DIGITAL PRESERVATION: STANDING THE TEST OF TIME

Priscilla Caplan
Florida Center for Library Automation
5830 NW 39th Avenue
Gainesville, Florida 32606

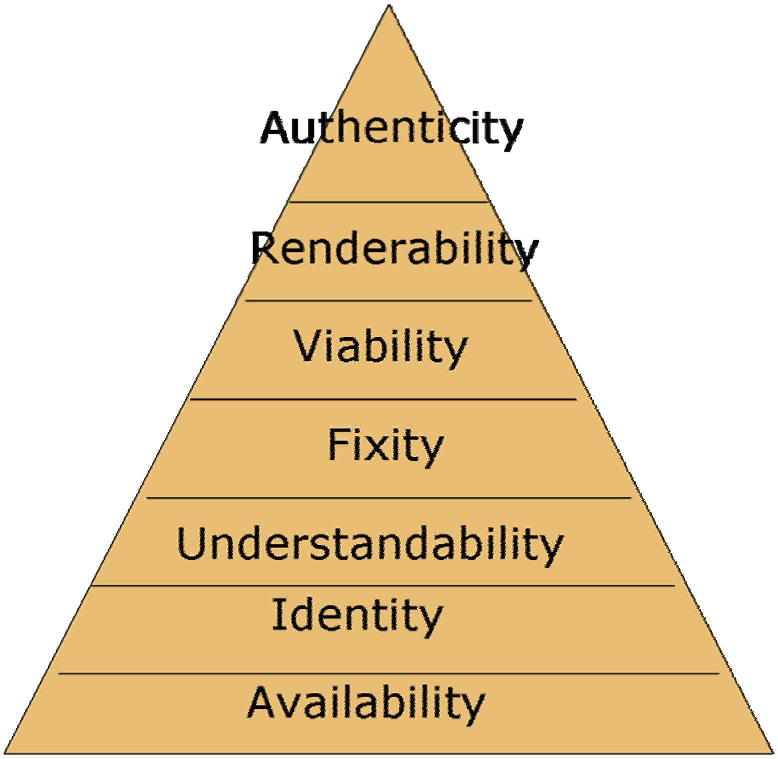
Abstract: The capture of digital materials, whether by web harvesting, deposit to an eprint repository, or other means, is a necessary first step in making those materials accessible over time. It is, however, only a first step, and additional actions must be taken to ensure the content is uncorrupted and remains usable as hardware and software platforms change. This paper provides an overview of the objectives of long-term preservation -- availability, identity, understandability, fixity, viability, renderability and authenticity – and methods for achieving them.

Keywords: Digital preservation

Digital preservation has been defined as the process of ensuring that a digital object remains accessible over the long-term. This raises some obvious questions, like what exactly is a digital object? What does it mean to be accessible? And how long is the long term? I'll beg the first two questions just now, but in the field the long-term is generally understood to be long enough for changes in technology to threaten the usability of the object.

One way to think about preservation is to imagine yourself some time in the future, needing some information that you know was recorded digitally in the past. What would you need to know? For example, imagine it's 2020 and for some reason you need to hear Nick Lowe sing "Maureen." Assuming you had the rights to do this, you'd certainly have to know how to get ahold of a copy of the audio file, what audio format it would be and what media it would be on. Then you would need a computer that could read the file from media and some software that could play it. If it was an old MP3 file, which by 2020 has been obsolete for years, you might have to reformat it into something your current software could play. And you might want to know for certain that what you're listening to really is Lowe's original recording and not a remix or digitally altered version.

These requirements can be formalized into a set of goals that the process of digital preservation is intended to ensure: availability, identity, understandability, fixity, viability, renderability and authenticity.



Availability can never be taken for granted. Papers and photographs can survive years of inattention and still be usable, but digital materials need active intervention to survive. There is no digital equivalent to that pile of old magazines in the attic. The first prerequisite for digital preservation is having a copy of the digital object in hand. This may be trivial, as when a library digitizes its own materials, or very difficult, when rights must be negotiated or content captured from the Web. It may involve organizing a cooperative effort to put materials into a shared repository such as the Aquatic Commons.

Similarly the object has to be identified and described sufficiently to be found in the future, and so future finders will know what they have. A persistent identifier should be assigned and linked to both the object and to a more detailed description if possible.

There is no descriptive metadata scheme specifically for preservation; any standard that fits the materials can be used.

Understandability is a tricky concept, because nothing is understandable to everyone. *Le Monde* is a cipher to me, because I don't speak French. Therefore the preservation community has adopted the concept of the "designated community." Each repository must define its designated community and attempt to ensure the future understandability of materials to that community, but not necessarily to others. Given that all digital files are strings of zeros and ones, understandability has both structural and semantic components. That is, it's important to know both that the first 16 bits of some data stream represent a positive integer, and that the integer represents a temperature reading in degrees Celsius. Sufficient documentation should be preserved to allow a future member of the designated community to both decode the file and understand the meaning.

Fixity is the quality of not being unintentionally altered or destroyed. Fixity can be threatened by insecure storage, transmission errors, or media degradation. It can be ensured by good security practices, checking checksums after transmission, and refreshing media periodically. Fixity is particularly important for digital objects, because unlike analog data, even a single bad bit can cause an entire file to be unusable.

Viability is the quality of being readable from media. Media obsolescence is a big problem, because technology changes so quickly. It is difficult today to find a microcomputer with a 5.25" floppy drive, and nearly impossible to find a drive for the 8" floppies that were ubiquitous in the 1970s. Even the 3.5" floppy has been largely superseded by flash and optical devices. Viability is easy to maintain by migrating data to more current media when necessary, but in order to do this the data has to be available and it has to be under a regime of active management. By the time you find those photo CDs you stashed away in a shoebox, it may be too late.

Renderability means that a digital object can be used in the way it was intended. Many types of objects are rendered by being displayed or played, but not all: SAS files, for example, are meant to be processed by a statistics application. Renderability is threatened by the obsolescence of software applications and the hardware that supports them. This is a more difficult problem than media obsolescence, and the topic of most preservation research. In some cases, old platforms can be emulated on more modern ones, a method well suited to video games and other interactive materials, while in other cases it might be possible to reformat (migrate) the content to a currently usable format.

Authenticity refers to the trustworthiness of the digital object. It is often defined rather anthropomorphically as the quality that an object is what it purports to be. Authenticity is enhanced if it is possible to verify the creator of the object and establish that the object has not been changed in an inadvertent or unauthorized way. Intentional changes, however, are allowed and may be unavoidable if the original format becomes unrenderable. Documentation of the creation and change history of a digital object is known as digital provenance, and if it is itself trustworthy, goes a long way towards verifying the authenticity of the object.

It is possible, and often advantageous, to distribute responsibility for digital preservation among different agencies and applications. The Aquatic Commons repository, for example, fulfills the functions of ensuring availability and identity. If contributors to the Aquatic Commons deposit documentation along with primary objects, it should contribute to future understandability as well.

The technical staff at the Florida Center for Library Automation (FCLA) who maintain the E-Prints software and storage for the Aquatic Commons are responsible for good security and data management practices that can ensure the fixity and viability of the content. Also, in a secondary way, they help ensure availability by providing backup and recovery facilities in the event the original files are lost.

Renderability and authenticity are more problematic. Sooner or later all digital file formats will become obsolete and require some preservation action. The extent of the problem can be limited by insisting on the use of more preservation-friendly formats, which generally means non-proprietary formats based on open standards with good documentation. Also, the more popular the format, and the more applications available to create and read it, the more likely it is that a good migration path will be available.

Some institutions are investing in building preservation systems. For the most part these are very large institutions like national libraries, or consortia of academic libraries. FCLA, for example, developed and runs the Florida Digital Archive for the use of all the libraries of the public university system of Florida. As the field matures, it is expected that consortially maintained facilities for digital preservation will become increasingly common, as will both for-profit and not for-profit third party repository services.