DEVELOPING A DIGITAL PRESERVATION FRAMEWORK AT THE UK NATIONAL ARCHIVES

Adrian Brown

The National Archives of the United Kingdom (TNA) was established in 2003, through the merging of two existing organisations. The Public Record Office was founded in 1838, with responsibility for preserving the records of central government and the courts. The Historical Manuscripts Commission was established a few years later, in 1869, and was responsible for maintaining the National Register of Archives, which catalogues other archival collections in the UK, both public and private. The merger of these two organisations created a single body, which is responsible for archives within England, Wales and the United Kingdom. As was the Public Record Office before, The National Archives is both a government department in its own right, and also an executive agency, which reports to the Department for Constitutional Affairs (DCA). The DCA is responsible for the judicial system in the UK, and for other constitutional issues, such as Freedom of Information.

TNA therefore has a very broad remit. We are responsible for preserving the records of central government and the courts, and for making them publicly available. We advise central government and the wider public sector on best practice in records management, and set standards. We also advise the owners, custodians and users of private archives on standards and best practice. Finally, we act as a portal to archival collections in the UK – we maintain catalogues, which are available on our website, to a huge number of public and private records

TNA has one of the largest archival collections in the world, which contains an unbroken span of records from the 11th century and before, to the present day, and includes some of our most important national documents, such as the Domesday Book. Our collections, which are primarily stored in our repositories at Kew, London, fill more than 180 km of shelving. Our more recent collections of electronic records are currently much smaller, but growing, and we currently store about 250 Terabytes of data, including both born-digital and digitised material.

TNA is responsible for administering the public records system in the UK, which covers all records created by public authorities, at both a national and local level. Our legal responsibilities are defined by the Public Records Act, 1954, although this has recently been amended by the Freedom of Information Act of 2000. All public records which have been determined to have permanent historical value must be transferred to TNA before they are 30 years old, unless special permission is agreed to retain them for longer periods. Traditionally, all records were also closed for 30 years, but, with Freedom of Information, this is no longer the case.

Over the last few years, we have seen many changes, as a result of the increasing use of electronic records – we are moving away from the traditional world of paper records, where our collections are measured in kilometres, to a new world where records are measured in Terabytes. There are a number of forces which are driving these changes. These include the e-Government agenda, and the Modernising Government initiative, Freedom of Information legislation, the demands of new audiences, and new technologies and, of course, costs.

In 1999 the government published a White Paper on 'Modernising Government', which included two key targets: the first was that all new records must be created and managed electronically by 2004, and the second was that all government services must be available online by 2005. TNA is responsible for monitoring the first of these targets, and the majority of government departments now have electronic records management systems in place.

Freedom of Information legislation came into effect in the UK on 2nd January 2005, and has led to major changes in how access to government records is regulated. It mandates that all government records are open on creation, unless a specific exemption, such as national security, applies. These exemptions must be re-examined, and justified, whenever a request for access to a record in made, and all requests must be answered within 20 working days. TNA receives the second-highest number of requests under Freedom of Information within UK government – last year we received and answered 5,500 requests.

The National Archives therefore faces a number of important new challenges. Since all new records are born-digital, we must have the capability to store and preserve electronic records, and to make them available to our users in electronic form. Our users also expect to be able to access more of our collections online, including our paper records, which means that we also have major programmes to digitise the most popular records in traditional formats, such as paper and parchment. TNA permanently preserves only a small percentage of the records created by government, but departments also need to keep many records for

significant periods of time for business purposes. For example, pension records need to be kept for at least 70 years, even though they will never be transferred to TNA. We therefore need to help government departments to sustain these semi-current records until either they are destroyed, or transferred to TNA. Although we have traditionally transferred paper records after 30 years, we need to transfer electronic records much sooner, typically after 5 years, if we are to be able to successfully preserve them. With the changes introduced by Freedom of Information, we need to manage requests for access to closed records. These changes are not just about technology – technology is one of the tools which we use but, more importantly, we need to change the business processes which we use, and bring about important cultural changes within government and the National Archives, to allow us to deal with electronic records successfully.

Another major challenge is the diversity of electronic formats which we must be able to preserve. These are not limited to the traditional 'Office' formats, such as word processed documents, but also include email, digital audio and video, databases, websites, Computer Aided Design models, virtual reality models, and software applications. Furthermore, we have little control over the formats used by record creators. Although we make recommendations on best practice, individual departments choose which software, and which formats to use. We already have example of all these formats in our collections, many of them from the records of Public Inquiries into major disasters. We need to find ways to preserve all of these formats, and make them available to our users, preferably over the Web.

As a result of these challenges, TNA has developed an approach to managing its electronic records. We have adopted a preservation strategy based on migration to new formats over time, coupled with the preservation of the original formats. We call each technical version of a record a 'manifestation' of that record, and we will preserve each manifestation which we create through migration, in order to be able to demonstrate the provenance of the record. As previously discussed, we must be able to accept any format. Although the volumes of records we are managing are currently quite small, we know that our storage systems must be scalable to manage many Terabytes and even Petabytes of data.

Cataloguing of records is also an important issue and, as with paper records, we expect the creating departments to catalogue their records prior to transfer. As the record creators, they understand the records better than anyone, and re therefore best placed to do this work. We must provide the means for departments to transfer their electronic records to us, and also have systems to provide access to those records, which must support both public access to open records, and secure access by departments to closed records. Finally, given the

enormous volumes of records we must process, we must seek to automate these processes wherever possible.

At The National Archives, we have not attempted to solve all of these issues at once, with a single system. Rather we have adopted an incremental approach, developing components of the overall system over time. The first system we developed was the National Digital Archive of Datasets, which was established in 1996 to preserve government databases. At the time, TNA did not have the necessary expertise in-house, so the work was contracted out to the University of London. This collection now contains more than 150 datasets, dating back to 1963, all of which are available online, in a standard format, which can be queried by users.

However, it soon became clear that TNA needed an in-house capability to preserve electronic records. In 2001, we therefore established a new Digital Preservation department, which is responsible for storing and preserving electronic records held by TNA. Our first task was to develop a Digital Archive, capable of securely storing records in any format, and this became operational in early 2003. In order to provide the necessary security, this system is air-gapped from our main network, and provides scalable storage for potentially 100s of Terabytes of data. The records are stored on LTO tape, in robotic tape libraries, with backups stored at an off-site location. The metadata is stored as XML files, and also in an Oracle database system. Having developed the means to store these records, our next priority was to provide public access. Our pilot presentation system, Electronic Records Online, was launched in 2005, and provides web access to the collections in the digital archive.

Another major priority was to begin addressing the long-term preservation of these many diverse formats. We felt that the first issue was to understand the nature of these formats, and their technical dependencies. We therefore began to develop PRONOM, our technical registry, which provides a knowledge base of technical information about file formats, software, operating systems, hardware, and the other technical components required to access electronic records. The first version of PRONOM was developed in 2002, and was only available internally. However, we quickly recognised that this information was more widely useful and so, in 2003, we made it freely available on the web. In 2005, we released a major new version which enabled us to store much more detailed technical information. This year, we also published a scheme of unique identifiers for the file formats stored in PRONOM, which provides a simple and unambiguous means of referencing the format in which a digital record is encoded.

Another important area of concern was government websites. We recognised that many websites were becoming increasingly important as records of how the government communicates with its citizens, and that we therefore needed to begin capturing these. We began by developing a selection policy for websites, based on the core functions of government. For each of these functions, we selected a representative sample of websites for repeated collection. The frequency with which each website is collected is based on the rate of content change, and the topicality and significance of the website. Currently 11 sites are collected on a weekly basis, and 53 are collected every six months. However, we can also add new sites in a flexible manner, to reflect current events. Our web archive was established in 2003, under contract with the Internet Archive in the United States. In 2005, we transferred the contract to the European Archive, based in Amsterdam. So far, we have collected over 15 million websites, comprising over 1.5 Terabytes of data, and we add to this at the rate of 0.25 million pages every week. Recently, we added 3,500 website previously collected by the Internet Archive, dating back to 1996.

We are also founder members of the UK Web Archiving Consortium, alongside the UK's national libraries, and various other national bodies. The Consortium is developing a national web archiving strategy, and a shared technical infrastructure for archiving websites. We use this infrastructure to collect other websites, such as those which we only collect once, or need to capture very quickly as a result of current events. During the 2-year pilot project, during which we have been using the PANDAS software developed by the National Library of Australia, we have collected over 1,000 websites.

These collections are all freely available online through our website.

Although we now have many elements of our electronic records system in place, there is much that we still need to do. In 2004 we therefore initiated a new programme called 'Seamless Flow', which is designed to develop end-to-end processes for managing electronic records, from appraisal and selection, through transfer, storage, and preservation, to delivery to users, and to automate as many of these processes as possible. This programme is due to be completed in 2008.

As previously discussed, we also need to help departments to sustain the records which they need to retain for business purposes, whether or not they ultimately come to TNA. We are therefore planning to develop an intermediate archive for government, which will allow all departments to preserve their semi-current records for the necessary period of time, from 5 to 100 years. This will be a shared service, to eliminate the wasteful duplication of

effort of every department developing their own system. It is possible that the passive preservation – the secure storage of records – will be contracted out to a commercial provider. However, the active preservation – the migration of records to new formats – will be provided by TNA, using the same services which we are developing for our own collections. This project is still at a very early stage, and we are currently trying to secure funding for 5 years, to begin building this service.

To provide a comparison with other international digital preservation programmes, it is useful to summarise the resources required to develop these various systems. The National Digital Archive of Datasets has cost US $14 million over the past 10 years. Our Digital Archive has so far cost US $12 million to develop. We are spending a further US $14 million on our Seamless Flow programme over a 3-year period. Our web archiving programme has cost a further US $1 million over the past 3 years, and we are now bidding for a further US $25 million to begin building our shared preservation service for government records. This means that we will have spent about US $66 million on digital preservation over the past 10 years. However, the most important resource to have is staff: we have a dedicated Digital Preservation Department of 12 full-time staff, but many other staff at TNA are also involved in aspects of the process, such as record selection, and delivery over the web.

In the course of developing these programmes, we have identified a number of major issues. We currently find it very difficult to accurately forecast the future volumes of records which ill be transferred to TNA. This can make it very hard to predict the storage capacities which we need to budget for each year. Developing new methods for describing and referencing electronic records, particularly the more complex objects, such as websites, has also been a very big challenge. Another important issue is how we undertake technology watch, which is the process whereby we identify important technological changes, and the impacts which these have on the records we store. As previously discussed, digital preservation is about processes more than technology, and achieving the necessary business and cultural changes within TNA and government departments has bee, and continues to be, a major challenge for us. Finally, there is the issue of sustaining semi-current records within departments.

However, we have also learned some important lessons from our work so far. Firstly, we have learned that it is important to think big, but start small. It is virtually impossible to solve all of the issue at once, and it is easy to be daunted by the scale of the problems. However, this should not prevent one from beginning to tackle them in stages.

It is also vitally important to collaborate, and build on the experiences of others. No one organisation has all the answers, or can solve all the problems, and we can all learn from the experience of others. However, it is also important to recognise that every organisation is different, with unique requirements, and that solutions which work for one institution may not be applicable to another. It is therefore important to identify your own requirements from the outset, and develop solutions which are specific to these.

Finally, it is essential to recognise the importance of business change. The business processes are much more important than the technology. You can achieve a lot with little technology, but the most expensive technology in the world will achieve little if it is not suited to your requirements, or underpinned by the necessary business change. The final message must be an optimistic one – it is possible to solve the problems of digital preservation, and a huge amount of progress is being made internationally in this area, which everyone can draw upon.