Web Archiving at National Libraries

Findings of Stakeholders' Consultation by the Internet Archive

Helen Hockx-Yu
Director of Global Web Services
Internet Archive

20 March 2016

Table of contents

Executive Summary

Internet Archive conducted a stakeholders' consultation exercise between November 2015 and March 2016, with the aim to understand current practices, and then review Internet Archive's current services in this light and explore new aspects for national libraries. This document reports on the consultation and summarises the findings.

Thirty organisations and individuals were consulted, representing national libraries, archives, researchers, independent consultants and web archiving service providers. All participants were interviewed, through face-to-face meetings, Skype calls or email. The interviews were semi-structured, using pre-defined as well as open questions. Some of the participants were interviewed for a 2nd time, and asked to comment on the outline of a possible new service, to understand to what extent the proposed service meets their requirements articulated earlier.

Many of the national libraries archive the web under a legal mandate, which requires them to crawl the entire country code top-level domain (ccTLD) and anything outside of it considered as in-scope (generally defined in ways that are complicated, inconsistent and technically not easy to implement). They also carry out selective crawls to build thematic collections. Access to content collected under the legal mandate is in general restricted to libraries' premises. Those currently without legal mandate feel they are not generally in a position to collect the entire top-level domain (TLD), but many of those continue to build smaller scale selective web collections.

Internet Archive is a provider of web archiving technologies and services, offering:

- Open source software for crawling and public access,
- A global web archiving service for the general public,
- Archive-It, a subscription service for creating, managing, accessing and storing web archive collections, and
- A tailored broad crawling services for national libraries.

Internet Archive has also collaborated with national libraries beyond these formal services, enabling national bodies to undertake web archiving locally by making available key software for web archiving. Many national libraries have made requests to Internet Archive for historical data extraction.

The main findings give an overview of the current practices of web archiving at national libraries, as well as a general impression of the progress in web archiving and specific feedback on Internet Archive's role and services. These findings are summarised and grouped into the following categories:

- Strategy and organisation
  National libraries choose to undertake web archiving in-house for various reasons. When web archiving becomes more important in a national library's strategy, many have wanted to own the activity and develop the capability in-house. This requires

integration of web archives with the library other collections and the traditional library practice for collection development. Budget cuts and lack of resources were observed at many national libraries, making it difficult to sustain the ongoing development of necessary tools for web archiving.

- Quality and comprehensiveness of collection
  There is a general frustration about the content gaps in the web archives. National libraries also have strong desires to collect the portion of Twitter, YouTube, Facebook and other social media which is considered as part of their respective national domain. They would also like to leverage web archiving as a complementary collecting tool for digital objects on the web and that are included in web archives such as eBooks, eJournals, music and maps.

- Access and research use
  National web archives are, in general, poorly used due to access restrictions. Many national libraries wish to support research use of their web archives, by engaging with researchers to understand requirements and eventually embedding web archive collections into the research process.

- Reflection on 20 years of web archiving
  While there is recognition of the progress in web archiving, there is also a general feeling that the community is stuck with a certain way of doing things without making any significant technological progress in the last ten years, and being outpaced by the fast evolving web.

- Perception and expectation of Internet Archive's services
  Aspects of Internet Archive's currently services are unknown or misperceived. Stakeholders wish for services that are complementary to what national libraries undertake locally and help them put in place better web archives. There is a strong expectation for the Internet Archive to lead the ongoing collaborative development of (especially) Heritrix and the Wayback software, and to improve the testing and documentation, not necessarily free of charge. There are also clearly expressed interests in services that can help libraries collect advanced content such as social media and embedded videos. A number of national libraries have expressed the need for a service supporting the use of key software including maintenance, support and new features.

The report also identifies a number of common challenges which at the same time represent opportunities for service development and collaboration. These are:

- Plug content gaps in web archive collections
- Integrate web archiving processes
- Integrate web archives with library infrastructure and workflow
- Leverage web archiving to save collecting effort
- Improve access and use

Internet Archive is actively reviewing its current services and exploring new elements to bring improvements. A vision of possible national library services will be presented at the General Assembly of the International Internet preservation Consortium (IIPC), on Wednesday 15 April 2016, by Internet Archive's Digital Librarian Brewster Kahle.

1. Introduction

Internet Archive started web archiving in 1996 and is today a major provider of web archiving software and services, which are used by hundreds of memory institutions and millions of end users across the world.

Internet Archive offers a range of web archiving services:

1. Global web archiving service. Internet Archive has undertaken various large-scale crawls since 1996, with the goal to archive the global web.
2. Archive-IT, a subscription service launched in February 2006 offering a fully hosted web application for creating, managing, accessing and storing web archive collections.
3. National libraries service, a crawling service launched since 2000 for national libraries / archives, collecting national web domains or building thematic collections, tailored to partners' requirements.

Internet Archive's Wayback Machine, hosted at https://archive.org/web/, is currently the largest and oldest web archive in the world. It provides access to Internet Archive's web archive collections, collected by the services described above. There are also a number of API services, offering machine-readable access to the Wayback Machine and Archive-It collections.

Between November 2015 and 2016, The Internet Archive conducted a stakeholders' consultation exercise, with the aim to review current services and develop new aspects for the national libraries service, addressing common challenges and requirements. This document reports on the consultation and summarises the findings.

Three principles, which underpin Internet Archive's overall activities, were used to guide the consultation and thinking of a new service. The idea is to enable each library to maintain control over its own digital collections, leverage work done by other libraries, and gain access to aggregated collections for broader and deeper access:

1. Collaborative collection development
2. Distributed preservation
3. Global and local access

2. Methodology

The consultation was conducted by Helen Hockx-Yu, who joined the Internet Archive in September 2015. The consultation was part of the process of evaluating Internet Archive's collaboration with national libraries, surveying their activities and interests in the areas of collecting born-digital materials. The goal is to propose adjustments to current services or

develop new services that are sustainable for all participants and help achieve their goals in building digital collections and access services. Before joining the Internet Archive, Ms. Hockx-Yu was Head of Web Archiving at the British Library between 2008-2015 and has been an active participant in the IIPC.

Thirty organisations and individuals were consulted, representing national libraries, archives, researchers, independent consultants and web archiving service providers. Among the national libraries and archives, some have used Internet Archive's services in the past, some are currently partners and some have not used the services at all. The Internet Archive would like thank the participants for being open with Ms Hockx-Yu through this process.

All participants were interviewed, through face-to-face meetings, Skype calls or email. The interviews were semi-structured, using pre-defined as well as open questions. Some of the earlier participants were interviewed for a 2nd time, and asked to comment on the outline of a possible new service, to understand to what extent the proposed new services meet their requirements articulated earlier.

Below is a list of the types of stakeholders who participated in the consultation:

Stakeholder category	Count
National Libraries	18
(National) Archives	2
National organisations / Consortia	2
Researchers	2
Independent consultants	3
Web archiving service provider	2
International organisations / consortia	1

Table 1. Break-down of stakeholders who took part in the consultation

The consultation should ideally involve more national libraries who currently do not do web archiving, and outside Europe. It however takes time to identify them and establish contact. We hope to pursue this in coming months to widen the sample of stakeholders consulted.

It should also be noted that the consultation exercise has not ended. Feedback and comments from stakeholders are still being sought till mid-May 2016. A key opportunity for obtaining feedback is the 2016 General Assembly of the International Internet Preservation Consortium (IIPC), which takes place in April 2016 in Reykjavik. Findings of the consultation will be reported to the community on 13 April 2016 at a panel entitled "Rethinking Web Archiving – Developing Services for National Libraries".[1]

3. Web archiving at national libraries

There are 246 national and state libraries in the world.[2] These are public bodies funded by the government each responsible for preserving their national heritage. Some countries have thematic or specialist national libraries in addition, e.g. Institut National de L'audiovisuel (INA) in France. Legal mandate, including dedicated legal deposit legislations and copyright exemption, exist in many countries to allow the national library to collect and provide access to the nation's published outputs.

In nearly 30 countries, and more likely to follow, digital and online publications have also become part of the scope of the longstanding legal mandate for national libraries.[3] For most national libraries in these countries, implementing Legal Deposit of web resources and harvesting at national scale have become a mainstream activity, supported by operational budget and staff. A common strategy is to crawl the Top Level Domain of a particular country (e.g. ".nz", called a ccTLD), as required by the mandate, and supplement this with additional focused and more frequent crawls of selected sites.

Just as with printed publications, access to content collected under the legal mandate is in general restricted to library premises. This requires users' physical presence and makes it hard for the national libraries to promote broad use of web archives for research. There is also a misalignment with users' expectations, especially because many websites were freely available online in the first place.

National libraries have a long tradition of collaborating through various federations, consortia and conferences. In the area of web archiving, 49 national libraries[4] are members of the International Internet Preservation Consortium (IIPC), which was established in 2003 to improve the tools, standards and best practices of web archiving while promoting international collaboration. Internet Archive is a founding member of the IIPC and has over the years actively collaborated with the national libraries.

Among the stakeholders being consulted, 19 are members of the IIPC.

19 of the 20 consulted libraries and archives are actively undertaking web archiving. The only national library that does not archive currently has run pilot projects in the past to archive the national web domain.[5]

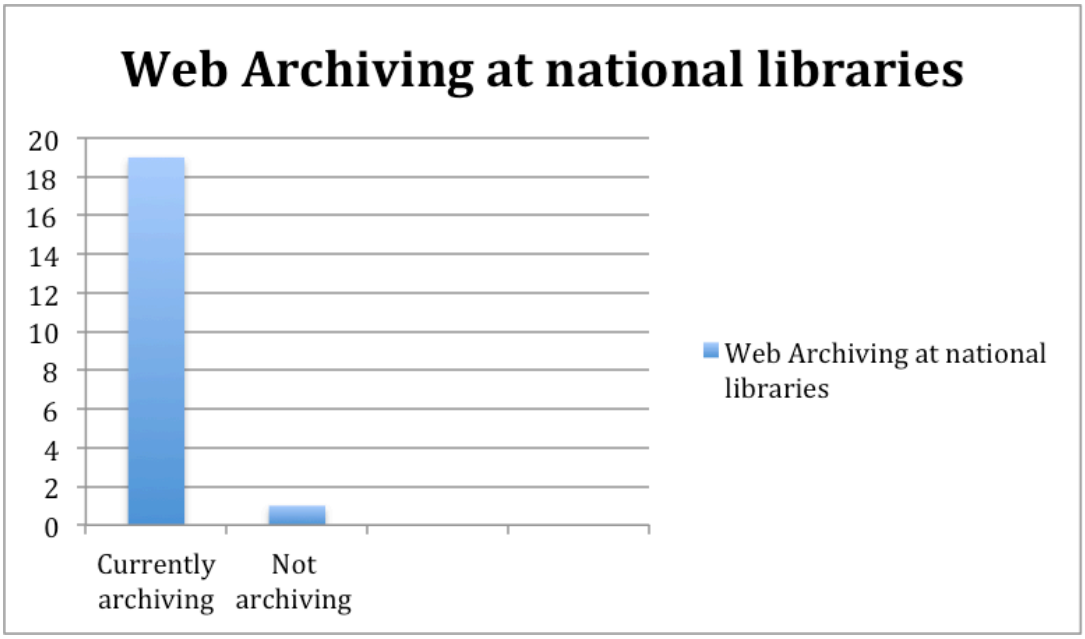

Figure 1. (Consulted) National libraries currently undertaking web archiving

14 of the 20 consulted libraries / archives archive the web with the support of an explicit legal mandate. At least two libraries are expecting legal mandate in the next couple of years.

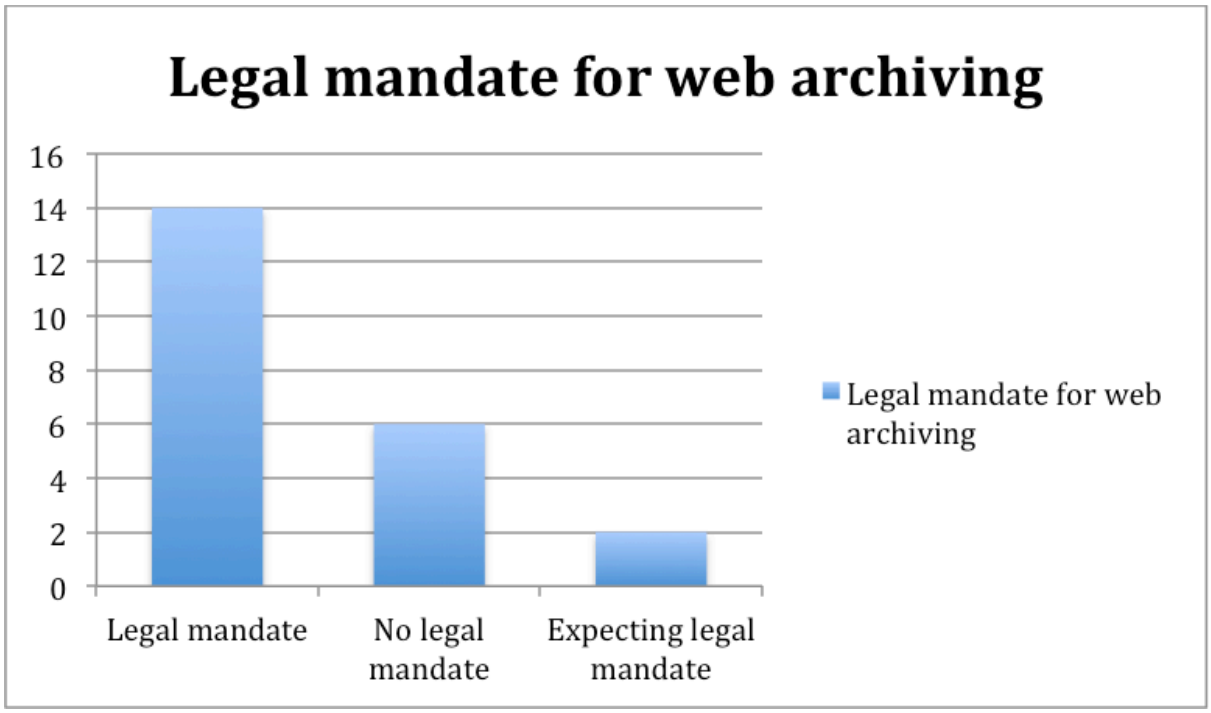

Figure 2. Number of National Libraries with Legal mandate for web archiving

Among the 19 libraries and archives that undertake web archiving, 13 conduct regular whole domain crawls, as well as selective crawls. 6 only archive the web selectively, building thematic collections, or periodically archive an organisation's web presence. Those archiving the web selectively are mostly libraries without the support of a national legal framework, with the exception of the National Library of Germany, which so far has only conducted an experimental domain crawl despite having Legal Mandate in place since 2006.

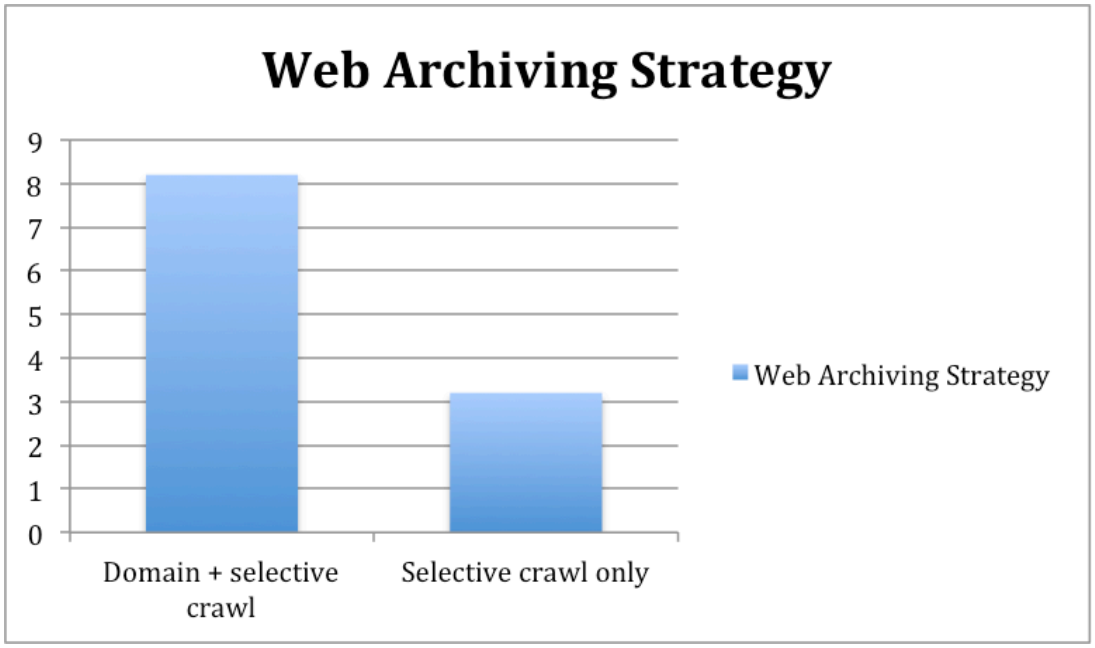

Figure 3. Number of National Libraries with different Web archiving strategies

14 libraries and archives undertake web archiving in-house. 5 use an external service provider, fully or partially, for their web archiving needs. Those undertaking web archiving in-house demonstrate a high concentration in the use of tools, i.e. Heritrix and Wayback, with the exception of INA, which developed its own suite of tools including archival file format to capture interactive and rich media websites through continuous crawling (as opposed to snapshotting). Many libraries also developed additional complementary tools to suit institutional archiving needs. The National Library of Australia has developed a high-level collection management tool called Bamboo, which brings together different crawls, regardless of the tools used to capture the content, so that these are indexed and managed in one place.[6]

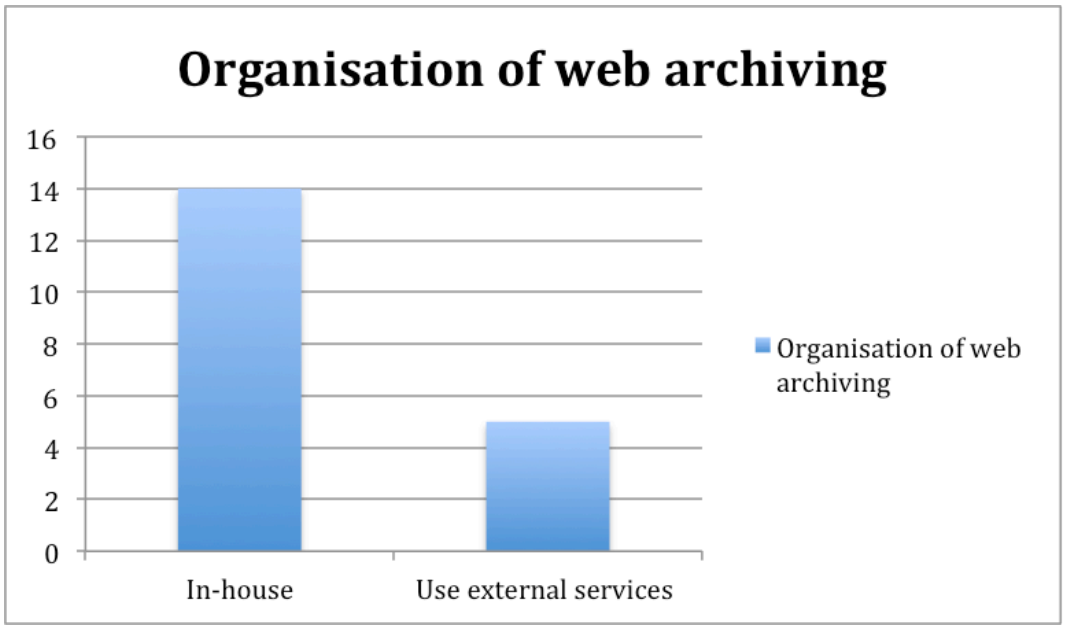

Figure 4. Organisation of web archiving

4. How national libraries use Internet Archive services

The Internet Archive has worked with national libraries and the Library of Congress since 1997 on policy, software development, and providing services to meet their needs. This section tries to represent how the Internet Archive's services are used by the national libraries.

4.1. Global web archiving service

The Internet Archive has been crawling the web since 1996, by taking ongoing snapshots of the web and several different crawl strategies. It also receives donations of crawl data, from companies, libraries, or through the Save Page Now feature. The goal is to build a broad and deep collection of publicly accessible web pages. Overall, about 1 billion web captures are collected each week.

The Internet Archive's position in Silicon Valley has helped it with technology and relationships to build its collections and to play a role in evolving law and policy aspects of digital libraries.

4.2 Wayback Machine

As a portal onto many of the web collections of the Internet Archive, the free Wayback Machine service has become a strong and established brand, benefiting the general public and researchers around the world, attracting 600,000+ visitors per day.

Various APIs also allow others to access and build services on top of the large historical collection.

Many national libraries use this service indirectly, by referring their users to the Wayback Machine and use material which the libraries themselves have not archived or which can only be accessed by being present in the library reading rooms.

URL-based search is currently the only access method for the Wayback Machine. This limits the potential of Internet Archive's vast collection of the historical web. Plans are already in place and the Internet Archive is doing work to improve the service. When completed in 2017, the next generation Wayback Machine will have more and better web pages that are easier to find.[7]

4.3. Archive-It

Several national libraries use Archive-It. It is a popular hosted web archiving service which allows partners to build (smaller scale) web archive collections. Its 400+ partner-base forms a critical mass required for a service that is supported and continuously improving. Archive-It 5.0[8] was released recently, with new features addressing partners' ongoing requirements. The general feedback on Archive-It is very positive.

The resulting thousands of curated collections can be full-text searched and annotated.

Partner institutions are encouraged to bulk download their collections for aid in preservation and alternative access services.

Archive-It is also developing a research service through projects and collaboration with researchers. Secondary datasets based on partners' collections can be requested for analytics and wider research use.

The current Archive-It partner-base has a strong North American geographical profile. There is potential for developing and increasing the partner-base in Europe and other parts of the world. Some useful features of Archive-It may not be widely known. The PDF-only feature, for example, is a great tool for collecting "official publications", which are materials issued for public use by federal, national, provincial or municipal governments and intergovernmental organisations, often published as PDF documents on websites. [9] National libraries have traditionally collected printed official publications and are required to collect the digital equivalent.

4.4. National library service

Among the consulted stakeholders, 4 currently use this bespoke crawling service and 4 were previous partners. The latter have since started crawling in-house with the Heritrix software from the Internet Archive. Legal deposit mandate was often the trigger for national libraries to carry out web archiving locally.

When the Internet Archive crawled a national domain on behalf of a partner, it greatly increased what was collected over its normal crawls. For instance the Internet Archive has so far captured 546 million unique .nz URLs for the National Library of New Zealand while only 273 million unique URLs were crawled via its other activities over all time.

4.5. Collaboration with national libraries

The Internet Archive has collaborated with national libraries beyond the formal services described above. The interaction and collaboration have focused on two areas:

4.5.1 Crawling and replay software

Internet Archive has played an important role in enabling national bodies to take on web archiving locally. The Heritrix software and Wayback software packages, which were developed and made available by the Internet Archive free of charge, are commonly used by national libraries and form critical elements of their web archiving infrastructure. Internet Archive currently develops and maintains Heritrix and contributes to the OpenWayback Project.[10]

4.5.2 Extraction of historical data

Most national libraries started web archiving much later than the Internet Archive. Once they start web archiving in-house, there is a desire to extract the historical data from the Wayback Machine, which relates to national libraries' respective Top Level Domains (TLDs), so that they can "have a complete collection".

5. Findings

The main findings of the consultation exercise, based on stakeholder response, are grouped and summarised below:

5.1. Strategy and organisation

While not ruling out the use of external web archiving services, many national libraries choose and will continue to do web archiving in-house, for various reasons:

- Legal requirement, explicit or perceived.
- Alignment with libraries' longstanding mission.
- Concern about service providers' longevity or concentration of power.
- Concern about security of data residing in an independent enterprise in a foreign country.
- Outsourcing is often only a transactional measure, a learning process to understand issues and prepare for in-house operation.

In general a national library thinks about web archiving because they expect legal mandate in the future. They typically start by running projects or experiments locally. When expecting web archiving to become more important in a national library's strategy, it is understandable that they would like to own the activity and develop the capability in-house.

Some national libraries partner with external service providers for web archiving. Reasons cited for choosing this strategy include:

- Lack of (human) resources
- Lack of technical skills
- Difficulty in maintaining the infrastructure required for domain crawl
- Wish to focus on curation and other things the libraries are best at doing

Web archiving is becoming an area of strategic importance for many national libraries. For those collecting under legal mandate, the aim is for web archiving to become a Business As Usual (BAU) activity, requiring integration of web archives with the library collections and the traditional library practice for collection development.

National libraries collecting the web under legal mandate typically run broad domain crawls in combination with selective archiving. Those currently working on or without legal mandate, are not in a position to collect the entire ccTLD, so may archive selectively (based on explicit or opt-out permissions). Both are in need of tools or services to manage the workflow including selection, scheduling, crawling, QA, description and access. Many develop these locally, on top of Heritrix and Wayback, but struggle to keep up with the effort required for on-going development.

Budget cuts and lack of resources were observed at many national libraries. This often means the choice between generic web archiving (e.g. a broad crawl of the national domain once a year) and the effort required to develop tools for and to archive advanced content.

5.2. Quality and comprehensiveness of collection

Deploying web archiving tools currently available would technically enable an organisation to undertake domain or selective archiving, capturing and replaying the static portion of the web and content which can be served by requesting a URL. However, the archiving technology is still not adequate to deal with the web in full, leaving certain types of content on the web out of reach. An increasingly bigger portion of the web is not being collected systematically, including content behind web forms and query interfaces, commonly known as the “deep web”, streaming media, web applications such as google-docs,content delivered over non-http protocols, and social media.

There is a general frustration about the content gaps in the web archives. National libraries have strong desires to collect the portion of Twitter, YouTube, Facebook and other social media which is considered as part of their respective national domain. Some see this leading to increased technological stagnation and eventually the failure of fulfilling national libraries’ mission of preserving cultural heritage.

For national libraries archiving the web under legal mandate, a common challenge is the scalable identification of in-scope content outside the country code top-level domains (ccTLDs). Depending on the specific territoriality definition in each country, National libraries currently use mixed manual and automatic methods for this but require scalable technical solutions for discovering such content on the Web. This is not only for the purpose of content collection, but also benchmarking, as the performance of national libraries is judged by the quality and comprehensiveness of their collections.

The legal mandate of national libraries covers a wide range of digital publications, e.g. eBooks, eJournals, music, maps. Many of these are on the web and are included in web archives but collected separately by national libraries. Some libraries have been experimenting or asking the question whether web archiving could be used as a complementary collecting tool for these objects. Many libraries expressed the desire for tools that would enable them to identify such contents within the web archives and process them in their own right to avoid duplication.

5.3. Access and research use

Most national libraries’ web archives are not publically available, especially the large scale national collections enabled by a legal mandate. These can only be accessed in the reading rooms or at premises controlled by the libraries. As a result, national libraries’ web archives are mostly “dark” and there is very little use of them. The perceived limited value of web archives sometimes leads to library management deciding to do the minimum and not invest more in web archiving.

Many national libraries wish to support research use of their web archives, by engaging with researchers to understand requirements and eventually embedding web archive collections into the research process.

Some national libraries have shown great creativity and made a lot of progress in engagement with the research community and developing research services, e.g. the British Library.

5.4. Reflection on 20 years of web archiving

There is recognition of the progress in web archiving, including the much better understood legal issues, the increasing number of practitioners and the emerging research use cases, based on the "big data" approach to take advantage of the uniqueness of web archives.

There is however also a general feeling that the community is stuck with a certain way of doing things without making any significant technological progress in the last ten years, and being outpaced by the evolving web.

Many expressed the desire for the IIPC to play a more active leadership role to advance web archiving.

5.5. Perception and expectation of Internet Archive services

There is a perception that national libraries' own domain crawls are deeper and more comprehensive (than the capture of national domains in the Wayback Machine). While this may be correct, a more appropriate comparison would be between a national library's own domain crawl and an Internet Archive's tailored crawl of a national domain. The latter not only covers a national domain in-depth and comprehensively, it also contains much content outside the ccTLD, based on analysis of historical and global crawl data, and automatic Geo-IP check.

Another misunderstood aspect of Internet Archive's National Library Service is that the contract crawls are limited to domain crawls. Internet Archive, on the contrary, has frequently worked with partners to build thematic collections, using automatic methods to suggest relevant seeds that complemented manual selection by curators.

A number of stakeholders think that it is timely for Internet Archive to start thinking about the national libraries' local situation and offering a service that goes beyond "data on disks". This is where the new service could make a difference but it is not something easy to build. The new service should be inclusive, focusing on offering something extra or complementary to what national libraries undertake locally. "It should not compete with what libraries want to do but to help them put in place better web archives".

Not all stakeholders think of the Internet Archive as a library. Its focus on paid service led to the perception of Internet Archive as a service provider or vendor. Some national libraries hope to collaborate more closely with the Internet Archive but need to understand more clearly the expectation and how to collaborate.

There are clearly expressed interests in services that can help libraries collect advanced content such as social media and embedded videos. A number of national libraries have expressed the need for a service supporting the use of key software including maintenance, support and new features.

There is a strong expectation for the Internet Archive to lead the ongoing collaborative development of (especially) Heritrix and Wayback, and to improve the testing and documentation, not necessarily free of charge.

Some stakeholders have encouraged the Internet Archive to depart from the centralised service model and move towards a distributed model where locally created and managed web archives are globally integrated.

Some stakeholders encouraged Internet Archive to enable “donated collections”, allowing researchers to archive and upload archived content in an easy way.

Many stakeholders are excited about Internet Archive’s plans for the Wayback Machine and confirmed that exposure of and access to “national collections” in Wayback Machine is a useful and desirable feature.

There is some confusion about the “Wayback Machine” and the “Wayback” software and both terms are used interchangeably. While the former is a public service that provides online access to Internet Archive’s web archive and the latter is a Java-based software, written by the Internet Archive and released in 2005, that replays archived web resources.

6. Common challenges and opportunities

Based on the findings, a set of common challenges can be identified which apply for all national libraries carrying out web archiving. These also represent opportunities for service development and collaboration.

6.1. Plug content gaps in web archives

As described in 5.2, a shared challenge for national libraries is to plug the gaps in web archives by collecting all in-scope content regardless of the formats. To achieve this, improvement of current tools and development of new tools will be required.

The identification of content outside ccTLDs needs to be done in a more confident and scalable manner. This is an area where those conducting global or broad domain crawls can collaborate.

6.2. Integrate web archiving processes

Most national libraries use a combined collecting strategy, conducting broad national domain level crawls as well as selective crawls of websites which are deemed curatorially important. Domain and selective archiving are in general run as separate processes, and in many cases, the separation applies to the user interface of the collections too. Internet Archive is no exception to this. The Wayback Machine and Archive-It collections are collected and accessed separately.[11]

The National Library of Australia (NLA) is an example of good practice which other libraries perhaps can learn from. They have been integrating all back-end and eventually also the front-end processes, aiming to bring together all web archiving effort regardless of how content is collected or who collected it. The Bamboo Collection Manager keeps track of all crawls, including domain crawls outsourced to the Internet Archive and Archive-It collections, as well as content crawled by the NLA (e.g. periodic crawls of the Australian Government domain .gov.au) and partners of the Pandora Web Archive.[12] By clicking one button, the archivist can import crawls and kick-off the indexing process. Bamboo also offers tools for collecting specific content, such as PDFs and YouTube video. NLA's approach allows great curatorial overview and control of library-wide web archiving activities.

6.3. Integrate web archives with library infrastructure and workflow

As web archiving becomes a mainstream activity at national libraries, there is a need to integrate web archives with the libraries' other printed and digital collections so that these can be managed and accessed together. This requires the web archives to be described, stored, discovered and accessed in a certain way based on each institution's chosen approach and the systems used to manage collections.

Effort can be observed in national libraries where web archiving is a more mature activity. Work is taking place to generate catalogue and metadata records for resource discovery, and Submission Information Packages (SIPs) for digital management and storage systems. Many libraries have expressed concerns about funding the ongoing development for this integration, due to budget cuts, and require technical skills. There is also little collaboration in this area. While each library has its own workflow for processing digital resources, there is a relatively small number of systems and standards in use. It makes sense to coordinate development and meet common needs.

6.4. Leverage web archiving to save collecting effort

Web archives contain different types of content such as ebooks, ejournals, music and maps. Many national libraries have traditionally collected these publications and are required to continue collecting their digital equivalents. Since many of these are embedded in websites and reside in web archives, it would be a significant saving if these are not collected separately.

6.5. Improve access and use

Access and use of web archives is an area where active development can be expected in the next few years. The Internet Archive has ambitious plans to develop better access to its web archive collection and support research use. National libraries may appear risk averse but they do have to deal with various degrees of legal issues, which means there is not much they can do in providing access to archived content. They however have intimate knowledge of the national web domains and can help the Internet Archive make sure that these are well captured as part of the global web archiving effort. In return, they themselves and users from their countries can benefit from the exposure of "national content" in the Wayback Machine. Opportunities also exist for access to metadata and the use of web archives as "big data", which does not

necessarily require access to the archived “text”. The British Library and the Royal Danish Library have already done work that leads this exploration.

7. Next Steps

Based on the feedback of stakeholders, the Internet Archive is actively reviewing its current services and scoping new elements to bring improvements. The intention is to build on and extend the current services, offering flexibility and more lifecycle support for national libraries. An outline of the new national library service will be presented at the IIPC GA by Internet Archive’s Digital Librarian Brewster Kahle, who is giving a keynote on Wednesday 15 April 2016.

Appendix 1. List of Stakeholders who were consulted

We are grateful to all who have taken the time to talk to us and participate in the consultation. Thank you.

National Libraries
National Library of New Zealand – Steve Knight
National Library of Australia – Paul Koerbin
National Library of Spain – Mar Perez
British Library – Andy Jackson, Richard Price, Jude England
National and University Library of Iceland - Kristinn Sigurðsson
Royal Library of Denmark - Birgit Nordsmark Henriksen
National Library of Finland - Esa-Pekka Keskitalo
National Library of Italy - Paola Puglisi
National Library of Estonia - Jaanus Kõuts
National Library of Singapore - Kenny Chan
National Library of France - Gildas Illien and Emmanuelle Berme
Institut National de L'audiovisuel - Claude Mussou and Thomas Drugeon
National Library of Japan - Masaki Shibata
National Library of Germany – Tobias Steinke
National Library of Austria - Michaela Mayr
National Library of Ireland – Joanna Finegan, Della Murphy and Geraldine Wilson
Portuguese Web Archive – Daniel Gomes
Library of Congress – Abbie Grotke

(National) Archives
UK National Archives - John Sheridan and Tom Storrar
UK Parliamentary Archive – Chris Fryer

National Organisations / Consortia
Dutch Digital Preservation Coalition – Marcel Ras
Digital Curation Centre UK – Kevin Ashley

Researchers
Jane Winters, Professor of Digital History and Head of Publications, Institute of Historical Research, University of London
Ian Milligan, Assistant Professor, University of Waterloo

Independent Consultants
Darryl Mead, previously Deputy Librarian of the National Library of Scotland
Peter Webster, previously Web Archiving Engagement Manager of the British Library and Programme and Communications Officer of the IIPC

Sean Martin, previously Head of Architecture and Development of the British Library

Web archiving service providers
Internet Memory Foundation – Julien Masanes
Archiefweb.eu – Marcel Privé

International Consortium
International Internet Preservation Consortium (IIPC) – Sabine Hartmann

---

[1] IIPC Web Archiving Conference 2016 schedule: http://netpreserve.org/sites/default/files/IIPC%20Web%20Archiving%20Conference%202016%20-%20schedule.pdf
[2] List of national libraries: https://en.wikipedia.org/wiki/List_of_national_and_state_libraries.
[3] Legal deposit: https://en.wikipedia.org/wiki/Legal_deposit.
[4] IIPC membership: http://netpreserve.org/about-us/members.
[5] The only library which doesn't do web archiving is disregarded in the remaining charts.
[6] Bamboo: https://github.com/nla/bamboo.
[7] Grant to Develop the Next Generation Wayback Machine: https://blog.archive.org/2015/10/21/grant-to-develop-the-next-generation-wayback-machine/.
[8] Archive-It 5.0 release notes: https://webarchive.jira.com/wiki/display/AITH/Archive-It+5.0+Release+Notes.
[9] Collecting PDFs: https://hhockx.wordpress.com/2015/10/26/collecting-pdfs/.
[10] The Open Wayback Project: http://netpreserve.org/openwayback.
[11] Archive-It content is merged into the Wayback Machine but the effort by partners in curating their Archive-It collections is not reflected. Internet Archive is currently running a project to expose provenance metadata and should shed light on the best way of integrating the two at the front end.
[12] Pandora Web Archive: http://pandora.nla.gov.au.