Wrestling with Shape-Shifters

Perspectives on Preserving Memory in the Digital Age

Kenneth Thibodeau

Abstract

Digital preservation is a difficult challenge due to the polymorphous character of digital information and the environment of ongoing, open-ended and multidimensional change in which it exists. The paper describes both aspects of the challenge and explores how multi-faceted and dynamic approaches to digital preservation in different circumstances can be articulated.

Author

Dr. Kenneth Thibodeau is an internationally recognized expert in electronic records and digital preservation. A senior guest scientist in the Information Technology Laboratory of the National Institute of Standards and Technology of the U.S., he previously directed the Center for Advanced Systems and Technology and the Electronic Records Archives Program at the National Archives and Records Administration in Washington. He also served as the Chief of Records Management at the National Institutes of Health and directed the Department of Defense Records Management Task Force, leading the development of the world's first standard for records management software. Fellow of the Society of American Archivists, Thibodeau won the Emmett Leahy Award and a Lifetime Achievement Award from the Archivist of the United States.

"Words strain,

Crack and sometimes break, under the burden,

Under the tension, slip, slide, perish,

Decay with imprecision, will not stay in place,

Will not stay still."[1]

The poet T.S. Eliot's passionate incantation of the difficulties of fixing memories in words might be appropriated to describe the difficulties of preserving memory in digital form. Like the raven in North American cultures or the fox of Japanese folklore, digital memory is a shape shifter that takes on very different forms, driven by two distinct causes: first the characteristics of digital information itself and second the environment of change that engulfs digital information objects. There is thus an inherent tension between digital information, which does not stay still, and

[1] T.S. Eliot, "Burnt Norton," *Collected Poems, 1909-1962* (New YOrk: Harcourt Brace and COmpany, 1963) 180.

digital preservation, which quintessentially seeks to keep things in place, without significant change.

1. Polymorphous Information

In contrast to information recorded on stone, clay tablets, paper, or other 'hard copy' media, digital information is polymorphic in several respects. First, digital data is not and cannot be affixed to a physical medium in a durable fashion. Its physical inscription changes every time it moves from computer memory to a storage medium or back, every time it is copied to a different storage medium, and whenever it is transmitted on a network. Digital preservation is not a process of preserving material things, but of transporting immaterial bit streams over time, using whatever storage media satisfy preservation needs for however long they are suitable. Most digital storage media are not long lasting, but have to be replaced after some time. The cause of this is the economics of the marketplace, rather than technological or physical constraints on possible storage media. Indeed, several digital storage media have been developed that should last for hundreds or even thousands of years, including microfilm,[2] metals, gold coated silicone,[3] and other formulations;[4] however, the longevity of the medium is outweighed by the fact that storage devices become obsolete within 5 or 10 years.[5] Obsolescence in the digital storage domain includes not only equipment that becomes increasingly difficult to maintain and media that wear out, degenerate, and become rare, but also the increasing expense of older storage technologies relative to newer alternatives because of exponential increases in storage density, improvements in data transfer rates, and significant decreases in purchase and operating costs.[6]

A second polymorphic characteristic of digital information is that the boundaries of a digital object can be difficult to determine. For example, web pages often include content that is not visible to the user or that is loaded into the page from external sources each time the page is viewed. External sources include links to other web pages, style sheets, graphic images, Java scripts, data about the person using the page, data elements extracted from databases, and others. Whenever any of these external sources changes, the content of the page changes accordingly, making it difficult to define what is the content of a web page we want to preserve. Moreover, parts of the content of a digital document may be subject to different ownership and control.[7] In order to preserve a web page, we have to define it as a finite object; that is, we have to apply

---

[2] Heather Brown, John Baker, Walter Cybulski, Andy Fenton, John Glover, Paul Negus, and Jonas Palm. "The role of microfilm in digital preservation," in *DCC Curation Reference Manual*, Digital Curation Center, April 2011.
http://www.dcc.ac.uk/resources/curation-reference-manual/microfilm/

[3] R. A. Stutz and B. C. Lamartine. "Durable High Density Data Storage," in *Fifth NASA Goddard Space Flight Center Conference on Mass Storage Systems and Technologies*, *College Park, Maryland, September 1966*, 409-419. http://storageconference.org/1996/papers.html/b2_3.pdf

[4] http://millenniata.com/technology/

[5] Michael C. Peterson, "Solving the Coming Archive Crisis," Storage Networking Industry Association, *SNIA Spring 2007 Technical Tutorials*. http://www.snia.org/education/tutorials/2007/spring.

[6] Chip Walter, "Insights: Kryder's Law," *Scientific American* (August 2005): 32-33. http://www.scientificamerican.com/article.cfm?id=kryders-law.

[7] John Patzakis and Brent Botta, "Authenticating Internet Web Pages as Evidence: a New Approach," *Next Generation Law and eDISCOVERY Tech Blog*, June 27, 2012. http://blog.x1discovery.com/

extrinsic criteria, cutting off at least some external sources of input in order to establish well defined boundaries, but these boundaries are not present in the web page itself.

Furthermore, although many web pages are transitory, many web sites have persisted for decades. The key to this survival is that they are dynamic. They evolve in response to changes in the enabling technologies and also to data about what does and does not work in achieving the purpose each web site is intended to serve. Any attempt to preserve such web sites as static objects looses this essential characteristic of the web site as an evolving entity.

These considerations bring us to a third polymorphic characteristic of digital memories: the relationship between what is stored and what is presented to a human can be both complex and variable. What is presented to a human, as a single object may comprise content drawn from many different data stores, as illustrated in the preceding description of web pages. Databases include rules, invisible to all but administrators that determine what specific data elements different classes of users, or even individual users, can access. Word processing files can contain content that their authors thought they had deleted. Just as something that appears to a human as a single document may be drawn from many data stores, one item of stored digital data may be part of many different objects. For example, a web page may contain links to many other pages, and each of those pages could also be referenced by many others. Conversely, different data can produce identical presentations. A textual document, for example, may be generated from a word processing file, the scanned image of a paper document, or as a report from a database. Thus, in preserving digital memories, we need to distinguish between the *data objects*, which are stored in computer systems, and the *presented objects*, which are derived from the data objects and are presented and at least potentially meaningful to people.

Another polymorphic property of digital memories is that data objects must be processed in order to be used. Moving the data between storage and presentation, or between transmission and presentation can involve changes in semantic, syntactic and apparent form. Even if the data remains intact in storage or transmission, processing for presentation can change or even corrupt the presented object. Furthermore, apart from any question of alteration or corruption, the same digital data can be rendered in different ways; for example, numeric data can be presented in tabular or graphic form. This ability of data objects to take on different shapes may not be merely an incidental possibility. It can be an essential characteristic of the memory we want to preserve. A clear advantage of digital imaging systems in science, medicine, and engineering, for example, is that they allow the data to be presented in a variety of ways. In addition, one of the most prominent aspects of the current digital environment is that much information is intended to be rendered on different types of devices, ranging from various mobile platforms through laptops and desktops to even wall sized and billboard displays. Besides further complicating the distinction between data objects and presented objects, this adaptable display capability contributes to ubiquitous computing, which is changing the role of information and communication technology (ICT) in human affairs.[8] Thus, preserving data objects is not sufficient for digital memory. We must also maintain the ability to process the data correctly and appropriately.

The polymorphism of digital information means that even the apparently basic issue of what is it that is to be preserved is not a given, but involves choice: should we preserve what was

[8] Adam Greenfield, *Everyware: the Dawning Age of Ubiquitous Computing*, (Berkeley, CA: New Riders, 2006).

displayed in a given instance or the data, structures, controls, and functionality that enabled the presentation, or both? In order to decide on appropriate choices, we have to consider not only the characteristics of the data objects and presented objects, but also the dynamic context in which digital information exists.

2. An Environment of Change

Digital preservation has a split personality: its object, memory, is from the past but its objective, access, is in the future. This schizophrenia is aggravated by the environment of ongoing, open ended and multidimensional change in which digital information exists.

Ongoing change has two faces, one looking forward, the other backwards. The forward face, technological progress, introduces frequent alterations in both hardware and software that can also include significant innovations or departures. The backward face of ongoing change is obsolescence: older products are no longer supported and become inoperable or unusable, so that, even if we can preserve the data objects, we may not be able process them or to reproduce the presented objects that the data represent. Even in cases where older technology could be maintained, improvements in price/performance of newer products impel us towards replacing it. Obsolescence has been a main focus of attention in digital preservation.[9] But even if obsolescence were not a factor, changing user expectations about access impel us to alter preservation tactics over time in order to take advantage of technological progress. Moreover, newer technologies may offer better options for preservation, making older preservation solutions themselves obsolete. Technological progress in itself should be anticipated and incorporated in planning for and carrying out digital preservation in order to enable use of the best current technology to preserve, examine, process, and communicate information from the past.

Change in ICT is not only ongoing; it is also open ended, with often surprising developments. The history of ICT since the mid twentieth century is one of repeated transformational changes, where the technology gains new capabilities; new functions are added; new classes of hardware and software are introduced; and even methods of producing and implementing technology change. Software paradigms have shifted from structured to object-oriented, to component-based, and to service-oriented approaches. The emerging paradigm of autonomic computing opens possibilities for additional, radical changes.[10]

Technological change has also expanded the varieties of information that ICT can handle. Computation was initially limited to numeric data. Over the last three decades, ICT's scope has grown to include more and more traditional forms of information, such as text, images, audio and motion video. And it has created new forms that cannot exist outside of the digital realm. Additionally, increases in speed and capacity have created new possibilities for processing and communicating information, greatly expanding possibilities for selecting, combining, analysing, and applying different types of information from disparate sources for a variety of purposes. All of this adds considerable diversity and complexity to the challenge of digital preservation.

---

[9] Donald Walters and John Garrett, *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information*, (Washington, D.C.: The Commission on Preservation and Access, 1996); Stewart Brand, "Escaping the Digital Dark Ages," *Library Journal*,124, no. 2 (1999): 46-48.

[10] Richard Murch. *Autonomic Computing*. (Englewood Cliffs, New Jersey: IBM Press, 2004).

Moreover, changes in one sphere can snowball into others. Web 2.0 flies in the face of traditional, pre-determined, and systematically controlled user interactions by enabling structure to emerge over time though use of free-form software tools.[11] The expansion of mobile computing has spawned the proliferation of "apps," which are substantially changing the end users' acquisition, use, and experience of software, while menacing corporate control of ICT resources.

Moreover, the environment of change is not limited to changes in the technology itself or the types of information it produces. Rather it is multidimensional. First, ICT changes the way we do things. Think, for example, of the differences it has enabled in the interactions between businesses and their customers or between citizens and governments. Second, ICT changes the things we do. For example, geo-positioning technology is enabling precise location tracking of individual vehicles, goods, and people, with substantial impacts on many commercial, governmental and social activities. Third, ICT changes who does what.[12] Prior to the growth of the Internet, for example, advertising was one-way dissemination function, but the possibilities the World Wide Web offers for active customer involvement via social computing has transformed advertising into a multidirectional form of communication in which individual and collective initiative by consumers can have rapid and decisive impact.[13] The memory of the digital age would be greatly impoverished, and probably falsified, if it does not take into account the additional dimensions of change precipitated by changing technology.

Indubitably, multidimensional and transformative changes will continue in the future, at least as long as ICT continues to change. Inevitably, the challenge of preserving digital memories, and therefore its complexity and difficulty, will evolve apace with changes in ICT and its impact.

3. A Plethora of Choices for Preservation

The polymorphous and metamorphosing characteristics of the challenge of preserving digital memories necessitate multi-faceted, diversified, and dynamic approaches to digital preservation: multi-faceted in order to deal with the polymorphism of digital information; diversified in order to accommodate varying requirements in different social, cultural, and institutional contexts; and dynamic to respond to continuing changes in ICT and its uses and in future user expectations and needs. We can elaborate approaches that take into account the immensity and difficulty of preserving digital memories and that are appropriate to different contexts by addressing three questions: what are you trying to preserve; why are you trying to preserve it; and how much preservation effort is required?

[11] Andrew McAfee. *What is Web/Enterprise 2.0.* http://www.youtube.com/watch?v=6xKSJfQh89k&feature=related

[12] Jay Rosen. "The People Formerly Known as the Audience," in *The Social Media Reader,* Michael Mandiberg, ed., (New York: NYU Press, 2012): 13-16.

[13] Shuai Yuan, Ahmad Zainal Abidin, Marc Sloan, and Jun Wang, "Internet Advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users," 2 Jul 2012. http://arxiv.org/pdf/1206.1754v2.pdf

3.1 What are you trying to preserve?

The question of what is to be preserved does not concern selecting things to be preserved, but determining what properties of those things have to survive in order to assert that they have in fact been preserved. Given the polymorphism of digital information, determining the properties that are essential to preserve can be complicated. The possibilities span a spectrum from the preservation of technology to the preservation of information. Between these extremes is the preservation of information artefacts created using the technology. Each alternative responds to different needs and entails different actions.

An obvious case where we would need to preserve information technology would be that of digital artworks that depend on unique technologies. At the other end of the spectrum, for example with statistical data, all we would need to preserve is the information because users could access and use the information with readily available hardware and software. In the middle would be classes like three dimensional models, where we would want to preserve functionality, such as the ability to rotate the model visually, that requires special technology, but where there are alternatives to the original technology used to produce the models.[14]

Obviously, determining where digital objects fall on the spectrum of preservation possibilities does not depend solely on the properties of the objects themselves. It is also a function of both the *object class technology* and the available *preservation technologies*. Object class technology includes both the original technology used to produce the data objects and technology currently available for that class of data objects. Preservation technologies are those created to maintain digital memories when the original technology is obsolete and there are no satisfactory alternatives outside of the preservation realm.

Illustrating how technology impacts where objects fall in the spectrum, the class of geographic information systems (GIS) is distinguished from mere geographic data by the ability to display data in a cartographic presentation having selected that data from many different types of data stored separately. In principle, as long as we can select data from the separate layers in a GIS and display them in map form, we do not need to preserve the original GIS software. If there were interoperability across GIS formats, GIS would be situated in the middle of the spectrum. However, most GIS depend on proprietary software, which often has features not present in other products in the same object class. If these features were deemed necessary to preserve, maintaining the technology that supports the unique functionality would be necessary, pushing GIS to the technology end of the spectrum.

A given class of digital objects could move across the spectrum, even from one end to the other. In the early days of word processing, for example, visual display technology was too crude to present text in different typefaces. In the 1980s displays were introduced that enabled text to be displayed on screen just as it would be printed on paper, but at that time the only way to preserve that presentation capability was to maintain the specific display technology. Today, however, we do not need any special technology to present text documents from the 1980s with their original formatting.

In sum, then, digital information objects fall at the 'preserve technology' end of the spectrum when the only way to ensure continuing access to them is to preserve the original

---

[14] Peter Bajcsy, Appraisal of 3D Data Conversions and Visualization Software Packages, January 21, 2009. http://www.archives.gov/applied-research/ncsa/

technology or some equivalent or surrogate, such as an emulator. Objects fall at the 'preserve information' end of the spectrum when, given their physical survival, they can be accessed using readily available current technology. Objects fall within the 'preserve information artefacts' range when they require specialized processing capabilities to render the data objects, but there are alternatives to using the original technology.

In some situations, more than one approach to digital preservation might be appropriate. In the case of interactive digital artworks, where the audience or spectators are involved in real time in the production or performance of the creative work, we face the alternatives of preserving the technology that makes the experience possible or somehow capturing the performance at some particular time and preserving that. In the first alternative, preserving the technology used in performing the work, we would not be preserving memories of specific happenings, but the capabilities that make such happenings possible. In the second, we would not be preserving the digital artwork, but a derivative product that not only does not include any of the technology of the artwork, but also is bereft of precisely what made the art interactive and creative. Basically, this is no different than the alternatives of preserving a written musical score and preserving a recording of a performance of the score.

We also face the option of preserving technology or preserving information in the case of websites where inputs from external sources constantly change what is presented. To enable people in the future to appreciate how users could interact with such a website, we would need to preserve the technology of the website, but that would provide no knowledge of what any users actually saw or might have seen on the website at any time in the past. For that, we would need to preserve snapshots of the website. We cannot decide which alternative is appropriate solely by considering the properties of the information and related technology. For that we have also to address the second question, the purpose of preservation.

3.2 Why are you trying to preserve it?

In addition to the properties of the information objects themselves and the state of related technology, determining the appropriate course of preservation actions depends on the purpose for which digital information is preserved. We can distinguish two basic reasons for preserving information: either for remembrance or for utilization. We preserve information for remembrance in order to provide the future with opportunities to gain knowledge of the past from materials produced or acquired at the time of which we wish to gain knowledge. We preserve information for utilization in order to enable future use of that information for purposes that are likely to be different from the purposes for which they were created or acquired. While remembrance and utilization are not essentially contrary to one another, they lead us to different preservation actions. Preserving for remembrance would lead us to maintain digital memories as pristine as possible. This goal is best served in many cases by preserving the original information technology. But preserving information embedded in specific technologies creates barriers to exploiting this information in the future. To optimize possibilities for utilizing preserved memories, we would want to reduce or eliminate dependencies on the hardware and software originally used to produce and/or retain the information.

Decisions on what we are preserving and why we want to preserve it are interrelated. This is illustrated in the case of preserving email and other communications on the Internet. In order to work globally, email has to be independent of specific hardware and software as well as relatively impervious to technological change. Thus, there is no need to preserve information technology to preserve email messages. Several initiatives have approached email as an artefact of technology, maintaining the organization of messages within individual users' accounts, because that is the way the technology is implemented.[15] However, if we focus on the value of the information in email as evidence of the conduct of human affairs, the emphasis shifts to the communications between and among individuals and groups of individuals. The threads of communication that evince and often enable important developments in human history are outside of the confines of individual users' accounts and even of the administrative domains of email implementations. To preserve the memory of events ranging from the Arab Spring[16] to the international outpouring of charity on behalf of an elderly school bus monitor who was abused by schoolboys in New York State this spring,[17] we need to preserve the connections among the messages, independently of the artefacts of the enabling technologies.

3.3 How much preservation effort is needed?

What is done to preserve digital memories also depends on how much effort is required. The amount of effort is proportional to the level of resources required to accomplish it. In most cases, resources will be the independent variable. Resource limits may have major impact in determining what is preserved and what preservation actions are carried out. Three sub-factors determine the amount of preservation effort required: quantity, variety, and range.

How much information? In the digital realm, quantity should be measured in both the volume of digital data to be preserved and the number of discrete objects the data comprise, and the probability of substantial growth in both parameters should be a major concern. Between 2006 and 2011, the amount of digital data produced worldwide doubled every two years, exceeding a trillion gigabytes in 2011, and it is expected to increase fifty times more by 2020. The number of files containing this data has increased even faster and is expected to increase seventy-five times by 2020.[18] These staggering numbers have both global and local implications: there should be significant benefits from international coordination to avoid wasting resources on duplicative efforts and to promote the development of technical capabilities that can be widely implemented; technical solutions developed for particular preservation challenges need to be scalable to accommodate projected growth; and repositories need to anticipate growth in data volumes and numbers of objects that will eclipse everything they have faced until now. The only exceptions would be closed collections, where there will be no further additions.

Given the probability of growth, the quantity of information to be preserved will probably become an increasingly important factor in deciding what gets preserved and how much effort is expended on any set of objects over any period of time. In cases of very valuable information resources, preservers may have to settle for merely ensuring the physical survival of data objects

---

[15] The Collaborative Electronic Records Project, http://siarchives.si.edu/cerp/

[16] Ekaterina Stepanova, "The Role of Information Communication Technologies in the 'Arab Spring' Implications Beyond the Region," George Washington University, PONARS Eurasia Policy Memo No. 159, May 2011. http://www.gwu.edu/~ieresgwu/assets/docs/ponars/pepm_159.pdf

[17] Rene Lynch, "Bullied school bus monitor calls it quits: She's retiring," Los Angeles Times, July 27, 2012. http://www.latimes.com/news/nation/nationnow/la-na-nn-bullied-school-bus-monitor-retires-20120727,0,6307218.story.

[18] John Gantz and David Reinsel, *Extracting Value from Chaos*, 2011. http://www.emc.com/digital_universe.

because there will not be resources and perhaps not even any technical possibilities for addressing other requirements, such as overcoming obsolescence or enhancing access.

What variety of digital memories is preserved? Preservation efforts will demand more resources and become more complex as the variety or heterogeneity of the information objects being preserved increases. In general, the greater the variety of objects being preserved, the greater the variety of preservation tactics that will be needed. Homo- and heterogeneity of digital memories are determined at several levels. At the highest level, they relate to the types of information; such as, text, image, audio, motion video, and so on. In addition, in the digital realm, any type of information can be represented in different ways; e.g., text may be encoded as characters or as images of printed documents, and graphic information may be represented by raster or vector data. So, for any type of information, we need to know what data types are used to express it in digital form. Going another level down, a given data type can be encoded in a variety of formats. For example, character encoded text may be in plain text, rich text, Hypertext Markup Language (HTML), eXtensible Markup Language (XML), portable document format (pdf), Microsoft Word, Apple Pages, and other formats.[19]

Furthermore, a digital object may be more or less complex. A digital document may consist entirely of textual information all encoded in a single format, but it might comprise several types of information; such as, text, photographs, graphic illustrations, and even audio. A Web page may include static text, data drawn dynamically from a database, images, applets that enable interaction with users, et al.

The variety of formats, and the complexity of objects are also likely to increase the variety of hardware and software necessary to support them. For example, preserving purely numeric data can be very simple; however, if the data are embedded in an object where the specific content is determined in real time from user input, it may be necessary to preserve the database management system used to manage the data, as well as technology needed to reproduce the corresponding presented objects.

What is the range of preservation efforts? Preservation efforts may be directed only at objects individually or extend to preserving relationships among objects. It is not a question of whether objects are related. Basically, all objects subject to a given preservation regime are related simply by virtue of being held in the same repository or being part of the same collection. Generically, we should distinguish cases where things are simply compiled or grouped together, with no intrinsic ordering or relationships among them, from combinations where there are relationships that must be preserved along with the individual objects. In the case of 'records' as defined in archival science; that is, documents produced or acquired and kept in the course of activity, there are essential relationships among records of a given activity; for example, between a letter and the response to it, and between a plan and documents produced in executing the plan and evaluating its success. If these archival links are lost, the possibility of reconstructing the activity on the basis of the evidence provided by the records is diminished.

For purposes of digital preservation, what matters is not the existence of relationships, but whether the relationships require preservation efforts in addition to those required for individual

[19] John C. Bennett. "A Framework of Data Types and Formats, And Issues Affecting the Long-term Preservation of Digital Material," *JISC/NPO Studies on the Preservation of Electronic Materials*. British Library and Information Report 50, Version 1.1. 1999. http://www.ukoln.ac.uk/services/papers/bl/jisc-npo50/bennet.html

objects. Consider a set of digital maps produced by scanning printed maps. There would certainly be relationships among the maps if they were all pages in a printed atlas, and we would need to ensure that both the relationship of parts to the whole and the sequence of maps in the atlas were preserved; however, these relationships could adequately be preserved in metadata. That would not require any special digital preservation efforts. In contrast, consider the maps that could be produced from a geographic information system. No amount of metadata would be sufficient to preserve a GIS. Many GIS contain such a rich store of data and provide so many options for displaying the data that it would not be possible even to enumerate the set of maps that could be produced using the system. Furthermore, the option of simply preserving each of the various data types or "layers" included in the GIS would not be sufficient because it would not preserve the essential ability to select data elements both within and across layers for composite display in cartographic form. A GIS, which normally consists of cartographic and attribute data, might also be linked to other types of information, such as scientific observations made at specified locations, or historic photographs taken at different times. Preserving such systems requires maintaining the links to such heterogeneous types of information and maintaining the ability to locate them correctly in presented objects.

4. Conclusion

Obviously, digital preservation constitutes an enormous and difficult challenge, one which we must attack less we fail to address important cultural, educational, scientific, social, governmental, and practical needs which depend on, or would benefit from, access to digital memories. The three questions of what, why and how much combine to form a framework for rational discourse on digital preservation, one that embraces the polymorphic character and metamorphosing context of digital information. The framework should guide an integrated and parallel consideration of the three questions because their answers will often be interdependent. This framework should be useful in a variety of contexts, ranging from articulating a theory of digital preservation; to developing a specialized bodies of knowledge and skills, such as digital curation and archival engineering; planning for and managing repositories; developing and implementing strategies for particular sets of information objects to be preserved, and defining the need for preservation technologies, guiding their development, and evaluating the relevance and adequacy of specific preservation techniques.