What Is Digital Preservation?

Defining Digital Preservation

Although there are any number of definitions of digital preservation in the literature, most agree that it is a set of activities aimed towards ensuring access to digital materials over time:

- "Digital preservation combines policies, strategies and actions that ensure access to information in digital formats over time."[1]
- "[Digital preservation] Refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary."[2]
- "Digital preservation is the series of actions and interventions required to ensure continued and reliable access to authentic digital objects for as long as they are deemed to be of value."[3]

Recently the emphasis on access has been questioned as potentially dangerous; publishers may be happy to have their content archived for preservation purposes only if it is not made routinely accessible. This may lead to new definitions referring to the usability rather than the accessibility of information. Ultimately, however, nothing is really usable if it can't be accessed for use, so preservation and access must continue to travel together, if not hand in hand.

In the United States digital preservation tends to be interpreted as the life-cycle management of materials from the point of their creation, while in the United Kingdom the term *digital curation* is used for life-cycle management while *digital preservation* is reserved for those activities specifically geared towards future accessibility.

Digital Preservation versus Digitization for Preservation

Some people find the difference between "digital preservation" and "digitization for preservation" confusing. Digitization for preservation is a concept that comes from the traditional field of analog preservation and conservation. In the 1990s a huge number of brittle books and newspapers were microfilmed with funding from the National Endowment for the Humanities and other grant programs. The intent was to preserve their information content and to make that content accessible without additional damage to fragile originals, but the effect was to limit access to only the most dedicated researchers. This was followed by a transitional time during which recommended practice included microfilming for preservation and digitizing for access. At this time, although it is still controversial, digitization alone is becoming an accepted approach for all preservation reformatting. The policy of "digitization for preservation" was endorsed by the Association of Research Libraries in July 2004.[4]

Digitization for preservation results in digital materials which must themselves be preserved. That is, digitization for preservation results in a need for digital preservation.

Goals of Preservation Activities

Since digital preservation is defined as a set of activities, it is most easily approached by asking what these activities are intended to accomplish. Although there is room for argument, a core set of goals that most would agree on includes ensuring the availability, identity, understandability, fixity, authenticity, viability, and renderability of digital information. In this section we'll look at each of these in turn.

Availability

It is a truism that you cannot preserve digital objects that you do not control. Depending on the materials and the circumstances, getting a copy of the objects may be trivial or quite difficult. A library that wants to preserve digital images it created in a scanning project is likely to have preservation masters in its possession, perhaps offloaded to tape or on DVDs. A library that wants to preserve the intellectual output of a university will have a much harder time and may need to work with faculty and administration to establish an institutional repository. Deposit agreements, licenses negotiated to provide a library with an archival copy, and contracts with publishers are all ways to get copies of published materials. Web archiving, a technique to gain control of Web-based content, is a major subdomain of digital preservation discussed further in Chapter 7.

Identity

The creation of descriptive metadata is usually thought of in the context of discovery and access, but it is also a preservation activity. If the end of digital preservation is long-term access and/or usability, the digital object must be described in sufficient detail to allow future access and/or use. Ideally, digital objects should be self-describing; that is, they should carry descriptive metadata within them. Many contemporary file formats, such as PDF and JPEG2000, support embedded metadata, but only if object creators take advantage of their capabilities. Whether the metadata is internal or external, there are no separate standards for descriptive metadata for preservation, and schema such as Dublin Core and MODS are commonly used by libraries. However, there is universal agreement that persistent identifiers are a critical element of descriptive metadata for preservation.

Understandability

The OAIS reference model, discussed in Chapter 3, mandates that a repository must ensure that the preserved information is independently understandable to its user community. For example, descriptive metadata may tell us that a dataset represents the results of a certain pre-election poll, but unless we have the codebook we won't know what questions were asked and how the answers were represented in the file. The repository is responsible for providing and preserving enough information, as metadata, documentation, and/or related objects, to enable future users to understand the preserved objects.

Fixity

Preservation systems must protect digital objects from unauthorized changes, whether deliberate or inadvertent. Industry-standard computer security regimes are the best defense against both malicious and careless behavior. These include virus protection, firewalls, tight authentication, intrusion detection, and immediate attention to security alerts. Media degradation can also cause bitstream corruption and is prevented by sound storage management practices, including climate control and media refreshment (copying data from one storage device to another). Fixity errors are detected by comparing message digests (more commonly called "checksums") calculated over the same file at different times. If checksums taken at different points in time are identical, the file has not been altered in between.

Authenticity

Often defined anthropomorphically as "the quality that an object is what it purports to be," *authenticity* means that the integrity of both the source and the content of an object can be verified. (It does not refer to the veracity of the object—an authentic document could be totally untrue.) A preservation program may not be able to guarantee that all the digital objects it handles are authentic, but preservation treatment should not compromise the authenticity of an object in any way. There must be policies and procedures to ensure data integrity (that is, that objects are not destroyed or altered in an unauthorized manner) and to ensure that the chain of custody and all authorized changes are documented. The event history pertaining to a digital object is known as its "digital provenance" and is a critical part of preservation metadata.

Viability

Viability is the quality of being readable from media. Media deterioration and media obsolescence are threats to viability, and both of these are experienced by most of us in our everyday lives, as anyone who has ever left a CD on the dashboard of his car can attest. Viability of digital files is easily ensured when the files are actively managed, as digital data, unlike analog data, can be copied without loss indefinitely. Files should be copied periodically to new media, and backup copies should be kept on different physical devices. However, ensuring viability for content that has been neglected can be a serious problem. Archives that once may have received cartons of personal papers may now be given shoeboxes full of obsolete floppy disks.

Renderability

Ensuring that a digital file is renderable (displayable, playable, or otherwise usable as appropriate) may be the heart of the digital preservation process. A file may be authentic, uncorrupted, and perfectly viable, but if the hardware or software required to render it is no longer available, the file is essentially unusable. Cornell's Digital Preservation Management Tutorial lists several reasons that file formats may become obsolete, including the failure of software upgrades to support legacy files, the format itself being superseded by another format, the format "take up" being so low that few software products support it, and the unavailability of supporting software because it has failed or been purchased by a competitor and withdrawn.[5] Many different strategies have been proposed to counter format obsolescence and ensure renderability, as we will see in Chapter 2.

These goals are depicted graphically in the "Preservation Pyramid" in Figure 1. Thinking about digital preservation in these terms can be particularly helpful when evaluating preservation tools or repository systems. Try to determine exactly which goals are supported by the application in question. It is often the case that multiple tools must be used together to ensure that all goals are addressed.

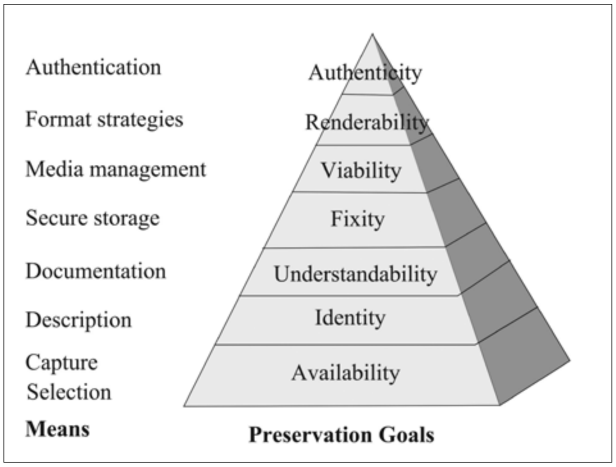

Figure 1
The Preservation Pyramid.

Notes

1. ALCTS Preservation and Reformatting Section Committee (PARS) Digital Preservation Discussion Group blog, http://blogs.ala.org/digipres.php (accessed Nov. 17, 2007).
2. Digital Preservation Coalition Web site, www.dpconline.org/graphics/intro/definitions.html (accessed Nov. 17, 2007).
3. JISC, "Digital Preservation Briefing Paper," Nov. 2006, www.jisc.ac.uk/publications/publications/pub_digipreservationbp.aspx (accessed Nov. 17, 2007).
4. Association of Research Libraries (ARL) Web site, www.arl.org/news/pr/digitization.shtml (accessed Nov. 17, 2007).
5. Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Problems, 3. Obsolescence and Physical Threats, www.library.cornell.edu/iris/tutorial/dpm/oldmedia/obsolescence1.html (accessed Nov. 17, 2007).