

## Selective Archiving of Web Resources

### A Study of Acquisition Costs at the National Library of Australia

Margaret Phillips  
Director,  
Digital Archiving,  
National Library of Australia

#### Introduction

During the past decade a small but growing number of national libraries have established Web archiving programs. These programs have taken one or more of four main approaches:

- selective archiving, for example, the archives of the National Libraries of Canada, Japan, and Australia<sup>1</sup>
- periodic harvesting of the country's entire Web domain, exemplified by the archives of the Nordic countries, including Sweden<sup>2</sup>
- thematic collecting, exemplified by the Library of Congress's MINERVA<sup>3</sup> collections of the Elections 2000 and 2002 and of September 11, 2001
- deposit collections, such as STORS at the State Library of Tasmania and the e-Depot<sup>4</sup> at the National Library of the Netherlands

The National Library of Australia chose the selective approach because of its perceived advantages:

- Each item in the Archive is assessed for quality and is functional to the fullest extent permitted by current technical capabilities.
- Each item in the Archive can be fully catalogued and therefore can become part of the national bibliography, and the bibliographic data can be shared. In the Library's own catalogue, Web resources are integrated with all other resources and users need look in just one place.
- Each item in the Archive can be made accessible via the Web to readers immediately because permission to do so has been negotiated with publishers.

- The properties of individual resources and classes of resources within the Archive are known to collection managers. This enhances our ability to develop methods and tools to collect them in the first place, store them, and provide access to them. This knowledge also appraises collection managers of the preservation strategies that will be required to keep the resources accessible for the long term.
- Sites that are inaccessible to harvesting robots can be identified and gathered using other methods as arranged with the publisher. This includes commercial titles, which may require a publisher-supplied password, and databases.

Despite these advantages, each of the archiving approaches has disadvantages, and the selective approach is no exception. It relies on library staff operating in an environment with very new information to judge what will be required for research in the future. It also takes a resource out of context, breaking its links to external resources. From the point of view of archive development and management, the biggest disadvantage is that the process is labour-intensive and the unit cost of each item collected is high.

## Background

We did not know enough about Web archiving to write a plan or strategy and to project its costs.

The National Library of Australia commenced experimental, selective Web archiving in 1996. At that time it was a very new activity and very little had been written about it. There was no one from which to learn. We did not know enough about Web archiving to write a plan or strategy and to project its costs. We could only proceed by taking small practical steps and learning as we went along. No funding had been received to undertake this new business, and we had to redirect collection development staff who showed interest in and aptitude for Web archiving away from more traditional library tasks to this new activity. For example, one person in the IT Division spent part of his time considering how these resources might be downloaded from the publishers' websites and their contents kept on the Library's server. We were obliged to use freely available software for harvesting and managing the downloaded files. Proceeding in this modest way, our early Web archiving costs were quite low and were largely hidden within existing staff budgets.

In the ensuing nine years, operations became much more sophisticated. The library established PANDORA, Australia's Web Archive,<sup>5</sup> as an ongoing operational archive, and nine other Australian libraries and cultural collecting agencies became partners, one by one. On the whole, growth was incremental and the institutions absorbed the increasing costs. A greater volume of archiving activity and the need to support partners contributing to the Archive from remote locations demanded a sophisticated technical infrastructure comprised of a delivery system, an archive management system, and storage. This meant quite substantial development costs, although these costs were still met through existing staff budgets in the Collection Development and Information Technology Divisions.

The Library could not find a suitable archive management system for purchase and therefore developed the PANDORA Digital Archiving System (PANDAS),<sup>6</sup> along with the PANDORA delivery system, in-house. The Library also purchased a Digital Object

Storage System (DOSS), which PANDORA shares with the Library's other digital collections.

The Library still has no additional funding to undertake this activity. In an odd way, this proved advantageous, as it forced us to fund the activity from the ongoing budget allocation from the Australian government, and there was no special short-term project funding that, when it came to an end, left the activity unsustainable.

One thing has not changed in the past nine years: the Library still has no additional funding to undertake this activity. In an odd way, this proved advantageous, as it forced us to fund the activity from the ongoing annual budget allocation from the Australian government, and there was no special short-term project funding that, when it came to an end, left the activity unsustainable.

### **The Pandora Archive Now**

The PANDORA Archive is a collection of significant Australian online publications and websites, developed by the National Library and its partners, that is stored, managed, and maintained centrally at the National Library in Canberra. As of April 30, 2005, the Archive contained 8,235 titles, growing at the rate of approximately 2,400 titles per year. These titles may consist of a single file, such as a text document in Portable Document Format

(pdf), or they may be complex Web objects, such as a large website, consisting of thousands of files in a variety of formats, including text, sound, image, or video. Many of the titles are re-gathered on a regular basis, creating a new "instance" of the title in the Archive, of which there are now 16,736. In building the Archive, the objective is to copy selected titles into the safety of the Archive and to provide access to them in perpetuity.

The Archive includes both static and dynamic online publications and websites and represents a wide range of publication types and formats used by publishers and creators on the Web. It includes online publications and websites that have now disappeared from the live Web and that are no longer available anywhere else.

Most of the titles in the Archive are freely available to anyone, anywhere in the world with an Internet connection. Access to a very small proportion of the Archive is restricted, usually for five years or less, for commercial reasons. These restricted titles can be consulted on a single PC in the Library's main reading room.

Access to the contents of the Archive can be obtained either via a hot link in the catalogue record for a particular resource or from subject and title lists on the PANDORA website. Commercial search engines, such as Google and Yahoo!, index the Archive to the title level.

### **Tasks involved**

The tasks that the staff of the National Library and other partners undertake as part of the PANDORA Web archiving program are determined by a number of factors, including policy decisions and local circumstances. These policy decisions and circumstances affect, in no small way, the cost of this program.

**PANDORA contributors place a high degree of emphasis on preserving the “look and feel” (appearance and functionality) of a publication or Web site, as well as its contents, to the greatest extent possible.**

PANDORA contributors place a high degree of emphasis on preserving the “look and feel” (appearance and functionality) of a publication or website, as well as its contents, to the greatest extent possible. Once the harvester has copied a resource to a server at the National Library, staff of contributing agencies check this copy for completeness and functionality before consigning it to the Archive for public access. This quality assurance process is very time-consuming and, therefore, expensive. It is, in fact, the most expensive aspect of the acquisition process.

Each title in the Archive is catalogued with a record in the National Library’s and other partners’ online catalogues, as well as with the National Bibliographic Database (a union catalogue of records of over 850 Australian libraries, with access provided by the Kinetica<sup>7</sup> service). This policy decision was made because it was considered important that the discovery of online resources be integrated with discovery of all of the Library’s other collections. It was also considered important that these significant publications, which are part of the national collection, should also be part of the national bibliography. This decision does, however, add to the cost of our Web archiving program.

Another significant contributor to staff task time is circumstances involving Australia’s legal deposit laws. At the Commonwealth level and also for most of the States, legal deposit legislation has not yet been extended to include online publications. Only the Northern Territory has recently passed legislation that unambiguously includes online publications in the legal deposit provisions. This means that all other partners, including the National Library, must seek permission from the publisher before copying a resource into the Archive and making it publicly available.

### **PANDORA staffing**

Building the PANDORA Archive, developing the applications on which it depends, maintaining its systems, and establishing expertise and planning for its long-term preservation involves staff from six branches in two divisions of the National Library.

Staff in the Digital Archiving Branch of the Collections Management Division are responsible for selecting and archiving content. The Applications Branch of the Information Technology Division develops and enhances the technical infrastructure. Website Services guides development of the user interfaces for both PANDORA and PANDAS. Business Systems Support maintains the production, testing, training, and evaluation systems and keeps them operating smoothly. Staff in the Preservation Services Branch of the Collections Management Division are responsible for maintaining long-term access to the contents of the Archive.

At the time the costing exercise described below was undertaken in 2004, the National Library’s contribution to the PANDORA Archive in full-time equivalents (FTE) was at the following levels:<sup>8</sup>

**Digital Archiving Branch:**

1 FTE EL2 (manager - librarian)

2 FTE APS6 (one supervisor and one special projects officer - librarians)

4 FTE APS5 (operational staff - librarians)

.17 FTE APS5 (technical problem solver – IT background)

**Information Technology Division:**

.2 FTE EL2 (two managers who spent 10 percent of their time each on system development and system maintenance)

.5 FTE APS6 (two staff who spent 25 percent of their time each on system development and system maintenance)

As mentioned above, staff from Preservation Services contribute at the equivalent of one full-time position, however, the cost of preservation was not included in this costing.

In preparation for the costing exercise, staff prepared a detailed flow chart of all tasks and processes required to acquire an instance for the Archive.

**The Cost of Acquiring Online Publications and Websites**

Although the Library has known that the unit cost of archiving (acquiring) an online publication is high compared to the acquisition of a printed book or serial, until recently we have had no precise information about the cost of Web archiving. In 2004, the Library decided to cost its legal deposit activities for serials and monographs and, since collecting Australian online publications is regarded as an extension of our legal deposit

responsibilities, it was decided to cost that activity as well.

**Scope of the costing**

The costing exercise examined the cost to the National Library of acquiring an “instance”<sup>9</sup> and adding it to the PANDORA Archive. The boundaries of what costs would be included and excluded were defined.

Only the direct costs incurred were included:

- staff costs for the Digital Archiving Section at the National Library
- the Digital Archiving Section’s share of administrative costs, such as travel, training, conference attendance, and office supplies (supplier costs)
- infrastructure development and maintenance costs, such as IT staff and hardware and software purchases

Costs that were excluded were:

- indirect costs, such as the provision of work stations to staff members
- lighting and building maintenance
- cost of preserving the contents of the Archive

Only costs incurred by the National Library were considered; costs that partner agencies assume in employing staff to contribute to the Archive were excluded.

## Methodology

In preparation for the costing exercise, staff prepared a detailed flow chart of all tasks and processes (that is, tasks undertaken by the Digital Archiving Section) required to acquire an instance for the Archive. Key cost drivers (activities) were determined. By a consensus process, staff estimated the average time in minutes that is spent per staff member on each driver each day. A working day contains 441 minutes.

- identification and selection – 30 minutes
- publisher contact, negotiating permission to archive the title, and filing of correspondence – 30 minutes
- gathering, quality assurance, and archiving instances – 210 minutes
- cataloguing – 81 minutes
- other activities (includes correspondence with indexing and abstracting agencies, reference enquiries based on the Archive, and Digital Archiving staff contribution to the development of PANDAS) – 60 minutes
- partner liaison and support - 30 minutes (activity not included in this costing)

Not all staff undertake all of these tasks, and some staff do more or less of them than others. For instance, the supervisor of the Section also carries out administrative tasks and staff supervision. This means he does less gathering, quality control, archiving, and cataloguing than the others.

Of the tasks listed above, the manager of the Digital Archiving Section undertakes only partner support/liasion and occasional contact with publishers. Most of her time is spent on administration, policy development, publicity, and liaison with other organisations involved in Web archiving. However, her salary is regarded as a necessary component of the Library's overall digital archiving costs and was included in the overall costing.

Some staff are involved in tasks that were completely out of scope. For instance, two librarians spend time each week on the reference desk in the Reading Room. This time was subtracted from the total time spent on archiving.

As the leading partner in PANDORA and the supplier of the technical infrastructure, the National Library has a support role in relation to all other partners, and this involves additional cost. All Digital Archiving Section staff spend varying amounts of time liaising with and providing technical support to partners. Operational staff spend an average of 30 minutes each per day in partner liaison and support, while the manager spends an average

of ten minutes per day. This activity was not included in this costing exercise but was reported separately.

### Calculating the costs

In the next stage of the costing exercise, an Excel-based costing methodology was designed to calculate activity costs per instance archived.

The salaries of Digital Archiving staff and daily time spent on each of the drivers by each staff member were entered into the Excel spreadsheet. The total number of minutes spent on each driver and the staff cost of each driver per day was then calculated.

Various supplier costs, as described earlier, were then added. Infrastructure development and maintenance costs were also included at this stage.

A total of 937 instances were archived by the National Library during July – October 2004, an average of 13 per working day.

### Acquisition costs

With all of this data entered, the spreadsheet calculated that each archived instance cost AUD\$178.68, excluding the activity of partner liaison and support. It also gave us information about the component breakdown of this cost:

- Digital Archiving staff cost per instance archived - AUD\$168.36
- Supplier costs per instance archived - AUD\$3.41
- Infrastructure development and maintenance costs per archived instance - AUD\$6.91

Here the stark reality of the high cost of the labour-intensive, selective approach to Web archiving is apparent. Staff costs comprise 94 per cent of the unit cost.

### Comparison with print

The stark reality of the high cost of the labour-intensive, selective approach to Web archiving is apparent. Staff costs comprise 94 per cent of the unit cost.

The high cost of acquiring Web publications compared to printed publications was also highlighted by comparing these costing results with similar ones undertaken at the same time for legal deposit printed monographs and serials.

- Cost of acquiring a legal deposit monograph - AUD\$43.77
- Cost of acquiring a legal deposit serial issue - AUD\$11.29

Attempting to compare these costs is a bit like trying to compare the unit cost of transporting watermelons, bananas, and grapes to market. They are not the same in a number of dimensions. An “instance,” the unit of measurement for the Web publication costing, is not the equivalent of either a monograph or a serial issue. Nevertheless, despite considerable differences in the nature of these publications and the processes involved in

acquiring them, it is quite clear that the cost of acquiring Web publications is substantially higher than for print publications.

It should be noted here that no purchase costs are involved in these figures. The Library receives its legal deposit printed materials free of charge, and the Web publications are harvested free of charge from the publishers' websites. These costs relate solely to the acquisition process. Shelf preparation of items was included in the costing. The subsequent costs of binding, stacks management, and collections care were not considered.

### Cost of particular activities

The National Library was also interested in analysing the cost of individual drivers (activities). The staff broke down the costs per instance as follows:

- identification and selection – AUD\$10.16
- publisher contact, negotiating permission to archive the title, and filing of correspondence – AUD\$10.34
- gathering, quality assurance and archiving of instances – \$71.09
- cataloguing – AUD\$27.42
- other activities (includes most of the manager's activities, correspondence with indexing and abstracting agencies, reference enquiries based on the Archive, and Digital Archiving staff contribution to the development of PANDAS) – AUD\$59.67

Note that this cataloguing cost is not very useful since it is a *per instance* cost, and resources are catalogued at the title level, not the instance level. Given that each title in the Archive has approximately two instances, a more realistic cataloguing cost is \$54.84.

### Possibilities for reducing costs

From this activity-level costing information we have been able to consider how we might reduce our costs. We could reduce our costs by changing our approach to Web archiving (changing our policies) or by finding ways to perform the tasks involved more efficiently

We are working with government publishers to supply metadata for online publications, which will be batch loaded into PANDAS for automatic harvesting of the described publications. In effect, the publishers will be identifying what will be archived in PANDORA by supplying the metadata for it.

or both. For the foreseeable future, we will maintain existing policies, for instance, of creating catalogue records for each title and of undertaking rigorous quality assurance. There are other opportunities for savings.

The identification and selection of titles is fundamental to the selective approach to archiving and, at first glance, it is difficult to see how we might reduce time spent on this activity. However, we are working with government publishers to supply metadata for online publications, which will be batch loaded into PANDAS for automatic harvesting of the described publications. In effect,

the publishers will be identifying what will be archived in PANDORA by supplying the metadata for it. PANDAS has yet to be enhanced to enable this batch harvesting and processing to take place, but we anticipate that once it is operational, the average cost of acquiring the publications of participating agencies will be significantly reduced. The metadata being supplied by a small number of publishers is already being automatically converted to MARC records for inclusion in the National Bibliographic Database and will be downloaded to the Library's online catalogue. This will help to reduce cataloguing costs.

Further calculations on the components of the most expensive driver (gathering, quality assurance, and archiving of instances) revealed that quality assurance comprised 86 percent of the cost. If we could identify reliable quality checking software to plug into PANDAS, then the cost savings could be worthwhile.

The Library is also lobbying the Australian government for the extension of legal deposit to online publications in order to obviate the necessity of obtaining permission from publishers to copy titles into the Archive. Staff time will be reduced after this comes to fruition.

Technology will improve and will enable us to reduce costs. The International Internet Preservation Consortium,<sup>10</sup> of which the National Library is an active participant, is working on a suite of Web archiving tools, including a harvester designed specifically by and for national libraries. This is expected to make our work more efficient.

## Conclusion

The costing calculation confirmed what we already knew – Web archiving is a costly business when compared to the acquisition of printed materials and that currently the per unit cost of selective archiving is particularly high because of its labour-intensive nature. This study has also given us useful additional information against which to evaluate our program. Are we really wedded to the policy of cataloguing each title in the Archive? Yes, but perhaps we can find less labour-intensive ways of creating the record. How committed are we to the very expensive quality assurance process? Very committed, but installing suitable checking software will be a high priority when it can be located.

Increasing sophistication of Web archiving technology will increase the efficiency of all Web archiving programs, including those using the currently labour-intensive selective approach, and lower unit costs over time.

## Acknowledgement

I would like to acknowledge the major contribution to this costing made by Nizam Yoosuf, Manager, Finance Branch, National Library of Australia. He designed the costing methodology and assisted with the interpretation of the results.

## Notes

<sup>1</sup>Library and Archives Canada. *Electronic Collection: a Virtual Collection of Monographs and Periodicals*. <http://www.collectionscanada.ca/electroniccollection/> (accessed 22 March 2005); National Diet Library of Japan. *Web Archiving Project (WARP)*. <http://warp.ndl.go.jp/> (accessed 22 March 2005); National Library of Australia.

PANDORA, *Australia's Web Archive*. <http://pandora.nla.gov.au/index.html> (accessed 22 March 2005).

<sup>2</sup>National Library of Sweden. *Kulturawa3*. <http://www.kb.se/kw3/> (accessed 22 March 2005).

<sup>3</sup>Library of Congress. *MINERVA*. <http://www.loc.gov/minerva/> (accessed 22 March 2005).

<sup>4</sup>State Library of Tasmania. *STORS: Long Term Storage of Tasmanian Electronic Documents*. <http://www.stors.tas.gov.au/> (accessed 22 March 2005); Oltmans, E. and H. van Wijngaarden. 2004. Digital preservation in practice: the e-Depot at the Koninklijke Bibliotheek. *Vine* 34 (1): 21-26.

<sup>5</sup>Information about PANDORA, Australia's Web Archive, and access to its contents are available at <http://pandora.nla.gov.au/index.html>.

<sup>6</sup>Information about PANDAS is available at <http://pandora.nla.gov.au/pandas.html>.

<sup>7</sup>National Library of Australia. *Kinetica*. <http://www.nla.gov.au/kinetica/> (accessed 22 March 2005).

<sup>8</sup>Levels and pay scales are explained in Attachment A – Salary Table of the *National Library of Australia Certified Agreement 2004-2007* available [here](#).

<sup>9</sup>An “instance” is a single gathering of a title. It includes the gathering of a monograph that has been archived once only, the first gathering of a serial or integrating title (for example, a website that changes over time), and all subsequent gatherings.

<sup>10</sup>*International Internet Preservation Consortium* <http://netpreserve.org/about/index.php> (accessed 22 March 2005).