Fleur Soper

Communications Officer
Digital Preservation, National Archives

The PRONOM File Format Registry

Records managers are increasingly faced with the challenge of preserving records in digital form, instead of on paper. After decades of preparing documents by computer, the practice of printing them out for filing is in decline, and in the case of government records the old ‘print to paper’ policy has been formally abandoned. At the same time, records have evolved from simple texts to complex assemblies of diverse elements, including embedded formulae, images and charts, with a corresponding proliferation of file formats. The introduction of Electronic Records Management software may help with the day-to-day management of records but does nothing to simplify the problem of long-term preservation, since the record elements are still held in their original formats and these formats become obsolete with alarming rapidity. The withdrawal of support for old versions of software and their formats is now being soothingly referred to by developers as “sunsetting” – but sunset comes round all too soon for the liking of anyone who has collections of computer records to preserve.

The National Archives is charged with the preservation of records in perpetuity and is addressing the problem of software obsolescence through PRONOM, a web-enabled database of information on file formats and their technical dependencies, including hardware, software and operating systems. PRONOM was launched on the Web in February 2004, and is freely available at www.records.pro.gov.uk/pronom/.

A global registry

Technical documentation of file formats is not easy to acquire, particularly after they become obsolete. The need to establish reliable, sustained repositories of file format specifications, documentation and related software has been recognised as an international issue, and a working group to develop a Global Digital Format Registry has been set up, including representatives of the national archives of the United Kingdom and the United States and several other major institutions. The concept is to create a global network of registries that can be shared by many institutions.

The National Archives developed PRONOM ahead of this initiative, starting in 2001, and we are working to ensure that the two programmes are complementary. We have developed not only our own prototype registry but also an initial collection of content. Our staff have undertaken intensive research and liaison with major software developers in order to create this initial data set; Microsoft and Adobe have been particularly helpful in providing information. The database currently holds details of about 550 file formats, 250 software products, and 100 vendors, and more are being added on a regular basis. We actively encourage the submission of new information for inclusion on PRONOM, and an online submission form is provided for this purpose.

Uses of the registry

Information about digital formats has many uses in a digital preservation programme. As old software products cease to be supported and become obsolete, preservation activity will be needed for records held in the formats that depend on those products. One strategy is the migration of records from obsolete formats to newer ones. Migration paths are already identified in PRONOM, and in future it will also provide a measure of the 'content invariance' of each migration path. This is especially important to maintain the authenticity of digital records. The content must not change, and the appearance of the record should stay the same too. Our intention is to define an objective and rigorous methodology for testing migration paths to measure content invariance, and to record the results in PRONOM. The registry will also tell you when a format is about to become obsolete, and act as a trigger for preservation action.

An example of the hazards of migration is the preservation of WordStar data files. Early DOS versions of WordStar used seven-bit ASCII characters, the eighth bit being used as a line wrap marker. When viewed by later products these characters are wrongly interpreted as eight-bit ASCII equivalents, and to achieve a successful migration it is necessary to strip out the marker bits from the WordStar files. Since line wrapping is handled differently in later products, the loss of the eighth bit normally makes no difference, and at worst causes the text to be adjusted to different margins. This example shows the part that detailed technical knowledge plays in implementing a workable migration strategy, and also the importance of keeping the original bit-streams.

Before you can preserve digital records you have to recognise what format they are in. Automatic file format identification is an important function of a format registry and this facility will be provided in future versions of PRONOM. When records are accessioned, after the bit-stream is copied to archival storage it is necessary to identify the format, and to test that the transfer has succeeded – that each digital object is complete and intact. In this way defective objects can be detected at the point of ingest and a replacement copy requested. The registry will not itself provide validation of records but will provide information to support the process.

Conclusion

Some paper records have survived despite long periods of neglect before they were received into our archives and libraries. If today's digital records are neglected they will not survive, because the ability to interpret old formats will be lost. This problem has been much discussed in the library and archives community, and pessimistic projections have been made of the loss of valuable records of the digital age. In fact the preservation of digital records is perfectly practical so long as early action is taken.

The correct interpretation of records has always required knowledge of the language in which they are written, and sometimes of other subjects too - mediaeval penmanship, for example. Fortunately enough of this knowledge has survived that we can make sense of most of the records that have come down to us. Modern technology has further complicated the problem of interpretation by making the viewing of records dependent on hardware and software environments whose own longevity is doubtful. Just as interpretation of the 1086 Domesday Book depends on the dictionaries and grammars for mediaeval Latin painstakingly compiled by longdead scholars, interpretation of contemporary electronic records in the future will only be possible if the necessary methods and tools are compiled, documented and preserved now. The initial Web release of PRONOM is a long step in that direction.