PANDORA – past, present and future
National web archiving in Australia

Transcript of a talk given by Dr Paul Koerbin, Manager Web Archiving, National Library of Australia at the Seminar Kebangsaan Sumber Electronik Di Malaysia 2012, at the Bayview Beach Resort, Penang on 6 December 2012.

[Slide 1]

PANDORA – past, present and future; or, national web archiving at in Australia.

A talk given at the Seminar Kebangsaan Sumber Electronik Di Malaysia 2012, Bayview Beach Resort, Penang, Malaysia, 6 December 2012.

[Slide 2]

I am very pleased to be invited to Malaysia and this conference. My thanks especially to Mazmin Binti Mat Akhir for initiating and managing my participation.

I have been invited to talk about the PANDORA Archive, which is the National Library of Australia's national web archiving initiative.

PANDORA was one of the world's first web archiving programs, being set up in 1996. So we have more than 15 years experience in this activity.

In this presentation I will briefly outline the history of our web archiving program, the current state of our web archiving activity and issues for future directions in web archiving.

In doing this I will highlight what we have learned from our experience, suggest what I think are some of the main current issues and objectives and in doing so I hope I will convey not only the challenges but also the importance of web archiving.

[Slide 3]

In order to provide some context and history I will address four questions:

1. Firstly, what is web archiving? – It is important to clearly understand what we are talking about.
2. Why, from the perspective of the NLA, it is important to do it?
3. How did we approach web archiving when we started?
4. And why we approached web archiving in the manner we have?

Then I'll quickly run through a timeline highlighting some of the major event milestones in our web archiving experience.

[Slide 4]

Firstly, what do I mean when I'm talking about web archiving?

The main point I wish to emphasise is that it is much more than just copying and storing data.

By 'web archiving' I am referring to, and mean, a sustainable commitment to number a number of processes and actions.

Web archiving involves:

1. Selecting or scoping what is collected – that is, collection development;
2. Acquiring the content from the web – mostly this is done through harvesting with a crawl robot;
3. Preserving what is collected – which in itself is a complex series of strategies and activities; and,
4. Providing access to the collection – which is the objective of collecting, maintaining and preserving the material.
5.

In summary, it is the action of determining and creating (and preserving) the heritage artefact out of the dynamic, ephemeral, living thing that is the web.

[Slide 5]

What it is that we are collecting?

Firstly, we are dealing with complex objects for the most part. Even a simple web page will be constituted of text, images, style elements and scripts. And of course the dynamic delivery of websites from content management systems and databases makes this all the more complex.

In the context I am discussing it is also important to understand that this is third party content – that is, I am talking about collecting websites or web documents created by others. In other words, as the collecting agency we have no control over the formats or systems used to create and deliver the content.

There is commonly also embedded content such as audio-visual material in a variety of formats – an additional problem for collecting and preservation.

The process of harvesting delivers, not the underlying database, but the outcome of an HTTP request – so essentially only a single browser view.

This does has a practical advantage of applying a normalising process to the acquisition of content – with dynamic content becoming static HTML.

Web archiving may involve deposit – though this is harder to deal with in many respects. It is also likely to become more and more necessary as more content, such as eBooks, are protected by TPMs that prevent acquisition through harvesting.

[Slide 6]

I would like to reflect for a moment on why we – the National Library of Australia – actually do web archiving.

Our National Library Act states that the function of the Library is to maintain and develop a comprehensive collection relating to Australia and Australian people. Moreover it is a function of the Library to make the national collection available in the national interest.

When we observed the emergence of the Internet in the mid-1990s it was evident that this was another important publishing medium we needed to deal with. And this was no great conceptual leap since we already had experience with non-print formats including manuscripts, pictures, microform – as well as a strong history of collecting ephemera.

As the major collecting institution in Australia we also had, and have, a responsibility to show national leadership in addressing new challenges.

We were also fortunate at the time to have people in the Library with the vision to recognise that collecting publications from the web was important for a comprehensive record of Australian culture and published expression.

[Slide 7]

Recognising the need for action is one thing – how to accomplish such action is another.

The approach taken by the Library from the start was to do what we could with the limited resources available to us. This meant proceeding and working at a scale that permitted outcomes; even though this means all problems and issues are not resolved from the outset.

So we began with a highly selective approach to what we would, and indeed could, collect.

We adopted a ‘proof-of-concept’ project approach – that is a practical approach – at first and developed that into established operational activity as soon as practicable.

We developed workflows, and an infrastructure to support them, in-house, because at that time there were no ‘off-the-shelf’ or open-source systems to be had.

We employed bright university students to do the technical development – because we could afford them; they were enthusiastic and creative; and they tended to stay for the duration of set project tasks.

Also, from very early on, we pursued a collaborative approach as far as we could; though this is limited to curatorial collaboration, not a collaboration of development skills or resources.

[Slide 8]

The reasons for the approach we took in establishing our web archiving program are largely pragmatic. It was a matter of doing what we could achieve at any given time with the resources available to us.

We took a staged approach: so, we worked out scoping and selecting guidelines before undertaking harvesting; we began harvesting before we had a full workflow system; we continued collecting and cataloguing content before we had a delivery system to make content accessible; we engaged curatorial participation before we had a web-based workflow system to manage remote contributors.

We had to work within the existing legal environment and constraints – specifically, the fact that we do not have legal deposit for digital materials. This meant we had to obtain individual publisher permissions – and it still does – in order to achieve the objective of delivering access to the archival content.

We do not have a Research & Development resource in the Library, so that meant that the best way for us to proceed and learn was through experience – a heuristic approach.

[Slide 9]

I would now like to quickly run through a timeline to highlight what I think are the major event milestones for web archiving at the NLA.

The first thing we did was establish a unit (the Electronic Unit) – in April 1996 to survey what was being published on the web in Australia and to develop selection guidelines.

In September 1996 there was a major restructure in the Library's technical services. We moved from separate sections for collection development, acquisition and cataloguing to multi-tasking work groups. The Electronic Unit (as the web archiving team was originally called) was aligned as a work unit with our Serials team as most of the Electronic Unit staff were originally serials cataloguers.

The name PANDORA – an acronym for Preserving and Accessing Networked Documentary Resources of Australia – was given to the project in November 1996.

The earliest collecting of content dates from late 1996 and into early 1997. By the middle of 1997 we had collected 30 titles or so – though they were not accessible to the public.

Online access to the PANDORA Archive began in May 1998. So, not right at the beginning of the project but not too long into it either.

[Slide 10]

By July 1998 the first PANDORA partner began participating. This was the State Library of Victoria.

There are now 11 participants (including the NLA). The last to join was the National Gallery of Australia in 2010.

A high priority for us was to develop a workflow system in order to achieve efficiency, particularly in managing the recording of selections and permission; and in order to allow librarian curators to run the harvests – that is scheduling them and scoping the harvest crawler.

We developed a workflow system called PANDAS – PANDORA Digital Archiving System, of which I speak about in some more detail later – and the first version was released in June 2001. A little over a year later in August 2002 an improved and upgraded second version was released.

In the meantime the profile of web archiving in the Library was on the rise and the Digital Archiving Branch was established.

In July 2003 the NLA was one of the founding members of the International Internet Preservation Consortium – a group of national libraries who, along with the Internet Archive, came together for the purpose of developing and sharing common standards, practices and tools for web archiving.

[Slide 11]

Up to this point the selective approach of PANDORA had been our only web archiving activity. However in 2005 we collaborated with the Internet Archive to do the first of our annual large scale harvests of the whole .au web domain. Earlier that year the IA had released its archival web crawler, Heritrix, to production level and so the time was right for us to take advantage of their expertise and capacity to undertake large scale crawling – a capacity we did not have in the Library.

In July 2007 the third version of PANDAS was released. This was a completely re-engineered version of the system with a much enhanced interface and workflow management.

In 2010 the PANDORA search index was moved to the Library's one search discovery service called Trove. While this represented some improvement and facilitated the searching of the web archive as part of broader searching of library resources, the interface remains less than ideal for the web archive.

Finally I would mention our Australian government web archive. We have of course collected a lot of government material – indeed more than 40% of the titles included in PANDORA are government publications. However, in 2010 we obtained a whole-of-government permission to collect federal government material. This has allowed us to undertake bulk collecting of federal government websites. We have done two collections so far and expect to make the collection accessible to public very soon.

[Slide 12]

This brings me to the current status of web archiving at the NLA.

I will say something about the organisational structure, staffing and the skills required; and the PANDORA participants.

I will also outline our current collecting and the extent of our collections.

And I will give a very quick overview of our workflows and the PANDAS workflow system.

[Slide 13]

The Web Archiving and Digital Preservation Branch is part of the Library's Collections Management Division. That is to say, while web archiving is a highly technical process in many respects, it is managed from the core operational area of the Library.

The Web Archiving Section consists of a strategic, policy and operational manager, an operational team leader (a senior librarian) and three web curators (librarians). We share our organisational structure with the Digital Preservation and Digital Collecting Support teams.

[Slide 14]

Web archiving is part of the Library's core business and staff require the skills to contribute to core operational tasks, as well has having specialist skills. Staff may be understood as librarians, cataloguers, web curators or web archivists.

The web archiving team perform all the tasks associated with collecting web materials including: selecting, negotiating permissions, acquiring the content (i.e. scoping and scheduling the harvesting), quality checking of harvested content, cataloguing and publishing (i.e. preparing content for access and display).

They need highly developed collection development skills with an interest in current affairs and what is published online – and how it is published. Moreover, they need the skill to select from the vast amount of material in accordance with the Library's objectives, priorities and resources.

They need experience in cataloguing using the major standards. And we are now in the process of moving towards the implementation of RDA.

Staff need a good aptitude for understanding the technical, such as the way websites are constructed and delivered, so they can scope the harvesting and undertake problem analysis and fixing.

More than anything, perhaps, they need to demonstrate engagement and initiative and self-learning to deal with rapid change and the many challenges that flow from this work. The work requires a proactive approach to skill development.

[Slide 15]

Currently 11 participants – that number includes the NLA – jointly curate the PANDORA Archive.

This includes most state and territory libraries, with the exception of the Australian Capital Territory library service and the state library of Tasmania. Tasmania has maintained its own web archive for nearly as long as PANDORA. The state libraries take responsibility for collecting resources that specifically relate to their state or local jurisdictions.

In addition to state libraries, four other heritage collecting institutions participate. These institutions take responsibility for developing the collection in their areas of expertise. These are the National Film and Sound Archive, The Australian War Memorial, the Australian Institute of Aboriginal and Torres Strait Islander Studies and the National Gallery of Australia.

[Slide 16]

We work together on the basis of a formal memorandum of understanding made between the NLA and each of the participants. This covers mutual responsibilities and adherence to common policies and procedures.

However, I would stress that each participant agency has its own selection guidelines and approaches to selection are not standardised across all participants.

Staff in PANDORA participant agencies undertake the same work as the web archiving librarians at the NLA including selection work, negotiation of permissions, scoping and scheduling harvests, quality assurance checking, cataloguing and preparing content for access.

The entire technical infrastructure along with the actual archival content is maintained centrally at the NLA in Canberra.

[Slide 17]

In talking about web archiving at the NLA I have mostly spoken about the PANDORA Archive – that is the selective web archiving for which we have been collecting content since the late 1990s and which we make fully accessible to the public.

I have already mentioned that we have also undertaken large scale harvests of the .au top level domain. We have done this since 2005 working in collaboration with the Internet Archive, who manage the harvests to our specification, index them and ship the content to us.

Since 2011 we have also done separate bulk harvests of Australian federal government websites based on a specific list of URLs. Management of this harvest is also outsourced to the Internet Archive.

These large scale collections are not yet available to the public although we should have access to the Australian Government Web Archive ready early in 2013.

I should also mention that the NLA has a strong commitment to collecting resources from the Asia Pacific region and has created a number of collections of web materials using the Internet Archive's Archive-It service. This is a web archives collecting and hosting service and, unlike the other collections, these collections are not maintained at the NLA.

[Slide 18]

This diagram shows the range of web harvesting collecting methods, from selective through 'themed', seed-list based harvests and scoped domain level harvest, to crawling the entire world wide web.

It shows that at the NLA, we are working to cover a range of these approaches – with the obvious exception of trying to do the whole web!

I don't have time to dwell on the relative merits and otherwise of the methods but some are outlined in the coloured boxes. The point being that different approaches produce different outcomes to suit the collecting objectives for the target material. In other words, we don't see all web content as the same in respect to how it is best or appropriately collected, and a range of approaches is the objective.

[Slide 19]

Certainly the different methods produce a different scale of collecting as can be seen by these statistics.

You can see that around 15 years of collecting for the selective PANDORA Archive has gathered around 4% of the amount of content collected from 7 whole domain harvests.

[Slide 20]

Here are the figures in chart form which show this comparison more clearly.

I would emphasise, however, that while collecting on an appropriate scale is an important objective, selective archiving also has advantages that I've already alluded to, such as ensuring quality content is collected at critical and specific times and as thoroughly as possible. And, working at that scale, providing access to the archive is more feasible.

[Slide 21]

I don't have time in this presentation to go through workflows in any detail. However I have mentioned our selective web archiving workflow management system, called PANDAS, a number of times.

So I will quickly run through the workflow that it actually manages.

[Slide 22]

This is a screenshot of the main screen of PANDAS – the PANDORA Digital Archiving (workflow) System.

It is divided into a logical series for workflow tasks using 'work trays'. These are grouped as Selection, Permission, Gather (covering the scheduling and scoping of the harvesting), Preserve (that is the quality checking process and the authorising the archiving) and Publish (which authorises making the archived content available to the public).

What you can see here is an Agency Administrator view for a participant agency – in this case the NLA. Each individual web curator normally works with their own personalised view for managing the work for which they are responsible.

The system also includes administrative functionality such as transferring work between curators and running statistical reports.

So the entire process from identification of websites and documents through to making archived version accessible is managed through this system. Not all tasks are completed in the system however. For example, cataloguing is done through another system but the completion of the cataloguing task is recorded here, so the workflow is managed.

[Slide 23]

The management of a selected resource is defined through the steps identified for the workflow which are represented in this diagram.

[Slide 24]

The PANDORA Archive is full-text indexed and can be searched through the NLA's single search discover service, Trove.

This screenshot shows how the search results are presented – in this case for a search on the word 'koala'.

You will see on the left side of the screen there are some facets by which the search results can be refined including subject keyword groupings, dates (which can be further refined to the year) and site types (based on file extension).

[Slide 25]

I will move on to some of the issues that I see affecting the direction of web archiving activity at the NLA now and into the near future.

This includes the environment – both technical and legal – and the organisation and the infrastructure required to sustain the task.

I will also touch on the issue of adding value to, and through, the process of web archiving.

[Slide 26]

From the outset there was a challenge of scale for collecting institutions and the scale has increased dramatically. Even to crawl and harvest the .au domain to collect 1 billion files takes nearly 2 months of continuous multi-threaded crawling.

The web is also more and more complex, particularly in the dynamic way content is delivered. This increases the technical challenges to collect in a way that is an accurate and functional representation of the original content as published.

The nature and form of publication itself has changed. Web content is as much communication as it is formal publishing. There are social media and crowd contributors – often anonymous. The very concept of publishing is challenged in many ways. Some content is intended to be ephemeral. There may be expectations (reasonable or not) that content can be 'un-published' as easily as it is published by removing it from the web. Such issues pose practical, conceptual and perhaps even ethical challenges for web archiving with its purpose of preservation and access in perpetuity.

In Australia we do not have legal deposit legislation that extends to digital content – yet. This is currently under review by the federal government. Even if the legal deposit provisions of the Copyright Act are extended to digital content, this may only provide efficiencies for the collection of content. We may still be left with complex and restrictive requirements in respect to providing access to the content and thus a compromise to the *sine qua non* purpose of preservation.

[Slide 27]

The most pressing issue for the NLA is the state of its digital library infrastructure. The replacement of a now aging infrastructure is the major priority for the Library over next 3 to 4 years. The program to mange this is called the Digital Library Infrastructure Replacement program (DLIR).

A consequence of early involvement with digital collecting is that now the systems and infrastructure are diverse, aging and unsustainable. So a major overhaul of digital collecting, management and preservation systems is underway and will involve digitisation systems, preservation infrastructure and a deposit system among other things.

Though not one of the first priorities, the web harvesting infrastructure and workflow system will be replaced. In preparation for this, we will be doing soft implementations to trial and gain experience with new web harvesting and delivery systems to better understand our requirements for the immediate future.

Likely changes to web archiving will include a greater emphasis on bulk thematic harvesting. This will also have an impact on workflows and the collaborative arrangements we have with PANDORA participants.

Organisational change will also be driven by the Library's increasing focus on digital collecting more broadly – by that I mean collecting through other means than web harvesting which is currently our primary digital collecting activity.

The Library recognises that we need to prepare for more collecting through deposit. This is driven by the growing amount of eBook publishing where content is encumbered with TPMs and delivered to devices and is not able to be harvested. The prospect of the extension of legal deposit to digital material makes deposit systems and workflows a high priority for us.

[Slide 28]

The future of web archiving I think is tied to access and discovery and adding value to the content we collect. In a country as large as Australia, remote networked access is particularly important.

Searching the archive is a fundamental requirement and we can currently provide this through our full text index searchable through the NLA's Trove discovery service. We also provide some browse paths, through subject listings on the PANDORA website, and we catalogue the websites and documents we archive.

These functionalities provide fundamental access but there are currently deficiencies. URL searching is not currently enabled for PANDORA, the subject listings are broad and not very useful; and cataloguing only applies to the selected entity – the 'title' – not to the many publications that may be gathered within that 'title'.

The future appears to be in adding to basic keyword searching and simple subject listings, discovery mechanisms that make use of techniques like visualisation. We have not yet embarked on this at the NLA. However, I would refer you to the UK Web Archive which has implemented some visualisations. I'm not sure that the visualisations implemented by the UK Web Archive, such as their tag clouds, n-gram search and 3D wall, really deliver on the promise of effective discovery, but they are interesting examples of this direction of discovery beyond text searching.

Large web archive collections also offer the prospect of some unique historic 'big data' for analysis – for research *per se* and perhaps to develop innovative discovery mechanism. The creation of APIs to allow researchers to work with this data is one path we may look at following – although, as ever, there are legal constraints as to what access may be provided and as to what use may be made of the data.

[Slide 29]

In summarising, I would like to mention some of the lessons learned from our experience and make some brief comments on what I think is the importance of web archiving.

In reflecting on the experience of web archiving at the NLA over 15 or 16 years, I will focus on three points.

Firstly, the importance of actually taking action. It is well to consider the long term implications of actions – but act we must.

The web is an ephemeral medium like no other. There is not the luxury of time – the window of opportunity to collect is always an unknown and so collecting needs to be opportunistic and timely.

Understand that in this area things will change and change rapidly. Some tasks may become easier, others will not. Not all problems can be solved in the planning. You need to gain experience and be agile in your approach.

[Slide 30]

The second point I would make is to approach web archiving in a way that is sustainable.

You need to understand that this is a long term commitment – that is the preservation objective after all.

Consider the implications of your choice of technology, infrastructure and approach. Is it sustainable and agile to respond effectively to change?

Choices the NLA made in regard to our infrastructure have resulted in it being a major task for us to move to new infrastructure. This is to a large extent because we started early, so the difficulties we now face have to be measured against the advantages of early progress.

So, be strategic in choosing the right approach for the purpose. Now there are open-source tools available and there are web archiving service providers (both commercial and not for profit) – a very different situation from when the NLA started out.

It is important to engage and support skilled, interested and self-motivated staff able and willing to deal with a challenging and dynamic activity.

[Slide 31]

Finally, focus on the purpose and on outcomes.

This means understanding that in collecting web content there is also the opportunity to add value to the content.

At its most basic value is added by the context created through purposeful curation and by the prospect of long-term access that web archiving delivers.

Embrace the responsibility that is involved in creating a chronology of historical cultural artefacts out of the chaotic and ephemeral world that is the web.

And promote what you do to demonstrate and articulate the value of web collections. Engage with others provided this advances your objectives. Collaboration and engagement with stakeholders may also help clarify those objectives.

[Slide 32]

I will end with just a few comments on what I see as the importance of web archiving.

In this context I would first make the point that it can be a difficult task to advocate the importance of web archiving.

The nature of the material means you need to collect it while it is still available on the original live website and it may not be obvious to some why it needs to be collected at all.

Online content is obviously highly variable in its apparent value – particularly long term value. It is also so much a part of our everyday life that some find it is easy to overlook or dismiss its potential long term value.

[Slide 33]

But, collecting web materials provides the only evidence of our cultural expression on the web.

The medium is highly vulnerable, in respect to content at least. Content can literally disappear without trace. There is no physical artefact or remnant for later, retrospective, collecting.

As a publishing medium that more and more people can engage in, collecting web materials allow us to understand our society over time in ways that have not been so possible in the past – provided we collect and preserve the record.

More and more ‘grey literature’ – that is, documents produced by entities that are not in the commercial business of publishing, that support and inform policy and research – are moving online only because of the cheap and convenient means of publishing afforded by the web.

Without web archiving we run the risk of allowing the proverbial ‘digital black hole’ to prevail in our cultural, social and intellectual memory.

[Slide 34]

Thank you.