the huge volume of new information being produced each year. Lyman and Varian (2003) estimate that in 2002 alone five Exabyte's of information, "equivalent in size to the information contained in half a million new libraries the size of the Library of Congress print collections," was produced and that 92% of that output was in digital format (Executive Summary, para. 2). Considering the short lifespan of many digital file formats and storage devices—sometimes as low as 1-5 years (Borgman, 2000, Rothenberg, 1995)—the task of maintaining these files while addressing the backlog of already obsolete documents is both expensive and daunting.

It is also significant to note the differences in the results of digitization efforts. Conway (2010) suggests that there is an important distinction between digital preservation and digitization for preservation—namely that digitization for preservation is the process of making hard copies of materials into digital copies that a user can search and access, while digital preservation is the process of storing and handling those files in a way that will ensure they are available in the long-term (p. 64). Like Rothenberg (2000), Conway asserts that, "preservation action should nearly always be taken in reference to use, rather than to the purely intrinsic value of an object" (2010, p. 64). In this view, the contents and accessibility rather than the form of the original are the most essential elements. Digitizing analog documents fits precisely into this category as a way to keep precious originals from being over-handled and to make the contents of those documents available to a wider audience via a DL interface; further, methods of preservation such as the migration of files to ensure currency make sense when the goal of DLs is to make sure that information can be used.

Authenticity of Digital Documents

Authenticity is a topic of increasing consideration in relation to digital documents (especially those that were born-digital) for a number of reasons, including the likelihood that researchers will need to know that digital documents are an accurate representation of the original as the author intended it to be seen (Smith, 2000; Levy, 2000; Bradley, 2005). Agreement over what such intentionality means, however, is a point of debate among librarians, archivists, and other information professionals. Lynch (2000) asserts that—especially in relation to multimedia documents—the viewing experience of users can vary widely depending on the hardware and software they use (i.e. issues of quality, color reproduction, and so forth); this "raises questions about how to define and measure authenticity and integrity" (p. 36). Certainly from this view it is arguable that a skipping, bucking video or sound recording is not what the author intended and therefore lacks both integrity (as in quality) and authenticity (as in faithful to the intended vision of the producer). Similarly, films and photographs sharpened or edited to reveal brighter, more saturated colors may be inauthentic, regardless of how such changes might improve the user's access to the information (such as revealing previously fuzzy details). Other definitions of authenticity, however, lean toward even stricter definitions.

Cullen (2000) questions whether the act of preservation implies that a document is authentic, and notes that, "more than one archivist has stated that the only sure means of preserving a digital object is to save a printed copy" (p. 3). He goes on to suggest, however, that altering digital information without a consideration for the authenticity of the original, in format as well as content, may be rendering the subsequent file useless in the long-term. This is a serious issue for DLs to consider since altering a format may help with file accessibility in the present, but ultimately render the digital document suspect. Options, then, for assuring the authenticity of a document rest largely with institutions that store and manage these objects. Bradley (2005) suggests that comparisons between documents stored in different locations can help DLs determine if a document has been altered in any way; this is important for both preventing data corruption and detecting purposeful changes. Write-only policies also prevent migrations or alterations that would otherwise go unnoted in the associated metadata of the digital document (p. 168). This metadata has, consequently, been suggested as a means of assisting in the verification of authenticity. Smith (2000) suggests that authenticity according to many different definitions can be ascribed through a series of metadata specifications about how a document should render, access requirements (hardware and software), records of authorship and/or ownership, preservation and migration histories, and version numbers (p. vii).

Similarly, Baudoin (2008) states, "the principle of digital preservation…is *to think archivally…"* (p. 558). This entails following and verifying the trail of ownership to ensure that the document is what it, or the associated metadata, states. From this view it doesn't matter what the collected form of the document is as long as any migrations have been recorded along the way. Essentially, a record of handling is more important to assure the authenticity than an attempt to contain the original; "the chain of custody, if it is documented from creation onward, provides the requisite audit trail and confers trustworthiness of the data" (Baudoin, 2008, p. 557). This trustworthiness is essential for both assuaging any lingering doubts from researchers as to the authenticity of the digital documents they access and in judging whether information, especially digitally-born, is worthy of future preservation efforts. Gladney (2009) cites the ability to read and use digital content as the creator intended and the ability to determine whether the information is trustworthy as two of the requirements of long-term digital preservation (p. 403). These requirements refer to a need for users accessing digital information to trust that it is accurate and complete.

Despite these concerns, when Bradley (2005) conducted a study of digital repositories, it was noted that, "while exceptions exist, ensuring the authenticity and integrity of digital resources seems to represent a low priority for many of the digital repositories surveyed" (p. 171). Reasons for this included a belief that researchers should verify their own sources or that the nature of the resources provided were not such that scholars would likely use the materials (172). However, this goes against the idea that digital libraries, much like physical libraries, provide trusted information sources. Users may, and arguably should, believe that the documents they are accessing have been carefully collected, preserved, assessed, and organized. Even in the face of volumes of digitally-born information, DLs must put forth substantial effort to verify documents or risk losing the trust of their users; as Bradley noted, the idea that researchers should check a digital document against the "original" undermines the concept of digital libraries without geographical barriers. This poses further problems when there is not a paper or film copy to begin with, as users are unlikely to have access to multiple stored files for the purpose of cross-checking authenticity, integrity, or completeness of a digital document (Bradley, 2005, p. 172).

Sustainability of Documents

Much attention has been paid to the concept of preserving digital material, but there is not widespread adoption of methods for doing so. This is undoubtedly due largely to differing ideas within the information science field about access and preservation needs, definitions of authenticity, and wide discrepancies in the number of staff and resources at any particular institution that can and should be poured into digital preservation. A number of strategies do exist, however, to help deal with the tide of finicky formats that characterize most digital library contents.

In terms of maintaining the authenticity of digital documents, one of the best methods for long-term storage involves digital cooperatives. One such example is the MetaArchive Cooperative, based on the Lots of Copies Keep Stuff Safe (LOCKSS) system. Originally, LOCKSS was developed at Stanford University for the purpose of journal archiving. Digital copies of documents are distributed across a network of servers housed at member institutions. Once information enters the system, programs constantly crosscheck the data against other copies to determine if any particular copy has been damaged. If the system detects a corrupted, truncated, or otherwise damaged copy it repairs it. In this way the integrity of all the copies as a primary source of accurate digital information is maintained (Howard, 2008; Horrell, 2007). The MetaArchive Cooperative uses a private LOCKSS network and as a group defines collection-level metadata policies and technical specifications; participating institutions retain autonomy over appropriate file format, arrangement of the contents, and the inclusion or exclusion of item level metadata (Howard, 2008, p. 16). This arrangement is geared only at preservation of digital documents for long-term use, however, and is not aimed at solving access issues. DLs must maintain documents on a local server in order to make them accessible to users; the archived copies are only requested in cases of data loss of the local copies. Despite addressing only preservation concerns (and by virtue access into the future) the Cooperative is a good, relatively cheap solution for DLs looking to store and protect archival copies of their documents.

Sustainability of digital documents (and by extension DLs), however, also depends on providing access to the contents in a readable format. Migration is one solution, albeit a controversial one depending on one's definition of an authentic document. Despite quibbles over whether material migrated from an obsolete word processing format to a newer version constitutes a derivative rather than authentic work, this is the method of choice for maintaining accessibility. Bradley (2007) suggests that functionality of documents is the core of sustainability and that access can be obtained by altering the files to work with newly developed software systems. Furthermore, storage media must also change to meet current hardware needs. (p. 151). A perfectly intact file is still obsolete if it resides on a floppy disk for which there are very few pieces of machinery still available to read the media. The impermanence of physical storage media through deterioration is a concern as well, but the idea of permanent digital storage media is rather unthinkable, given the pace of technological change. As Bradley suggests, manufacturers are quick to abandon support for older technologies in the face of new and better products (p.153).

Migration is a logical solution for DLs since it meets the primary requirements of providing access to digital information. If it is handled responsibly—transfers noted in the associated metadata, careful conversion to avoid losing data, etc.—it is a technique that can assure both usability and trustworthiness of the document. However, Bradley (2007) cites a 1998 assertion by Rothenberg that multiple transfers of information have the capacity to damage a digital document in several ways. Corruption is always an issue when files are transferred to new formats as the byte streams can become convoluted—partial or total data loss would be possible. Beyond technical concerns however, Bradley (2007) cites more ethical objections from Rothenberg and his peers regarding the content of a digital document. Namely, authenticity and integrity again come into play as assertions are made that alterations to the "look and feel," or a loss of "significant properties" are major concerns with document migration (p. 153). Instead, the idea of housing software capable of running old files on new technology systems (emulation) was favored, with the notion that leaving an unaltered byte stream was a better preservation strategy (p. 154). This is a logical solution, however, it is one that is likely to be cumbersome and prohibitively expensive for DLs as a long-term, widespread preservation strategy. Instead, it may be more reasonable to make such decisions based on a particular case where emulation (or even stockpiling original, outdated playback equipment) would be necessary. In fact, the example of films used earlier comes to mind in this latter instance. It might be desirable to have digitized versions of 8mm films that are loaded online for viewing through a DL interface, however this would not eclipse the need for some researchers to view the media in its original format. The original would have physical qualities—sound, color, detail, hand markings, splices—that would themselves have value. As such, it would be wise for a repository to stock an appropriate projector to meet these specific needs.

Additional Access Issues

Borgman (2000) suggests that while many physical documents have research value based on characteristics such as paper, ink, and binding material, digital surrogates are sufficient for most users and libraries are increasingly digitizing materials to improve access. She notes that the American Memory Project DL receives more visitors than the physical Library of Congress, underscoring the idea that easy (often remote) access is key when digitizing or collecting digital-born works (p. 45). As discussed, preservation and access often go together when dealing with digital materials. However, copyright is another significant factor that can hinder the ability of DLs to provide widespread access to digital materials, even when such materials are readily available within a collection. In fact, Akmon (2010) states that mass digitization as a method of meeting user needs for digital, remotely accessible content is important, but that “copyright is frequently noted [by archives] as a significant obstacle to these efforts…” (p. 45). Strategies for dealing with copyright include avoiding messy rights issues, declaring fair use, or obtaining permission. In the former strategy, the obvious issue is creating an unbalanced DL collection, with the latter two strategies carrying huge costs (real or potential) for DLs (p. 46). In the subsequent study of non-response rates when librarians and archivists attempted to gain rights permission to digitize and display documents, Akmon noted that a prevalent approach was to consider all non-responses as a negative answer in order to avoid any risk of expensive legal trouble. The question that arose, since most inquiries went unanswered, was whether appropriate resources were being devoted to this area if only a small portion of the collections were able to be displayed online (p. 63). This creates an interesting dilemma that will need to be resolved as more collections are slated for inclusion in DLs and more information becomes publicly accessible to a wide audience via the Internet.

Conclusion

If digital libraries are to play an increasingly significant role in the future of information access, information professionals must find ways to address long-term preservation problems that seriously threaten the ability of users to access information. A future where most information is stored digitally and accessed via a DL is certainly plausible, as is the notion that much of what is currently produced may be lost. Some would even suggest that a lack of digital preservation efforts will ultimately result in a kind of digital dark age where much of the record of current human history has or will be erased (Kuny, 1997; Brand, 1999; University of Illinois, 2008). While differences exist in the definitions that various practitioners assign to digital document authenticity, Gladney (2009) and Baudoin (2008) are convinced that the future of preservation is intimately tied to establishing and maintaining a record of document provenance. This view of document preservation provides the most usable approach for DLs as it accounts for access and authenticity, migration and record keeping. If trustworthiness of information is key, so to is providing widely useable document formats that can be accessed offsite without the use of special systems or equipment. Preservation will remain a formidable issue for the foreseeable future, carrying with it problems of cost, copyright, selection, long-term storage, metadata standards, and so forth. It is an issue, however, that must addressed with the utmost care and immediacy within the field. This is a major part of the future of digital library endeavors because without preservation, there will be little left in the way of information for users to access.

References

Akmon, D. (2010). Only with your permission: how rights holders respond (or don't respond) to requests to display archival materials online. *Archival Science*, *10*, 45-64. doi: DOI 10.1007/s10502-010-9116-z

Baudoin, P. (2008). The principle of digital preservation. *The Serials Librarian*, *54*(4), 556-559. doi: 10.1080/03615260802291212

Borgman, C. (2000). *From Gutenberg to the global information infrastructure: Access to information in the networked world.*. Cambridge, MA: MIT Press.

Bradley, K. (2007). Defining Digital Sustainability. *Library Trends, 56*(1), 148–163.

Bradley, R. (2005). Digital authenticity and integrity: Digital cultural heritage documents as research resources. *Portal: Libraries and the Academy, 5*(2), 165-175. *doi: 10.1353/pla.2005.0018*

Brand, S. (1999). Escaping the digital dark age. Library Journal, 124(2), 46-49. Retrieved from *www.rense.com/general38/escap.htm*

Conway, P. (2010). Preservation in the age of Google: Digitization, digital preservation, and dilemmas. *The Library Quarterly*, *80*(1), 61-79. doi: 10.1086/648463

Cullen, C. T. (2000). Authentication of digital objects: Lessons from a Historian's Research. *In Authenticity in a Digital Environment (pp. 1-7). Washington D.C.:* Council on Library and Information Resources. doi: 10.1.1.36.5694

Dougherty, W. C. (2009). Preservation of digital assets: One approach. *The Journal of Academic Librarianship*, *35*(4), 599-602. doi: doi:10.1016/j.acalib.2009.08.008

Gladney, H. M. (2009). Long-term preservation of digital records: Trustworthy digital objects. *The American Archivist*, *72*(2), 401-435. Retrieved from http://www2.archivists.org/american-archivist

Horrell, J. L. (2008). Converting and preserving the scholarly record: An overview. *Library Resources & Technical Services*, *52*(1), 27-32.

Howard, R. (2008). Preservation perspectives: Preserving digital information. *Kentucky Libraries*, *72*(3), 16-17. Retrieved from Library Lit & Inf Full Text database.

Kuny, T. (1997, August). A digital dark ages? Challenges in the preservation of electronic information. *Proceedings of the 63RD IFLA Council and General Conference*, archive.ifla.org/IV/ifla63/63kuny1.pdf

Levy, D. M. (2000). Where's Waldo? Reflections on copies and authenticity in a digital environment. *In Authenticity in a Digital Environment (pp. 24-31). Washington D.C.:* Council on Library and Information Resources. doi: 10.1.1.36.5694

Lyman, P. & Varian, H. R. (2003). *How Much Information.* Retrieved from http://www.sims.berkeley.edu/how-much-info-2003

Lynch, C. (2000). Authenticity and integrity in the digital environment: An exploratory analysis of the central role of trust. *In Authenticity in a Digital Environment (pp. 32-50). Washington D.C.:* Council on Library and Information Resources. doi: 10.1.1.36.5694

Rothenberg, J. (1995). Ensuring the longevity of digital documents. *Scientific American,* 272, 42-47. doi:10.1038/scientificamerican0195-42.

Rothenberg, J. (2000). Preserving authentic digital information. *In Authenticity in a Digital Environment (pp. 32-50). Washington D.C.:* Council on Library and Information Resources. doi: 10.1.1.36.5694

Smith, A. (2000). Authenticity in perspective. *In Authenticity in a Digital Environment (pp. 69-*

75). Washington D.C.: Council on Library and Information Resources. doi: 10.1.1.36.5694

University of Illinois at Urbana-Champaign (2008, October 29). Digital Dark Age May Doom Some Data. *Science Daily*. Retrieved from http://www.sciencedaily.com/releases/2008/10/081027174646.htm