Active Preservation

Robert Sharpe[1] and Adrian Brown[2]

[1] Tessella, 26 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire OX14 3YS, United Kingdom
[2] Parliamentary Archives, Houses of Parliament, London, SW1A 0PW
(formerly The National Archives, Kew, Richmond, Surrey, TW9 4DU)

Abstract. In order to perform long-term digital preservation it is necessary to be (i) understand the technology of the material being stored, (ii) be able to decide whether this technology is obsolete (and, if so, what to do about it) and (iii) perform verifiable actions to remove the causes of this obsolescence (e.g., via format migration). This demonstration will show a real-life solution for dealing with these challenges. It is based off pioneering work performed mainly in conjunction with the UK National Archives' Seamless Flow programme and the Planets project[1] and is now deployed in a variety of national libraries and archives around the world.

Keywords: Digital Archiving and Preservation, Characterization, Preservation Planning, Migration.

1 Introduction

Modern libraries receive a large quantity of digital material. This needs to be preserved and yet suffers from a very different form of degradation risk than traditional material. This risk takes at least two forms:

- Storage media degradation. This problem can be mitigated by retrieving the content from volatile media to a central storage location and then applying an appropriate backup regime and policy of regular integrity checking.
- The ability to view digital objects requires a technical environment including information on the file format(s) involved, application software that can render these formats, an operating system that can support the application software and hardware on which to run the operating system. In other words, the ability to use digital objects depends on a stack of technical components with each component in this stack often rapidly becoming obsolete. There are three main approaches to dealing with this issue:
    - the "museum" approach which attempts to preserve all the components;
    - the emulation approach which accepts that some technical components need to change but attempts to preserve others (e.g., by

[1] The Planets project is co-financed by the European Union's Sixth Framework Programme for Research and Technological Development (FP6).

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 465–468, 2009.

preserving the original format, application software and operating system and emulating this operating system on new hardware);
- the migration approach which transforms the digital object to a new format and uses new application software running on an appropriate operating system / hardware platform.

The Open Archival Information Standard (OAIS) [1] has provided a conceptual framework for repository systems and a useful language to enable practical discussion of the problem. In particular, the standard makes a distinction between Information Objects (the conceptual entity that needs to be preserved) and Digital Objects (the physical entity that is initially created or is created as a result of some subsequent activity).

The Planets project, has built up a three-fold approach to addressing digital preservation [2]:

- Preservation Characterization. The need to characterize both Information Objects and Digital Objects to both determine the most appropriate actions to take and to provide a basis for validating those actions.
- Preservation Planning. The need to assess the preservation needs of digital material based on characterization information and plan any appropriate action.
- Preservation Action. The need to perform this action including verification of the resulting migration (or emulation).

We call this approach "Active Preservation" and distinguish it from "Passive Preservation" (the steps needed to preserve the original). In this demonstration a practical "Active Preservation" framework is introduced. This includes the ability to plug tools into this framework to deal with some formats and the ability to extend the supported toolset to both deal with other formats and improve existing tools.

2 Characterization

2.1 Technical Characterization

The first step towards preserving material that is ingested into the archive is to ensure that the files that constitute such material are technically characterized. This uses a framework created as part of the Planets project and involves four main steps:

- Identification of the format of every file using DROID [3, 4].
- Validation of that identification using a format-specific tool (e.g., Jhove [5]).
- Extraction of key properties about each file using a format-specific tool.
- The identification of embedded bytestreams within each file if it is a container format (e.g., bytestreams within a ZIP file). This, too, uses a format-specific tool. If a new bytestream is identified, it is then characterized in turn.

The framework allows for extension of new tools as they become available: they simply need to be wrapped. The PRONOM database service [3, 6] is used as a source of information for what to do at each step for each format after initial identification

and is automatically queried by the framework using web services. For example, PRONOM holds a prioritized list of identification, validation and property extraction tools for each format and also, importantly, determines which properties to measure and assigns each such property a unique identifier for future comparison.

2.2 Conceptual Characterization

Material then goes through a second stage of characterization that divides Information Objects into atomic conceptual units of preservation called "components". For example, a web site might be divided into its constituent documents (e.g., web pages and PDF documents), images etc. that are too numerous to be catalogued and described by humans. The characteristics of each of these components are measured by aggregating the properties of its constituent files. This is an important step since these components form the atomic units of migration whilst allowing the number or structure of files that manifest them to vary depending on the technology of the day.

Now that characterization has been completed, the material can be ingested into an archive.

3 Preservation Planning

At some time in the future, material may have become obsolete. The system uses PRONOM to monitor this obsolescence using a risk-based system. This allows each format to be assigned a risk score based on configurable criteria. This can be queried (either via a user interface or an automated web service) to determine which formats are at risk. It is also possible to specify a risk associated with format properties (e.g., Word documents with track changes on or containing macros might be considered to be a bigger risk than those without). The output of this process is thus a list of all the formats (or formats and property combinations) that are currently at risk.

This list can then be used to automatically search an archive to find out all the Information Objects whose current technological manifestation is at risk. These can then be dealt with one by one (or in a batch process). For each such Information Object manifestation, the system can ask PRONOM to determine the optimum migration pathway (and the tool to use). PRONOM also holds a list of all the component properties that should be invariant under a migration. The framework can accept a configurable degree of tolerance (i.e. allow for controlled loss of significant characteristics if this is unavoidable). Hence, the framework can also be used to create presentation copies (e.g., lower resolution images for transport over the web) in a controlled manner (i.e. with a known and measurable degree of degradation).

4 Preservation Action: Migration

Once preservation planning has been completed, the next step is then to carry out the migration. This involves running the selected tool and then re-characterizing the output to both discover the technical characteristics of the new files created and to check that all the components are still present, the relationships between them are intact and that the list of properties described above have indeed remained invariant. In addition,

there is the option to perform further tool-specific validation tests (e.g., for image migration compare the color distribution of the before and after images).

Once this process has been completed, the new manifestation can be ingested into the archive and, if appropriate, the old one marked as inactive so it does not seed further migrations (although the new one might at some time in the future).

5 Demonstration

The features described above will be demonstrated live showing the complete lifecycle of ingest, characterization, migration and object download using a range of Information Objects and Digital Objects that are relevant to the real data held by Libraries and Archives. This is based on software currently deployed at 7 libraries and archives.

6 Conclusion

Hence, the end result is a fully-automated digital preservation workflow provided in a framework that allows the addition of further characterization and migration tools as needed while allowing librarians and archivists to control detailed policy information and workflows.

References

1. Reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-B-1 Blue Book (January 2002), http://public.ccsds.org/publications/archive/650x0b1.pdf
2. Farquhar, A., Hockx-Yu: Planets: Integrated Services for Digital Preservation. Int. Journal of Digital Curation 2(2), 88–99 (2007)
3. Brown, A.: Automating preservation: New developments in the PRONOM service. RLG DigiNews 9(2) (2005)
4. DROID home page, http://droid.sourceforge.net
5. Jhove home page, http://hul.harvard.edu/jhove
6. PRONOM home page, http://www.nationalarchives.gov.uk/pronom