Submitted on: 13.06.2017

Getting Started in Web Archiving

Abigail Grotke
Library Services, Digital Collections Management and Services Division, Library of Congress, Washington, D.C., United States.
E-mail address: abgr@loc.gov

Abstract:

This purpose of this paper is to provide general information about how organizations can get started in web archiving, for both those who are developing new web archiving programs and for libraries that are just beginning to explore the possibilities. The paper includes an overview of considerations when establishing a web archiving program, including typical approaches that national libraries take when preserving the web. These include: collection development, legal issues, tools and approaches, staffing, and whether to do work in-house or outsource some or most of the work. The paper will introduce the International Internet Preservation Consortium and the benefits of collaboration when building web archives.

Keywords: web archiving, legal deposit, collaboration

1 BACKGROUND

In the more than twenty five years since the World Wide Web was invented, it has been woven into everyday life—a platform by which a huge number of individuals and more traditional publishers distribute information and communicate with one another around the world. While the size of the web can be difficult to articulate, it is generally understood that it is large and ever-changing and that content is continuously added and removed. With so much global cultural heritage being documented online, librarians, archivists, and others are increasingly becoming aware of the need to preserve this valuable resource for future generations.

In 1996, Brewster Kahle founded both Alexa Internet, Inc. (a commercial web traffic and analysis service) and the Internet Archive, a nonprofit digital library, and began archiving the internet. The results of the web crawls performed by Alexa for its commercial purposes were donated to the Internet Archive's web collection, and so began an effort to document the web and preserve it for future researchers. The crawler made snapshots of as much of the web as it could find, downloaded the content, and eventually made it accessible via a tool called the

Wayback Machine. Sites were visited again and again, to collect new versions so that comparisons of a particular site were possible, and changes could be documented.

About the same time that the Internet Archive began crawling the web, national libraries and archives around the world also began to see the importance of preserving this global resource. Many countries even sought changes in their laws to allow them (or mandate them) to collect and preserve the output of their nation's creativity. The National Library of Australia was among the first to partake in web archiving at a national library; the Nordic Web Archive---a project started in 2000 between the National Libraries of Denmark, Finland, Iceland, Norway and Sweden---developed a toolset for accessing web archive collections and is an early example of collaborative web archiving efforts involving a number of national libraries. And in 2003, the International Internet Preservation Consortium was formed, with eleven national libraries and the Internet Archive. Many national libraries have established programs today, still others are beginning to consider archiving portions of the web to meet legal deposit requirements or to simply collect the digital output of their own country's citizens.

Although the Internet Archive captures a lot of the web, no one institution can collect an archival replica of the whole web at the frequency and depth needed to reflect the true evolution of society, government, and culture online. A hybrid approach is needed to ensure that a representative sample of the web is preserved, including the complimentary approaches of broad crawls such as those made by the Internet Archive, paired with deep, curated collections by domain, theme or site, tackled by other cultural heritage organizations, such as members of the IIPC. Many types of organizations archive the web, not just national libraries. Archives, universities, museums, and government organizations are all involved, plus corporations and others preserving their own content.

For institutions that are just starting out a number of factors are worth considering, as well as various social, technological and legal challenges that organizations are typically faced with. Web archiving is technically challenging, but in the experience of IIPC members, many of the initial challenges can be in integrating web archiving into the traditional library processes and training staff on how to work with this material. Beyond determining what to preserve, some considerations include: staffing and organization, legal situations or mandates, types of approaches to crawling and the tools used, and access. This paper will provide a brief introduction to these topics.

2 GETTING STARTED

2.1 Why Start a Web Archiving program?

Even though it might appear that Internet Archive and others are archiving much of the web, focused collecting by cultural heritage institutions is important – it allows for more depth and coverage, and more frequent and comprehensive captures of important web content, such as news and government sites. Reasons for embarking on web archiving vary, but typically national libraries are interested in preserving their national domains, particularly as legal deposit laws are modified to include electronic materials. Many are also interested in collecting representative samples of web content created and published in their countries to supplement more traditional collecting of print and electronic materials, as materials that used to be published in print are increasingly found only online. Websites are also archived to document significant events, for capturing the records of government and businesses, and to capture the creative output of citizens.

2.2 Establishing Collection Policies and Determining What to Collect

One of the biggest questions facing organizations when they begin web archiving is determining exactly what to collect. With such a mass of data being produced on the web, and so many other organizations archiving portions of the web, it is important for each organization to determine their collecting strategy when it comes to archiving the web. Legal deposit or other legal considerations may drive some aspects of this (see Section 2.3).

There are a few distinct approaches to web archiving that your library should understand as you begin: bulk or domain harvesting and selective, thematic, or event-based harvesting, for example. An example of bulk archiving is the approach that the Internet Archive takes, trying to archive as much of the public web as possible based on an automated approach. Many national libraries focus on top-level domain crawling—archiving all the web sites identified by their URLs as hosted in a particular country. For instance, Iceland and France's country domains are easily identified (.is and .fr are fairly good indicators), and the sizes of those domains are manageable enough for them to do a crawl of them once or twice a year. However, identifying what makes up the "United States web" is nearly impossible (there is no one domain that covers all websites published in the United States), so the Library of Congress instead takes a selective approach to archiving around certain themes and events. This approach is also used at other national libraries who might not have legal deposit laws, or who want to complement their domain crawls with more focused collections around themes or events.

If selective archiving is part of your plan, it is important to decide on what themes, subjects, or types of sites your organization will archive. Discussing selection and collection policies can help focus the activity, which is particularly important if resources are limited. If you are just starting to think about web archiving, this area is likely not covered by your organization's existing collection development policies. Some IIPC members have created new policies (http://netpreserve.org/collection-development-policies) or supplemental guidelines to help curatorial staff make decisions about what to select for archiving. For instance, the Library of Congress guidelines state:

> *The Library of Congress will acquire through web harvesting selected websites and their multi-format contents for use by the U.S. Congress, researchers, and the general public. The Library will define the attributes for selection, preserve the web content that reflects the original user experience and provide access to archived copies of the harvested material. The sites of Legislative Branch agencies, U.S. House and Senate offices and committees, and U.S. national election campaigns will be acquired comprehensively. For other categories of web sites, only representative sites will be chosen, primarily as part of collections on a defined topic, theme or event.*

At the Library of Congress the topics, themes, and events are determined by Recommending Officers who are responsible for collecting traditional and digital materials for the Library's collections.

Many national libraries that are just starting out begin with archiving elections in their regions of the world, since campaign websites or other materials surrounding elections are particularly ephemeral – once the election cycle is complete, campaign materials are likely to disappear from the web. This can be a good pilot project activity to get started with, because it provides a focused, time-bounded project that can be evaluated before proceeding with

other collecting efforts. This type of collection can also show the value of preserving at-risk content as such websites change quickly and even disappear.

Be aware that other organizations in your country might be archiving portions of the web that your library might be interested in as well, so investigating what your organization's responsibilities and interests are can be helpful (although some duplication of effort is not a bad thing, as the approaches or strategies each take can vary). In some countries, national libraries have worked with other archiving organizations to coordinate or share preservation responsibilities. For instance, in France, the Bibliothèque nationale de France preserves most of the French web under their legal deposit program, however INA (the national institute for radio and television, which is responsible for preserving the audiovisual heritage of France) collects sites related to audiovisual communications (primarily radio and TV). This distinction between their collections was determined in a decree published in December 2011, which specifies both selection and communication procedures for web archives collections. In the United States, the Library of Congress works with a number of other federal agencies and archiving organizations on collaborative projects to preserve government websites.

2.3 Understand the Legal Situation

International and national legislation and other local and institutional policies often have a profound impact on what web content can be archived and made accessible for research use at cultural heritage institutions, so legal situations must be discussed as organizations plan their web archiving activities. National libraries and other web archiving organizations face different legal frameworks: some are awaiting legislation, others have legislation that covers web archiving, or other legal doctrines (such as fair use, in some countries) that permit or mandate web archiving. Example scenarios some International Internet Preservation Consortium (IIPC) organizations have experienced include:

- Laws covering some types of electronic content do not always extend to websites.
- Identifying what falls within the scope of a domain can be a challenge. Content produced by a country's citizens might appear on domains outside that country.
- Access to archived content is not guaranteed under legal deposit laws. Some laws may only allow researchers to use the archives on the library premises, in some cases because of privacy laws or concerns. There is hope that some may extend the concept of "premises" to include other branches or partner libraries, to allow for broader access.
- Legal deposit laws may allow institutions to ask for passwords and technical information for subscription content or other material that cannot be collected by automatic harvesting; and/or in some cases publishers can deposit files directly rather than the institution using harvesting.

When legal frameworks are unclear or absent specific legislation, many web archiving organizations follow a permissions-based approach. Seeking permission or notifying site owners can be a viable option in order to archive and preserve web content. This approach may allow site owners to opt out of archiving and/or public access to their archived content.

However, this approach has its own set of challenges:

- If seeking permission when archiving, there can be challenges with a lack of response from site owners. It's not that websites often deny permission, it is more a lack of

response to attempts to contact them. When permission is not granted, the result can be patchy, unbalanced collections.
- The level of effort required to notify site owners or obtain permission (sometimes tremendous) can overwhelm staff resources. Some organizations, such as the Library of Congress, have developed in-house tools to manage the permissions process and responses from site owners.

Without a legal mandate, risk assessments and fair use analysis may allow some organizations to collect without extensive permissions processes. For instance, in the United States, some organizations follow the Association of Research Library's Code of Best Practices in Fair Use (http://www.arl.org/focus-areas/copyright-ip/fair-use/code-of-best-practices) when evaluating approaches to take with web archiving.

2.4 Identify Organizational Resources

No matter the size of the organization, libraries just starting out must investigate and identify what staff resources are available for web archiving activities and where these activities fit within their library organizational structure.

How the web archiving staff are organized, and who specifically will be involved at startup will of course depend on a number of factors: How will the crawling be performed? Will staff work full- or part-time? Will some of the work be outsourced? Will the archive be accessible to researchers or to the general public? How will researchers access the archive? Will the web archives be cataloged or metadata generated to support access? In any case, web archiving typically involves a mixture of curatorial staff who determine what content gets archived, and technical staff who perform various aspects of the work, such as preparing the nominated sites for crawling, initiating crawls to capture the content, data processing, quality assurance, and preparation for access.

IIPC members report that some of the more successful programs have cross-organisation, cross-skilled teams working together to deliver a service: curatorial, reference, and cataloguing staff working with technologists, programmers, and web designers to accomplish program goals. In some large organizations, it has proven helpful to have one strong technical lead or program manager who can bridge the gap between the curatorial staff and technical staff.

Although it may not be feasible to establish a cross-skilled team when first starting out, it is helpful to recognize that web archiving requires a number of disciplines coming together in ways not typical in traditional library acquisition processes. Organizations establishing new programs are wise to consult with a variety of staff in planning web archiving activities, including legal staff to help guide policy decisions; strong curatorial staff to determine scope of those activities and technical staff to help determine best practices and tools that will be required.

2.5 Determine an Approach for Capturing Content

Availability and expertise of staff will likely drive decisions about whether web archiving using in-house staff and tools and services, or performed by staff using externally hosted services, or completely by external vendors or contractors. Some organizations start out using

external services but then as internal expertise is gained and more staff is available, transition to internal crawling operations.

Many organizations that are just getting started decide to outsource much of the work, unless there are technically savvy staff members available to manage the crawls and an infrastructure in place to store large quantities of data. Outsourcing or collaboration on projects allows organizations to gain experience and learn more while setting up internal infrastructure to manage in-house web archiving projects.

There have been many of changes in web technologies, in the tools used to archive the web, and in the community that has created and preserved born-digital content since the Internet Archive began in the 1990s. Most of the IIPC members use web crawlers that generate WARC files, an ISO standard for web preservation (https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml). Many are using open source crawling and access tools, including Heritrix (https://github.com/internetarchive/heritrix3) a web crawler that was designed by Internet Archive and libraries to archive the internet, and OpenWayback (https://github.com/iipc/openwayback/wiki), an open-source tool that allows playback of web archive content, which is supported by members of the IIPC. Various tools available and in use by IIPC members can be found at http://netpreserve.org/web-archiving/tools-and-software.

Despite many advancements in crawler tools, there continue to be some limitations as to what can be archived. The crawler technology lags behind the technology of the current web. For example, Heritrix is currently unable to archive streaming media, "deep web" or database content requiring user input. Social media present obvious challenges as new services are developed and used by more and more content producers, and it's often an important piece of the web presence of organizations or individuals that we are trying to archive. IIPC members and others in the web archiving community continue to work together to develop new tools to solve some of the more critical challenges.

It is important to identify tools that can help with selection or workflow management as well. Many of the web archiving services available provide ways to manage different processes such as nomination of URLs, permissions, crawling, quality review, and description. One curator tool that many national libraries performing domain and selective harvests have used is NetarchiveSuite (https://sbforge.org/display/NAS/Releases+and+downloads). This tool was developed by netarchive.dk and is used by a number of IIPC members to manage their web archiving work.

For those starting out, it's never been easier to test out tools and explore how web archiving works, thanks to efforts like Webrecorder (https://webrecorder.io/), a tool created by Rhizome and funded by the Andrew W. Mellon Foundation and the James S. and John L. Knight Foundation, that allows users to create interactive recordings of websites and a platform to make them available, and Archive-IT (https://archive-it.org/), a web archiving service provided by the Internet Archive and in use by many smaller cultural heritage institutions, to name a few. Services and tools like these allow institutions to build collections and experiment with capture processes and workflow even if programs are not fully realized, and without a huge investment in local technical infrastructure.

2.6 Consider Access

Researcher access sometimes is an afterthought in the frenzy of just trying to capture the content before it goes away. But in starting up a new web archiving program, libraries will want to think about how researchers will access their web archives, and how the web archives will integrate with other digital collections available for research use.

If your library has a legal deposit law or policies you must adhere to, you will need to determine if it allows for open researcher access, such as the National Library of Iceland (http://vefsafn.is/index.php?page=english) and the Portuguese Web Archive (http://arquivo.pt/?l=en), or if access is to be limited to library premises. Some IIPC members must restrict access to their web archives to onsite or restricted use only, even if the websites archived were once or are still publicly available, such as at the National Library of Norway (http://www.nb.no/fag/nasjonalbibliotekets-samling/nettdokumenter2) and the Danish web archive, netarchive.dk. In some instances, when seeking permission to perform web archiving, permission is also requested to provide offsite access; if permission is not received, that content might not be available to researchers outside the library. Some cultural heritage organization must embargo content prior to making it available for researcher access, such as the Library of Congress (loc.gov/websites). The embargo times vary but a year to six months is typical if this approach is used. In some cases, metadata about the archives, or derivative datasets, can be shared more widely even if the content might be more restricted.

It is also helpful to plan what sort of access you can provide for research, in terms of technology, description, and search, no matter the restrictions in place. Typically URL search via software such as OpenWayback is provided. Some libraries provide additional metadata, such as item-level or collection-level catalog records, that can be searched or to allow better browsing of sites. Visualizations and bulk data downloads are perhaps realized at later stages of a web archiving program, but if your library is already doing this type of access with other digital collections, it might be useful to think about how web archives could be made available in this way. The British Library has done a lot of work in this area; explore their site at https://www.bl.uk/collection-guides/uk-web-archive to learn more. Their SHINE tool, a prototype historical search engine, was developed in conjunction with researchers.

It is also helpful to think about how the web archives will be integrated (or not) with existing digital collections. Integrated search, combining web archives with other digital objects, can be a challenge and may take some time and coordination with other digital library staff at your library.

IIPC has worked in recent years with researchers to develop tools and methods for working with web archives. Those starting out in web archiving may be interested in the work of Web Archives for Historians (https://webarchivehistorians.org/) and Archives Unleashed (http://archivesunleashed.com/).

3 COLLABORATION IS KEY – JOIN THE IIPC

With all these considerations, it is clear that one institution cannot archive the web alone. That is why many who are archiving the web collaborate with partner libraries, archives, and other organizations around the globe. As more and more national libraries and organizations

begin web archiving, it is helpful to share experiences and learn from others who have been engaged with web archiving for some time.

The IIPC, as it was formed in 2003 by national libraries and the Internet Archive, has acknowledged the importance of international collaboration for the preservation of internet content. The goals of the consortium have included collecting a rich body of internet content from around the world and fostering the development and use of common tools, techniques, and standards that enable the creation of international archives.

Today, the consortium has over 50 member organizations that meet regularly, virtually and in person, to discuss and solve issues related to access, harvesting, and preservation. Members develop collaborative collections together; convene around development and support of open-source tools, sponsor training and workshops for members, and holding conferences for members and the general public. The IIPC encourages any interested organizations to join the IIPC (http://netpreserve.org/join-iipc) as they begin to establish their own programs.

Acknowledgments

Thanks to IIPC members Andy Jackson (British Library), Jefferson Bailey (Internet Archive), Olga Holownia (British Library and IIPC PCO), Emmanuelle Bermès (BnF) and Gina Jones (Library of Congress), for their assistance with this paper.

References

The Archive-It Team Internet Archive (2013) Web Archiving Life Cycle Model. Retreived from: http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf

Archives Unleashed (n.d.) Retrieved June 12, 2017, from: http://archivesunleashed.com/

Hallgrímsson, Þ. and Bang, S. (2003) Nordic Web Archive. Retrieved from: http://nwatoolset.sourceforge.net/docs/nwa@ecdl2003.pdf

International Internet Preservation Consortium, http://Netpreserve.org

The National Archives (2011) Web Archiving Guidance. Retrieved from: https://nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf

National Digital Stewardship Alliance (2017) Web Archiving in the United States: A 2016 Survey. Retrieved from: http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf

Web Archives for Historians (n.d.) Retrieved June 12, 2017, from: https://webarchivehistorians.org/