Long-Term Management and Preservation

Digital assets are inherently fragile and are threatened by media instability and format and hardware obsolescence. There are many fundamental problems that can imperil digital information, for instance, that practices undertaken to solve short-term problems—compression or encryption, for instance—may result in an inability to "unscramble" information in the long term; that digital works are often complex, and tracking their interrelations and determining their boundaries over the long term are likely to be difficult; that a lack of clarity about whose responsibility it is to preserve digital material leaves it vulnerable to falling through the cracks and becoming inaccessible to future generations; and that translating digital information into new environments often entails some change in meaning.[6] This means that it is vital to develop a preservation strategy at the very beginning of the life cycle of a digital image collection if it is to be retained as useful and valuable in the long term. Oya Y. Rieger has identified four goals for **digital preservation**: (1) bit identity, ensuring files are not corrupted and are secured from unauthorized use and undocumented alteration; (2) technical context, maintaining interactions with the wider digital environment; (3) provenance, maintaining a record of the content's origin and history; and (4) references and usability, ensuring users can easily locate, retrieve, and use the digital image collection indefinitely. [7]

The key to digital preservation is the establishment of a managed environment. The default fate of analog objects is, arguably, to survive (think of cuneiform tablets or papyrus scrolls), but without persistent and regular intervention it is the fate of digital works to perish. Digital preservation necessitates a paradigm shift—from one where we subject objects to one-time or occasional conservation treatments then leave them, perhaps for decades, in a temperature- and humidity-controlled warehouse—to a new approach where we, for example, periodically review each work and copy it onto a new storage medium, in all likelihood more than once per decade. Digital works require ongoing management. All current digital preservation strategies are flawed, or at best speculative, and thus a broad-based strategy is the best current safeguard of any investment in digital imaging. Over time it will be necessary to be vigilant as to both the condition of the data and technological trends and to be prepared to reassess policies accordingly. It will also be essential to have a long-term commitment to staffing, continuous quality control, and hardware, storage, and software upgrades.

The primary preservation strategy is to practice standards-driven imaging. This means, first, creating digital image collections in standard file formats at a high enough quality to be worth preserving, and second, that sufficient documentation is captured to ensure that the images will continue to be usable, meaning that all necessary metadata is recorded in standard data structures and formats. One complication here is that it is as yet unclear exactly what all the necessary metadata for digital images is; some commentators are concerned that too little metadata is captured, others that too much is. The RLG preservation metadata elements are intended to capture the minimal information needed to preserve a digital image. Various groups have developed broader protocols or frameworks for digital preservation, such as the OAIS model discussed below.

The secondary preservation strategy is redundant storage: images and metadata should be copied as soon after they are created as is practicable. Multiple copies of assets should be stored on different media (most commonly, hard disks; magnetic tape, used for most automatic backup procedures; and optical media such as CD-ROMs) and in separate geographic locations; one of the most common causes of data loss is fire or water damage to storage devices in local mishaps or disasters. All media should be kept in secure archival conditions, with appropriate humidity, light, and temperature controls, in order to prolong their viable existence; additionally, all networked information should be protected by security protocols.

Such redundancy is in accordance with the principle that "lots of copies keep stuff safe," formalized by the LOCKSS system designed at Stanford University Libraries to safeguard Web journals. Refreshing, or the periodic duplication of files in the same format to combat media decay, damage, or obsolescence, essentially extends this principle. As yet, no robust preservation medium for digital data that is practical for general use has emerged. Because of the availability of analog media of proven longevity, some researchers suggest a hybrid approach to preservation, in which digital material is derived from analog material (such as photochemical intermediaries) or, alternatively, analog backups or copies of digital material are created. (In this context it is interesting to note the Rosetta Project, which aims to provide a near-permanent archive of one thousand

languages by recording them—as script readable with a powerful microscope rather than binary code—on micro-etched nickel disks with a two-thousand-year life expectancy.[8])

Migration, the periodic updating of files by resaving them in new formats so they can be read by new software, is where preservation starts to become more problematic. Reformatting allows files to continue to be read after their original format becomes defunct, but it involves transforming or changing the original data, and continued transformation risks introducing unacceptable information loss, corruption, and possible loss of functionality. One suggested method of mitigating this problem is **technology preservation**, which involves preserving the complete technical environment necessary to access files in their original format, including operating systems, original application software, media drives, and so forth. However, it is unlikely that this approach will allow the maintenance of viable systems over long periods of time.

Emulation takes the alternative approach of using software to simulate an original computer environment so that "old" files can be read correctly, presuming that their bit streams have been preserved. Emulation is a common practice in contemporary operating systems—for instance, code can be written to make programs designed for IBM-compatible or Wintel personal computers run on Macintoshes as they would in their native environment or to make programs designed for previous versions of an operating system function in newer versions. Research into emulation as a preservation strategy is ongoing. The initial research direction sought to emulate the entire original hardware and software environment and the functionalities it offered. More recent approaches have suggested that it is more economically feasible to use emulation to provide a viewing mechanism only, and that some loss of functionality is acceptable.

Emulation holds out the seductive possibility of preserving the original "look and feel" of archival data, but its large-scale practicality remains to be demonstrated. However, it has already had some success as a tool used in **digital archaeology**—the various methods and processes undertaken to recover data from damaged or obsolete media or hardware, defunct formats, or corrupted files when other preservation strategies have failed or perhaps never been attempted. Emulation was used to revive the BBC's Digital Domesday Book, an extremely ambitious project stored on 1980s-era interactive videodiscs that became inaccessible within fifteen years, in 2002. (The original Domesday Book, a record of William the Conqueror's survey of England compiled in 1086 by Norman monks, remains in good condition.) Emulation has also been moderately successful in reviving obsolete arcade videogames, re-creating solely through software the experience created by their original hardware-specific environments.

Re-creation is a concept developed in the world of born-digital multimedia or installation art. It postulates that if artists can describe their work in a way that is independent of any platform or medium, it will be possible to re-create it once its current medium becomes extinct. Such a description would require the development of a standard way of describing digital art analogous to musical notation.

All or some combination of these strategies can be carried out in-house, transferred to a third party such as a commercial data warehouse service, or done in collaboration with other groups and institutions through commercial or nonprofit resource-sharing initiatives. Resource sharing may be the only practical way to conduct preservation for many institutions in the long term. Examples of such initiatives include the OCLC (Online Computer Library Center, Inc.) digital archival service and the UK-based AHDS (Arts and Humanities Data Service) data deposit service, both of which provide long-term management, access, and preservation of digital assets. Transferring risk and responsibility to a third party, however, does not by itself guarantee preservation—the third party must be reliable and likely to continue in existence. *Trusted Digital Repositories: Attributes and Responsibilities*, a report written by RLG-OCLC in 2002, describes some of the characteristics that would be required in such a storage facility.

The OAIS reference model can potentially provide a common conceptual framework for the preservation and access of digital information, and thus a common ground for discussion, collaboration, and research in these areas. The model distills the entire life cycle of digital objects, from ingest through storage and display, down to a fundamental set of functions, relationships, and processes. It rests upon the central concept of "information packages," meaning the data or bit stream itself and the "representation information" that allows the interpretation of the bit stream as meaningful information. These may be regarded as analogous to the concepts of data and metadata.[9]

In reality, no one yet knows what the best preservation strategy or combination of strategies will be. Whichever is chosen, it will be necessary to run regular—annual or biannual—checks on data integrity and media stability and to be prepared to enter into a migration program within five or so years. It is advisable to retain original files over the long term if this is possible, but this will make further demands upon management and storage capacity. Master files should be afforded the maximum possible protection. Constant vigilance and the consistent use of open standards and system-independent formats, where possible, will be the best guarantee of the long-term viability of a digital image collection.