WEB ARCHIVING IN A WEB 2.0 WORLD

Edgar Crook

Web Archiving & Digital Preservation Branch
National Library of Australia
Canberra
ecrook@nla.gov.au

ABSTRACT

The National Library of Australia is the lead institution for digital archiving and preservation in Australia. Its PANDORA Archive has been the repository for archived web resources in Australia for over ten years and is a mature but continually developing system. The archival management system PANDAS that underpins the Archive, is as of 2007, in its third major revision. Other web archiving activities also now include annual Australian Domain Harvests and the usage of Archive-It, both of which are conducted in conjunction with the Internet Archive.

This paper discusses the current state of web archiving in Australia, and how libraries are adapting their services in recognition of the expanding role that online material plays in their collections. For many years it was considered that archiving could only ever completely capture a small, albeit representative, sample of the Internet. Today the gap between what is available and what can be archived is decreasing. But as our archives and our archiving abilities increase, we are still confronted by new technologies and web 2.0 applications. Using as an example the 2007 Federal Election in which a large number of interactive sites such as Kevin07, MySpace and YouTube were archived the paper will show how Australian web archivers continue to adapt to and meet new challenges.

HEADINGS TO BE IN UPPER AND LOWER CASE

INTRODUCTION

This paper discusses web archiving in Australia, as the National Library of Australia is the lead player in this field and the author works within it, it will have a strong emphasis on what is happening in that institution. In saying this I would also like to acknowledge that there are other web archiving projects in Australia such as Tasmania's Our Digital Island (http://odi.statelibrary.tas.gov.au/) or here in the Northern Territory the Territory Stories (http://www.territorystories.nt.gov.au/) service.

The National Library currently undertakes web archiving using three different methodologies. Selective archiving within PANDORA - Australia's web archive, and in conjunction with the Internet Archive contracted whole domain harvesting and utilisation of their Archive-IT service. In this way we are working towards creating a comprehensive collection of Australian online publications. However, as new technological challenges arise the Library has to continue to adapt and plan for the future, amending its collection scope and forging new partnerships to continue this important work.

THE 3 ARCHIVING METHODOLOGIES

The PANDORA Archive (http://pandora.nla.gov.au) has been archiving Australian web publications since 1996. In that time it has established itself as an internationally respected repository. The main achievements of PANDORA have been to forge a network comprising nine Australian archiving participants, comprising all the mainland state libraries as well as AIATSIS, The National Film and Sound Archive and the Australian War Memorial; the creation of PANDAS our archival workflow system; a system to persistently identify our archived content and our relationships with a range of indexing and abstracting agencies as well as a myriad of Australian publishers

As at 1 July 2008 the Archive contained 19,307 titles containing 53,112,080 files which amounts to 2.2TB of data.

The PANDORA Archive has achieved the following objectives:

- a world class archive of selected Australian online publications, such as electronic journals, government publications, and web sites of research or cultural significance;
- policy, procedures and selection guidelines for the collection and provision of long term access to items in the Archive;
- a collaborative national approach to the archiving and long-term preservation of Australian online publications, involving the participation of State libraries and other cultural institutions
- a digital archiving system (PANDAS) to streamline the gathering and loading of publications into the Archive, store information about them, and manage public access to them.
- a scheme for persistently naming all objects in the Archive and a resolution service to them;
- arrangements with indexing and abstracting agencies under which publications that they are indexing or abstracting are archived in PANDORA and persistent identifiers are provided for citation and persistent access;
- and contents as at 1 July 2008 totalling 19,307 titles containing 53,112,080 files which amounts to 2.2TB of data.

Within the Archive there is a wide range of publications, half of which have been archived from state and federal government websites and the other half reflecting the full diversity of Australian research and culture. The type of publication within the Archive can be a single pdf document or a complete organisation's website containing thousands of files. There are also blogs, podcasts and videos.

Directed selective harvesting of web publications that are considered to be of long-term cultural or research value has necessarily meant that only a relatively small proportion of the Australian domain has been archived. The Library understanding this, has from 2005, contracted the Internet Archive (http://www.archive.org) to do annual directed crawls to gather what it can of the .au domain. These crawls which run for about one month each year have gathered amounts of data that dwarfs the contents of PANDORA, in 2007 for instance the whole domain harvest picked up 18 TB of data in a month whilst PANDORA gathered 2 TB of data in 11 years. The 2008 gather is expected to crawl a billion files.

Domain Harvest Date	2005	2006	2007
Unique documents (files) crawled	185,549,662	596,238,990	516,064,820
Total documents (files) crawled	189,824,119	621,664,876	523,510,945
Hosts	811,523	1,260,553	1,247,614
Raw data size	6.69 TB	19.04 TB	18.47 TB
Compressed ARC file size	4.52 TB	10.48 TB	10.18 TB

Koerbin, P, The Australian web domain harvests: a preliminary quantitative analysis of the archive data, NLA, 2008 - http://pandora.nla.gov.au/documents/auscrawls.pdf

The harvests which are conducted using Heritrix (http://crawler.archive.org/) are large but are not completely comprehensive as they only gather for a month a year (and a lot can come and go on the Internet in that space of time), they obey robots.txt rules, and although Heritrix has powerful functionality there are still websites that are technically difficult for it to crawl. Whatever the shortcomings, the whole domain harvests do gather to a significant degree and thus the resulting large amount of data collected makes any attempt at quality assessment of individual websites difficult. So unlike PANDORA where we can identify problems within each title and perform quality assurance work – gathering and fixing pages where necessary – here this is not possible. Another drawback is that again unlike PANDORA (where we gain individual publisher's permission to archive) here this is also unfeasible. Therefore due to current Australian copyright law wherein online publications are not included within legal deposit, we are unable currently to publicly display what has been preserved. That is not to say that no work is being done on this archive, as academics are working on this data and their research we think will be valuable.

Archive-It is a hosted web archive service provided by the Internet Archive. The first (and only thus far) Australian organisation to use this service has been the Asian Collections section of the National Library (http://www.nla.gov.au/asian/asianwebarchive.html).

It is used to gather collections of overseas websites recording particular social and political events, as it is not expected that any other regional organisation will fulfil this role. The hosted option was chosen as it was perceived that it could be a quick and easy

way of gathering and storing collections, and not requiring technical skills or big ((tranches)???) of staff time. This has proved only in part to be true as it was quickly noticed that to successfully create collections takes far more curatorial time than was initially envisioned. Selection of which websites to crawl is an often misunderstood activity and can take up surprisingly large amounts of time.

Whilst using Archive-IT has benefits in that you do not need to worry about hosting and preserving gathered content, there are also some major drawbacks about not being in control of the archival and display functions. Archive-IT allows for a set of your chosen seed urls to be gathered, however once the gatherer has gone out and gathered, or not gathered, your files, there is no way to manually fix up any broken or missing content, as can be done with your own system. Similarly there is no real control or ownership of the display process, so that for instance a link to a seed URL which has failed to gather will still appear within a collection. Another drawback is that if you discontinue your annual subscription to the service your collections are dispersed back into the general Internet Archive pool of content. Notwithstanding these issues the Library plans to continue to archive using this method.

Sites archived with Archive-IT for Asian Collections - National Library of Australia

Papua New Guinea Government and Research Websites

Selected Papua New Guinean governmental and significant research institute web sites archived from 2008. Some 2008 archived sites were not current at the time of capture.

Thailand Elections 2007

Selected International intergovernmental and Thai web sites related to the general elections of Thailand 2007.

Cambodian National Election 2008

Selected governmental, political parties and media websites related to the Cambodian National Election 2008.

Burmese Uprising 2007

Selected international web sites related to Burmese monks uprising in Sept-Oct 2007

Lao PDR Government and NGO Websites

Selected Lao PDR governmental and NGO web sites archived from 2008.

Papua New Guinea Elections 2007

Selected international intergovernmental and Papua New Guinean web sites related to the Parliamentary elections of Papua New Guinea 2007. Includes a snapshot of the PNG

Electoral Commission website, Papua New Guinea election petitions, the Report of the Commonwealth Pacific Islands Forum Election Assessment Team and a media statement from Transparency International (PNG).

East Timor Elections 2007

Selected international intergovernmental and East Timorese websites related to the presidential and parliamentary elections of East Timor 2007. Includes voter education and political advertising material.

Indonesian Islamic organisations 2007

Selected Indonesian web sites

GATHERING FILES

When PANDORA began life many websites could not be gathered, due to the early inabilities of web gathering software. Websites in basic HTML were able to be gathered, but sites using such simple things as frames initially posed many problems. But once one problem was solved another was raised. For web archivists therefore the amazing development of the Internet, its JavaScript, applets, Cascading Style Sheets, Shockwave Flash and a myriad of other file and style formats has been just one damn thing after another. While most file types problems have been overcome, multimedia content remains an issue. From Real Player files and now to podcasts we have had to contend not with the complexities of the files themselves, but with their delivery systems. This is particularly the case now with video.

The collection for the 2007 federal election was the biggest thus far attempted. The National Library was responsible for archiving all national resources to do with the election, including party websites, lobby groups, some candidate's websites, blogs, videos and media websites. The state libraries were responsible for collecting the

Candidate and party and local media websites in their respective states. In total over 350 websites were archived by the National Library and its partners, many of these sites were gathered multiple times to capture changing content. The biggest challenge in archiving this election was the large number of videos, which were not a problem in themselves, but were made so by the same delivery mechanism and embedding technologies that make them so useful for users.

We took different approaches to archiving videos depending on the nature of the website. For general websites where there were single videos on a webpage we downloaded the video files separately using various freely available downloading tools (web harvesters generally cannot automatically gather videos) and then using file converters changed the files from .flv to into something more end user-friendly like the Mpeg format. Where there were a number of videos linked from one webpage the videos were retained in their original format and an flv player installed into the gathered files, so that the files could be rendered for users as easily as in the live website. When we gathered the YouTube election website (http://nla.gov.au/nla.arc-76644 consisting of over 700 Videos) we called upon the technical expertise in our IT department and were able to extract from the live site the URIs for the videos, download them and do the necessary changes to the harvested web pages.

None of these processes was quick and all required a fair amount of technical skill including the ability to recode the archived pages with the changes we had to make to make the videos playable within the Archive.

Due to the election we were also called upon to preserve the previous government's online record. We had made plans in anticipation of this (as we had for previous elections) and had archived all the government ministerial websites whilst they were in caretaker mode just prior to the election date. This was done as it was suspected that with a possible change in government there may be a general deletion of content from the Internet, as this had occurred in other jurisdictions. This foresight was rewarded as some government departments' web publications, particularly those with changed portfolios, were removed from public view.

GATHERING DIRECTIONS

Given the National Library's .au domain harvesting and PANDORA selective archiving we can now relatively safely say that a significant amount of Australian publications or websites are being archived. Though to what extent this is a comprehensive or complete collection we cannot say. We do know however that there remain large gaps.

We are not comprehensively archiving websites produced by Australians outside of the .au domain (although hopefully the Internet Archive will get much of this). Outside of the traditional publications, but in some ways more important, we are also not gathering the vast amount of individually produced creative content which is hosted on video, photograph and art hosting websites, blogs, virtual worlds and social networking sites.

There are attempts at gathering some of this content but they are small directed projects. One such is a recently curated Australian dance collection, which seeks to collect together some examples of Australian dancing as posted on various websites and video hosting sites.

Although the National Library has made arrangements with Flickr, and we have received archiving permission from MySpace and YouTube, the Library has only archived or collected tiny portions of these resources. Other online resources where there is Australian activity such as virtual worlds (Second Life etc.) and social networks (Facebook, Bebo etc.) are also not archived. The prime reason for this is that the content generally has copyright and privacy features that would disallow it, or because of the nature of the resource, which places it outside of the public Internet.

Where we as librarians can individually also make a difference to preserve our online heritage is by taking a responsibility where we can within our organisations or parent organisations to make sure that what is put up on our organisation's websites is preserved or maintained. The movement, particularly noticeable in government and academia, of publications going from print to online continues. We have found is that it is not possible to trust or expect these online publications to remain available on publisher's websites in the long (or even short) term. Universities are now compelled to create digital repositories for their intellectual output and in this way their publications are remaining accessible. While we might expect government websites to retain public access to their publications, experience shows this is not always the case. Therefore if a publication is valuable to your collection or users it is wise for each of you to make some attempt at action to keep it available in the long term.

FUTURE DIRECTIONS

In what is an ever growing Internet landscape there are always new technologies and newly identified gaps in our gathering that are in need of addressing. Web archiving therefore is never it seems going to become an area whereby established practices or protocols in collection development will ever be fully established or maintained. We will always need to be identifying and gathering content and not waiting for it to come in to us. The web is too dynamic, the technology so changeable, the number of publishers so vast that it is unlikely that we will ever be able to establish as did our print forbears a physical deposit style system that will be scalable.

When the National Library first began web archiving, there were few tools and few institutions out there which we could work with or learn from. Consequently we had to devise our own systems and tools. To this end the National Library built in-house its own archival management system - PANDAS. This system is now in its third and we think final stage as the Library can no longer independently sustain such an investment in development. However, now that web archiving is an established practice in a range of international institutions, there are partners with which we can share the load. The International Internet Preservation Consortium (IIPC) which contains as members all the leading web archiving national libraries, including the National Library of Australia, and other related bodies is leading this development. In this way the Library can continue its leading role in web archiving by adopting, adapting and developing tools with partner libraries and institutions.

BRIEF BIOGRAPHY OF PRESENTER

Edgar Crook has worked in the National Library of Australia since 1999. Since 2000 he has worked on PANDORA: Australia's Web Archive. Previous to this he worked in ACT public libraries.