Operational Challenges and Innovation for National Web Archiving

Paul Koerbin[1]

Abstract

This paper takes a long-view perspective of the interaction of innovation and operational objectives in the development of a national web archiving program at the National Library of Australia. In looking at this program over its twenty year history it is possible to discern an approach that is based on pragmatic outcomes strategically focused on operational workflows and access. While this approach has served the Library in developing and sustaining one of the earliest and longest active web archiving programs in the world, early successful outcomes have also produced longer term implications and constraints for keeping the program vital and fit for the purpose of collecting content from a dynamic, changing and expanding web. This paper covers developments in the three pillars of web archiving – collecting, preservation and access – as well as issues associated with bringing substantial legacy data into a program focused on the future. It suggests that a pragmatic and operational focused program remains a viable context for innovation.

Introduction

It is a notable fact that the systematic archiving of the World Wide Web (the web) – or more precisely selective parts of the web – has now been underway for more than twenty years. Moreover, programs taking on the task of preserving the material published on the web were established a mere five years or so after the appearance

[1] Assistant Director Web Archiving and Government Publications, National Library of Australia.

of the web itself as a public medium[2]. The establishment of strategic programs to preserve access to the web should, *per se*, be recognised as a significant innovation particularly in the context of the library world.

Acknowledgement must be given in the first place to the Internet Archive established in 1996 by Brewster Kahle with the ambition of archiving the entire web – an ambition that could only be realised by someone as visionary and entrepreneurial as Kahle. Nevertheless, in the same year that Kahle began realising his vision with the Internet Archive, the National Library of Australia, with more modest ambitions, also began working to establish an archive of Australia web publishing. If innovation is understood as a new idea and way of doing something with strategic objectives, this undertaking certainly meets this definition. Innovation in operational practice needs to be built on sustainable infrastructure and objectives and so it is not by chance that the earliest web archiving programs – with the exception of the Internet Archiving proving the rule – were established by national collecting institutions (specifically national libraries). The strategic objective of web archiving is necessarily predicated upon the sustainability of a viable program.

In discussing operational challenges and innovation for national web archiving in the Australian context, more than twenty years after the National Library established the PANDORA Web Archive, I will take a historical perspective. This is not to conclude or even suggest that the 'job is done' but rather to consider the processes of pragmatism that accompany innovation in order to achieve sustainable operational

[2] The first quarter century of the web and web archives as a source of history are the subject of two publications forthcoming in 2017. The first is *The Web as History* edited by Niels Brügger and Ralph Schroeder, UCL Press (due March 2017) and *Web 25: histories from the first 25 years of the World Wide Web*, Peter Lang (forthcoming) also edited by Brügger which includes a case study of the PANDORA Archive and the web archive as artefact by Paul Koerbin.

objectives. For the National Library web archiving is not a theoretical or research exercise, but an operational matter of fulfilling its statutory function and purpose to comprehensively collect, preserve (and provide ongoing access to) the nation's documentary heritage.

Early development: understanding web archiving and the implications of workflow innovation

When the National Library formerly began its web archiving program in 1996 there were no established models to follow and no systems designed specifically for the task. This situation resulted in the early pioneers of web archiving pursuing different models. The Internet Archive originally obtained content passed on from the web data and analytics service Alexa that Kahle founded in early 1996. The Swedish Royal Library founded 'Kulturarw3' in September 1996 and took an approach of collecting the entire Swedish web domain using harvester technology. In Australia the National Library, perhaps not entirely surprisingly given the time and context, took a very bibliographic, very library-centric, approach to the task[3].

While also visionary, though perhaps less ambitious than programs such those of the Internet Archive or Kulturarw3, the selective and bibliocentric approach of the National Library permitted, even fostered, innovation in a strategic and operational way. In practice this involved working through and implementing procedures and

[3] Major early established national web archiving initiatives and their approaches were recorded and documented by another National Library of Australia initiative, the PADI (Preserving Access to Digital Information) portal, see: http://pandora.nla.gov.au/nph-wb/19991025130000/http://www.nla.gov.au/padi/index.html. The PADI website was launched in January 1997 to bring together information about the emerging issue of digital preservation. Originally 'PADI' stood for 'preserving Australia's digital information' however it soon became evident that such a portal was of international interest.

workflows directed towards the strategic objectives of collection, preservation and access as it became possible to do so. In other words, an approach that was not dependent upon solving (or even understanding) all the issues before implementation. It was an experiential approach with an imperative of becoming operational as soon as practicable. So, the decision to undertake a selective approach to web archiving was made in the context of utilising the limited available resources and technologies to achieve operational outcomes quickly.

At this point we should consider what it is we mean by ‘web archiving’. The term is now long established and commonly used among the practitioners, though it is not without its ambiguity, not the least because people will have various ideas of the common meaning of the word ‘archiving’. Such understanding can range from simply backing up content on physical media to store somewhere; to retaining content in accordance with statutory requirements established under archival legislation. For the purposes of web archiving in practice it must be understood as encompassing a range of processes with the strategic (and *sine qua non*) objective of long term access. Such a process requires establishing policies, workflows and systems constituting a significant commitment of resources and a fundamental attention to sustainability. Web archiving is a long term project dependent upon the long term delivery of the objective of access. Fundamentally, a serious web archiving project requires a collection development policy and procedures; acquisition mechanisms; metadata extraction and creation; description and metadata management; quality checking processes; preservation policy, procedures and systems; and discovery and access services including the indexing of very large amounts of data[4].

It may be argued that the National Library's early work and most significant achievement was in the development of workflow procedures and the system to support them. The impact and implications of these early workflow and system achievements (both enabling and constraining) remain with the web archiving program to this day. So, the consequences of such innovation deserve as much consideration as the significant though relatively small number of websites and online publications actually collected in the period from 1996 up until mid-2001 when the first full workflow system was implemented[5].

The PANDORA web archive was conceived as a collaborative project inviting state libraries and other cultural collecting agencies to contribute curatorial expertise and resource – thus assuming responsibilities for their jurisdictions or specialist collecting areas[6]. To accomplish this, a shared workflow system was required when at the time, the late 1990s, no such system existed. While such a system was extensively scoped and modelled and specifications were developed, workflow processes were

[4] Niels Brügger, perhaps the most prolific scholarly writer on web archiving, succinctly suggests a definition for web archiving as the "deliberate and purposive preservation of web material", Brügger, N. (2010), 'The future of web history' in N. Brügger (ed.) *Web history*. New York: Peter Lang.

[5] Issues associated with identifying early collected web content in the PANDORA Archive are discussed in the 2013 National Library of Australia blog post by Paul Koerbin titled *What is the oldest website?* See: http://webarchive.nla.gov.au/gov/20140803161826/https://www.nla.gov.au/australias-web-archives/2013/05/03/what-is-the-oldest-website-and-will-an-artefact-do (viewed 14 November 2016).

[6] Over the life of the PANDORA Archive collaborative partners have included the state libraries of New South Wales, Queensland, Victoria, South Australia and Western Australia; the Northern Territory Library; the National Film and Sound Archives, the Australian War Memorial, the Australian Institute for Aboriginal and Torres Strait Islander Studies and the National Gallery of Australia. However many of these organisations have found it difficult to maintain the resources and staff skills to sustain this work. The NFSA ceased as a PANDORA participant in 2014 and the NTL is currently not contributing. Some other organisations contribute on a very small scale. The National Library alone is responsible for around 65% of all the content collected and 50% of the archived instances; while all other participating agencies together have contributed the remaining 35% of the data and the other 50% of the instances.

tried and implemented to the extent that was practical at the time. So scoping and selection of target content was pursued before it could actually be collected; harvesting was managed using a variety of tools – indexing software and various offline browser software – before the functionality for delivering access to the archived content was achieved. This new collecting territory was entered in an exploratory and practical way. The outcomes of this approach were early realised achievements in terms of collecting and access.

The workflow system that was developed and first implemented in June 2001 was and has remained the principal innovation upon which the two decade long program of web archiving by the National Library operates. A second version of this system, named PANDAS (PANDORA Digital Archiving System), was released in August 2002 and a completely re-engineered third version – being technically more robust with greatly improved workflow management for individual curators – was implemented in July 2007, thus completing a decade of workflow system development. As the PANDAS system was developed using the Apple WebObjects environment it was not readily open for others to use though the Library did manage to provide the British Library with the code for PANDAS version 2 which enabled the then established United Kingdom web archiving consortium to get their web archiving program underway in 2004. By this time however the International Internet Preservation Consortium (IIPC – initially a collaborative initiative of national libraries and the Internet Archive) had been established and work was already underway to develop an open source and standards based web archiving workflow system[7].

[7] This application is called the Web Curator Tool and was an initiative of the International Internet Preservation Consortium (itself established in 2003) and developed collaboratively by the National Library of New Zealand and the British Library in 2006. See http://webcurator.sourceforge.net (viewed 31 January 2017).

The National Library was perhaps already becoming the victim of its own operational success with PANDAS. The system effectively manages the entire acquisition and delivery workflow from selection to rights management; through harvest scoping and scheduling; to quality checking and fixing; to delivery and access restrictions[8]. This is achieved through a web interface that allows for contributing partner agencies and individual curators to manage their respective workflows through the one system. Given the genesis of PANDORA, it is also not surprising that the workflow is entirely 'bibliographic' in character. Websites (in their entirety or in part), web pages, web documents are all characterised by a curator applied 'title'. Material is collected in discrete units, catalogued and listed on the PANDORA website and accessed through 'title entry pages'. The content is full text indexed and searchable; however, the descriptive records for the archived 'titles' appear in the Trove books and journal zones[9] while full text and URL indexed content is discoverable through a dedicated web archives zone[10].

There are obvious problems and constraints with such an approach particularly as the collected material has become less and less bibliographic – even less publication-like – over time. The web is now an interactive and transactional space

[8] A significant constraint of the holistic nature of PANDAS is that individual functional modules, such as the delivery functionality, are not modular and cannot be developed separately from the whole system.

[9] The inherently uneasy fit of archived websites with a bibliographic approach is demonstrated by the fact that websites archived in PANDORA are provided with MARC records catalogued using the format of 'integrating resources' which consequently appear in the Trove 'Journals, articles and data sets' zone. It might surprise (and certainly not be intuitive to) the non-librarian searcher that websites are ostensibly considered analogous to a journal.

[10] The business of the organisation running the web archiving program will naturally influence the approach to access and the emphasis given to types of metadata, whether that is descriptive cataloguing metadata or indexing metadata. The National Library of New Zealand, for example, embeds access to its web archive resources squarely within its national catalogue; conversely the Portuguese Web Archive (Arquivo.pt) is run by the Portuguese Foundation for Science and Technology, a body responsible for digital research infrastructure, and the web archive benefits from considerable research activity with access through a search portal rather than a bibliographic catalogue.

and a communication medium as much as it continues as a medium for disseminating publications. This is accompanied by a relentless increase in the amount of material online that manifests in a constant ‘present’ whereby material is potentially never static and never necessarily artefactual – content can disappear (become ‘unpublished’) without trace as quickly and as efficiently it was published[11]. In this environment a highly engineered, long established (entrenched), workflow rich web archiving system becomes a constraint as well as being a significant asset. While highly effective in the timely collection of targeted significant (and especially publication-like) material on a small scale, this model has become less effective in meeting the challenges of collecting and describing Australia’s online cultural heritage in a more comprehensive way.

Collecting web archive content

While the first decade of collecting online Australian content using the PANDORA selective model – that is up to the mid-2000s – never really provided the prospect of truly comprehensive collecting, the early web (before social media) certainly seemed within conceptual grasp if the necessary resources were forthcoming. The explosion of growth of the web and its increasing complexity in delivery (through server-side, dynamic applications) and the turn towards social media and interaction certainly challenged the prospect that a selective model and system, no matter how good, could ever fully deliver on the Library’s collecting objective.

[11] These issues are explored in the 2014 National Library of Australia blog post by Paul Koerbin titled *Web archiving – an antidote to ‘present shock’?* See https://www.nla.gov.au/blogs/web-archiving/2014/03/18/web-archiving-an-antidote-to-present-shock (viewed 31 January 2017).

In 2005 the Internet Archive released the first production version of its purpose-built web archiving crawl robot Heritrix. This tool is currently the standard though it increasingly has its limitation[12]. The main advantage of Heritrix is its purpose built support for bulk, large scale harvesting since it was designed to meet the large scale collecting requirements of the Internet Archive. With their new harvester in production the Internet Archive – a not-for-profit enterprise – was open for business and the Library contracted its first whole Australian domain crawl in 2005. The Library has continued to contract an archival crawl of the '.au' top level domain from the Internet Archive annually. This content is retained at the National Library and currently amounts to around 400 terabytes of data.

Domain harvesting is, *prima facie*, the most efficient and effective means of collecting content given that most of the work is done by the harvest robot. However, while the crawl can be scoped it is a very blunt instrument for collecting content that is dynamic and time dependent in its importance and does not make any discrimination in respect to the value of the content[13]. An Australian domain harvest, run by the Internet Archive using multiple servers will still take in the order of eight weeks continuous crawling to collect something like 700 million to 1 billion files or around 50 terabytes of data. Thus the oft characterised 'snapshot' has a very wide exposure time indeed! There is virtually no control over the specific timing for collecting content from any given website. Moreover there is little practical means of reacting to the collecting in order to quality check the harvesting and take action to improve the harvest (as is possible with selective harvesting).

[12] New generation harvesting technology that incorporates browser like functionality to supplement link crawling is in development including at the Internet Archive.

[13] For some researchers such an approach may be desirable because the discriminations of curators are presumably removed. However, the technical and temporal constraints of large scale harvesting do add other significant if less obvious discriminations to what is ultimately collected and represented in the archive.

While the Library has been collecting harvests of the .au domain since 2005 little has been done with the content other than custodial, bit-level, preservation. Domain harvesting is largely opportunistic collecting, conveniently (if not necessarily helpfully) leaving the complex problems of indexing and providing access to the vast amount of preserved content to a later time.

I have already suggested that the success of the PANDORA model also imposes a degree of constraint in terms of innovation. For limited scale selective archiving the process delivers; and the workflows are now entirely established within the National Library and supported by a highly engineered and fit-for-purpose management system. Thus the challenge for the Library is to find the means to develop its operation and systems to be able to increase the scale of its collecting to something that approaches the comprehensive coverage demanded by its statutory function.

Like the term ‘web archiving’, the term ‘comprehensive’ needs to be defined in the context of web archiving. Collecting web content of a web domain can never be comprehensive in any absolute sense to mean collecting everything published on the web Even if it were possible to collect the entire .au domain the time it takes to collect a ‘snapshot’ falls far short of collecting the dynamic, persistent present nature of the multi-dimensional web including the chronological dimension that ultimately defines the web archive[14]. The Library has had to grapple with an ostensible chasm between the objective and the reality; not the least so as to functionally express its

[14] And this is not even taking into consideration that web content relevant to a national collection will not be exclusively found on its country code top level domain (ccTLD). Much content that is clearly of Australian national interest and relevance is published outside the .au ccTLD.

collection development policy. For the Library's most recent collection development policy, introduced in 2016, the comprehensive objective is framed as a curated practice of collecting – extensive where possible but selective and representative when necessary to provide a comprehensible and interpretable expression of the elusive whole.

> "For all published Australian materials, the Library's collecting aims to be sufficiently comprehensive that researchers interrogating the collection could extrapolate an understanding of the entirety of Australian published output. When possible, comprehensive collection of Australian publications is achieved through acquiring a copy of every published work. When it is not possible to collect every work, a combination of comprehensive and representative collecting approaches are used, minimising collection gaps in subject matter, format type, author group or other category of material."[15]

In order for the Library to increase its representative collection of web materials – and certainly to fill in the gap between the selective and the annual domain harvests – it was necessary to look beyond the PANDORA infrastructure. As in so much of what drives the Library in the innovation space it was the opportunity for providing access that pushed new developments. A new project emerged focused on bulk collecting of Commonwealth Government material. This will be discussed below under the 'recent developments' section, but first we need to consider preservation and access that along with collecting make up the pillars of web archiving.

[15] CDP – What We Collect: Australian Published Collections, see: http://www.nla.gov.au/collection-development-policy/what-we-collect-australian-published-collections (viewed on 14 November 2016).

Preserving web archive content

The National Library's collection of web content represents a significant and unique asset. As with physical materials the value of the collection depends upon the application of preservation processes. One of the signifying qualities of a national web archiving program is the commitment to and application of a planned and sustainable preservation strategy. While the challenges of digital preservation, particularly in respect to very large and complex sets of data such as the collection of web content, are beyond the scope of this paper it should be noted that the Library does have bit-level preservation in place for this content; however, web archive content is yet to be ingested into its digital preservation management system, Preservica.

Web content represents particular challenges for digital preservation, much of which derives from the fact that the original digital content is not created by the Library. In collecting broadly (and technically largely indiscriminately) from the web the Library does not control or standardise the formats used. Moreover, web pages represent complex items constituted of many files, scripts and media to form the published entity. A web page is a multitude and confluence of contextualised relationships that need to be retained to preserve the entity. Managing this on the scale and detail required remains one of the great challenges for digital preservation.

Much of the digital preservation effort to date has been in planning and preparing for the eventuality of obsolescence and the breakdown of access. Content exists in the web archive collection that is already subject to these failures, though remarkably the larger amount remains accessible. This may be attributable in part to the limitations of collecting – some of the more problematic formats may never have been able to be harvested – as well as due to the normalising effect of harvesting which reduces dynamic content to static web pages.

One innovation driven by an intended operational outcome by the Library has been the development of 'preservation intent statements'. The development of these statements was a process to engage collection managers with preservation specialists to identify the significant characteristics of the web collection. Digital preservation specialists have been discussing and debating the concept of significant properties of digital objects for the purpose of preservation for years (or decades). The complexity and difficulty of this objective often defeats useful operational outcomes. The purpose of developing the concept of preservation intent statements was to situate the preservation concept within the operational collection managers' purview in a high level, conceptual, but hopefully also a practical way[16].

The purpose of preservation is to ensure – to the extent that is feasible – the technical accomplishment of access. Access, from the preservation perspective, is about being able to render the content now and into the future in a manner faithful to the original; or at least in a manner consistent with identified constraints and objectives, as for example are outlined in the preservation intent statements.

[16] The rationale and process of developing preservation intent statements is discussed in: Webb, C. Pearson, D. & Koerbin, P. (2013). 'Oh, you wanted us to preserve that?!' Statements of preservation intent for the National Library of Australia's digital collections. D-Lib, 19(1-2). Available at: http://www.dlib.org/dlib/january13/webb/01webb.html (viewed 14 November 2016).

Accessing web archive content

Building collections of web content, even when the statistics become large enough to impress, and developing and committing to sustainable preservation strategies does not complete the objective of such an enterprise. The objective of collecting and long-term sustainable preservation – that is, of 'web archiving' – must be access. Access, however, is comprised of more than just being able to render the content; it extends to and depends upon descriptive metadata, indexing and the experience of user focused discovery.

A defining characteristic of the National Library's web archiving program has been the emphasis placed upon access. Somewhat counterintuitively the constraint of legal deposit legislation in Australia, which up until February 2016 was not applicable to online publications (indeed not applicable to digital formats at all), promoted an access focused web archiving program. The lack of legal deposit warrant for collecting contributed to the Library's decision to build a web archiving program that was selective and permissions-based. Permission to provide access was sought at the same time as permission to collect and preserve the content. The corollary associated with this approach was that if the permission to provide access was not forthcoming then the online publication was not collected as part of the PANDORA web archive program.

When the Library began collecting whole .au domain harvests from 2005 the foundation of the PANDORA selective archiving approach was somewhat undermined since this involved large scale collecting without any regime of access (including description, indexing and discovery). The passing of legislation in the Federal Parliament that reframed the Commonwealth legal deposit provision in the *Copyright Act 1968*, extending the coverage of the provisions to digital materials including online publication, further challenged the PANDORA selective permissions based approach[17]. The Library now had a clearer mandate to collect online materials without the encumbrance of a seeking prior permission for the collecting process; indeed publishers are obliged to deposit online publications when requested to do so by the Director-General of the Library (which includes a request made automatically by a harvest robot user agent). However legal deposit provisions in the *Copyright Act 1968* only cover the delivery of content to the Library and do not extend to the provision of access. This is left to other aspects of the *Act*. Taken together, the large amount of content already collected through domain harvesting and the new mandate to collect online material as part of a legal deposit process required the Library to revisit its approach to access to web archive content. These conditions – large scale bulk harvesting and legal deposit collecting – do provide impetus to the disengagement of access from collecting in the operational workflow.

There are operational implications from this push to separate access from discovery for the Library's web archiving program since the workflow system used for PANDORA locks the full workflow into a single system. The extension of legal deposit required only some minor changes to the PANDAS system to permit completing collecting activity without the previously required permission licence. However, it also set in motion a significant redevelopment of access functionality to

[17] See sections 195CA to 195CJ of the *Copyright Act 1968* relating to the delivery of material to the National Library of Australia at http://www.austlii.edu.au/au/legis/cth/consol_act/ca1968133/ (viewed on 31 January 2017).

the Library's web archive collections, specifically in respect to discovery and the display of archived content.

Recent developments: the Australian Government Web Archive and the Bamboo Collection Management Tool.

The Library's approach in continuing to develop its web archiving program has followed a similar pattern to its initial development: that is to undertake small, staged practical steps and incremental implementation of innovations. An example of this is the Australian Government Web Archive (AGWA).

The AGWA developed from a prototype application built initially to provide access to Commonwealth Government material already collected through the cooperative arrangements with the Internet Archive[18]. This provided a substantial but manageable set of data to prototype a delivery application for web content based on what were now international standards; that is, using the WARC archival file format and Wayback delivery tool. Around this the Library developed a prototype public interface as a demonstration of the potential direction to move beyond the aging PANDORA delivery interface. This was deemed very successful and the prototype was released into production in March 2014.

The success of the development of the AGWA public interface consequently enabled other developments in regard to the collecting of content. While AGWA did not employ the PANDORA workflow management system (PANDAS), it was a simple

[18] The impetus for this was the approval of whole-of-government arrangements in May 2010 that permitted the Library to collect, preserve and make accessible online Commonwealth Government content.

matter to deploy the Heritrix harvester to begin running harvests of Commonwealth Government material in-house. The Library did not (and still does not) have a curator tool like PANDAS to run Heritrix and so harvests are run using the native Heritrix console. While this is limiting, since harvests can only be initiated by someone with the requisite technical knowledge of XML files and regular expressions to configure crawls and there is no scheduling tool, it allowed the usefulness and the profile of the AGWA service to increase rapidly by being able to add new content more efficiently than relying only on the annual harvests from the Internet Archive. Thus there was considerable operational gain for relatively little resource commitment.

Subsequently, the ability to run in-house bulk seed harvests and deliver them through a new fit-for-purpose discovery and delivery application formed the basis for more developments. Tools and applications were developed to meet immediate operational needs. In a development environment given the working title of ‘Bamboo’ the Library prototyped collection management tools to manage the indexing of bulk harvested content, including providing the non-technical operational curator the ability to import Heritrix harvested content and initiate both URL (CDX) and full text (Solr) indexing. This infrastructure was also developed to filter the indexing of content from domain harvest collections so that it can be delivered through the AGWA portal. Thus the coverage of the AGWA service was able to be extended back to 1996 using Internet Archive supplied data while new content continues to be added through bulk harvests. The seed lists for the bulk harvests are also managed through a tool maintained in the Bamboo environment.

With the AGWA service rapidly integrated into the operational environment of the Library's web archiving program further innovation opportunities emerged. The bulk of the content is added to the AGWA through both the annual domain harvest contracted from the Internet Archive and three in-house harvests run during the year. With a back-end collection management tool in place to initiate indexing of content and import it into production, an opportunity was identified to provide staff collecting government publications the ability to collect individual online documents in real-time. A simple PDF harvester tool, christened 'Butterflynet', was added to the Bamboo environment to allow staff to collect online documents and import them into AGWA with the single click of a button[19].

The infrastructure to allow real-time collecting of simple content with Butterflynet was followed by the implementation of a version of the Webrecorder[20], a headless-browser harvester that allows the curator to browse through webpages collecting the page content, including embedded content and pass it to the AGWA in real-time. This functionality provides a valuable adjunct to the more usual crawler web harvesting since more complex and covert JavaScript can often be discovered and collected. In the current deployment Webrecorder works as a useful patching tool for curators doing manual quality checking.

[19] There was a strong business case for this quite simple application because government material is increasingly moving from print to online. Previously when collections staff identified content they would need to try and pursue print copies (that may not be available) or wait until the next scheduled bulk harvest to collect the publication at which point it may have disappeared. This also meant managing a workflow spread over a long period of time with no workflow system support. Now staff can identify online content, collect it and add it to the public AGWA collection with the click of a button and complete the descriptive cataloguing task with no delay.

[20] A tool developed by Ilya Kreymer, see: https://github.com/webrecorder/webrecorder (viewed 31 January 2017).

The development of AGWA with its new standards based harvesting and infrastructure including a fit-for-purpose discovery and delivery interface distinct from and not connected to the PANDORA service (itself consisting of an uneasy partnering of the PANDORA delivery interface and the National Library's Trove indexing and discovery service) exacerbated the central problem arising from operational focused innovation. That is, the proliferation of collections and systems to deliver a common strategic outcome that do not naturally connect with each other. The Library had multiple web archive collections built over two decades collected by various means and maintained in different formats. The PANDORA collection included content collected by the Harvest indexer in the late 1990s (for the earliest PANDORA content) and content collected by various other harvesting mechanism, mostly 'offline browsers' such as WebZip and, since 2002, HTTrack. Since the PANDORA infrastructure also served as a *de facto* deposit system prior to the Library's development of its eDeposit service[21], PANDORA also contains content ingested directly. In addition, the Library maintains domain harvest content supplied by the Internet Archive collected through various means over two decades including content originally obtained from Alexa and content harvested by various version of the Heritrix harvester. To this was then added content collected in-house since September 2013 by the Library to populate the AGWA collection.

It became of increasing concern to the Library that content contained in the PANDORA web archive and the AGWA[22] could not be searched, discovered and accessed through a single portal. Moreover, while Trove indexed PANDORA content

[21] This service was implemented in February 2016 to support publisher initiated deposit of digital publications as required under the legal deposit provisions of the *Copyright Act 1968* that came into effect at that time.
[22] That is, the Library's two publicly accessible web archive collections. There is no public access to the collection of .au domain harvests.

the delivery mechanism was still embedded in the PANDORA management system infrastructure and content was displayed, not through Trove but through the aging PANDORA web interface.

Much of the recent development in respect to the Library's web archiving program has consequently been focused on the back-end infrastructure and management of the content with the objective of being able to build a common index and a single discovery and delivery system. A more efficient CDX database to replace the cumbersome CDX file native to the Wayback delivery tool was built and deployed in the Bamboo development environment. This service, OutbackCDX[23], uses a database to manage the indexing thereby allowing incremental updating of the index and a providing a significantly more efficient management of the indexing process than the large single CDX file. In addition, a substantial amount of work was required to map PANDORA content (which was collected by offline browser technology that commonly alters the file names) to the original URLs so that they could be rendered in a consistent and connected way with content harvested using Heritrix.

This back-end work was the prerequisite for developing a new single index and discovery interface for all the Library's web archive collections. Work on this interface, to be incorporated into the Library's single discovery service Trove, became a major development project for the Library in the second half of 2016. Business issues concerning access to the large amount of content collected through whole domain

[23] OutbackCDX, originally called 'tinycdxserver', a remote resource index server, was developed by Alex Osborne and it has begun to be used by other major web archiving programs around the world including The British Library, see https://github.com/nla/outbackcdx (viewed 31 January 2017). Acknowledgement is due to Alex Osborne who is the driving force and virtually sole IT engineer behind the current development of the Library's web archiving system infrastructure discussed here. Acknowledgment should also be made to Dr Mark Pearson who did the original technical engineering development of the AGWA application.

harvest remain to be resolved, however this new interface which is expected to be implemented in the first half of 2017 will provide a new integrated interface to PANDORA legacy content, AGWA content and new content to be collected through existing (and yet to be developed) collecting systems and workflows.

Critical future work must involve returning attention to where the Library's web archiving program started. That is, redeveloping a collecting workflow infrastructure that is capable of continuing to manage the highly developed selective collecting workflows of the PANDORA model while supporting more efficient bulk (seed list based and themed) harvesting.

Conclusion

That it is possible to reflect on twenty years of national web archiving in Australia is due to the National Library's approach to innovation and development focused on outcomes and driven by the objective to provide access to its collections. The innovations managed by the Library have perhaps been small in their parts but represent a major achievement in their totality constituting one of the world's pioneering web archiving programs. Small achievable developments deliver but they are not without consequence. While the program is strategic in its conception and objectives the technical development has often been driven by operational need and opportunity. This does not always make for ideal outcomes and in the case of the web archiving program this has resulted in the proliferation of disparate systems for collection and for discovery. Nevertheless, innovation driven by operational needs is a practical and necessary response to the complex and resource straining challenges to collect, preserve and provide ongoing access to the dynamic, unpredictable and expanding medium of online publication and communication.