The PRONOM Service

A technical registry to support long-term preservation

Introduction

Maintaining access to digital objects depends upon a complex set of inter-related technical components, such as file formats, operating systems, character encoding schemes, and software tools. It is now widely recognised that technical registries, providing impartial, definitive information about the technical characteristics and dependencies of digital objects are an essential prerequisite for long-term preservation. This paper discusses current initiatives, with particular reference to the ongoing development of the PRONOM[1] service by the UK National Archives (TNA).

The need for registries

Electronic records pose many challenges for archivists, but these arise from a single underlying issue: access to a digital object is entirely dependent on technology. A file in a given format requires software to decode and display it; that software in turn requires a specific combination of hardware, operating systems, and other software to run. Equally, the storage media on which the file is stored requires its own combination of hardware, software and operating system in order to be accessed. Understanding this complex network of technical dependencies lies at the heart of any archiving programme for electronic records.

The technologies on which electronic records are so utterly dependent are constantly evolving: existing technologies are redeveloped in new versions, or become obsolete, and entirely new technologies emerge to replace them. This can happen at a very rapid rate, with new versions of software products being released on an annual basis. The challenge for the archivist is not only to understand the nature of these technical dependencies, but also to continually monitor changes which threaten the continued accessibility of electronic records.

In order to meet these challenges, TNA has developed a technical registry system called PRONOM. PRONOM is a resource for anyone requiring impartial and definitive information about the file formats, software products, operating systems, hardware components, and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value. PRONOM has been freely-available on the web since 2003 and the latest version, PRONOM 4, introduces a number of significant enhancements, including a substantial revision of the data model, which allows us to record detailed technical information about specific file formats, including links to the formal specifications where possible. For the first time, users can search for specific file formats, and retrieve a wide range of information on everything from

supported byte orders and compression methods, to copyright and patent restrictions.

Registry services may be required to support a number of different preservation activities, including technical characterisation of digital objects, preservation planning, and preservation actions such as migration. The role of a registry may also be more extensive than simply acting as a passive repository of information; TNA is developing PRONOM as an active, and indeed pro-active, component to enable the automation of its preservation services.

Supporting preservation

The active preservation of digital objects requires three major functions, operating in a cycle. Firstly, we need to understand the nature of the digital objects being stored. Once this is known we must identify and plan the preservation actions, such as format migration, which may be required, and when such actions need to be undertaken. Finally, the preservation plan must be enacted, and the results validated. Technical registries have a central role in supporting these three functions of characterisation, preservation planning, and migration.

TNA has recently initiated a major programme called Seamless Flow, which will integrate and automate processes for managing electronic records throughout their lifecycle, from creation, appraisal, selection and transfer from government departments, to preservation and dissemination by TNA. Central to this programme is the development of an active preservation capability, which will provide characterisation, preservation planning, and migration functions. These functions, and the role of the PRONOM registry in supporting them, are discussed in detail in the remainder of this section.

Characterisation

This underpins all subsequent preservation activities: if we don’t understand the precise technical characteristics of a digital object, we cannot hope to preserve it or make it accessible. The TNA characterisation methodology comprises three discrete stages: identification, validation, and property extraction. Identification determines the precise format of the object (e.g. PNG 1.0). Validation checks that the object is well-formed and valid against its formal specification (e.g. the W3C specification for PNG 1.0). Property extraction measures those properties of the object which are significant to its long-term preservation. These may be generic properties of the object format, or properties specific to a singular object, and can be divided into two categories:

- **Representation characteristics:** These are characteristics deriving from the particular technical manifestation of an object, which define the technical dependencies upon which access depends, and therefore determine the available preservation options. These will include both explicit characteristics, such as the file format specification to which the object conforms, character encoding schemes and compression algorithms employed, and implicit characteristics, such as the technical environment required to render the object, including hardware, software and operating system dependencies.
- **Inherent characteristics:** These are characteristics deriving from the underlying nature of the object itself, rather than any specific technical manifestation, which define the properties which must be preserved over time, and across multiple technical manifestations, in order to maintain the authenticity of the object. These may relate to the form of the object (e.g. the resolution of a raster image, the sample rate of an audio recording, or the fonts used in a document), or to its content (e.g. the date on which an email was received).

In both cases, these properties may be explicitly or implicitly identifiable, and may be derivable either from the object itself, or from an external source. For example, the compression algorithm employed in a raster image may be identified by a specific value within the object bitstream, or may be established by default through identification of the file format. Equally, an XML document may require an external schema to allow certain properties to be determined.

TNA is developing a characterisation service which provides these functions via a modular design, allowing combinations of new and existing tools to be utilised via a standard interface. The first such tool, to provide identification services, has been developed as part of the release of PRONOM 4. DROID (Digital Record Object IDentification) performs automated batch identification of file formats, using internal and external signatures to identify and report the specific file format versions of digital files. DROID first attempts to match a file against a list of internal signatures - specific patterns of bytes which can be used to identify a format. These signatures are expressed as sequences of hexadecimal values, and can also incorporate wildcard operators, providing a very flexible and expressive syntax. A match against an internal signature will result in a positive identification of the format. As a secondary method, DROID also attempts to match any external signatures, which are currently limited to file extensions, although support for other types, such as Macintosh data forks may be added in future. However, any identification based purely on an external signature is accorded a much lower priority.

The signatures are stored in an XML signature file, generated from information recorded in the PRONOM technical registry. Currently, this provides over 130 internal signatures and over 600 external signatures. New and updated signatures will be regularly added to PRONOM, and users can configure DROID to automatically download updated signature files via web services. Full documentation of the signature syntax and XML schemas are being made available on the PRONOM website.

DROID is designed to support batch processing of large numbers of files. It allows files and folders to be selected from a file system for identification, and saved as a list in XML or CSV format. After the identification process has been run, the results can be output in XML, CSV or printer-friendly (HTML) formats, for further processing or import into the appropriate preservation metadata scheme.

DROID is being made freely-available to download from the TNA website and, being written entirely in Java, is fully platform-independent. It provides both a graphical user interface and a command-line interface, for ease of integration with other systems.

For validation and property extraction, we intend to use existing third-party tools wherever possible, such as the JHOVE format validator[2] developed by Harvard University, and the National Library of New Zealand's metadata extractor[3], will be used.

PRONOM Unique Identifiers

In parallel to these developments, TNA will be implementing an extensible scheme of PRONOM Unique Identifiers (PUIDs), which will provide persistent, unique and unambiguous identifiers for file formats. Such identifiers are fundamental to the exchange and management of electronic objects, by allowing human or automated user agents to unambiguously identify, and share that identification of, the encoding format of an object. This is a virtue both of the inherent uniqueness of the identifier, and of its binding to a definitive description of the format in a file format registry, such as PRONOM. No existing, universally-applicable system provides for this. UNIX 'magic numbers' and Macintosh data-forks do provide some of this functionality, but the same is not true within Microsoft DOS or Windows environments. The three-character file extension is neither standardised nor unique, and is interpreted differently by different environments. Equally, the IANA MIME-type scheme does not provide sufficient granularity or coverage to satisfy the requirements for unique identifiers. The PUID scheme has been developed for the single purpose of providing such identifiers.

The new scheme of PUIDs has been adopted as the recommended encoding scheme for describing file formats in the latest version of the e-Government Metadata Standard[4]. This will mean that, for the first time, a consistent, persistent and highly-granular scheme for describing file formats will be in use across the UK government.

In order to allow PUIDs to be expressed as Uniform Resource Identifiers (URIs), and make those identifiers available for public use in Web-based description technologies, it is intended to register the PUID scheme as a namespace under the *info* URI scheme[5]. This scheme has been developed by the library and publishing communities to facilitate the referencing by URIs of information assets which have identifiers in public namespaces but have no

representation within the URI allocation. It provides a simple URI registration mechanism to support the referencing of public information assets in advance of any possible subsequent URI scheme or URN namespace application. A typical PUID might be expressed as an *info* URI as follows:

```
info:pronom/fmt/42
```

In future, the scheme may be extended to include other technical components described by PRONOM, such as operating systems and codecs.

Preservation planning

Preservation planning forms the decision-making heart of any preservation service. Its role is to identify and monitor technological changes and their potential impacts on the electronic records stored in a repository, and to develop preservation strategies to mitigate the impact of these technological changes. TNA's long-term preservation strategy is based on object migration, coupled with retention of all records in their original formats. Our system will therefore focus on the development of migration pathways for the automatic conversion of electronic records to new formats as required for preservation or presentation purposes.

The TNA approach to preservation planning comprises four main functions:

- **Risk assessment:** Every object will be subject to a risk assessment at the point of ingest into the repository. This risk assessment will be based upon a set of standard criteria, chosen as key indicators of the current risks posed to the continuing accessibility of the object. These criteria can be divided into generic and specific risk factors. Generic risk factors are common to all objects in a given format, such as the degree of public disclosure of the format specification, or the current diversity of software support, and can be directly calculated by reference to information stored in PRONOM. Specific risk factors are those which relate to a single instance of an object, such as the presence of macros in a Word document, or the use of external fonts in a PDF file. These will need to be identified using characterisation tools.

  The result of the risk assessment will be used to determine the urgency of preservation action: a low risk may simply indicate that the risk assessment should be recalculated at a future date, whereas a high risk would trigger immediate action.

- **Technology watch:** Technology watch describes the process of continually monitoring technological change, which will result in updates to the content of the technical registry. These updates may change the risk criteria described above. For example, the cessation of support for a particular software product might alter the risk associated with formats supported by that software. PRONOM already records

information about the product support lifecycle for the software tools required to create or render electronic records and, in PRONOM 4, this support information has been extended to cover other technical components such as file formats. This information will play a vital role in making decisions about when to migrate.

- **Impact assessment:** This process analyses the results of risk assessments and technology watch, to determine their actual impact upon the objects stored in the repository. For example, a change to a risk criterion caused by technology watch will require new risk assessments to be undertaken on all affected objects. Equally, objects which were previously assessed as low risk will need to be periodically re-assessed: the impact assessment function will determine when this is the case, and identify the affected objects.

- **Migration pathway generation:** The final stage of preservation planning is to determine the detailed preservation action required. This will take the form of a migration pathway, describing the precise steps necessary to migrate an object from one technical manifestation to another. A migration pathway may therefore be defined in terms of a sequence of migration tools, together with any necessary configuration parameters. Information about migration tools is recorded in PRONOM, enabling migration pathways to be defined with specific reference to the registry. In addition, information about the formats which particular software can render and create will form the basis for identifying potential migration pathways.

  Each potential migration pathway will then need to be rigorously tested, and those certified as suitable for use will then be recorded in the registry for future use.

Migration

The final stage in the preservation process is to actually perform the selected migration process. In an operational environment, automation is the only viable approach to this, and TNA's work is therefore focused on the development of a framework for controlling automated migration services. Specific tools which have been tested and approved to provide these services will then be deployed within this framework as required. We will seek to use existing conversion tools wherever possible. The migration service will export electronic records previously identified as requiring migration from our digital repository, automatically migrate them using the selected migration pathway, and accession the migrated records into the digital repository as new manifestations.

The documentation of these preservation actions will perform a vital role in establishing the continuing authenticity of electronic records, and TNA is developing a comprehensive data model to enable multiple manifestations of records, and the migration pathways which link them, to be documented and

managed within our digital repository. This model also takes into account the issue of dependencies between digital objects. This is of particular importance, given that many electronic records are compound objects, composed of many inter-related files, such as websites, or office documents with linked or embedded content. It is essential to understand the nature of these dependencies, in order to predict the full impact of a preservation action such as migration. To give a simple example, the migration of images in a web page from GIF to PNG will also necessitate updating of the HTML image references. Thus, a migration pathway may actually comprise a complex set of format conversion processes, emendation processes, and management of the associations between objects.

Finally, the results of the migration must be validated. For real-world use, involving batch migration of large numbers of objects, such validation will need to be automated. We have already seen how the significant properties of an object can be derived automatically, using characterisation tools. TNA's validation methodology will therefore involve the characterisation of newly-migrated objects, and automated comparison of its significant properties with those of the source object. For example, this might verify that the sampling frequency, bit rate, number of channels, and duration of an audio recording remain constant before and after migration.

Other initiatives

The need for technical registries to support digital preservation is widely recognised. Other planned future registry services include the Global Digital Format Registry[6], an initiative being co-ordinated by the Digital Libraries Federation, with international participation including TNA. The GDFR would provide a federated global network of format registries and it is anticipated that, if realised, PRONOM would become a major node within this network. Significant effort has therefore been made to ensure that the GDFR and PRONOM data models are closely aligned, to support future interoperability. Funding is currently being sought from the Mellon Foundation to develop a prototype GDFR registry service.

The potential for external systems to exploit PRONOM services is already being demonstrated through a JISC-funded project with Southampton University. The PRESERV[7] project is integrating the DROID tool into a new ingest module for the Eprints[8] digital repository system, which is used by over 130 archives worldwide.

The future

Although PRONOM is central to TNA's own digital preservation services, TNA is also committed to making it available as a resource for the wider digital preservation community. This includes UK government departments, which may be required to sustain digital records over long periods for business purposes, and other digital repositories around the world. Recognition of

PRONOM's wider utility prompted the decision to make it available on the web on 2003, and TNA now intends to develop a number of mechanisms to support the automated remote usage of PRONOM services. We plan to create a number of interfaces to allow the exposure of registry content for remote querying and retrieval, such as REST and SOAP. We are also investigating the possibility of exposing PRONOM as a metadata repository to support the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[9].

TNA will also shortly be implementing the PUID scheme, and developing the service to resolve PUID URIs to the corresponding records in the PRONOM database.

However, the most significant developments will take place within TNA's Seamless Flow programme, as we build the characterisation, preservation planning, and migration services described previously. As these are developed, the PRONOM technical registry will truly take its place at the very heart of the digital preservation process.

Adrian Brown
Head of Digital Preservation
The National Archives
Kew, Richmond
Surrey, TW9 4DU
UK

---

[1] See http://www.nationalarchives.gov.uk/pronom/
[2] See http://hul.harvard.edu/jhove/
[3] See http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction
[4] See http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=872
[5] See http://info-uri.info/
[6] See http://hul.harvard.edu/gdfr/
[7] See http://preserv.eprints.org/
[8] See http://www.eprints.org/
[9] See http://www.openarchives.org/