Chapter 2

Preservation Practices

Preservation Strategies

Preservation strategies can be thought of as falling into two categories. The methods in the first category address the goals of fixity and viability, and include techniques such as copying data to new media of the same type (media refreshment), copying data to newer media (media migration), and maintaining multiple, frequently verified copies of data. These activities are often referred to as "bit-level" or "passive" preservation. Because they are part of sound data management practices and are not specific to digital preservation *per se*, they are not described here in any detail.

The methods in the second category attempt to address the goals of renderability and authenticity, and are unique to the preservation realm. Format migration and emulation are often touted as the two main, even competing, digital preservation strategies. In fact, a number of different strategies are available to preservationists, and multiple approaches are often used together to good effect. When strategies addressing renderability are employed, it is called "full" or "active" preservation.

Technology Preservation

Often called the "computer museum approach," technology preservation is familiar to anyone who still owns a record player for listening to vinyl LPs. If a format depends upon a particular combination of hardware and software for rendering (for example, an old Wang word processing system), it should be possible to preserve at least a few working examples of the obsolete platform. Technology preservation is generally considered an interim approach at best, because it is not scalable and because old computer systems can't be kept running indefinitely. However, preserving old technologies, such as early-generation video game consoles, can provide historical information about the genuine behaviors of obsolete applications and thus provide valuable information for writing emulators.

Emulation

Emulation involves the use of hardware and/or software (emulators) that allow computer instructions written for one platform to be run on another platform. Emulation has been in use in the computer industry for years to extend the life of programs written for earlier models of machine. (In the mid-1960s, IBM offered System 360 mainframe customers a microcode emulator for the popular but superseded 1401.) Today emulation is widespread, particularly to allow programs written for one microcomputer operating system to run on any other, and to keep old video games usable on modern machines. Emulation as a strategy for long-term digital preservation, however, is still largely experimental.

Universal Virtual Machine

One of the problems with simple emulation is that modern computer technology is a moving target—not only does an emulator have to be written for each obsolete platform, but emulators must be updated or rewritten as current platforms change. The Universal Virtual Machine (UVM) addresses this issue by providing an intermediate layer between the emulator and the current platform, isolating the emulator from these technology changes. Although the UVM itself may require updating or rewriting for new

platforms, it is presumably less work to update one UVM than dozens of emulators.

Universal Virtual Computer

Raymond Lorrie of IBM expanded the concept of a Universal Virtual Machine to that of a Universal Virtual Computer (UVC) for preservation. In this approach, files of a given format are translated to a simpler Logical Data View by a "decoder" program written to run on the Universal Virtual Computer. The original file, the Logical Data View, and a schema describing the Logical Data View are all archived together. In the future, files in the format can be rendered by first building a UVC emulator to run on then-current hardware, then executing the decoder to generate the Logical Data View, and finally writing a viewer to render the Logical Data View according to the schema. The Koninklijke Bibliotheek (National Library of the Netherlands) has been a leader in exploring the use of the UVC in a production preservation environment.

Format Migration

Format migration, also called "forward migration," creates a version of a source file in a different format that is considered to be a successor format. This is routinely done by common desktop applications such as Microsoft Word or Excel, which can open a file written by an earlier version of the program and save it in the current format. In some cases the successor to one format may be an entirely different format, as, for example, PDF can be considered the successor to PostScript. One concern about the use of format migration for digital preservation is the likely need for successive migrations over time. Since any format transformation could potentially lose or even add information, it is possible that successive migrations would accumulate errors leading to results less and less like the original. A counter strategy is to save the original and write programs to migrate directly from the original to the current format.

Format Normalization

There is consensus that some formats are more "preservable" than others (see Life-Cycle Management of Materials, below). The process of normalization creates a version of a source object in a preferred format while maintaining the essential properties of the original. For example, textual documents in proprietary word processing formats could be converted to Rich Text Format or to an open XML-based format. Some preservation systems, particularly those designed for archival materials, normalize all incoming documents on ingest. This has the advantage that there are far fewer formats for the repository to support and maintain over time. A disadvantage is that normalization can be lossy, and unless the original is also preserved, the initial decision as to what properties must be maintained is critical.

Deciding which preservation strategies to employ, when to take action, and how to evaluate the action taken is the heart of preservation planning. As the suite of preservation strategies has matured, preservation planning has become the focus of a number of research and development efforts. One example is the European Union PLANETS project, which is developing PLATO, a package of preservation-planning tools that includes risk assessment and decision support modules.

Readings

- Jeff Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation,* Council on Library and Information Resources, Jan. 1999, www.clir.org/pubs/abstract/pub77.html. An early call to action by an influential proponent of emulation.
- Kenneth Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," in: CLIR, *The State of Digital Preservation: An International Perspective. Conference Proceedings,* July 2002, www.clir.org/pubs/reports/pub107/thibodeau.html. Exactly what the title says.
- "Digital Preservation Strategies," Preserving Access to Digital Information (PADI) Web site, www.nla.gov.au/padi/topics/18.html. An annotated webliography of resources.

Life-Cycle Management of Materials

Life-cycle management takes a proactive approach to preservation by actively managing each stage in the life of the digital object, and by taking preservation actions as early in the life cycle as possible. Life-cycle management emphasizes the creation, appraisal, documentation, and re-use of the object.

Creation

There are a number of circumstances under which the custodial institution has some control over the creation of a digital object, either directly (as in the case of a retrospective digitization project) or indirectly (for example, through guidelines for submissions to an institutional repository). In this case both the choice of file format and the selection of format options are important preservation decisions.

There is a clear consensus that some formats are inherently more sustainable (likely to be long-lived) and preservable (amenable to preservation actions) than others. Considerations include

- the extent to which the format is open and nonproprietary
- how well documented the format is, and whether the documentation is publicly available
- whether files in the format are self-documenting, and allow embedded metadata
- how widely adopted the format is, and how many different vendors and software applications support it
- how transparent (easily understandable with basic tools) the format is

The Library of Congress discusses these and other sustainability factors in its excellent online guide. Applying such criteria, however, is always a judgment call. For example, how does one weigh an open, well-documented, transparent, but little-used format against a well-documented, proprietary format with worldwide adoption?

Library of Congress guide to sustainability factors
www.digitalpreservation.gov/formats/sustain/sustain.shtml

Even within a single format there can be wide variation in how a file is created. A rule of thumb is that digital objects should be as true to their analog equivalents as possible, which generally means using a high resolution or sampling rate. Preservation-worthy objects should not be encrypted or compressed with proprietary or lossy compression schemes. They should be as self-contained as possible and not rely on external dependencies, meaning, for example, that content should be embedded rather than linked to. Technical and descriptive metadata should be embedded within the file itself to the extent allowed by the format.

For most formats, it is up to the creator of the digital file to determine and select the optimal options for longevity. In the case of Adobe's Portable Document Format (PDF), best practices for sustainability have been codified and standardized as PDF/A, a subset or profile of the larger PDF format. The standard is of considerable interest to business and government as well as memory organizations. PDF/A disallows embedded audio, video, and JavaScript, and requires embedded fonts, metadata, and device-independent color information. Encryption and hyperlinks are prohibited. Starting with version 7.0.7, Adobe Acrobat Professional supports conversion directly to PDF/A format and offers a "preflight" feature to validate files against the standard.

The time of resource creation is also the best time to document certain information about the resource. When the resource is created as part of a retrospective digitization project–for example, converting a set of analog audiotapes to digital format–information such as who did the conversion; when it took place; what equipment, specifications, and benchmarks were used; how quality control was performed; and other project-wide details can be easily captured. This may also be the time to record some items of descriptive metadata and copyright information.

Selection and Appraisal

Most collecting institutions such as libraries and museums have well-documented collection-development policies that define collection scope, collection goals, and criteria for selection and retention. Digital resources, however, raise different issues for appraisal, retention, and preservation decisions, and are rarely well handled under traditional collection policies. One major difference is the impact of abundance: an institution may be able to acquire more digital records, Web pages, datasets, etc., than it would ever be able to curate and preserve over time. It is neither possible nor desirable to retain all data indefinitely. Practical considerations, such as whether the institution is capable of preserving a particular data type, may have to take precedence over abstract collection goals.

Another major difference between digital and nondigital materials is that decision points must occur earlier in the life cycle of the resource. Records managers are well aware that records retention and disposition schedules developed for paper materials must be modified for their electronic equivalents. Earlier decision points are required, as are earlier interventions to keep digital data usable. Emerging best practice takes a risk-management approach and focuses first on high-risk, high-consequence materials.

The Decision Tree for Selection of Digital Materials for Long-Term Retention is an example of a tool to help curators assess whether or not to assume responsibility for long-term retention and preservation of specific digital materials.[1] It can also be used to help develop or test an institution's own selection policies. Decisions are based on criteria related to selection policy, rights, technical costs, and metadata costs.

Use and Reuse

In the life-cycle approach to digital preservation, there is an emphasis on the use and reuse of digital materials as the heart of digital curation. In the stage of active use, active curation adds value to digital assets, which can be extended by annotations, aggregations, and linkages, or distilled through extractions or summarizations. This stage may be followed by storage and preservation,

but this is not necessarily a terminus but a step along the way towards active reuse. For digital materials to support reuse, their integrity and authenticity must be demonstrably maintained.

The LIFE (Life Cycle Information for E-Literature) project of the British Library and University College London developed a methodology to model the life cycle of digital objects and calculate the costs of preserving them over any period of time.[2] An updated version of the model, incorporating the suggestions of early implementers, was released in 2007.[3] The model considers costs of creation or purchase, acquisition, ingest, description, bit-level preservation, content preservation, and access.

Readings

- Maxine K. Sits, ed., *Handbook for Digital Projects: A Management Tool for Preservation and Access,* Northeast Document Conservation Center, 2000, http://nedcc.org/oldnedccsite/digital/dighome.htm. A life cycle–oriented guide directed at digitization projects, based on an early School for Scanning workshop.
- Digital Curation Centre, *Digital Curation Manual,* www.dcc.ac.uk/resource/curation-manual. This manual is being issued in installments and will ultimately be a very comprehensive resource.
- Digital Preservation Coalition, *Preservation Management of Digital Material: The Handbook,* www.dpconline.org/graphics/handbook. A good-practice guide to life-cycle management designed for the Web. Links are checked and updated weekly.

Preservation and Intellectual Property Rights

All preservation activities take place within the context of intellectual property law. Preservation strategies can involve actions such as making multiple copies of a digital object, making a new version of an object in a different format, recreating all or part of an object, reverse engineering, or rendering an object (repeatedly) using an emulator.

Any or all of these actions could be forbidden under law, as copyright law grants the copyright owner the exclusive right to make copies of a work, publicly perform it, or create derivative works.

If the copyright holder can be found, the safest course is to obtain permission to undertake preservation actions. At this time, however, there is no consistent vocabulary for expressing the rights needed for preservation. Some repositories take a granular approach, listing all of the copying, reformatting, and access activities that might occur, while others simply ask for "the right to preserve" the material.

In the United States, if the copyright holder cannot be determined or cannot be located, permission might be claimed under fair use or the Digital Millennium Copyright Act (DMCA). Fair use might be construed to allow copying for preservation for noncommercial purposes but is always a judgment call that can be determined only in the courts. The DMCA allows up to three preservation copies of certain materials to be made under certain conditions, one of which is that the original is "damaged, deteriorating, lost, or stolen, or if the existing format in which the work is stored has become obsolete." Most preservationists would argue that when the original is missing, damaged, or obsolete is the absolute worst time to initiate preservation actions. Other problems with DMCA include the fact that it applies only to libraries and archives (is an institutional repository a library?) and its requirement that libraries/archives own a legal copy of the object being preserved (unlikely for harvested Web sites). Moreover, the DMCA explicitly prohibits circumvention of access controls if the digital object is technologically protected, even for preservation.

Readings

- Peter B. Hirtle, "Digital Preservation and Copyright," 2003, http://fairuse.stanford.edu/commentary_and_analysis/2003_11_hirtle.html. Clear, concise, and very helpful; highly recommended.
- Catherine Ayer and Adrienne Muir, "The Right to Preserve: The Rights Issues of Digital Preservation," *D-Lib Magazine* 10, no. 3 (March 2004), www.dlib.org/dlib/march04/ayre/03ayre.html. Intellectual property rights from a more international perspective.
- Karen Coyle, "Rights in the PREMIS Data Model," 2007, www.loc.gov/standards/premis/Rights-in-the-PREMIS-Data-Model.pdf. Commissioned to address preservation metadata, this report also discusses the general context of rights for preservation.

Notes

1. Digital Preservation Coalition, "Decision Tree for Selection of Digital Materials for Long-Term Retention", March 8, 2006, www.dpconline.org/graphics/handbook/dec-tree.html (accessed Nov. 17, 2007).
2. LIFE Web site, www.life.ac.uk (accessed Nov. 17, 2007).
3. Paul Wheatley et. al., "The Life Model v1.1," Oct. 2007, http://eprints.ucl.ac.uk/archive/00004831/01/4831.pdf (accessed Nov. 17, 2007).