PANDORA at the Crossroads-lssues and Future Directions.

Jasmine Cameron, Judith Pearce
National Library of Australia

Introduction

PANDORA stands for Preserving and Accessing Networked Documentary Resources of Australia. This is the project name given to the work that has been undertaken by the National Library of Australia to develop an electronic archive of Australian publications on the Internet. Work on the project, which commenced in earnest at the beginning of 1997, has concentrated on two strands of activity. These are the development of a working 'proof-of-concept' archive and the development of a series of documents that provide a conceptual framework for a permanent electronic archive.

The National Library of Australia has envisaged from the beginning of the PANDORA Project that the knowledge gained from the project would form the basis of a much broader strategy for the creation of a National Collection of Electronic Publications. Australia has a long history of national co-operation in the areas of collecting and provision of access to information, and the National Collection of Electronic Publications will involve co-operation with other major libraries and national collecting bodies, with a view to sharing the responsibility for preserving and providing future access to Australian electronic publications.

Background

The National Library of Australia has a statutory obligation to collect and preserve Australia's printed heritage and it regards the care of electronic publications as a logical extension of this mandate. The PANDORA Project is the first step in developing a strategy for the collection and preservation of Australia's documentary heritage as it is represented through publication on the Internet. The project's two key objectives are: to develop and test policy and procedures for the acquisition, preservation and provision of long term access to Australian information published on the Internet *and* to test the feasibility and determine the cost of establishing a National Collection of Electronic Publications.

Work towards achieving these two objectives has proceeded on two levels, one a purely theoretical level and the other a very practical level. This approach has yielded benefits to the project because both of these streams of work have informed and shaped each other. On the practical level the National Library of Australia has developed a 'proof-of-concept' archive that contains over 200 titles and is growing at the rate of approximately 10 titles a month. Policy and procedures have been developed for each step in the process including

- scanning the Internet and selecting titles for the archive
- liaising with creators for permission to archive their titles and for additional information about the frequency of update and format of their title
- cataloguing the title onto the National Bibliographic Database[1] and providing a hotlink from the PURL in the catalogue record to the entry screen in the archive
- capturing the title on a regular basis using a modified version of Harvest software and creating entry screens for each title in the archive with access to the individual issue within the archive.

[1] The National Bibliographic Database is a shared cataloguing database and a union catalogue of the holdings of Australian libraries.

Policy has also been developed for the management of commercial pay-per-view or subscription titles within the archive.

Work has also proceeded on the development of a conceptual framework for a permanent electronic archive and is described in two key documents; the PANDORA Business Process Model andthe PANDORA Logical Data Model. The Business Process Model outlines the business directions and principles on which the development of the PANDORA 'proof-of concept' archive has been based. The Logical Data Model defines the data elements in the archive, the relationship of these elements to each other and to external data. Extensive work has been undertaken on the definition of metadata needed to describe and manage titles within the archive and this work forms part of the Logical Data Model. These two documents are available on the PANDORA Home Page at ‘http//www.nla.gov.au/pandora’. Work is currently progressing on a much broader document that will be completed in August. This document, which may be released as a Request For Information, describes the National Library of Australia's requirements for a digital object management system which will meet the needs not only of PANDORA but the Library's other digital collections. This work is being carried out as part of the Library’s Digital Services Project.

PANDORA Business Principles

Several key business principles have been incorporated into the design and management of the PANDORA 'proof-of-concept archive' It is important to stress that most of these business principles are based on practical decisions and as the project has evolved so has the Library's thinking begun to broaden in relation to many of these principles.

Selectivity

From the beginning of the project, the principle of selectivity has formed the basis of the PANDORA selection guidelines. The National Library of Australia is also selective to an extent in its acquisition under legal deposit of Australian print publications and relies on the State libraries to collect material at a local level. For example, the Library does not collect publications such as school magazines and local club newsletters. The Library’s policy in relation to the collection of electronic publications is intended to be the same as that for print, and although it is currently more restricted than our print collecting it is proposed that in the future the Library’s collecting of on-line publications broaden to match the print policy.

Unlike our colleagues at the Royal Swedish Library, no attempt has been made to capture the entire Australian domain. Selection guidelines determine a range of publications to be archived, from scholarly titles to those representing popular culture and the use of the Internet by Australians in general. This approach has its merits including the ability to exercise a degree of control over the quality of what you have archived and to provide access to what you have archived. On the other hand to regularly scan the Internet to select individual titles for archiving is resource intensive and valuable information may be missed and therefore lost. The National Library of Australia believes there is merit in both approaches and while continuing to use a selective model, the Library may also experiment in the future with 'snapshots' of defined segments of the Australian domain.

A decision was also made very early on that titles with print equivalents would not be selected for the archive. This decision was made because it reduced the large amount of information that would be eligible for selection to a manageable amount and it was reasoned that Australian titles in print were already being collected and preserved as part of the Library's legal deposit role. However, this decision is also under review because it is readily acknowledged that electronic titles with print equivalents can often vary in nature and content from their print counterparts. Future ease of access to electronic versus printed information is also an issue.

Access

Following on from the Library's selective approach to archiving titles in PANDORA it was considered important to let other Australian libraries, and indeed libraries internationally, know which titles the National Library of Australia has undertaken responsibility for archiving. This has been done by creating a catalogue record on the National Bibliographic Database for each title in the archive. Catalogue records created on the National Bibliographic Database for titles in the PANDORA archive are also downloaded into the Library's Online Public Access Catalogue. This is considered the best mechanism currently available for providing integrated access to information in any format held in the Library's collection.

The issue of resource discovery relating to the actual content in the archive below the title level has not been addressed in the development of the 'proof-of-concept' archive. However, the facility for capturing and/or generating Dublin Core compliant metadata which is indexing the content of publications captured for the archive, and posting this metadata to a designated metadata repository, has been included in the Request For Information referred to earlier. It is recognised that in the future the PANDORA archive will contain a large number of titles that exist nowhere else and that access to the content of the archive will be an important issue. The Library’s Australian Public Affairs Information Service which indexes selected Australian printed journals in the Library’s collection, is indexing selected Australian electronic journals. This indexing service could also be expanded in the future to routinely index Australian on-line publications.

Management of commercial publications

The Commonwealth of Australia Copyright Act, which covers legal deposit, does not currently include the legal deposit of electronic publications, either in physical format or on-line. However, the Copyright Act is under review and it is anticipated that legal deposit will be extended to cover electronic publications for preservation purposes. The Library is lobbying to have the concept of legal deposit extended to electronic publications and has made formal submissions to this effect to the Copyright Law Review Committee. In the event that electronic publications are covered by legal deposit, the Library will not, at this stage, be negotiating to provide access for remote users to current issues of online commercial titles through the PANDORA archive. PANDORA is first and foremost an archive and it is expected that users will visit the publisher’s site for all currently available material.

The difficulty for the PANDORA project at the moment is that the use of the Internet for publishing in the Australian domain is still very much restricted to gratis and nonprofit publications so that the Library's thinking on the issues surrounding the management of commercial on-line titles has not been tested. In view of the fact that Internet publications in the Australian domain are not yet subject to legal deposit, the PANDORA project has developed a 'Voluntary Deposit Deed' for on-line publications. This deed closely mirrors the Voluntary Deposit Deed used by the Library when seeking physical format electronic publications. Publishers will be asked to nominate from a standard list of timeframes the period for which they wish their publication to be suppressed from the public domain. Publishers will also be asked to agree to allow gratis access to current information to on-site users. On-site access to electronic publications is seen as a parallel to the on-site use of printed material received on legal deposit.

To date the National Library of Australia has negotiated successfully with only one Australian Internet publisher for the voluntary deposit of their commercial literary and reference 'monographs' in the PANDORA archive. Although a voluntary deposit deed has not yet been signed, the publisher has agreed to allow the Library to provide on-site access to their titles in the archive. A timeframe for future access to the titles by remote users has not yet been agreed. The Library does not see a role for itself as a middle-man for commercial Internet publishers by levying, on behalf of these publishers, a fee-for-use of commercial publications held in the PANDORA archive. One of the first business principles established was that the archive is a secondary resource, for use when issues are no longer available on the publisher's site. The Library is approaching its management of commercial titles in the archive on this basis.

The Library's thinking on this issue may well have to be modified in the future as major Australian publishers move into the Internet publishing market. We are looking closely at arrangements such as those made by the Royal Dutch Library with major commercial publishers like Kluwer and Elsevier. The Royal Dutch Library pays a license fee to these publishers in return for being able to provide access to their on-line publications to both the Royal Dutch Library's registered on-site and remote users. The Royal Dutch Library is taking responsibility for archiving these electronic journals in the same manner as the PANDORA archive.

Legal Deposit and Copyright

Legal deposit

As mentioned above, electronic publications are not yet covered by legal deposit although the Library anticipates that this will be included as part of the current revision of the Copyright Act. The broader issue of how to best filter and select titles on legal deposit for the archive is yet to be resolved. One method is to request that Australian Internet publishers register their titles with the Library so that these titles can be scanned and selections made from the registry. This registry could then form a valuable resource, doubling as a national bibliographic listing of Australia on-line titles. The value of maintaining such a listing, and the value in monitoring what could be a large number of publications registered for legal deposit, has yet to be fully debated in the Library. What is certain, however, is that coverage of electronic titles by legal deposit will require the Library to establish a new set of relationships with Australian publishers on the Internet and to review some of its present PANDORA principles and procedures.

Copyright

It is somewhat ironic that the electronic age with all its vast potential has brought with it a set of restrictions that often make the provision of access more limited than that for printed material. Copyright is a complex issue in the online environment, particularly where multi-media web sites are concerned. There may often be many more creators involved with an on-line publication than is the case with print. There may be different authors of text, images, software and so on. In the case of electronic journals it is not unusual for the copyright to be held by authors of individual articles.

The PANDORA project has approached the issue of copyright by including a general copyright warning, plus a link to the publisher's own copyright statement, on the entry screen for each title in the archive. Under the current copyright reform agenda, the Australian government has announced recently that a new, broad-based, technology-neutral right of communication to the general public will be introduced. This right will apply to information made available on the Internet and other on-line services. This right will be subject to exceptions for fair dealing, libraries and educational institutions.

Standards

The National Library of Australia has a leadership role within the Australian library community in the development of standards. This involves the Library in a wide range of activities including representation on key Australian standards bodies and international working groups such as the Z39.50 Implementors' Group and the Dublin Core group. The PANDORA project has a particular interest in the development of metadata standards for resource discovery, permanent naming conventions, standards for preservation metadata and Internet publishing conventions.

Although not directly involved in standards work, through its contact with Internet publishers PANDORA plays a role in creating awareness of the benefits of generating metadata for resource discovery and of the role of permanent naming for stability of links to documents on the Internet. The ability in the future to provide adequate access to documents in the PANDORA archive depends to a large extent on the development of standards and publishing conventions in many areas. The Library is currently a partner in a project known as Metaweb which is developing metadata element sets, user tools and indexing services to promote the use of metadata. The project has also looked at the concept of a national metadata repository. The Library has also set up a PURL Resolver Service based on OCLC software, although it is recognised that PURLs are only an interim solution for permanent naming. PANDORA has encouraged publishers to assign PURLs to their documents and by way of example routinely assigns PURLS to titles in the archive.

Within the PANDORA project a large amount of work has been done on identifying the key data elements of the PANDORA archival management system. It is anticipated that this work will be compatible with future preservation metadata standards. PANDORA is also interested in, but has not been directly involved with, the development of Internet publishing conventions. A wide range of actions, for example non-standard file structures and file names, undertaken by Internet publishers affects the ability of the PANDORA project to satisfactorily capture and maintain the look and feel of titles in the archive.

Technical requirements

The PANDORA project has reached the stage where the 'proof-of-concept' archive needs to be underpinned by a robust software and hardware platform suitable for the establishment of a permanent electronic archive. PANDORA's technical infrastructure requirements form part of the Library's new Digital Services Project referred to earlier. It is hoped that solutions will emerge following the completion of a Request For Information later in the year. The two key elements which will assist the progress of the PANDORA project are the identification of a suitable 'gathering' or document capture mechanism, and the identification of a suitable digital object management system which will facilitate version control, rights management and authentication within the archive.

Gathering

At the commencement of the PANDORA project it was decided that it would be better to pro-actively capture documents for preservation in the archive than to rely on publishers to 'push' information to the archive. This decision was partly influenced by the fact that the majority of Internet publishers in Australia were new to publishing and did not have established relationships with the Library. It seemed a lot easier to say “we will come and get it” rather than to rely on publishers to send the information to us, and particularly in situations where a regular capture schedule is necessary. The PANDORA project currently uses a modified version of Harvest software to capture publications for the archive. However, Harvest is essentially designed for indexing and is not suitable for use in the medium to long-term as a capture mechanism.

Finding an alternative to Harvest is a top priority for the PANDORA project. The difficulty in this endeavour is that few other institutions require software for this exact purpose and it may be that we will have to settle for a modified version of a more sophisticated indexing software. Another option would be to develop, preferably in collaboration with another national library (or libraries), software specifically for this purpose. Towards the end of 1997 the PANDORA project began to investigate alternative capture strategies. This was spurred on by the emergence of publications structured as databases which operate on creating data for the user on the fly. It had become obvious that ‘pull’ technology would not cope with this new publishing format and as a result experimentation with publishers ‘pushing’ their publications to the PANDORA archive has commenced. While PANDORA still intends to use gathering software to capture many of its publications, a mixture of both push and pull technology will be used in the future.

Management of documents in the archive

The PANDORA ‘proof-of-concept’ archive has grown to the stage where it is critical that a digital object management system be implemented to facilitate the wide range of functions associated with the selection and management of the titles in the archive including the negotiation status of the publication, procedures for the capture of copies for the archive including an automated capturing schedule, updating the archive, restrictions on access, version control, retaining the ‘look and feel’ of a publication and authentication.

The Library is relying on developments in the commercial sector to provide solutions to some of the broader business issues facing the PANDORA project. For example, authentication is an issue for most people doing business on the Internet and solutions are beginning to emerge for general use such as encryption, time stamping, watermarking and digital signatures. For PANDORA, the aim is to provide a record of Australia’s documentary heritage and that must be an accurate record. The documents that are archived must remain unchanged and true to the original. It is anticipated that the current work on encryption and other authentication methods which will facilitate the verification of the content of a publication will provide an answer to the issue of authentication that is suitable for PANDORA.

Look and feel

One of the important PANDORA business principles is to retain, as far as possible, the look and feel of a document in the archive. Preserving the integrity of a document selected for archiving is considered important for the accuracy and completeness of Australia’s future historical record. However, it is recognised that in some cases this may simply be impossible. The technical issues surrounding the retention of the look and feel of a publication are numerous. It requires the PANDORA archive to maintain a software repository and to capture copies of the software supporting the viewing of publications selected for the archive, if not already in the software repository. In the future, it means that this supporting software will need to be migrated to a new format. And all this activity will have to be tracked and managed within a digital object management system.

The Australian National Collection of Electronic Publications

From the beginning of the PANDORA project the National Library of Australia has envisaged that the task of managing and preserving electronic publications would be taken up nationally once the Library has completed its developmental work and a viable set of policies and procedures are in place. The Library is looking to establish the National Collection of Electronic Publications in co-operation with other collecting institutions such as the other Australian deposit libraries, all of which have responsibility for collecting and preserving Australia’s electronic documentary heritage, whether originally published on-line, or in digitised or physical format. It is anticipated that in the long term the National Collection of Electronic Publications will operate within a distributed framework for the selection, description, acquisition and provision of long term access to all information in any electronic format.

In order to develop the Library’s thinking in this area and to test the procedures and policies developed to date within the PANDORA framework, two of the Australian deposit libraries, the State Library of Victoria and the State Library of New South Wales, have agreed to participate in the PANDORA project. The State Library of Victoria will take responsibility for all the procedures from selecting and cataloguing, to liaising with the publisher for permission to archive, and creation of entry screens in the archive. The National Library of Australia will continue to capture and store the publications within the PANDORA archive until such time as the State Library of Victoria has the capacity to undertake its own capture and storage of publications. The critical issue here for the Library is the identification, or development, of a robust archive management system which will minimize as far as possible the work entailed in supporting a National Collection of Electronic Publications. The current thinking on the national co-operative model envisages that the technical framework will consist of distributed digital object servers separately maintained by each collecting body, with unified access provided by the catalogue entries in the National Bibliographic Database. This will be the primary strategy for providing access to the information in the PANDORA archive at the whole item level.

Resource and costing model

Producing a resource and costing model has proven to be a difficult exercise and has not been completed satisfactorily to date. This is largely due to the fact that the PANDORA project is still very much in a research and development phase, with policies and procedures evolving and maturing. We know how we would like things to work, in the sense that we would like to ‘automate’ as much of the work being done manually as we can. However, the extent to which this will possible over the next few years depends on advances in software development in areas like the capture of digital objects, and digital object management systems. Procedures currently being followed in the PANDORA project are interim until something more sophisticated is available.

As procedures are developed for the PANDORA project they are taken up and implemented within the Library's ‘Electronic Unit’, which currently consists of four librarians who do all the tasks from regular scanning and selection of titles from the Internet, liaising with the publishers, cataloguing, devising a capture schedule, version comparisons and creation of title entry screens for each new title selected for the archive. It is currently estimated that it takes one librarian a full working day of seven and a half hours to undertake all of the procedures linked to selecting and capturing a title for the archive, including all the steps from regularly scanning the Internet to the creation of an individual listing of issues attached to each title in the archive. On-going maintenance per title is obviously less time consuming. In addition to these four staff, there is a unit manager who assists with the research and development of PANDORA and a full-time Information Technology staff member who works with Harvest and solves technical problems associated with the capture and storage of publications in PANDORA.

Those closely involved with the PANDORA project feel that they have virtually replicated a whole new acquisition and control system, equivalent in some ways to the work originally required to devise and implement the Library's legal deposit acquisition program. The Library has yet to clearly quantify the likely on-going cost of routinely selecting, acquiring and providing access to on-line Australian publications. The likely cost of a robust software and hardware platform to underpin the archive is still unknown, as is the amount of in-house or co-operative software development the Library may have to undertake on this project.

International collaboration

The National Library of Australia sees great benefit in collaborating with other national libraries or institutions working in the field of digital management or archiving. The Library believes that other national libraries are natural partners and is actively seeking to share information and undertake co-operative research and development with other national libraries. International collaboration is in fact the next logical step in the development of the PANDORA project. Although each national library may approach the actual management of electronic publications in a different way there are a range of key issues where the simple exchange of ideas and information, or the development of agreed principles, would provide mutual benefits for all national libraries.

In order to make progress towards collaboration, the National Library of Australia intends to write an issues paper for discussion amongst national libraries with an active interest in electronic archiving. The paper will attempt to describe the status of work currently being undertaken by national libraries in this area, and to distill those issues where some sort of common approach or agreed principles may help influence key developments such as permanent naming, or even the development of software and hardware platforms.