National Web Archiving in Australia: Representing the Comprehensive

Paul Koerbin

It's an impossible task but we started anyway!
(Dr Marie-Louise Ayres, Director-General of the National Library of Australia, [Easton 2019]).

Abstract National libraries have been at the forefront of web archiving since the activity commenced in the mid-1990s. This effort is built upon and sustained by their long-term strategic focus, curatorial experience and mandate to collect a nation's documentary heritage. Nevertheless, their specific legal remit, resources and strategic priorities will affect the objectives and the outcomes of national web archiving programmes. The National Library of Australia's web archiving programme, being among the earliest established and longest sustained activities, provides a case study on the origin and building of a practical approach to comprehensive national collecting and access.

1 Introduction

In what is now more than a quarter of a century of active collecting and preserving web content, it should be no surprise that national heritage collecting institutions, and more specifically national libraries, have been at the vanguard of web archiving programmes. National web archiving has been a driver for web archiving initiatives because the core function of national libraries is to collect and preserve their national documentary heritage in a manner that usually aims to be comprehensive and consequently inclusive of online publication. Moreover, national libraries have the strategic and operational history and expertise in collecting other published (and unpublished) materials on a national scale. These institutions, especially those supported by legal deposit legislation, are in function and mandate focused on the

P. Koerbin (✉)
National Library of Australia, Parkes Place, Canberra, Australia
e-mail: pkoerbin@nla.gov.au

D. Gomes et al. (eds.), *The Past Web*,
https://doi.org/10.1007/978-3-030-63291-5_3

long-term maintenance and sustainability of their national collections. Web archiving more than any other collecting is an activity dependent upon a commitment to sustainability. Collecting essentially ephemeral and intangible digital artefacts commits the institution to the considerable resource required for digital preservation and access along with the concomitant complexities and uncertainties that require strategic, sustainable policy development and programme management.[1]

Web archiving, approaching 2020, is now a much broader-based activity with a range of collecting and research institutions engaged, many with objectives to serve more narrowly defined and not necessarily national audiences. In 2003, the International Internet Preservation Consortium (IIPC) was established, 7 years after the earliest web archiving programmes began. The IIPC was originally constituted of the national libraries of France, Great Britain, Finland, Iceland, Canada, Denmark, Sweden, Norway and Australia together with the United States Library of Congress, the National Central Library of Florence and the ambitious and visionary Internet Archive (the only non-state organisation). By 2019, the IIPC membership had expanded to include numerous research institutions and academic libraries, yet national libraries still made up around two-thirds (67%) of the member organisations. The newer members of the web archiving community from the research and academic sector add a necessary vitality for innovation, research and development; nevertheless, it remains true that collecting content remains, as Winters (2019, p. 83) states, "with a few exceptions ... conducted on a national basis by major national institutions, in keeping with well-established missions to preserve national cultural heritage".

The National Library of Australia (NLA) began its web archiving programme in 1996. The programme, which was given the name PANDORA—originally an initialism for Preserving and Accessing Networked Documentary Resources of Australia[2]—grew out of existing and established library operations, specifically the acquisition and cataloguing of serial publications. The conceptualisation of the Web at that seminal time was largely as a publishing medium and thus readily understood as an extension of existing collecting paradigms. The web archiving programme was based in the collections management and description area of the Library, not the information technology area. This had an important impact on the NLA's approach to web archiving because it meant process and procedure

[1]For a monograph-length collection covering a broad range of aspects relating to national domain collecting and preservation, see Brügger and Laursen (2019). That volume includes only passing reference to the National Library of Australia's web archiving activities—one of the earliest established long-maintained web archiving programmes. It is in this context that this chapter therefore focuses specifically on the Australian experience. On business planning for web archives, see Koerbin (2010).

[2]This initialism is no longer promoted and PANDORA is merely used as a branding for the collaborative selective web archiving programme that forms one part of the NLA's broader web archiving strategy. If the designation was formulated today, it would likely read "Digital Online" in place of "Documentary". In retrospect, the use of the term "documentary", linking the resources to be preserved to the NLA's statutory function to comprehensively collect Australia's documentary heritage, rather than to format and medium, is itself instructive.

was the driving factor rather than the development of technologies. Consequently, the first tasks undertaken were to select and catalogue Australian online journal publications. The selective approach operating in the context of general collecting objectives meant that online materials that were also available in print formats were not selected for archiving. This approach presented an understanding of comprehensive not as the collecting of the entire web publishing medium per se, but as web archiving supplementing the established collecting of print publishing and supporting a broader concept of the national collection.[3]

This nascent approach to web archiving engendered a sense that comprehensive collecting, as it was then understood, was possible. The technological and resource challenges were not ignored but they did not drive nor, more importantly, hinder application to take up the task. Even pioneer visionaries such as Brewster Kahle and Peter Lyman, who recognised the Web as a new medium for cultural expression that in its early stages would imitate the forms of existing media, characterised the Web as essentially a cultural *artefact* (Lyman and Kahle 1998). While stressing the need for a technological response at that time, the characterisation was still in terms of documents, publishers and libraries, concepts that national collecting institutions were best equipped to tackle. Since future developments that would change the character of the Web—Web 2.0, social media, interactivity—were, self-evidently, yet to emerge, this was not entirely a naïve perception and the objective of comprehensive collecting not outlandish. Conceptualising the Web in terms of existing publishing media gave impetus to a programme like that at the NLA precisely because it presented as achievable and sustainable within existing institutional infrastructures and resources. Reinforcing this point, the NLA has conducted its web archiving programme of a quarter of a century within its established operating model and without having received any additional resources specifically for the web archiving programme. Thus, there is the critical need to make the activity incremental, operational and directed towards delivering strategic outcomes such as open access.

2 Comprehensive Collecting

While the NLA's web archiving programme (operating as PANDORA) began as a selective approach to collecting web content—and was the only approach adopted by the Library until 2005 when country-level domain harvesting began—it was still conceptually part of a comprehensive collecting strategy. Selective web archiving was not pursued as a rejection of the domain harvesting approach but, rather, as a practical step towards collecting web content as soon as possible with available resources and low-cost infrastructure development. The NLA in

[3]For detail concerning the establishment and early activities of the NLA's web archiving programme, PANDORA, see: Cathro et al. (2001); Phillips (2002); and Webb (2001).

fact began exploring options for domain harvesting soon after establishing an operational selective web archiving programme. Later the Library adopted the approach of contracting the Internet Archive to undertake custom scoped .au level domain harvests to supply to the Library for preservation, indexing and access. The practicality of this latter arrangement became cogent in 2005 when the Internet Archive released its purpose-built archival web harvester, Heritrix, and funding arrangements within the NLA made capital funding available to purchase domain harvest collections. This offered a way forward for the Library to increase the scale of collecting by outsourcing the harvesting of the Australian web domain, which could then be purchased as a collection item. Large-scale domain harvesting was not established in-house since operational budgets remained un-supplemented and unavailable for the required infrastructure and expertise.

Like all national collecting institutions, the NLA's functions are writ in legislation, specifically the *National Library Act* (1960), which includes the mandate to maintain and develop a comprehensive collection of library materials relating to Australia and Australians and to make the collection available "to the most advantageous use of the collection in the national interest". Unlike many other jurisdictions, enabling legislation requiring publishers to deliver material to the library, generally known as "legal deposit" provisions, are not contained in the *National Library Act* but rather in the Australian Commonwealth *Copyright Act* (1968), legislation over which the Library had little influence for change. The original act only clearly applied the legal deposit requirements to print format library materials and the difficulty in influencing change meant that legal deposit in Australia was not extended to digital materials (including, specifically, online materials) until March 2016.[4] Thus, while the Library's establishing legislation framed its comprehensive collecting function, legal deposit legislation constrained comprehensive collecting for the first 20 years of the web archiving programme's operation because it did not extend to online content.[5]

Since legislation not only enables but may also constrain collecting, an institution's approach to risk in managing and extending its collecting within the legislative framework becomes important. Prior to the extension of legal deposit to online materials, the NLA's primary web archiving activity involved curated selection which was only pursued to a collecting and archiving stage when explicit permissions—copyright licences—were received from publishers. Permission negotiations were conducted by email, so they were usually quick and mostly resulted in agreement, at least where responses could be obtained. In the context of a publications-focused approach to collecting online materials, the permissions based, selective regime was largely successful. However, under this regime, significant

[4]Fortuitously, though untypically, this was a time when the Attorney-General responsible for the *Copyright Act* also had portfolio responsibility for the National Library—a situation that facilitated the progress of legislative changes.

[5]Copyright and legal deposit issues in the early years of the PANDORA web archiving programme are briefly outlined in Phillips and Koerbin (2004) and in more detail by Gatenby (2002).

material that was identified and selected for collection would not be pursued if permission was not forthcoming, as was the case with a seminal election campaign website called jeff.com.au in 1999.[6] As Winters (2019, p. 76) rightly states, "inconsistency of selection and capture is thus not accidental but central to the nature" of essentially patchwork collections. While Winters was referring specifically to a patchwork of various collections (in the UK), this temporal and artefactual patchwork exists within the very collections themselves.

It was common to hear speakers at conferences and forums in the early years of web archiving talk about "time travel" in terms holding out the prospect of a future when we could choose to surf the Web as it was at any point in time—as if all archived content would continue to exist in its original form, context, completeness and functionality. This idea persists to some degree as the ultimate objective. However, such virtual time travel is dependent upon the time coordinates of the artefacts that are collected from the Web—and like oral culture, websites do not exist as artefacts until and unless collected[7]—and thus upon the date-stamp that becomes a defining dimension of the artefact. The technical processes of collecting online content necessarily limit what is represented in the archival collection, so that even the largest scale collecting remains selective, especially when considering the timing and frequency of collecting and the period of time over which the "snapshot" is harvested. There is a curatorial effect (or bias) on collecting no matter what scale of effort is achieved.

When the NLA extended its web archiving programme to domain-level collecting in 2005 (and in-house bulk collecting of government websites from 2011), it was to build scale within the collection and to address curatorial biases and blind spots. It also represented a willingness to manage risk since the legislation at the time was ambiguous in its warrant for such an expansion of web collecting and preservation. While exemption under the *Copyright Act* (s. 200AB)[8] exists to allow libraries to conduct their business and function as a library, in accordance with Article 13 of the TRIPS Agreement,[9] it does not have the same clarity as legal deposit. As well as the objective to increase the scale of collecting, the move to domain harvesting addressed the Library's need to understand the scale and nature of the content published on the .au domain using the empirical evidence of the harvested content. Domain harvesting was also pursued as soon as feasible as a necessary step in developing in-house experience in managing large-scale web archive content.

Collecting the websites published on the country code top-level domain (ccTLD) is an obvious and relatively straightforward objective for national collecting. The

[6]For a detailed discussion of this particular case and other constraints in relation to collecting content for the PANDORA web archive, see Koerbin (2017).

[7]See Koerbin (2017).

[8]For a useful overview of these provisions, see the Australian Libraries Copyright Committee (2008).

[9]See the World Trade Organization's Agreement on Trade-Related Aspects of Intellectual Property Rights (1995).

ccTLD represents published material clearly identified with a country, at least for larger nations, since some smaller nations offer their ccTLD as one of the few international commercial assets they have. For nations like Australia undertaking national web archiving, collecting the ccTLD is central to any approach to comprehensive collecting. It is relatively easy, since scoping covers all content appearing on the .au domain and can readily include embedded content, that is, the links on pages that bring in content, whether on the .au domain or not, that are essential for an accurate rendering of the webpage. Nor is it required to compile a complete seed list of published websites, for example from domain registrations that do not necessarily accurately reflect what is published, since many domains may be registered but never used. In collecting the published record it is what actually exists on the .au domain that is of primary interest since that is what forms (or formed) part of the social and cultural discourse. Consequently, a substantial representative URL list will serve to seed the harvest that is then scoped to follow and include any content found on the .au domain through the harvest crawl process. This process, if run for sufficient time, is the best evidence of the ccTLD, though never a complete record since the harvest must be terminated at some point and is never really completed. Typically, the annual harvests of the .au domain contracted by the NLA, and run by the Internet Archive, run for around 2 months, collecting a representative 800 million to 1 billion documents and data in the order of 60–70 terabytes.

Domain harvesting, as suggested above, supplements and to some degree balances out the curatorial biases of selective archiving, since, within the scope of the harvest, the robots collect without discrimination. Nevertheless, this process, while a critical element of a comprehensive approach to collecting national web content, has some significant limitations. What is collected is dependent upon what the robot identifies and is actually able to successfully harvest. To mitigate risk, large harvests generally follow robots.txt disallow rules and, consequently, much content may not be collected. Other content resists harvesting because of technical complexity or server configurations. Moreover, the scale of domain harvesting means it is very difficult—for both resource and technical reasons—to collect content frequently and in a timely (or time-specific) manner. Domain harvesting is efficient but not entirely effective per se.

Perhaps the principal limitation of ccTLD harvesting is that there is a large amount of content that is intellectually within scope for the national collection that is not published on the .au domain. Not only a large number of personal websites such as blogs that are published on services on international domains, but many Australian-based or Australian-focused organisations, businesses, online news sites, even academic institutions and government bodies have websites on non .au domains.[10] In addition, the modern Australian citizen's online world does not stop at the jurisdictional borders but ranges wherever it may through the borderless Web and of course into social media, where the concepts and boundaries of publication

[10]Webster (2019) suggests that as much as a third of the UK Web exists outside the ccTLD, that is, those hosts located in the UK but not on the .uk domain.

and communication are blurred or non-existent. International publishing platforms and social media services often offer the simplest access to online expression but can be the most challenging formats for web archivists both technically and in jurisdictional terms.

The warrant of legal deposit has jurisdictional boundaries and a collection representing a nation and its people is fundamentally constrained by the reach of this remit. While the nature of the Web itself does not sit entirely comfortably with archiving along national jurisdictions—and the mission of the Internet Archive indeed does try to overleap that constraint—national institutions by the fiat of this same remit remain the driving, sustaining and responsible organisations for the task. When the practical outcomes of the comprehensive collecting objective are considered, we must recognise that it is not about collecting all and every resource, in every form at every point in time, but rather using available methodologies, technologies and warrant to collect in sufficient scale and time to provide an intelligible representation of the whole.

3 Comprehensive Access

In March 2019, the NLA released its entire web archive holdings to open public access under the banner of the Australian Web Archive (AWA) through its Trove discovery service. The AWA includes content from the selective curated PANDORA Archive and the bulk harvested Australian Government Web Archive (AGWA).[11] Both these collections were already searchable and publicly available, PANDORA since 1998 and AGWA since 2014. Most significantly, the AWA also made accessible the entire content of the Australian domain harvests collected since 2005 as well as older .au content from 1996 to 2004 obtained from the Internet Archive. The Australian domain harvests comprise around 85% of the entire corpus; thus, the release of AWA through Trove was a significant step towards providing comprehensive access to the web archive.[12]

As stated above, the NLA's statutory function is not only to collect and preserve Australia's documentary heritage but also to make it available for use in the public interest. The purpose of collecting and preservation is only truly realised through access, and access is the real test of the national institution's commitment to the task since it exposes the organisation to a greater amount of legal and reputational risk. The extent to which the institution is able or willing to expose its web archive content is not only determined (and constrained) by legislation, but also by the degree of organisational aversion to risk.

[11] The Australian Government Web Archive was a prototype web archiving programme established by the NLA in 2011 with the objective to introduce in-house some infrastructure capacity for larger scale bulk harvesting and in doing so also to comprehensively collect Australian Government websites.

[12] At the time of its public release in March 2019, the Australian Web Archive consisted of around 8 billion documents amounting to a little over 600 terabytes of data.

Building on a two-decade history of providing access to its selective web archive, the NLA's approach to the expanded AWA was to provide open search access to content through both full text and URL indexing. This gives primacy to the content as documents as encountered by individual users and does not treat the corpus as essentially a dataset. Moreover, facilitating the personal encounter between user and intellectual content for which the Library was neither the creator nor first publisher may yet be considered an act of publication. In that context, actions to ameliorate risk associated with privacy, copyright and potentially illegal, defamatory or offensive material—including takedown processes—were implemented.

In exposing the complete web archive collection, the NLA took a number of actions to reduce reputational and legal risks including significant work on search results ranking, including Bayesian analysis of content, to push possible offensive content down the results rankings. This is preferable to a censorship approach but it should also be understood as bringing further curatorial bias to the collection as presented to the user. Another important action to mitigate risk is to limit unintentional exposure to the archive. While the entire web archive is openly accessible through the Trove Australian Web Archive portal, content is exposed neither to search engines nor to the Trove single discovery function that interrogates multiple collections in a single search. Thus, accessing the archive has to be a conscious and intentional act by the user.

In providing open access to the web archive content, the NLA does not identify or privilege any particular target user group. In the spirit of its legislated function, access is provided for the use of all Australians generally. This, certainly, means that the specific needs of potentially high-value research users are not necessarily met. For example, at the time of writing, there is no API available to researchers to interrogate the web archive metadata or content.

Like the Library's other collections of printed and digital materials relating to Australia and Australians, the national web archive serves to represent Australian culture to the world. As an online collection accessible over the Web, the AWA—like the Library's digitised collections—is at the forefront of how the Library presents Australian culture. The opportunity for such a collection to be curated with value-adding pathways, collections and analytics is considerable since metadata is collected or created at various stages of the curation process. The NLA has done little to exploit this opportunity while focusing on establishing the basic search facility. While the opportunity to build additional research access tools, curated pathways and analysis exists, the significant human and technical resources required to provide these value-added services continue to be a constraint for an organisation with its business committed to many services. This is not simply a technical matter but goes to the core of corporate planning and priorities.

The NLA's approach to developing its web archiving, from collecting to maintaining to providing fully open access, has been described by the Library's Director-General Marie-Louise Ayres as "radical incrementalism"—that is, taking the small and achievable steps, learning and evaluating along the way, and working within inevitable constraints that nevertheless lead, over time, to profound change. This allows the organisation "to achieve goals that would have previously sounded

too big, ambitious or risky ... setting a course and then sticking to it for the long term" (Easton 2019).

4 Conclusions

National web archiving programmes are long-standing, prominent and critical components in the international efforts to preserve the Web because the national libraries (and archives) that establish and maintain them are statutory institutions with the functional mandate to collect national documentary heritage. Moreover, they bring long experience and established expertise in collecting materials in a variety of formats, often supported by legal deposit legislation (or, if not, an approach resembling the purpose of legal deposit), and a sustainable vision supported by institutional robustness. Certainly, the National Library of Australia's web archiving programme could be characterised in these terms. Each national institution will, of course, have its own history, culture, resource priorities and structures that shape the nature of their individual web archiving programmes. For the NLA, whose web archiving programme spans the entire historical period of web archiving activity commencing in the mid-1990s, comprehensive collecting and comprehensive access have continued to be the objective, though how comprehensiveness has been conceptualised has necessarily changed over time. It may have been to see collecting web resources as completing gaps in the broader national collection; or achieving national scale by collecting the entire .au web domain; or understanding that the remit to collect published material should also include social media. Comprehensive national collecting is really a process of turning a statutory function into strategic objectives to collect from the elusive and protean Web the artefacts created and shaped by the technical, resource and legal constraints. In other words, national web archiving is fundamentally a strategic attempt to collect a functional representation of a comprehensive national collection of online publication. The success or otherwise of such attempts will ultimately depend on the engagement of the user with the collection. For the user seeking a specific document, this perspective may be different from a researcher looking to interrogate the archive as a coherent and complete dataset.

References

Agreement on Trade-Related Aspects of Intellectual Property Rights (1995) https://www.wto.org/english/docs_e/legal_e/27-trips.pdf. Accessed 14 November 2019

Australian Libraries Copyright Committee (2008) A user's guide to flexible dealing provisions for libraries, educational institutions and cultural institutions. Section 200AB of the Copyright Act 1968 (Cth). Australian Libraries Copyright Committee and the Australian Digital Alliance, Kingston, ACT

Brügger N, Laursen D (2019) The historical web and digital humanities: the case of national web domains. Routledge, London

Cathro W, Webb C, Whiting J (2001) Archiving the web: the PANDORA Archive at the National Library of Australia. Paper presented at the Preserving the Present for the Future, Web Archiving Conference, Copenhagen, 18–19 June 2001

Easton S (2019) Australia's top librarian tells how the National Library fosters a culture of in-house innovation. In two words: 'radical incrementalism'. The Mandarin. https://www.themandarin.com.au/110303-australias-top-librarian-tells-how-the-national-library-fosters-a-culture-of-in-house-innovation-in-two-words-radical-incrementalism/

Gatenby P (2002) Legal deposit, electronic publications and digital archiving: the National Library of Australia's experience. A paper presented at the 68th IFLA General Conference and Council, Glasgow, 2002. http://pandora.nla.gov.au/pan/21336/20080620-0137/www.nla.gov.au/nla/staffpaper/2002/gatenby1.html. Accessed 14 November 2019

Koerbin P (2010) Issues in business planning for archival collections of web materials. In: Collier M (ed) Business planning for digital libraries: international approaches. Leuven University Press, Leuven, pp 101–111

Koerbin P (2017) Revisiting the world wide web as artefact: case studies in archiving small data for the National Library of Australia's PANDORA archive. In: Brügger N (ed) Web 25: histories from the first 25 years of the world wide web. Peter Lang, New York, pp 191–206

Lyman P, Kahle B (1998) Archiving digital cultural artifacts: organizing and agenda for action. D-Lib Magazine, 4(7/8). http://www.dlib.org/dlib/july98/07lyman.html. Accessed 14 November 2019

Phillips M (2002) Archiving the web: the national collection of Australian online publications. Paper presented at the International Symposium on Web Archiving, National Diet Library, Tokyo, 30 January 2002

Phillips M, Koerbin P (2004) PANDORA, Australia's web archive: how much metadata is enough? J Internet Catalog 7(2):19–33

Webb C (2001) The National Library of Australia's digital preservation agenda. RLG DigiNews: 5(1). http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file1108.html#feature1. Accessed 14 November 2019

Webster P (2019) Understanding the limitations of the ccTLD as a proxy for the national web: lessons from cross-border religion in the northern Irish web sphere. In: Brügger N, Laursen D (eds) The historical web and digital humanities: the case of national web domains. Routledge, London, pp 110–123

Winters J (2019) Negotiating the archives of the UK web space. In: Brügger N, Laursen D (eds) The historical web and digital humanities: the case of national web domains. Routledge, London, pp 75–88