PRESERVING TO ENSURE ACCESS

Adrian Brown

Since 1999, the UK government has established a major e-Government agenda, and has developed a detailed framework of policies to support this. The most important policy is the e-Government Interoperability Framework, which defines standards and procedures which all government departments must follow, to support interoperability between all e-Government systems. The framework also contains a number of more detailed technical policies. There is a standard for metadata in e-Government, which is based on the Dublin Core metadata standard. Alongside this, the Integrated Public Sector Vocabulary provides a standard authority file for the use of these metadata elements. The Government Data Standards Catalogue defines data standards for particular specialised uses, such as postal addresses. The Technical Standards Catalogue defines detailed technical standards to be used. Wherever possible, these are open standards, and include XML for information exchange, and XHTML for web pages. XML is a core technology for data exchange with UK government, and a standard library of database schemas has been developed to support interoperability. The Framework also advocates the use of open source software as an alternative to proprietary software, although its use is not mandatory. Equally, software developed by government departments can be released under an open source license.

The UK government is also developing the next generation of the national information infrastructure which is required to support UK research. The Department of Trade and Industry is coordinating this funding, which will cover the period 2007-2014. As part of this process, a number of expert working groups have been established, to advice on funding priorities. Once such group, on which TNA is represented, is advising on funding priorities for digital preservation and Curation. The group has recommended funding for this area of c. US $300 million, although we wait to see if this will be accepted. This work will provide much of the basic infrastructure and common services required to support preservation in the UK public sector.

To further support e-Government, The National Archives has developed a range of standards for electronic records management. In 2002, we developed a set of functional requirements for electronic records management systems, including a metadata standard,

which has been incorporated into the wider e-Government Metadata Standard. We have a programme to evaluate Electronic Records Management software, and certify whether or not it complies with our functional requirements. We are currently contributing to the development of the next version of the MoReq standard, which is likely to become the European standard for Electronic Records Management systems. Once this has been finalised, it will replace our current standard.

We have also defined requirements for sustaining electronic records over time. These describe the requirements for sustaining authentic records for long time periods, for use by government departments. Finally, we publish a range of Electronic Records Toolkits, which provide detailed guidance on particular aspects of electronic records management, such as email management or appraisal of electronic records. All of these standards are freely available on our website.

e-Government brings many challenges for The National Archives. We are already considering the next generation of Electronic Records Management Systems, and the impact which these will have. These will provide much more advanced functionality, which will be more closely integrated with standard operating systems and desktop applications, and will be much less visible to users. We face the challenge of supporting government to sustain its electronic records for business purposes, once they have ceased being in active use – we refer to these as semi-current records. Some of these records will be selected for transfer to The National Archives, and we must develop methods to allow this transfer to happen at the appropriate time. We also have the challenge of developing the means to preserve these records for the long term, and to deliver them to our users.

To do so, we have begun a major programme called 'Seamless Flow'. This will develop end-to-end processes for managing electronic records at The National Archives, beginning with appraisal and selection, and moving through transfer and preservation, to delivery to our users. The approach we are taking is modular – we will not deliver a single system which provides all of these functions, but rather develop that functionality incrementally over the next two years. The systems must be scalable, as the volumes of electronic records we must manage increase over time. Although we will be undertaking a considerable amount of software development, Seamless Flow is as much about processes as systems – we have spent a lot of time developing the workflow, and the operating procedures which staff will use. However, we are seeking to achieve the highest possible level of automation – it is simply not feasible to undertake many of these processes manually, given the volumes of records.

The business case for undertaking the Seamless Flow programme is based on a number of important factors. The volumes of records we will receive are beyond our current capacity to process. The introduction of Freedom of Information legislation has changed the means by which access to records is regulated. There is an increased demand for online access to our records, and our current systems cannot scale up to meet these demands.

The Seamless Flow programme is composed of a number of separate projects, covering different aspects of the electronic records workflow. The first project addresses how records are created and managed in government departments. The next deals with appraisal and selection of records. Another project is developing systems for transferring records to The National Archives. Two further projects are addressing the long-term storage and preservation of records. The final projects deal with resource discovery, and the delivery of records to users. Each of these projects is described in more detail below.

The first project is addressing the creation and management of records within government departments, which we refer to as semi-current records. It is addressing the requirements for sustaining those records for business purposes, until they can either be destroyed, or transferred to The National Archives. It has also developed a custodial policy for electronic records, which describes when departments must transfer records, and the point at which those records come under the custody of The National Archives. This is particularly important under Freedom of Information legislation, since it is the custodian of the record which is responsible for answering requests for information. Clarifying changes in custody is therefore vital. As part of this, we are developing standard transfer agreements with departments, which describe the types of records which will be transferred, and when this will happen.

Another project is examining how we should carry out appraisal and selection for electronic records. Our traditional approach, in which reviewers perform file-by-file manual review of records, is not feasible when we need to appraise thousands or millions of electronic documents. We have therefore developed a macro-appraisal methodology, where records are appraised at a higher level than before. We are now testing this approach with a number of departments, and developing an appraisal toolkit to support departments.

Once records have been appraised and selected for transfer, we must be able to facilitate the actual transfer process. The records will be catalogued by departments prior to transfer, with support from our own records management staff. The records and metadata will then be transferred to The National Archives, using either physical media or online transfer.

Once the records have been received, we will undertake a number of 'pre-accession' processes, before we formally accept custody. The records are quarantined, while we perform anti-virus checks, validation, and quality control on the records and cataloguing. Only after these processes have been completed are the records formally accessioned into the Digital Archive. We currently have to perform many of these processes manually, but we are currently developing a pilot system to automate them, which will be launched in April 2007.

The next project provides for the secure storage of the records which have been accessioned. It is based on enhancement of our existing digital archive system, and includes improved functionality for checking the integrity of records, to safeguard against accidental corruption or deliberate alteration, disaster recovery, and refreshment onto new storage media over time. The new system will also provide improved support for managing different technical manifestations of electronic records, which we create by migrating them to new formats over time, to ensure their continued accessibility.

This active preservation is being addresses by a separate project, called Technology Watch. This project is developing new tools to technically characterise electronic records, for example to identify the file format used. We have already developed a software tool, called DROID, to perform this function, which is freely available from our website. Once we have this information, we can plan the appropriate preservation strategies to use. Our principle preservation strategy is to migrate records to new formats over time, to ensure their continued accessibility, and we are developing systems which will enable us to identify and test the appropriate methods for performing these conversions. Once we have developed a suitable plan, we need to enact this, and we are therefore developing other systems to actually perform the automated migration of records to new formats.

Some records need to be redacted, by securely removing certain information from the record. We are therefore developing and testing tools which will allow us to do this, and to ensure that the information we remove cannot be retrieved by users.

All of these systems will be supported by our technical registry, called PRONOM, which is already freely available in our website. PRONOM contains the detailed technical information about file formats, software tools etc., which we need to be able to perform processes such as characterisation, preservation planning, and format migration. We are also using PRONOM to provide advice to government departments on the types of file formats which they should use. We have developed criteria for selecting formats which can be more easily preserved. These criteria include the openness of the format specification, how widely adopted the format is, and how well supported it is by current software, the ease with which

the format can be converted to other formats, and whether there are any intellectual property issues.

Another project is addressing how our users can discover the electronic resources which we make available. We are currently finalising a new metadata scheme and data model for cataloguing our electronic records, and a new system for referencing those records. We have also just launched a new global search engine on our website. This uses the Autonomy search engine, and allows our users to search not only our web pages, but also 11 separate catalogue databases for our collections. However, although providing the means to search for records is vital, we consider the ability to browse through the record hierarchy to be equally important, allowing users to discover records within their context.

Once a user has successfully found the record they wish to access, we must deliver it to them. We need to develop two separate presentation systems: a web delivery system to allow the public to view our open records, and a separate secure system to allow government departments to access their closed records. The first version of the open delivery system, Electronic Records Online, was released in 2005, and a new version, which includes the ability to search the records using our new search engine, is currently under development. The system for delivering closed records will be developed as a future version of Electronic Records Online.

The systems which we are developing in Seamless Flow are based on the OAIS model. In 2005, we undertook a joint research project with the UK Data Archive, to assess our respective levels of compliance with the OAIS reference model and the METS metadata standard. Our conclusion was that the TNA system was fully compliant with the high-level OAIS model, and largely compliant at the lower levels. However, we identified a number of areas which the OAIS model does not address, or where it does not fit well with our requirements. For example, it is widely recognised that OAIS does not yet address the detailed processes for active preservation. It also does not cover security issue, with regard to user access to records. However, we have found OAIS to be very useful as a baseline model for our systems. The results of this research have now been fully published, and are available online.

We are also very interested in the Trusted Digital Repositories concept, being developed by RLG and NARA. This is based on the OAIS model, and is developing a practical methodology for assessing the capabilities of a digital archive against a standard benchmark, and for accrediting it as meeting this standard. We have reviewed the draft

accreditation checklist, and are considering how TNA could use this approach within its role as an inspection body for other archives in the UK.

Our systems are not being developed in isolation, and we engage in a good deal of collaborative research with many partners. We have just been awarded US $13 million in European funding for a 4-year project, called PLANETS, to develop new systems and methodologies for digital preservation. We are part of a 15-member consortium, which includes many European national libraries and archives, universities, and commercial partners which include Microsoft and IBM. The project will develop practical preservation services for characterising digital objects, preservation planning, and preservation actions such as migration and emulation. We will also develop an interoperability framework for these services, and a testbed environment.

In 2005, Harvard University was awarded a grant of US $600,000 over 2 years by the Andrew W. Mellon Foundation, to develop a prototype Global Digital Format Registry. The aim is to establish a global network of interoperable registries, containing detailed technical information about file formats. Harvard will develop a prototype registry, and protocols and mechanisms to support information exchange with similar registries which may be established elsewhere. TNA's PRONOM registry fulfils a similar function, being the first file format registry to be established, and we will be working with Harvard to contribute to the design of the GDFR, and to ensure that PRONOM is interoperable with the Harvard registry.

TNA is also working with the British Library, and the universities at Southampton and Oxford on a two-year research project, to develop preservation services for institutional repositories. These repositories are used by many academic institutions to store the results of academic research. TNA's main contribution is to provide preservation services through PRONOM. In the first stage, we have integrated our DROID automatic file format identification tool with the Eprints repository software, to provide automated characterisation of digital records when they are ingested into the repository.

TNA is also working to support digital preservation in local authority archives and other smaller archives, such as business and personal archives. We provide a growing range of practical advice and guidance on specific topics, such as digital image formats, image compression algorithms, and care and handling of removable storage media, which is available on our website. We are developing standard requirements and models for digital preservation systems, based on the OAIS model, to support the procurement of such systems by smaller institutions. And we are contributing to a growing range of training courses and

workshops in the field of digital preservation and electronic records management, to help develop new skills within the UK information management profession.

To summarise, TNA is: developing systems, processes and standards to support electronic records management and digital preservation in central government, undertaking collaborative research projects to inform this work, and developing a range of advice, guidance and training based on this work, to support electronic records management and digital preservation across the UK.