An Open Source Tool for Migrating Digital Records for Long-term Preservation

Andrew Wilson
National Archives of Australia
Canberra, ACT, Australia

Abstract

The National Archives of Australia, as the archives and records authority for the Government of Australia, has a requirement to ensure that high value digital records created through the business activity of Australian Government agencies are accessible indefinitely. However, indefinite preservation is extremely difficult when dealing with digital records encoded in proprietary data formats.

Over the last three years the National Archives of Australia has been developing an in-house digital preservation program built around the eXtensible Markup Language (XML) and open source software. This approach focuses on archival data formats as the key to ensuring long-term accessibility of 'born digital' records. To achieve its preservation objectives the National Archives is developing or adopting a range of open data formats in XML which will be used as schema to transform original digital objects into XML data formats. To support the transformation process the National Archives is developing a software application, known as 'Xena', which will carry out the transformation process, which we call 'normalisation', and will also be able to render the XML data format into a viewable 'performance' of the original digital object, when required by researchers.

The paper will describe the software application being developed and tested by the National Archives.

Introduction

The National Archives of Australia (NAA) is the archives and records management organisation of the Australian Government. NAA cares for valuable Australian government records and makes them available for present and future generations to use. One of our major roles is to develop recordkeeping standards to help government to be accountable to the public, ensuring that evidence is available to support people's rights and entitlements, and that future generations will have a meaningful record of the past.

The records in our collection trace the events and decisions that shaped Australia. We hold the papers of Governors-General, Prime Ministers and Ministers. We have Cabinet documents, Royal Commission files and departmental records on defence, immigration, security and intelligence, naturalisation, and many other issues involving the federal government.

The main focus of our collection is records created since the formation of the Commonwealth of Australia in 1901. We also have some nineteenth-century records relating to functions that were transferred by the colonies to the Commonwealth government, including shipping and post offices.

Since the ultimate objective of the National Archives is to ensure that significant records of the Australian government are available for future generations, the preservation function of the organisation is of great significance to the Archives and to its clients. The rapid spread of computer systems into the everyday work of government organisations over the last two decades has dramatically altered the way in which work is done and how information is communicated and shared. The National Archives of Australia has been putting a great deal of work into changing its preservation methods to cope with the multitude of records in the wide range of digital formats that it will take into custody.

The Problem of Digital Document Formats

Every day tens of thousands of Australian Government public servants create digital documents in a wide variety of different formats [a document format is the rules by which the information within a document is organised. Software applications then use these rules to process a document and to construct a human-understandable rendition of the document]. Many of these document formats are proprietary, ie. they have been created by software companies for use with their own particular programs only, and the rules of the format are closely guarded corporate secrets. As a result, other software companies do not have access to these rules, and their programs can't 'read' documents created according to these rules. This is the reason why many digital documents can only be opened, and their contents properly displayed, by the software application in which they were created.

As a result, many digital documents are at great long-term threat: if the original creating application is no longer available (or its price is too prohibitive for users to acquire the appropriate license) access to the digital document might be lost forever. Both individual agencies' corporate memory and the archival resources of the Australian Government are thus placed at risk.

At the National Archives of Australia our recommended solution to this problem is to convert important digital documents into open, fully documented formats that can be understood by a wide variety of software applications on a wide variety of computer platforms. Using open formats means that digital documents are more accessible today and stand more chance of remaining accessible over the decades to come. The National Archives' approach is based on the use of XML (eXtensible Markup Language) document formats as the archival data formats of choice.

XML is our format of choice for a number of reasons:

- XML is a widely used, fully documented standard for structuring digital documents;
- the XML specification is freely available and frees the organisation from dependence on particular IT vendors;
- XML can be used as a technology base indefinitely;
- creation of additional XML data formats to meet the preservation needs of many record types is not technically challenging.

An Open Source Software Tool: Xena

In order to convert records from the wide variety of formats received by the Archives into the XML, the National Archives has spent the last 2½ years on a digital preservation project, among other things developing a custom made software tool to carry out the conversion from source data formats into archival data formats, a process the Archives refers to as 'normalisation'. The tool, **Xena** (**X**ML **E**lectronic **N**ormalising of **A**rchives), is a software program that converts digital documents from their original (usually proprietary) data format into the selected open, fully documented, formats used for archival preservation by the National Archives. Some of these archival formats were especially created by the National Archives of Australia for long-term preservation. Details of these formats can be accessed from the Digital Preservation pages on the website of the National Archives of Australia at http://www.naa.gov.au/recordkeeping/preservation/digital/xml_data_formats.html. Other formats used by Xena are fully documented, open, industry standards that any software implementer can use free of charge:

- **Portable Network Graphics (PNG),** a bitmap image format commonly used for images on the World Wide Web. PNG images are accessible on any platform that supports recent (ie, version 4.0+) web browsers.
- **Hypertext Markup Language (HTML),** a text formatting and document linking language used to construct web pages on the World Wide Web. HTML documents are accessible on any platform that supports web browsers.
- **Open Office XML Format (OO XML)**, a XML-based document format used by the OpenOffice.org and StarOffice applications for structuring word processed documents, spreadsheets, vector drawings, presentation slideshows, charts, and mathematical formulae. OpenOffice.org has similar capabilities to Microsoft Office and is available for Windows as well as Linux, Sun Solaris, Mac OSX, and other Unix-compatible platforms.

Although the National Archives of Australia has developed Xena as a digital preservation software application for its own internal use, it believes that Xena may be a useful tool for many other individuals and organisations as well. For example:

- *Other archival institutions* may find Xena useful in developing their own in-house digital preservation programs;
- *Government agencies and other organisations* may find it useful to integrate Xena into their own records management systems so that they can convert digital records to standard formats at the point of capture into their records systems and/or can batch convert existing corporate records repositories into standard formats for long-term accessibility and preservation;
- *Individuals* may find Xena useful when they need to move documents across computing platforms, send documents to others, post documents to the World Wide Web, or when they change to new versions of software programs that do not support old formats.

In short, Xena is a tool that allows archival institutions, organisations, and individuals to leave behind the restrictions of closed, proprietary document formats and to place their digital documents and digital records in open, documented, and accessible formats where important business information can be accessed from a wide range of applications on a wide range of computing platforms.

Why Open Source?

To meet the Archives' values of comprehensive, equitable and sustainable access to the Australian government's archival resources the digital preservation project developed a number of principles to underpin the approach to preserving digital records. An important aspect of these principles is that the Archives should not rely on proprietary software solutions to 'normalise' and re-present the digital records. Reliance on proprietary software applications would mean that each time software changed both the Archives and researchers would have to purchase new versions of the software. Using proprietary software would also mean that preserved digital records would have to be migrated at much more frequent intervals.

For these reasons the digital preservation project decided to adopt an open source software approach. This approach meets the organisational commitment to providing free long-term access to digital records for our clients. The open source approach also ensures that access to the digital records we preserve is in our hands and is not dependant on a private sector company's intellectual property rights, as it would be if our approach used proprietary software applications.

The Xena application developed over the last 24-30 months makes use of available open source technologies. The terms of the open source licence mean that the Archives does not sell Xena. Although other developers may make use of the source code of Xena, neither are they able to profit from the work the Archives has done on Xena. The digital preservation project believes that this approach enables the Archives to retain control over the technologies we rely on to preserve and access digital records. By providing Xena at no cost, we will encourage government agencies to normalise records themselves, which in the long-term could help reduce the resource commitment our approach entails.

It is important to emphasise that Xena is only a part of the NAA digital preservation tool suite. It is designed to be implemented in conjunction with a digital preservation workflow. Many design elements have been included that make it relate specifically to the NAA digital preservation workflow. As such, Xena is targeted at a very specific audience. It is best implemented as an administrative component of a recordkeeping system, rather than a user desktop application. Furthermore, any agency or archival institution using it will almost certainly need to customise it to suite their specific needs.

Through its involvement in the recently formed Australasian Digital Recordkeeping Initiative (ADRI), the National Archives is actively collaborating with other archival institutions to build common approaches to digital recordkeeping in general and digital preservation in particular. Although any user can download and change Xena without reference to the National Archives, we welcome comment on Xena, its development, and the possibility of collaboration from other organisations and individuals.

Xena Functionality

Xena is able to convert Microsoft Office documents from Word, Excel, and PowerPoint into a OO XML.

Documents converted to OpenOffice.org XML retain the formatting and functionality of the original Microsoft Office documents, with the ability to chose between keeping and deleting embedded metadata. These documents are fully editable in OpenOffice.org. Not all conversions, however, are perfect.

Xena can also convert:

- **Email messages** from Microsoft Outlook, either individually or entire .pst data files, and any email program capable of exporting MBOX mailboxes or individual messages in the RFC822 or MIME format (for instance, email programs such as Netscape Communicator, Eudora, or Apple Mail). Xena can also convert email messages that have been captured into TOWER Software's TRIM Captura digital record and document management system. All email messages are converted into the emailmessage XML format, created by the National Archives of Australia.
- **Delimited text files** exported from database applications (such as Microsoft Office and FileMakerPro) or recordsets from JDBC-compliant relational databases (such as MySql, Microsoft SQL Server, Oracle). All such documents are converted into the dataset and database XML formats, created by the National Archives of Australia.
- **Bitmap images** in many different image formats can be converted into the PNG image format.
- **Every version of HTML** can be converted to the new XML-based version of HTML, XHTML.
- **Entire websites** can be automatically downloaded and individual documents converted into appropriate formats with no manual intervention using Xena's batch processing module.

As described above, normalisation is the conversion of a digital record from its original format into an archival XML equivalent. The digital preservation workflow currently being tested by the National Archives also uses Xena to create an XML wrapped base64 version of the record. This process is referred to as XML wrapping.

The base64 version, which is also referred to as the original bitstream version, enables the Archives to conserve the digital record in its original source format as a binary object. This is important if source records cannot be normalised immediately on receipt. Furthermore, having the source bitstream available means that records can be normalised again using different normalisation paths if necessary, or using a different or changed normaliser for a particular source data format. It is not possible to reverse the normalisation process, whereas it is possible to reverse the XML wrapping process, converting the record from base64 back to the original data format. Wrapping also enables the Archives to preserve the authenticity of digital records by allowing the Archives to provide users with the source record as originally received by the organisation.

Xena creates the normalised version of a digital record by converting the original data object into an XML format (see above). This format will vary, depending on the format of the original record, and what is determined as the essence of records in particular data formats. In all cases the XML format will be open source.

The Design of Xena

Xena is a Java based open source application developed by the National Archives of Australia (NAA) under the GNU Public License (GPL). Xena requires the Java 2 version 1.4 runtime environment or better, and OpenOffice.org version 1.1 to run. As it is Java based, Xena can run on any operating system that can run Java and has been successfully used under Windows 98, NT, 2000, and XP, Linux (Red Hat), and Mac OSX (note that there are some restrictions on OpenOffice.org functionality in Mac OSX.3).

The total Java environment is fully documented and specified. As anyone can build a Java runtime environment, and Java is widely used, it is highly likely to remain viable

for many years. If it ever does become obsolete, there is a very high degree of probability that there will be clear and effective migration paths to replace any application running on a Java platform. This increases the ability of Xena to avoid operating system and hardware obsolescence. If you can run Java, or migrate from a Java platform, you can run Xena.

As an open source application, the source code for Xena is publicly available. This means that while the National Archives will continue to make a significant investment in enhancing and maintaining Xena, other organisations and users can adapt Xena to their needs. The National Archives hopes that this will lead to a community of Xena users and developers which will be able improve Xena and expand its capabilities to deal with new electronic record formats and electronic recordkeeping environments.

Technical Design

Xena is designed around a central core program, with plugins used to normalise digital records. It is these plugins that allow it to expand to deal with new data formats.

Each plugin can normalise one or more data formats. The Image plugin, for instance, is designed to normalise most image files (such as JPEG, TIFF, GIF and BMP) into PNG image files. The plaintext plugin, on the other hand, is designed to normalise unformatted text files from different platforms (such as Windows, Unix, and MacOS) that use different character sets. As the need arises to normalise new data formats, new plugins can be developed. Because Xena is open source, anyone can develop new plugins, which is especially important for rare or specialised data formats.

Some plugins rely on third party software. The Office plugin relies on OpenOffice.org, an open source office productivity suite. This reliance on third party software may be for the actual normalisation process, as with the Office plugin, or it may be for some pre-normalisation treatment; for example, PDF files. The path for normalising PDF is to convert it to a directory of images using Adobe Acrobat Writer, then to normalise the directory of images using the multipage plugin which does not require third party software.

How Xena Preserves Digital Records

In this section some architectural aspects of how Xena preserves digital records are described. It covers the basic concepts underlying the normalisation of digital records and their display, expressed in the programming of the Xena plugins.

The Xena modules that normalise data into XML are called **normalisers**. Each plugin carries one or more normalisers. A normaliser accepts a data object, which is usually a single digital file such as an image or a Word document, and converts it into an XML version of that data object, depending on the programming of the normaliser. The output is in the form of a XenaFileType.

Xena plugins include guesser modules to decide which normalisers to use. These modules check if a data object conforms to a particular type – such as a JPEG image, or plain text document – based on numerous factors, including file extension and the data it contains. Xena passes the data object by each guesser in turn. The guessers respond with the probability that the data object is of a particular data format. Xena then examines these results, and offers the user what it thinks is the most appropriate normaliser as the first option. Thus, when normalising a single data object, users are offered a list of file type names such as String, Binary and HTML, so that the actual file type of the document guessed by Xena can be confirmed by the user.

FileTypes are conceptual notions that represent a particular data format. XenaFileTypes are a subset of FileTypes representing the Xena XML formats. A FileType can have multiple normalisers, each producing a different XML version of the original data object.

Once a data object has been normalised and saved to a file, the in-built Xena guessers will recognise it again as a XenaFileType, and Xena will offer the user the correct primary choice. The Xena viewers will then present the XenaFileType to the user. A XenaFileType may have multiple views; for example, a preserved image file may be displayed within Xena as an image. It can also be viewed as raw XML, or in the form of an XML tree. More complex file types, such as multimedia files, will have more views available.

Conclusion

At the National Archives we've been working towards creating a long-term viable digital preservation approach for a while now, and we believe our conceptual approach and the tools we have been developing offer a useful and practical approach to the preservation needs of digital records.

The most important point about the NAA approach is that we believe that ***data formats provide the key*** to long-term preservation of electronic records. The best way to ensure that the digital data in your electronic records is accessible into the future is to put them into open, fully documented data formats that are accessible by many different applications on many different platforms. That belief in the importance of data formats has led the Archives to the development of a software processing platform that will allow the Archives to maintain valuable government records in long-term archival data formats.

We have released version 1.0 of our Xena tool to the archives and records community and will be continuing to improve the software and extend the range of data formats that Xena can normalise. Currently the Archives is working on a normaliser for Microsoft project files. The whole development process with Xena has been and is a really exciting project for the Archives because it provides us with a very real and concrete tool for converting some document types into XML and then viewing those document types as well. From the second quarter of 2005 the National Archives will be accepting into its custody transfers of records from Australian Government agencies and will be using Xena to normalise these records. The normalised records will be

stored in our recently constructed digital repository and will eventually be available to researchers through the National Archives search rooms around Australia. Information about the progress of the Archives' digital preservation approach and the software platform (including Xena) will be regularly updated and available from the website of the National Archives of Australia at: http://www.naa.gov.au/recordkeeping/preservation/digital/summary.html .

Biography

Andrew Wilson is a Director in the Digital Government Branch of the National Archives of Australia. He is currently managing a project to develop the capability of the National Archives to manage digital records from transfer to access. Prior to that he managed the project developing an approach to the long-term preservation of digital records. Andrew is currently a member of the Dublin Core Usage Board and co-chair of the DCMI Preservation working group. He also represents the National Archives on the RLG/OCLC PREMIS working group.