The International Journal of Digital Curation
Issue 1, Volume 2 | 2007

Developing Practical Approaches to Active Preservation

Adrian Brown,
Head of Digital Preservation,
The National Archives, UK

June 2007

Abstract

The National Archives is developing a range of practical solutions to the active preservation of electronic records, using an extensible service-oriented architecture and a central technical registry (PRONOM). This paper describes TNA's methodologies for characterisation, preservation planning, and preservation action, the technologies being adopted to implement them, and the role of PRONOM in supporting these services. It describes how this approach fits with international research programmes, and the types of preservation service which TNA may be able to provide externally in the future.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

The National Archives (TNA) has been actively collecting, preserving, and making available electronic records for nearly 10 years. From the initial establishment of a contracted-out service to preserve government datasets in 1997,[1] TNA's activities have expanded with the deployment of its own digital archive in 2003,[2] the world's first publicly available format registry (PRONOM) in 2004,[3] a web archiving programme, also in 2004,[4] and a pilot web presentation system (Electronic Records Online) in 2005.[5] TNA's approach to digital preservation is founded on two fundamental activities: *passive preservation*, which provides secure storage, and *active preservation*, which ensures the continued accessibility of the stored records over time, and across changing technologies.

TNA's Digital Archive already provides a passive preservation capability which is scalable to petabyte levels. With an ever-increasing variety of electronic records to preserve, the next major challenge for TNA was therefore to develop an active preservation capability. Based on the existing PRONOM registry, this work is being taken forward as part of the wider Seamless Flow Programme, which is developing end-to-end processes for managing electronic records.[6] However, TNA is also a participant in a number of national and international research projects and, perhaps most notably, is leading research on characterisation within the EU-funded Planets Project.[7] We are therefore working to ensure that the results of Seamless Flow are both compatible with, and contribute to, the services which will be developed by Planets.

TNA is also required to support broader sustainability requirements for electronic records across UK government, and it is intended that many of the preservation services developed for in-house use should also be made available as shared services within the UK public sector and beyond. The first components of TNA's active preservation framework, including publicly accessible services, will be delivered in 2007. Further elements are scheduled to be developed within Seamless Flow and Planets before 2010.

1 See www.ndad.nationalarchives.gov.uk/
2 See www.nationalarchives.gov.uk/preservation/digitalarchive/
3 See www.nationalarchives.gov.uk/pronom/
4 See www.nationalarchives.gov.uk/preservation/webarchive/
5 See www.nationalarchives.gov.uk/ero/
6 See www.nationalarchives.gov.uk/electronicrecords/seamless_flow/
7 See www.planets-project.eu/

The Active Preservation Framework

TNA's active preservation methodology comprises three main functions operating in a cycle, supported by a central technical registry, as illustrated in Figure 1 below:

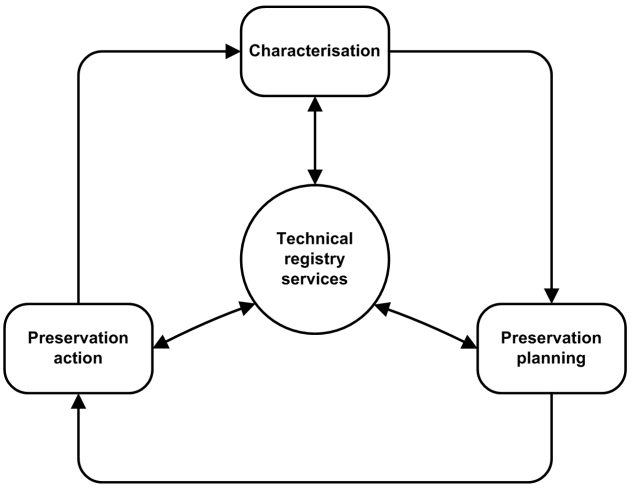

Figure 1 The active preservation cycle

The *characterisation* function measures the properties of digital objects which are significant to their long-term preservation; *preservation planning* determines the appropriate preservation actions to be undertaken; and *preservation action* enacts the results of preservation planning, transforming the objects to ensure their continued accessibility. The technical registry provides the knowledge base required to support these three functions.

The active preservation system is being developed using a service-oriented architecture, whereby the various functions are exposed as web services. A hierarchy of services has been defined, to offer maximum flexibility. Thus, for example, the 'characterise manifestation' service can be called, but it is also possible to call directly the individual component services, such as 'identify files', 'validate files' and 'extract file properties'. Each web service call typically takes as parameters the location of a set of files to be processed, and the location of the accompanying XML metadata. The framework is being developed using Java J2EE, to support platform independence, and is designed to allow the re-use of third-party tools, through a standard interface. Simple wrappers are being written which map the APIs of existing tools to this interface, allowing them to be deployed within the framework. A workflow engine will control the invoking of active preservation services. This workflow is responsible for managing all processing of electronic records, from selection and transfer through to delivery to users. Work has already begun on developing and configuring the workflow system, initially to support the transfer of new records to TNA but, by the end of 2007, to control the entire preservation process.

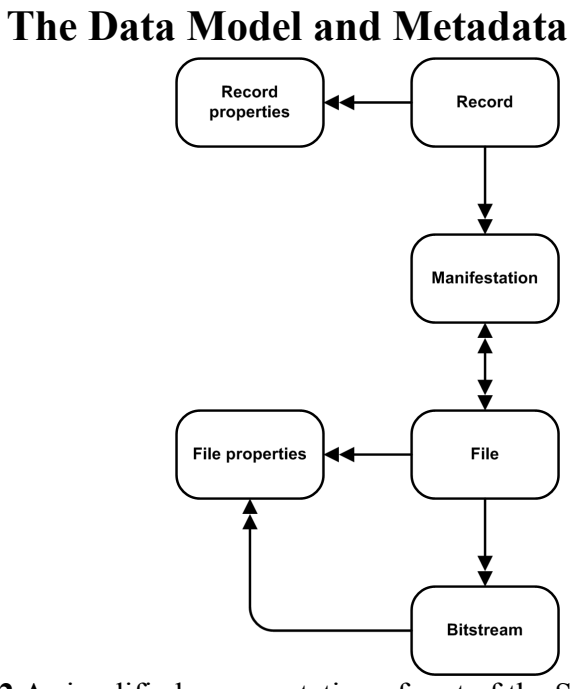

Figure 2 A simplified representation of part of the Seamless Flow data model

A robust and flexible data model and accompanying metadata scheme are critical components of any preservation service. The TNA model was originally developed in 2003 and, as part of Seamless Flow, we have taken the opportunity to implement substantial revisions, particularly relating to the modelling of preservation processes and significant properties. A detailed description of the data model is beyond the scope of this paper. However, Figure 2 provides a very simplified overview of some of the key entities.

The scheme was developed prior to the emergence of the PREMIS (PREservation Metadata Implementation Strategies) standard (PREMIS Working Group, 2005). However, TNA has reviewed and mapped the scheme against PREMIS: this has been useful in identifying much common ground, as well as some important differences, such as the way PREMIS addresses significant properties. The record entity describes the conceptual electronic records being managed, and corresponds to the PREMIS information object. Manifestations describe the particular technical instantiations of that conceptual record, and are equivalent to the PREMIS concept of representations. A given manifestation is composed of one or more files and bitstreams, which are defined in accordance with PREMIS. Both re-cords and files may have properties associated with them, the uses of which are elaborated in Preservation Planning below.

The scheme is expressed as a set of XML schemas, representing the key entities in the data model. This modularity allows flexibility in the processing and exchange of metadata between systems, and also supports simple extensibility, a key requirement for TNA. The metadata scheme is a key component of the generic interfaces which TNA has developed as part of its service-oriented architecture. The metadata defined in the data model is stored in a Generic Metadata Management System (GMMS), a database which acts as the authoritative source for all metadata. A snapshot of the metadata for each record is also stored in the Digital Object Store (DOS) at the point of ingest, for additional security.

Characterisation

Characterisation is an essential precursor to preservation. It provides the information required to make preservation planning decisions about digital objects, and to validate the results of preservation actions. The characterisation function comprises three distinct processes, which can each be applied either to an entire manifestation or individual files.

Identification

This process identifies the precise file format version of a file and updates the file metadata to record this and the characterisation event. TNA currently uses DROID as its identification tool,[8] and describes formats in terms of PRONOM Unique Identifiers (PUIDs), which relate to detailed format records in the PRONOM registry (Brown, 2006). The first version of DROID was released by TNA in 2005, and an enhanced version was released under an open source licence in 2006. A further major release is planned for 2007, and will provide improved performance and accuracy of identification.

Validation

The validation process determines whether a file is valid with regard to the format specification determined in the identification stage. It allows for two levels of validation, as follows:

- Well formed: the file is syntactically correct with regard to the specification.
- Valid: The file is well formed and conforms to additional semantic constraints.

PRONOM records the validation tools which are available, and the formats which they can validate. Based on the identification results, the characterisation framework can automatically deploy the appropriate validation tool and process the file.

Property Extraction

The property extraction process measures the properties of a digital object which are significant to its preservation. These can be categorised as:

- Technical properties, which are associated with files and inform the preservation planning process. The most basic of these is the format of the object, but they may include any property relating to the technical representation, such as the compression algorithm used in a digital image, or the nature of macros contained in an office document.
- Inherent properties, which relate to the underlying records to be preserved. These properties are independent of any particular technical representation, and represent the qualities of the object which have been identified as essential to preserve over time, and across preservation actions such as migration. Such properties might include the time duration of a video clip, or the textual content of a word-processed document.

TNA is implementing a flexible and extensible approach to defining and measuring significant properties. It is widely recognised that this is a very complex problem, and that very little research has so far been undertaken in this area. The InterPares Project is one exception, and has been applying the traditional archival science of diplomatics to the digital environment for many years (see Duranti &

8 See http://droid.sourceforge.net/

MacNeil, 1996). TNA, in partnership with the AHDS, has recently secured funding from JISC to undertake a two-year research project to develop and test methodologies for describing, measuring and comparing the significant properties of a wide range of digital object types. InSPECT (Investigating the Significant Properties of Electronic Content over Time) will form the basis for developing a standard and robust approach to significant properties at TNA.[9] In parallel, TNA is also collaborating with NARA to compare the very similar approach being taken within its ERA programme.[10]

In the first instance, TNA will use Harvard University's JHOVE tool for format validation and property extraction.[11] The open source Java POI library is also being investigated as a possible tool for validating and extracting properties from Microsoft Office documents, while the Java JAXP API may be used for generic validation of any XML-based format against its schema. A lightweight wrapper will be written for each tool, which will enable it to be invoked by the TNA characterisation framework, and will translate the resultant output into the TNA XML metadata schema.

Characterisation Framework

The characterisation process will be controlled through a Characterisation Framework, forming one element of the broader Active Preservation Framework described above. This incorporates a Process Control component, which initiates, allocates resources to, and manages the various sub-processes involved. A screenshot of the prototype system is illustrated below:

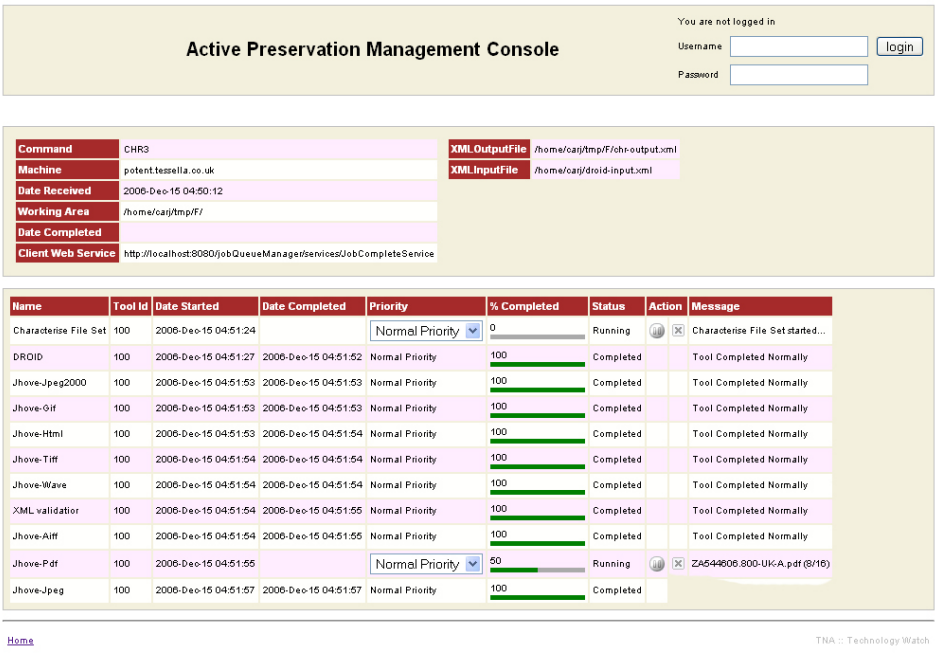

Figure 3 Prototype Characterisation Framework in operation

Here, a characterisation job has been initiated through a web service call, specifying the location of a set of files and their accompanying metadata. The Framework firstly initiates the Identification service, which uses DROID. Upon completion, the Framework queries PRONOM to discover appropriate validation and

9 See http://ahds.ac.uk/about/projects/inspect/
10 See www.archives.gov/era/
11 See http://hul.harvard.edu/jhove/

property extraction tools for each identified format, and initiates each tool. In the example, various JHOVE modules have been implemented as tools, and are at various stages of completion. Once all jobs have been completed, the Framework will output updated metadata for each object. The management console screen illustrated also allows a user to pause, restart and cancel jobs, and configure the system.

Preservation Planning

The preservation planning framework is the decision-making core of the active preservation system. It determines what preservation actions should be applied to which objects, and the appropriate time to apply them. Preservation actions are required to mitigate internal or external events which threaten the continued accessibility of a digital object, and TNA has therefore developed a risk-based approach to preservation planning. The TNA data model supports the storage of multiple manifestations of a record, which may be driven variously by the needs of preservation and presentation. However, the preservation planning function must also support this differentiation, by allowing the definition of separate sets of risk factors for each case.

The TNA risk assessment methodology also considers two distinct types of risk: format risks are generic to a particular format, whereas instance risks are specific to an individual digital object. For the sake of simplicity, TNA is currently limiting the consideration of risk to the file format – risks arising from other elements of the representation network, such as software and hardware, can effectively be aggregated at the format level. Format risks are calculated and recorded in PRONOM, using information about key format properties. Typical properties which might be used to calculate format risks include the number of software tools available which support that format, and the openness of the format. Instance risks are calculated from the measurement of specific properties of the file, using a similar method. Examples of file properties which might affect instance risk include the presence of macros in an office document, or the compression algorithm used in an image file.

The preservation and presentation risks for each object are first calculated during the ingest process, immediately following characterisation. If an object's preservation or presentation risk score is above the designated threshold value then this automatically triggers the generation of the appropriate preservation or presentation plan, and subsequent execution of the migration process.

The process of technology watch, whereby changes in the technological landscape which impact on the continued accessibility of digital objects are monitored, can therefore be defined in very concrete terms: technological changes are captured as updates to the content of PRONOM. Any such update which affects the properties of a format initiates an automatic recalculation of the associated risk score. If this causes the risk score to rise above the risk threshold, it triggers the impact assessment service, which interrogates the GMMS to identify all stored objects affected by the change. If any such objects are identified, these are then passed to the final stage of the process – the generation of the appropriate preservation plan.

The preservation plan is the ultimate output of the planning process. It comprises a migration pathway, defining the precise set of preservation actions to be undertaken, and the list of digital objects upon which that pathway should be executed. The

PRONOM registry will provide support in two key areas. Firstly, it will identify the most appropriate target format for migration, i.e. the one which supports the designated significant properties of the source at the lowest possible risk. Secondly, PRONOM contains information about the software tools which are capable of reading and writing each format, and can therefore identify the possible set of migration pathways between a given source and target format.

The candidate migration pathways must then be tested, and the results validated using the defined significant properties. The outcome of these experiments will be the selection of the preferred migration pathway, and the certification of that pathway as suitable for use in the given scenario.

Preservation Action

The creation of an approved preservation plan triggers the preservation action function. The objects specified in the plan are automatically exported from the DOS, and processed in accordance with the migration pathway.

In common with the other parts of the active preservation framework, the preservation planning framework will provide a lightweight service for automatically deploying preservation action tools. Execution comprises three steps. Firstly, the migration tools specified in the migration pathway will be automatically deployed to convert the designated files and generate a new manifestation. The migrated files will then be characterised, using the standard characterisation process. Finally, they will be validated by comparing the characterisation results of the migrated manifestation with those of the original.

The PRONOM Technical Registry

TNA's PRONOM technical registry will form the core of the active preservation system. PRONOM already provides a freely available online source of technical information about file formats, and the software tools required to support and process them. Within Seamless Flow and Planets, PRONOM is being enhanced to provide a knowledge base which will support characterisation, preservation planning and preservation action. PRONOM already supports characterisation, as the source of the signatures used by DROID, and through the allocation and resolution of PUIDs. It also offers some degree of support for preservation planning, and for the selection of tools for characterisation and preservation action. In the near future, these roles will be enhanced to support full automation.

At present, PRONOM is primarily designed for human use, but it is recognised that support for machine-to-machine interaction will be essential to drive automated processes. TNA is therefore developing SOAP and REST interfaces to PRONOM, which will allow its functionality to be directly accessed by software systems. Where appropriate, these interfaces will be made available externally, allowing anyone to utilise PRONOM services such as risk assessments and identification of appropriate migration pathways. TNA is also developing a persistent resolution service for PRONOM Unique Identifiers. This will enable PUIDs to be resolved to the relevant PRONOM record in either human-readable XHTML or machine-readable XML.

A number of other major institutions have expressed interest in using PRONOM

as a key element of their own preservation systems, and TNA is working with them to develop an appropriate framework for using PRONOM, including data sharing and jointly funded future development. More widely, TNA is investigating how best to support demand for the wider adoption of PRONOM, both in terms of the content and the technologies. As part of the JISC-funded PRESERV Project,[12] TNA has been working with Southampton University, Oxford University and the British Library to integrate DROID and PRONOM with the Eprints digital repository software. This has not only enabled the automatic identification of file formats on ingest to the repository, but has also provided a format profiling component to the Registry of Open Access Repositories (ROAR).[13]

Other organisations are also active in the development of various types of technical registry. The Digital Curation Centre has been working for some time on its Representation Information Registry Repository,[14] and has been collaborating with TNA on compatibility and data sharing issues. TNA is also actively participating in the Global Digital Format Registry Project,[15] which aims to develop a global network of file format registries. Although narrower in scope than PRONOM, GDFR has the potential to address some key issues relating to governance, persistence and interoperability. If the project bears fruit, it is anticipated that PRONOM will participate as one of the first nodes in the GDFR network.

Conclusions

By the end of 2007, TNA will have established the core components required to provide automated active preservation for its collections on an operational basis. This framework should provide a practical basis for future extension, and allow TNA to expand the range of preservation tools and services which are available to support external digital repositories.

References

Brown, A. (2006). *The PRONOM PUID Scheme: A Scheme of Persistent Unique Identifiers for Representation Information* (Digital Preservation Technical Paper 2). London: The National Archives.

Duranti, L., & MacNeil, H. (1996). The Protection of the Integrity of Electronic Records. An Overview of the UBC-MAS Research Project. *Archivaria*, *42,* 46-67.

PREMIS Working Group. (2005). *Data Dictionary for Preservation Metadata.* Online Computer Library Center and RLG.

12 See http://preserv.eprints.org
13 See http://archives.eprints.org/
14 See http://registry.dcc.ac.uk/omar/
15 See http://hul.harvard.edu/gdfr/