Web Archiving Environmental Scan

Harvard Library Report
January 2016

Prepared by Gail Truman

HARVARD LIBRARY

Prepared by Gail Truman, Truman Technologies
Reviewed by Andrea Goethals, Harvard Library and Abigail Bordeaux, Library Technology Services, Harvard University

This report was produced with the generous support of the Arcadia Fund.

Citation:
Truman, Gail. 2016. Web Archiving Environmental Scan. Harvard Library Report.

Table of Contents

Executive Summary

In this environmental scan, made possible by the generous support of the Arcadia Fund, we document web archiving programs from 23 institutions from around the world and report on researcher use of – and impediments to working with – web archives. The collective size of these web archiving collections is approximately 3.3 petabytes, with the smallest collection size under one terabyte (TB) and the largest close to 800TB. The longest-running programs are over 15 years old; the youngest started in 2015. The scan does not include corporations or for-profit institutions archiving their web presence for business reasons, but focuses on cultural memory institutions – libraries, museums and archives – collecting for researcher and historian use. For each section in this report we identify common concerns and questions to be addressed and present them as opportunities for future research and investment.

Through engagement with 23 institutions with web archiving programs, two service providers and four web archive researchers, along with independent research, the environmental scan uncovered 22 opportunities for future research and development. At a high level these opportunities fall under four themes: (1) *increase communication and collaboration, (2) focus on "smart" technical development, (3) focus on training and skills development, and (4) build local capacity.*

- Our investigation into current practices in web archiving reveals the need to radically *increase communication and collaboration*. Of the 22 opportunities identified for future exploration, 13 fall under this theme, making *increase communication and collaboration* the number one theme (note, some opportunities fall under more than one theme).
- *Focus on "smart" technical development* ranks as the second most popular theme, showing up in eight of the 22 opportunities.
- *Training and skills development* ranks as the third most prevalent theme - it surfaces in six of the 22 opportunities for future exploration.
- *Build local capacity* relates to augmenting an institution's resources and proficiency in the area of web archiving and related services. It ranks fourth in themes with four of the 22 opportunities falling here.

Our findings, outlined through each section, are described in more detail in *Findings and Opportunities for Future Research and Development*. To summarize these findings by theme:

Increase communication and collaboration

More communication is needed not only among the librarians and archivists who build and steward collections of archived websites, but also between these stewards and the historians, data scientists, and researchers who use their collections. We see opportunity for increasing communication and collaboration

across these different types of collectors and users. Absent sufficient communication, institutions today lack insight into the web archive collection decisions and practices of others, and this shortcoming can result in duplication or gaps in coverage and siloed collections. These siloed collections are often hard to find or access by other institutions – or by the researchers whose work now must include web-based resources to support their research questions. Insufficient training and experience using web archives and the lack of adequate description of the collections exacerbate this problem. The result is that researchers cannot easily discover and use archived web content together with related non-web content or easily study web archives through large-scale data mining. We offer some suggestions for greater communication and collaboration with the web archiving researcher community, to gather researcher feedback on requirements and impediments to the use of web archives. From a discovery perspective, we see an opportunity to communicate across web archiving institutions to investigate whether Memento should be adopted more broadly as part of their discovery infrastructure. In this document's tools section we identify the need to communicate across institutions and researchers to gather requirements for the next generation of tools that need to be developed.

Focus on "smart" technical development

Several institutions advocated for the creation of software tools to assist with aspects of web archiving, which prompted an exploration of the tools currently available for various aspects of the web archiving lifecycle. For this, the lifecycle of traditional library practice for collection development was used (pre-accession, accession, process, preserve, access) and then subdivided into more granular areas, some of them specific to web archiving. This breakdown plus the entire list of 77 tools is shown in Appendix C.

The tool categorization by function along the web archiving lifecycle revealed that tools seemingly serve some functions very well – such as capture and analysis. However, many of these capture and analysis tools are very specific to narrow types of media (e.g. capturing tweets) or support for particular types of analysis (e.g. link analysis). And, depending on the sites or research of interest, multiple different types of capture and analysis tools may be needed. However these tools were not designed to be used together or in modular ways. To address this, we see opportunities to develop an API framework that could allow these tools to be used together more easily and swapped out as needs or technologies change. Continuing with APIs, we identify the opportunity to work with service providers to help reduce the risk of reliance on them, where APIs could help transfer content as needed.

The scan surfaced many tool development opportunities that would make it easier for researchers to use web archives. We recommend that tools be developed (leveraging existing tools where possible) to make researcher analysis of big data found in web archives easier, and we recommend establishing a standard for describing the curatorial decisions behind collecting web archives, so that there is consistent and machine-actionable information for researchers. Also from a tool development perspective we suggest that a collection development and nomination tool be developed to enable rapid collection development decisions, possibly building on one or more current tools. Service providers such as Archive-It make some tools accessible to non-programmers but there is an opportunity for service providers to offer hosting and

support service to the many other web archiving tools that still require institutions to have local programming skills and/or IT support.

Focus on training and skills development

Because the collection and research of archived websites is a fairly new activity, it is not surprising that the scan revealed a skills gap that could be addressed by the development and provision of new training programs. Web archiving programs would benefit if institutions focused on training and skills development for existing staff new to web archiving, or that have an ancillary role in the collecting of websites, for example catalogers. We also see the need to train researchers with skills they need to analyze big data found in web archives, plus suggest the need to train content hosting sites on the importance of supporting libraries and archives in their efforts to archive their content. Finally we see an opportunity to conduct outreach and education to website developers to provide guidance on creating sites that can be more easily archived and described by web archiving practitioners.

Build local capacity

With the possible exception of the national libraries, the scan revealed that institutions have been slow to dedicate staff to web archiving. Our research suggests that in order to stay abreast of latest developments, best practices, and consequently fully engage in the community, institutions dedicate a full-time staff person to work in web archiving. With increasing need to support researcher use cases, we suggest exploring how institutions can augment the Archive-It service and provide their own local support to researchers. For service providers, we see an opportunity to increase their local capacity in the area of providing computing and software tools and infrastructure for those institutions lacking their own onsite infrastructure, and also for them to develop more offerings around the available tools.

Introduction

Websites are an integral part of contemporary publication and dissemination of information, and as more and more primary source material is published exclusively to the web, the capture and preservation of this ever-growing and ever-changing, dynamic content has become a necessity to support researcher access and institutional needs. Today's research libraries and archives recognize website archiving ("web archiving") as an essential component of their collecting practices, and various programs to archive portions of the web have been developed around the world, from within national archives to individual institutions.

To meet website acquisition goals, many institutions rely on the expertise of external web archiving services; others, with in-house staff, have developed their own web archiving services. Regardless of the approach, the rate at which textual, visual, and audio information is being produced and shared via the web, combined with the complexity and specialized skills and infrastructure needed for web archiving

processes today – from capture through quality assurance, description, and eventual discovery, to access and analysis by researchers – poses significant resource and technical challenges for all concerned.

Harvard Library sponsored an environmental scan to explore and document current web archiving programs (and institutions desiring a similar capacity) to **identify common concerns, needs, and expectations in the collection and provision of web archives to users; the provision and maintenance of web archiving infrastructure and services; and the use of web archives by researchers**. **The ultimate goal of the survey is to identify opportunities for future collaborative exploration**. Information provided by this scan will help Harvard Library and other US-based institutions prepare for a grant on collaborative web archiving, and will be shared internationally to inform research and development priorities.

This environmental scan is not the first investigation into these areas. Other surveys over recent years have provided valuable information about the landscape of web archiving activities, such as:

- The National Digital Stewardship Alliance (NDSA)'s Web Archiving in the United States. A 2013 Survey[1]
- NDSA Web Archiving Survey Report, 2012[2]
- North Carolina State University (NCSU) social media scan, 2015[3]
- A Survey on Web Archiving Initiatives, Portugal, 2011[4]
- Use of the New Zealand Web Archive[5]
- Researcher Engagement with Web Archives, 2010 (Dougherty, M)

While there may be overlapping areas covered within these reports and surveys, each examines a particular subtopic or geographical region in relation to web archiving practices. The NDSA surveys are focused on the USA; the NCSU scan is focused on other areas of social media (such as Twitter) and does not include use cases or details about individual institutions; the Portuguese study examined 42 global web archiving programs reporting only on the staffing and size (size in terabytes) of each institution's collections; and the Dougherty/JISC study focuses solely on the uses and needs of individual researchers. Other more narrowly-focused surveys, such as the IIPC working group surveys, address targeted informational needs.

[1] http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_USWebArchivingSurvey_2013.pdf
[2] http://www.digitalpreservation.gov/ndsa/working_groups/documents/ndsa_web_archiving_survey_report_2012.pdf
[3] https://www.lib.ncsu.edu/social-media-archives-toolkit/environment
[4] http://sobre.arquivo.pt/about-the-archive/publications-1/documents/a-survey-on-web-archiving-initiatives
5 http://natlib.govt.nz/librarians/reports-and-research/use-of-the-nz-web-archive

Methods Used

Over the course of five months in 2015, Truman Technologies conducted research to inform this environmental scan. Research methods included in-person and remote interviews and emails with web archiving practitioners, independent web research, and participation in working groups and relevant conferences. Interviews were semi-structured, with questions used to guide the conversation (Appendix D). A profile template was developed to capture a set of common information gathered during the interview process, or via email interaction when an interview was not possible. The 23 web archiving profiles are shown in Appendix B. Interviewees included a mix of archivists, library and museum curators, web scientists, researchers, and service providers from around the world. Institutions and people were selected from participants and members of the International Internet Preservation Consortium (IIPC), the Web Archiving Roundtable at the Society of American Archivists (SAA), the Internet Archive's Archive-It Partner Community, Ivy Plus institutions, Working with Internet archives for REsearch (Ruters/WIRE Group), and the Research infrastructure for the Study of Archived Web materials (RESAW).

The following table categorizes the institution or individual stakeholders consulted for the environmental scan:

Category	Count
National Library	10
University Library	9
Museum/Art Research Library	4
Individual Researcher/ User[6]	4
Service Provider	2

Table 1: Breakdown of stakeholders for environmental scan

Current Practice and Plans

In this section we look at how institutions are providing and maintaining their web archiving service and identify the main challenges and gaps. We also look at how the institutions are integrating their web archives within their own library collections and with collections of other institutions. We investigate the tools that have been developed to address various functional needs across the lifecycle of web archiving (from capture to access and analysis by researchers), identifying gaps and opportunities for new research

[6] Includes Columbia University librarian Pamela Graham who is conducting research into researcher use of web archives

and development. Through examining current practices and tools we hope to surface which areas are well covered by community and shared practice and areas where collaborative efforts could address gap areas.

Staffing

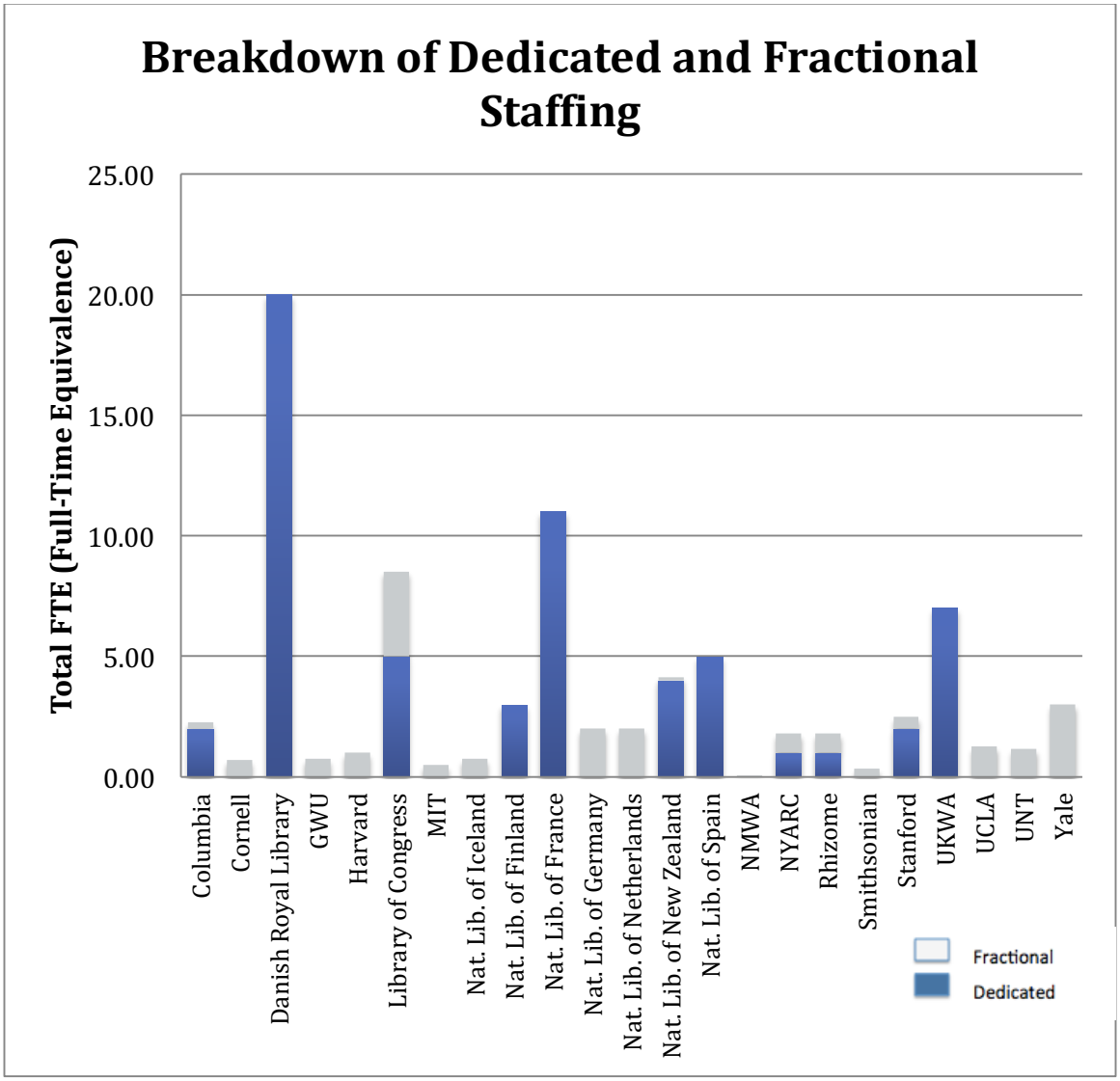

Figure 1: Each participating institution's number of dedicated web archive staff is shown in blue, with any additional part-time staff or the relevant portion of labor from non-dedicated staff shown in grey. The total FTE (Full-Time Equivalence) is the number of dedicated and fractional staff, where part-time staff counts as less than 1 FTE. For the purposes of making the graph, a part-time employee was counted as ½ FTE unless indicated otherwise in survey responses.

Staffing levels vary considerably across organizations, with 14 of the 23 institutions surveyed responding that they have one or fewer dedicated full-time employees for their web archiving projects. Of these 14, most are university, museum or art research libraries. Of the 10 national libraries surveyed, all but one (Iceland) has two or more full-time equivalent staff focused on web archiving, with six of them reporting four or more people and Denmark (the Danish Royal Library and State and Local Library combined)

reporting an impressive 20 people.

Some of the institutions that reported few dedicated resources – like the Smithsonian Institution Archives – rely on a combination of full-time people with other duties, part-time employees, seasonal interns, and volunteers to manage web projects. Others – like Yale University – rely on existing personnel to manage web archival content as it is relevant to their particular department, but lack any specialized employees whose sole responsibilities are for web content. Indeed, the survey data indicate it is quite common for web archiving responsibilities to be assumed by existing curatorial or department staff.

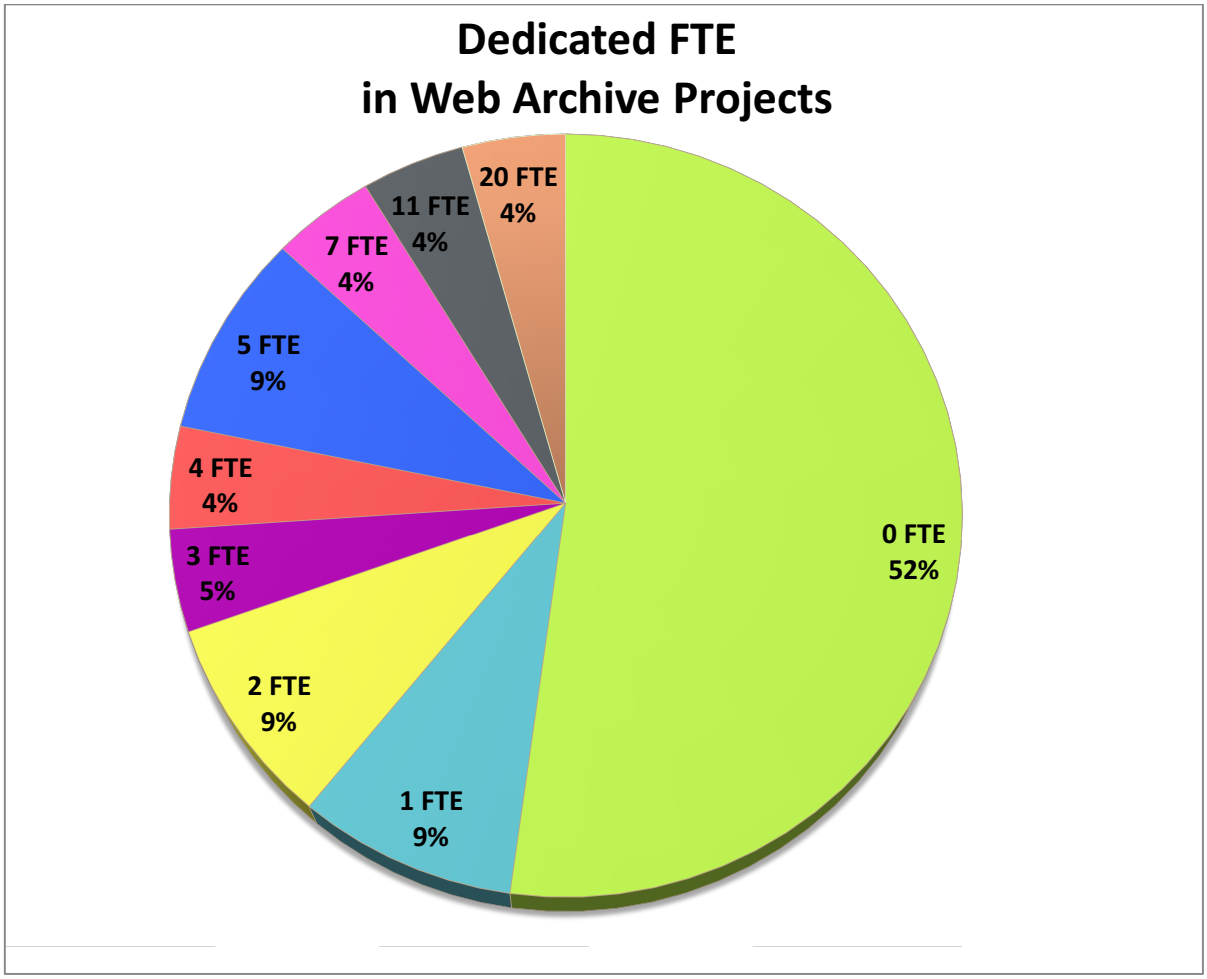

Figure 2: More than half of participants report having no dedicated full-time staff for their web archive projects. 39% have two or more dedicated staff with France and Denmark's programs having 11 and 20 full-time staff respectively.

The New York Art Resources Consortium (NYARC) has one full-time project manager and four paid interns who each work one day per week on Quality Assurance (QA) work, while staff at each of the NYARC libraries contributes to cataloging and collection development by nominating sites. Sumitra Duncan of The Frick Collection at NYARC points out that most programs don't have staff dedicated to program management or development, and that she is in a fortunate position whereby she spends most of her time on NYARC's web archiving program (Duncan).

Our survey's finding that almost two-thirds (61%) of web archiving institutions surveyed dedicate 1 or fewer employees is actually quite optimistic, as the National Digital Stewardship Alliance (NDSA)'s 2013 study found that "81% [of US organizations with web archiving initiatives] devote half or less of the equivalent of one full-time (FTE) staff person's time," and that the average value is a staggering 1/4 of a full-time employee's time (Bailey et al., Web Archiving in the United States 8).

However, adequate staffing is vital to a successful web archiving venture and will disproportionately affect institutions with less funding for web archiving, leaving them unable to participate fully. As Daniel Chudnov from George Washington University remarked:

> *If you don't have a dedicated full-time employee, how do you get involved in the community, and do these institutions [with limited staff] struggle to stay abreast of where the field is going and evolving?*
> *Chudnov*

Heather Slania from the National Museum of Women in the Arts (NMWA) points out:

> *The trick is making web archiving more accessible to those libraries, archives, and museums who don't have large IT departments or positions devoted to this. I'd want all of these institutions to be able to say 'yes I can be a part of this' or 'I can at least archive our own websites.'*
> *Slania*

Massachusetts Institute of Technology (MIT) currently has 0.5 staff focused on web archiving (1/2 FTE fellow) and has identified staffing as a priority area for funding. MIT is considering centralizing a web services manager position to help coordinate distributed activities (Appendix B: MIT Profile). Nicholas Taylor of Stanford University identified training and education as key components to any web archiving program. Stanford currently has 2.5 web archiving staff (1 full time services manager, 1 full time developer and fractions of curator and metadata staff time):

> *Web content collecting is new for many curators, let alone institutions. There's a vast amount of material that could be archived, and largely unsystematic methods to assess what's been or is being archived. In light of these challenges, training, education, and clarifying service processes and responsibilities are essential.*
> *Taylor, Appendix B: Stanford Profile*

Opportunity #1: Dedicate full-time staff to work in web archiving so that institutions can stay abreast of latest developments, best practices and fully engage in the web archiving community.

Opportunity #2: Conduct outreach, training and professional development for existing staff, particularly those working with more traditional collections, such as print, who are being asked to collect web archives.

Location in Organization

The commitment to saving and maintaining records for public and/or academic use is common to both libraries and archives. While both have more experience with physical and other digital materials, institutions of both kinds have made efforts to collect web data just as they would for any other set of material, as the historical value and research potential of web data continues to grow.

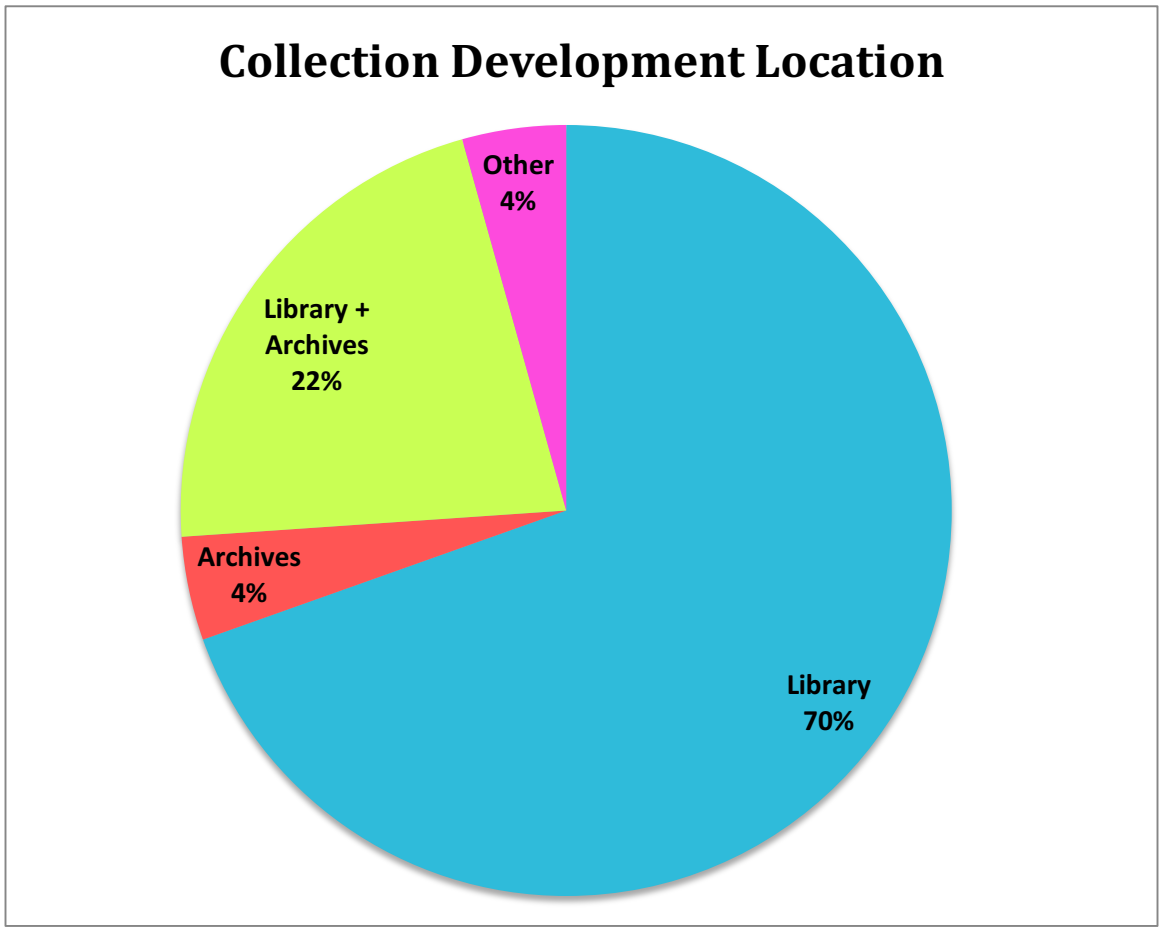

Figure 3: The majority of the 23 web archiving institutions surveyed report that their web archive collections are developed in libraries. This is unsurprising given the proportion of national and university libraries interviewed as compared to archives, museums, and other types of organizations.

But libraries and archives often have different policies and these policies (such as preservation priorities and methods, as well as security measures) affect how the data is managed once it is collected. Anna Perricci of Columbia University Libraries illustrates the division between the two groups:

> *The location of the web archiving work makes a difference. I'm most closely affiliated with our program, which is firmly planted in the Libraries. Web archiving staff is based in the Original and Special Materials Cataloging department (OSMC), which is organizationally distinct from*

> *the archives (though OSMC does advise the archives on metadata issues more broadly). We do web collecting on behalf of the archives, for example materials about Columbia itself and websites belonging to human rights organization whose papers we hold. With Ivy Plus we are working with subject librarians, not curators associated with archives. With my involvement with SAA, I do have some sense of how archivists are collecting and using web archives. That said, I think the most advanced programs in the US right now, including the Library of Congress, the University of North Texas, Stanford, NYARC, and Columbia, are decidedly based in the library.*
> *Perricci, personal interview*

When asked what she views as the distinctions between how things work in archives versus how libraries are approaching things, Perricci suggests the following:

> *A question to consider: are web archives a corpus in their own right, are they part of a library collection, or are they a part of a larger set of archival records (i.e. an author's personal papers)? How web archives fit into a larger collection strategy can vary between libraries and archives, but the distinctions between how web archiving is done in each venue is less clear than it is with either physical or other born digital materials. For the most part, librarians use Archive-It and archivists use Archive-It, but the goals for web collecting can vary based on the mission and needs of each institution or unit.*
>
> *If one handles an archived website as a record what are the implications in terms of how it is described, arranged, appraised, stored, and included in planning for long-term preservation? When we have better use cases archives and libraries might alter their collecting patterns, and more clearly define how description and access should be adapted to meet the needs of specific stakeholders. I'd defer to an archivist on how local workflows for web collecting are emerging at her or his institution but on a higher level I think it is productive to at least think about an archival approach versus a bibliographic approach.*
> *Perricci, Message to Truman Technologies, LLC.*

Not every web archiving initiative is housed under either a library or an archive. Researchers who want to study web-based material sometimes create their own collections. And, as Pamela Graham of Columbia University pointed out, big data web analysis has not been a typical library concern in the past (Graham). She gave the example of a faculty member who plans to use their Human Rights web archive collection in a technology and communication class, to explore networking or language use among groups of people. This analysis requires creating subsets of the data to extract URL link data and language-related metadata. These subsets, or derivatives, of the data are being made available by Archive-It Researcher Services[7] and

[7] https://archive-it.org/blog/post/launching-archive-it-research-services-part-1/

Graham wonders how a library should best work with end users to make these data more accessible to them and to provide support for these users.

Rhizome is the one institution surveyed that does not house its web archiving program under the library or archives. Rhizome, a not-for-profit born-digital art institution, supports and provides a platform for new media art, including games, software and interdisciplinary projects with online elements ("Rhizome organization"). Rhizome has been building a high-fidelity archive of this web-based art material since 1998 and is very focused on the preservation and future rendering of content, using emulation techniques and exploring new digital preservation methods.

We would be remiss if we overlooked web archiving programs based in for-profit corporations, where the service location is likely to be in records management, legal compliance, or corporate branding. Although not the focus of this environmental scan, numerous for-profit, multi-national companies are web archiving due to legal eDiscovery, records management and compliance requirements, or for marketing support. For these use cases, for-fee, fully-managed services such as Hanzo Archives[8] are sometimes deployed. Also of interest is Washington State's designation of a particular vendor – PageFreezer – to do their web archiving for state agencies.[9]

While the institutions we surveyed were mostly libraries overseeing web archiving programs, it's important to acknowledge the ability and desire for other parties to compile specific collections to meet their own individual or corporate/institutional needs.

Opportunity #3: Increase communication and collaboration across types of collectors since they might collect in different areas or for different reasons. See also Memberships and Collaborations section.

Memberships and Collaborations

The environmental scan participants were asked to identify the collaborations and organizations of which they are members. Given that the institutions were originally selected for the scan from these particular groups, it is hardly surprising that the five most prominent groups and collaborations for web archiving identified by interviewees are:

1. International Internet Preservation Consortium (IIPC)[10]
    - The 49 IIPC members are organizations from over 45 countries, including national, university, and regional libraries and archives. Membership fees start at 2,000 Euros. Its annual conference tends to be hosted in the United States every third year.
2. Society of American Archivists (SAA) Web Archiving Roundtable[11]
    - The SAA is focused on the archivist community. Its web archiving email discussion group and community includes about 924 individuals, who have an opportunity to meet

[8] http://www.hanzoarchives.com/
[9] https://www.pagefreezer.com/government/washington-state-public-record-laws-for-website-socialmedia/
[10] http://www.netpreserve.org/
[11] http://www2.archivists.org/groups/web-archiving-roundtable

during the SAA annual meeting. Membership in the roundtable and email distribution list is free to both paying SAA members and non members.

3. Internet Archive Archive-It Partner Community[12]
    - With just over 400 partner organizations in 48 U.S. states and 16 countries worldwide, the Archive-It community is very active. Partners meet regularly at the SAA annual meeting and throughout the year at regional events. Membership comes from being a subscriber to the Archive-It service.
4. Ivy Plus[13]
    - Ivy Plus is a small group of the Ivy League plus additional US-based universities. Ivy Plus collaborative web archive collection development includes the Contemporary Composers Web Archive (CCWA) and the Collaborative Architecture, Urbanism and Sustainability Web Archive (CAUSEWAY) pilot web collections. The group is currently exploring additional collaborative web archiving activities.
5. Art Libraries Society of North America (ARLIS/NA)[14]
    - The ARLIS/NA annual meeting agenda now includes web archiving birds of a feather, plus presentations about web archiving relevant to art libraries.

And two collaborations of relevance to researcher use of web archives cited by the researchers surveyed:

6. Working with Internet archives for REsearch (Rutgers/WIRE Group)
    - Rutgers WIRE/Group began as a workshop hosted by a research team of scholars from Rutgers University, Northeastern University, and the Internet Archive. Its website for the 2014 workshop identifies 24 participants from around the globe[15]. The group is planning two hackathons in 2016 and uses its online site, archivehub.rutgers.edu, as a data-hosting site.
7. The REsearch infrastructure for the Study of Archived Web materials (RESAW network)
    - The RESAW network consists of major European national libraries as well as leading research communities studying web archives. At the time of this writing, the RESAW network had 47 members, all but one (Loyola University Chicago) from outside of the USA.[16]

The Library of Congress and Smithsonian Institute both belong to the federal web archiving group which meets monthly but does not have formal governance at this point (Neubert, email).

[12] https://www.archive-it.org/
[13] https://library.columbia.edu/content/dam/librarywebsecure/behind_the_scenes/web_resource_collection/CUWARCpres_Perrricci_2015-corrected.pdf
[14] https://www.arlisna.org/
[15] https://wp.comminfo.rutgers.edu/nsfia/
[16] http://resaw.eu/participants/

Organization /Community	Geographic Coverage	Institutional or Individual	Cost to Join	Main Activities	Main Focus
IIPC	International, predominant European	Institutional	2,000 Euro and up	• Standards dev. • Tool maintenance • Best practices	Mainly large institutions and national archives
Archive-It Partners	International, mostly USA	Institutional /Dept.	Subscription. Partner events are open to all	• Crawling service • Researcher services	Institutions of all sizes and types
SAA Web Archiving RT	USA	Individual	None	• Discussion • Best practices	Individual Archivists
Ivy Plus	USA (small)	Individual	None (due to Mellon support)	• Collaborative collections	Selected large universities
ARLIS/NA	North America	Individual	$50 and up	• Discussion • Best practices	Art museums and libraries
RESAW	International (small)	Individual	None	• Tool development • Best practices	Researchers, computer scientists
Rutgers/ WIRE Group	International (small)	Individual	None	• Discussion	Researchers, computer scientists

Table 2: Summary of organization or community membership

When looking at these collaborations to study membership overlap as shown in the bubble diagram, we see that no overarching, single collaboration covers all the types of practitioners involved in web archiving.

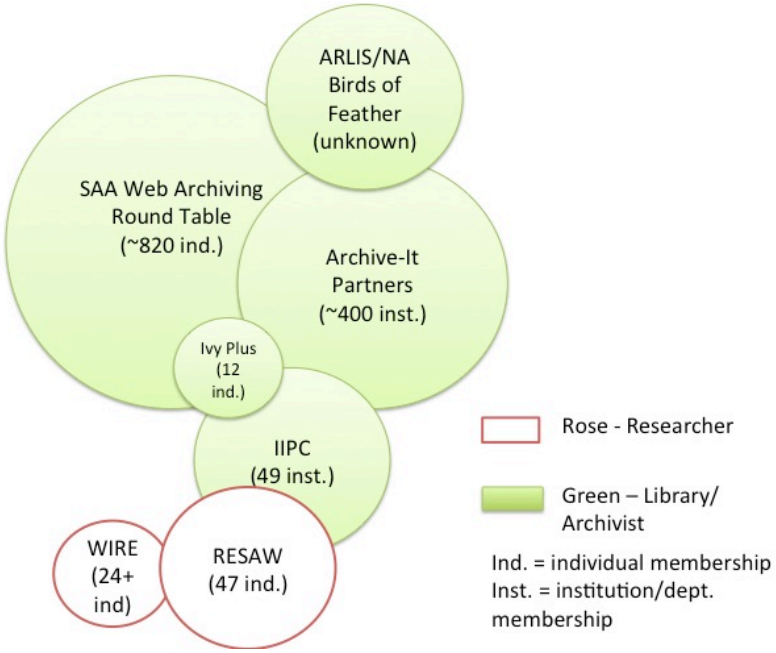

Figure 4: Showing how membership in organizations and communities overlap. Sizes and membership type shown in parenthesis.

By looking at the membership of collaborations and affiliations that serve as meeting places for web archivists or researchers, we see that, with the exception of RESAW/IIPC, there is very little overlap between researcher collaborations and those of librarians and archivists. The overlap of RESAW and IIPC is due to the many European national libraries who are members of both. Indeed, RESAW has a decidedly non-USA focus with just one listed participant located in the USA (Loyola University, Chicago). This small degree of overlap suggests that while researchers are discussing web archiving among themselves, there might not be sufficient dialogue with those whose role it is to steward and manage web archives. Absent communication and interaction, collections will be developed without input from the research user community.

Pamela Graham of Columbia University explained how their Human Rights web archive collection is being used in class activities and by researchers. She noted that big data analysis of web archives doesn't represent everyone and that it is not a typical library issue, suggesting we will need to "foster better networks/communities around this" (Graham).

Facilitating discussion, community and outreach among the practitioners today should result in more productive and mainstream use of web archives moving forward. One idea put forth (Graham) was to emulate the UK British Library's bursary award program[17] that funded researchers to use web archives and meet monthly to discuss findings and experiences. Such a program could get researchers comfortable with using web archives, foster activity and feedback in a project that others could learn from and provide librarians and archivists with greater insights about how researchers want to work with the assets they steward.

Opportunity #4: A funded collaboration program (bursary award, for example) to support researcher use of web archives by gathering feedback on requirements and impediments to the use of web archives.

Opportunity #5: Leverage the membership overlap between RESAW and European IIPC membership to facilitate formal researcher/librarian/archivist collaboration projects. Such collaborations' goals might include to:
- facilitate understanding of how researchers want to use web-based content in their research
- determine how to provide web archives to researchers, including APIs and other means of access
- determine description and metadata that would help validate research samples
- determine how to provide support services to researchers

Collection Development

Perhaps the most important information gathered in the environmental scan relates to the web content collected and archived by the various stakeholders' programs.

[17] http://buddah.projects.history.ac.uk/news/bursaries/

Collection development focus

Collection development for institutions surveyed falls primarily into four areas:

1. National domain crawls (e.g. .dk, .de, .is. .uk, .nz)
2. Institutional website crawls (e.g. for university or museum websites and their social media presence)
3. Thematic or topical crawls to enhance existing collection development or special collections (e.g. Human Rights[18] and 2012 Olympic and Paralympic Games[19]), or to capture current events unfolding via web-based communication (e.g. Hurricane Katrina[20])
4. Researcher-led, purpose-built collections created within the scope of a research question

The collecting focus for national and institutional domain crawls is by and large relatively unambiguous when defining the scope of "what should be crawled" where the scope is the country's domain extension (.nz or .fr) or the institution's web URL (e.g. institution.edu, companyname.com). However, the National Library of New Zealand spoke of copyright implications with newer domains (e.g. kiwi) that might not necessarily be hosted in New Zealand and where ambiguity makes it unclear whose legal and geographic responsibility it is to collect them (Knight, Steve).

Collections based on specific themes or topics require more scoping effort to determine which sites are to be collected (i.e. the list of starting URLs, collectively referred to as the seed list), as well as the depth and frequency of site captures and media types to capture.

Researchers' use of web-based material and their need for specialized datasets to support their research have resulted in several web archives built specifically for data mining (e.g. the Stanford University WebBase project out of the computer science department, which no longer crawls for new sites but provides a large corpus of data for researcher analysis; and the webarchives.ca portal to the Archive-It collection of the University of Toronto, created by Dr. Ian Mulligan which encompasses websites from 50 political parties and political interest groups over a 10-year span).

Collaborative collection development and overlap

While there are some examples of large-scale collaborations to create thematic collections, several interviewees voiced frustration at the lack of coordination and communication around collection development, and the resulting fragmentation of collections or potential duplication of effort. Indeed, defining collection strategies for collections typically is very manual and rarely communicated outside of the collection or collecting institution, opening the door for collection overlap or gaps. As Jason Kovari of Cornell University, Stephen Abrams of the California Digital Library (CDL), and Kari Smith of MIT put it:

> *How are we facilitating understanding of who's collecting what? Is there a shared knowledge base? There is concern among many when discussing web archiving that so much content*

[18] http://hrwa.cul.columbia.edu/
[19] http://www.webarchive.org.uk/ukwa/target/720896/source/search/
[20] https://archive-it.org/collections/174

exists – how can institutions capture enough while limiting scope to a manageable selection? And if others are focusing on a similar area, where is our place? How we know whether we are duplicating effort?
Kovari

What does it mean to do collaborative collection development? The strain that web archiving puts on any one institution, however large it is, is enormous. We can't quantify it quite yet, but we strongly suspect there is probably a great amount of duplicate effort because no one knows what everyone else is doing. So is there a way we could put more transparency into the collection development activity and stretch our resources that way?
Abrams

What is the redundancy of web archives, especially if they're topical? How will we even know when and what each other is collecting? But we know the crawls are going to be different, the times are going to be different, and so even if the content is extensively the same content, it's unlikely that it actually will be the same information or data.
Smith

As Smith points out, even with overlapping collection seed lists, the time of capture and the pages/depth of each crawl will yield different information. This overlap is something that Daniel Chudnov of George Washington (GW) Libraries and Gina Jones of Library of Congress view as acceptable:

We need to facilitate broader capture. It already feels like we lost the first 10 years of the web so we know what lost looks like now…. Having as many [web archiving] partners as possible in the mix is okay, even if it looks like they all may be competing.
Chudnov

People crawl for different reasons and they also crawl at different frequencies and depths, so overlap is not necessarily a bad thing.
Jones

While all the interviewees did not necessarily view overlap and duplication as a bad thing, their comments suggest a need for greater transparency about what each web archiving program is collecting. Michael Neubert of Library of Congress said in an email, what is needed is more understanding for institutions doing crawls about the completeness of crawls they observe at other institutions along with more information shared about who is crawling what. And, as Andrea Goethals of Harvard Library pointed out in an email, it's also the case that the curatorial decisions made during collection aren't exposed, or possibly even documented. If these curatorial decisions were exposed, collecting institutions could make informed decisions about whether or not to crawl sites already in others' collections.

Heather Slania of the National Museum of Women in the Arts (NMWA) noted that few art libraries are doing web archiving today and that there would be a great deal of duplication if they did. She would like to see a collaborative collection development effort that combines those with robust budgets and staffing and those with smaller resources (Appendix B: NMWA Profile).

Coordinating collections

Currently there is no easy way for a web archiving program to know what is, or is not, collected by other programs. Several interviewees (Duncan, Knight, Thurman) spoke specifically to the challenges of reading room-only web archive collections. Reading room-only access is typically required for institutions with holdings that may not, for legal deposit and copyright reasons, be shared outside of their institution. For these collections it is particularly important for the collecting institution, where legally possible, to minimally share collection-level holdings information with the broader community.

> *A better tool for viewing who is collecting content in different subject areas would be useful, especially with the challenges of those European libraries who cannot make public what they have in their collection. I'd like to see a way to have a better sense collectively of what is covered and I feel this effort would be well served by IIPC spearheading this and having people voluntarily provide this information in a centralized place. Almost as if people sign up for collection stewardship of certain areas with agreed upon levels, so it is a registry in some sense. Something similar with Archive-It partners would also be useful.*
> *Duncan, Sumitra*

During the IIPC preservation working group "find a room" meeting at iPRES 2015 in Chapel Hill, the attendees floated the idea of the IIPC overseeing a collection-level registry of the IIPC member institutions, since so many have collections that are hidden from public access.

Opportunity #6: Institutional web archiving programs become transparent about holdings, indicating what material each has, terms of use, preservation commitment, plus curatorial decisions made for each capture.

Opportunity #7: Develop a collection development tool (e.g. registry or directory) to expose holdings information to researchers and other collecting institutions even if the content is viewable only in on-site reading rooms.

Facilitating broader capture

Collaborative collection development is essential where large-scale capturing poses too much of a challenge for any single group to handle, and for capturing historically significant events unfolding in real-time, such as major international events, disasters and volatile political situations where websites are at risk of being taken down. Nonetheless, regardless of efforts to capture as much as possible, there remain portions of the web that currently cannot be archived due to technical, legal and/or business reasons. This section briefly explores these technical and legal hindrances.

Technical challenges to broader capture

Many of the technical challenges of web capture (and later display) are due to dynamic content – meaning that some or all of the web page is generated at run-time by a program executing either on the client or on the server. New advances in crawler technologies (such as Umbra,[21] PhantomJS,[22] Webrecorder.io[23] and the new "Brozzler"[24] from Internet Archive are helping capture this content but the ever-evolving nature of the web means that the live Web and Internet technology will always be ahead of the capture tools. Lynda Schmitz Fuhrig of the Smithsonian stressed the importance of Internet Archive and its Archive-It service to keep working at improving capture for dynamic content and to be open to other tools that they might be able to roll in as well. The Tools section in this report identifies a seeming abundance of capture tools, but there are even more that are not included in this list, bringing the quantity of capture tools to well over double those of most other areas of the web archiving life cycle. This reveals the large amount of effort that has been spent (and that will be required on an ongoing basis as web technology continues to progress) to try to capture content originating on the web.

Kari Smith of MIT suggests that broader capture of these rich and technically-challenging sites might benefit from educating website creators:

> *What interests me is how can we talk to people about good practice in [website] creation. How do we help people create really rich websites, web applications, web-based art, whatever it is, and still be able to capture and archive this material into the future? This includes things like embedding metadata in headers, which is kind of old school but people aren't always doing this anymore. How do we get people to embed titles, etc., so we can use that embedded metadata going forward?*
> *Smith*

MIT is testing tools that help assess the readiness of a website for archiving (such as Archiveready.com[25] and Wappalyzer[26] Chrome extension).

Opportunity #8: Conduct outreach and education to website developers to provide guidance on creating sites that can be more easily archived and described by web archiving practitioners.

Legal or business challenges to broader capture

From a legal perspective, site or technology owners may add restrictions that prevent capture of their content. Here is an area that the community of web archivists can possibly influence if they can bring

[21] https://webarchive.jira.com/wiki/display/ARIH/Introduction+to+Umbra
[22] http://phantomjs.org/
[23] https://webrecorder.io/
[24] The Brozzler tool is not yet public as it is still in development, but was described by Jefferson Bailey at University of Michigan's Web Archives conference (Bailey, Jefferson. Proc. of Web Archives 2015)
[25] http://archiveready.com/
[26] https://wappalyzer.com/

enough pressure to bear on these organizations. One interviewee (Steve Knight of the National Library of New Zealand) suggested that the community should put pressure on the news- or social-media and technology companies who have made it difficult technically or legally to capture their sites. He suggests approaching the International Federation of Library Associations (IFLA) to spearhead this effort, saying:

> *Possibly have the IIPC talk to them [IFLA] … any big tech companies you approach at the right level would probably help us – but we don't have our leg in the door. The narrative needs to be much stronger about things being lost. People think YouTube will be there forever, while providing and exit strategy is the job of archives and libraries. This needs to be part of the narrative with the tech world.*
> *Knight*

Opportunity #9: IIPC, or similar large international organization, attempts to educate and influence tech company content hosting sites (e.g. Google/YouTube) on the importance of supporting libraries and archives in their efforts to archive their content (even if the content cannot be made immediately available to researchers).

Discovery

As the amount of content being created for the Internet increases, and is captured by web archiving programs, so does the challenge of discovery and delivery in these key areas:

- Within an institution – discovery and delivery across web archive collections
- Within an institution – discovery and delivery across *both* web archive *and* other types of collections
- Across multiple institutions – discovery and delivery across web archive and other types of collections

Within an institution, discovery and delivery across web archive collections

Institutions that maintain collections at the Archive-It service can have their collections searched using both full-text search and metadata search (at the collection, seed, and document level). The results include any metadata that institutions have added for the Title, Description, Collection, Partner, Publisher, Creator, Subject, and Seed URL fields. Searches can be conducted from within the Archive-It.org site or from landing pages and search boxes from within an institution's own website.

At the Library of Congress, websites are cataloged using preliminary keyword, title, and subject extracted to create Metadata Object Description Schema (MODS) records.[27] A Lucene-backed discovery interface searches across the MODS records both within and across their archived web collections (Grotke, Abigail).

[27] http://www.loc.gov/webarchiving/cataloging.html

Although these discovery and access mechanisms may appear sufficient, actual use cases, like those cited by Skip Kendall of Harvard Library, illustrate the limitations of finding exact web pages using today's search capabilities:

> *We need to continually refine how to deliver searches and other methods of finding things in web archives since it's so hard to do. For example, how can I know what harvest will be responsive to my search if I'm looking for a certain faculty member that I know was here from a certain date range, but I'm not sure which harvest his information will be in? Search is a big, complicated issue due to the scale of the content and the infrastructure needed to support it.*
> *Kendall*

Susanne Belovari of University of Illinois echoes this frustration. She explains many web archives offer only URL-based access and others have inconsistent metadata against which to search:

> *The largest web archives, the Wayback, as well as several other smaller IIPC members can only offer search access via urls which presupposes that future researchers will know these urls. But how would they? In contrast to uncountable directory style inventories offering multiple and easy access points to analog materials (think of phone directories as a basic example), just a handful of very limited Internet directories exist, often only available for local reference and not even shared through interlibrary loan. The many contemporary online directories such as 'White Pages' are not yet captured and time stamped as future access points.*
> *Belovari*

Even those archives that also offer keyword search can be a problem:

> *A lack of good metadata and authority control is a critical issue for researchers who reasonably assume that national web archives use good authority control for particular government offices to give one example. In reality, this is frequently not yet the case and unless researchers know and search under four or five different names for a particular office, they will not find relevant sites.*
> *Belovari*

Within an institution, discovery and delivery across web archive and other types of collections

Given the growing diversity of formats in library and archive collections, including content originating on the web, it is only natural for users to expect that these resources be discoverable without having to use format-specific catalogs or portals. As Stephen Abrams from CDL explains:

> *If you're interested in a topic you look at the library catalog, then you go look at the digital library portal, and now you have the Archive-It portal. Search should be search, and it should*

> *be giving you access to materials independent of their form, or the modality of their acquisition, and so forth.*
> *Abrams*

Institutions are adding bibliographic records to catalogs and finding aids to help users discover web archive collection material. At the Library of Congress, in addition to a MODS records for each seed-level site, a collection-level MARC record is available in the Library of Congress Online Catalog so that the collection can be found along with other library materials. The New York Art Resources Consortium (NYARC) has taken this a step further and has integrated its web archives with the Ex Libris Primo discovery service to simultaneously serve archived websites along with other resources. This may be the first instance of integration of Archive-It collections using the OpenSearch API in a commercial discovery layer. Sumitra Duncan explained:

> *We worked closely with Ex Libris so that our Archive-It collections are pulled directly into their system for discovery. As far as we know, it's the first implementation of this and is based on full text indexing done by Archive-It. When a search is done in Primo, results from Archive-It are displayed dynamically in the Primo interface. The results appear just as they do in the native Archive-It interface; Primo does not rank the results independently, It is essentially a view of the Archive-It indexing in Primo, but we can control how many results display in that interface.*
> *Duncan*

Across multiple institutions, discovery and delivery across both web archive and other collections

Discovery and delivery of web archives across institutions remains hit or miss, exacerbated by many national web archiving programs (and some smaller institutions' programs) that have legal deposit rulings or other copyright or privacy laws that affect the institution's ability to make them publicly accessible. The collections affected are only accessible from within a "deposit library" reading room (as is the case with many European national web archiving programs) or from within the institution's domain (as is the case with the UNT.edu web archives which require the user to have the UNT.edu email address to login and access them).

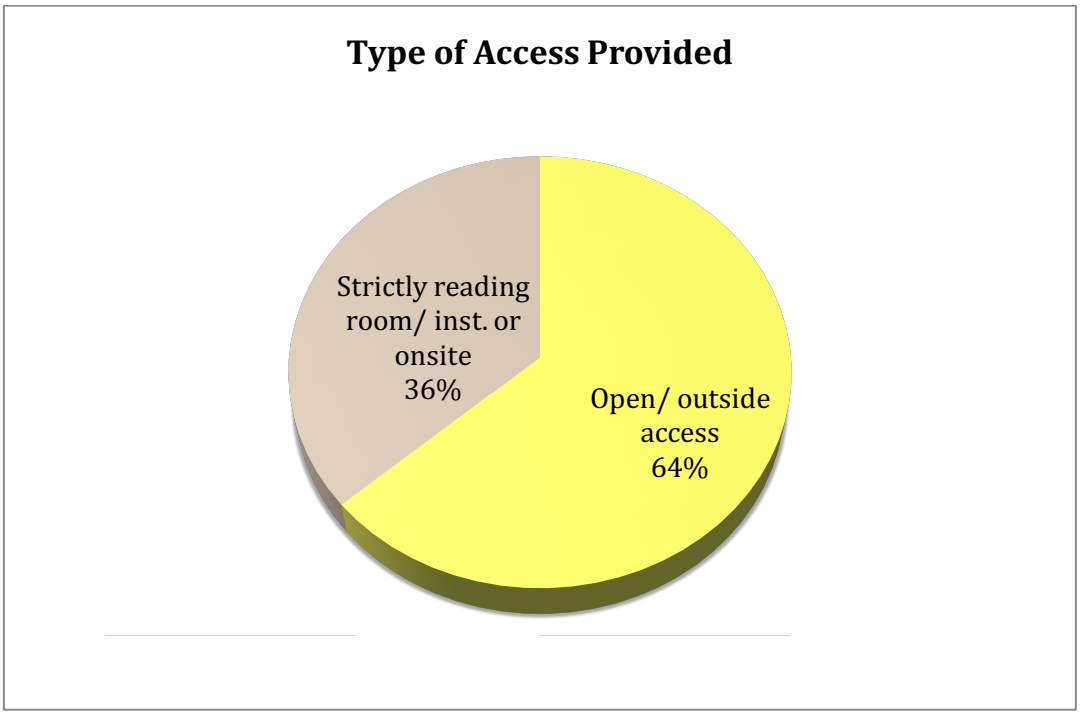

Figure 5: Showing user access provided to web archiving collections. Over one third interviewed had restricted access citing copyright legal reasons.

Absent any communication to explain what collections and sites are being archived by these institutions, it is impossible to search across them, or even to know that a trip to their institution will allow you to view the content you seek:

> *Lots of institutions have resources only available in reading rooms so identifying and getting permission is key. If someone else has a good copy it would be good to link to it. Is there value in them archiving it too, including responsibility for long-term preservation? We're all busy doing our own thing and we need more communication. For example, I am looking at a French national library website that I know has been archived but I cannot find the archived version. Can they at least provide basic metadata that would not be copyrightable to allow us to find out what is in there?*
> *Knight, Steve*

As for those sites not impacted by legal restrictions, they still remain difficult to search across. Pamela Graham of Columbia University envisions the perfect environment where there are tools that knit together certain collections and make it easier to use and discover web archives (Graham). Alex Thurman of Columbia University Libraries pointed out that the archiveit.org access point to its (approximately 3,500) public partner collections is currently the best place for users to perform full-text searches against a large accumulation of web archive collections, and more could be done to highlight archiveit.org as a major research resource. Columbia has tried to leverage this resource with the experimental search extension feature of its locally built access portal to its Human Rights Web Archive collection.:

> *With Search Extension you do a search and get hits back, but if you want to extend the search to other external reference sources you can do this. You can follow a search into the entire Archive-It collection. It's a natural use case – a user has run a search against their stuff and wants to expand to ALL Archive-It partners, following the principle that users don't care where it came from.*
> *Thurman*

But for institutional collections not pooled together, Thurman prefers the Memento Framework,[28] which offers a way to discover archived websites housed at institutions across the globe – leaving the content where it is.

> *I think using Memento is certainly a more cost effective way to "unite" existing far-flung web archives than to literally consolidate all the web archive WARC files in one place for access.*
> *Thurman*

However, as Andrea Goethals of Harvard Library pointed out in an email, Memento is still just a URL-based search tool and does not support metadata searches with facets. This means that users would need to know the exact URLs in advance in order to find the archived web pages. And as Thurman and Duncan both pointed out, Memento-based search interfaces must either be configured to query public collections only (thus excluding many reading-room only collections) or to show all known capture dates, even those whose archived content can't be directly accessed online, at the risk of frustrating users.

The Library of Congress expressed interested in tools or collaborations that will increase awareness and use of its web archiving collections, such as broader adoption of Memento Time Travel,[29] or a registry site or central hub (Appendix B: LC Profile). The idea of a registry (which was covered in the preceding Collection Development section) came up in several conversations. And although "not another registry please!" was voiced by one (un-named) interviewee, with the right amount of buy-in and support – and upkeep – it could help solve this problem.

Opportunity #10: Investigate Memento further, for example conduct user studies, to see if more web archiving institutions should adopt it as part of their discovery infrastructure.

Building on the need to support discovery across collections of different media types within an organization, the end goal of libraries should be to facilitate discovery of web archives in catalog and finding aids outside their own institution, similar to what NYARC has accomplished.

[28] http://www.mementoweb.org/guide/quick-intro/
[29] http://timetravel.mementoweb.org/

Tools

Several institutions advocated for the creation of software tools to assist with aspects of web archiving, which prompted an exploration of the tools currently available for various aspects of the web archiving lifecycle. For this, the lifecycle of traditional library practice for collection development was used (pre-accession, accession, process, preserve, access) and then subdivided into more granular areas, some of them specific to web archiving. This breakdown plus the entire list of tools is shown in Appendix C. The tools list was compiled from the following sources of information:

- The tools and software identified on the IIPC website[30]
- Presentations given during the Curator Tools Fair at the IIPC GA 2014,[31] the IIPC GA 2015,[32] and the Columbia University Web Archiving Collaboration: New Tools and Models 2015[33]
- Interviews and profile data from each of the participants (Appendix B)
- Independent web research and conference attendance

Because the list of tools attempts to be as inclusive and thorough as possible, the tools range from mature, robust applications (such as the Web Curator Tool, Heritrix, and Solr) to beta-release tools, scripts and browser extensions that are particularly prominent in the area of capture and analysis tools. We stopped searching for additional tools at 77 in count, but acknowledge there are others that have not made it into the list (which will by necessity need to be a living document if it is to be useful moving forward).

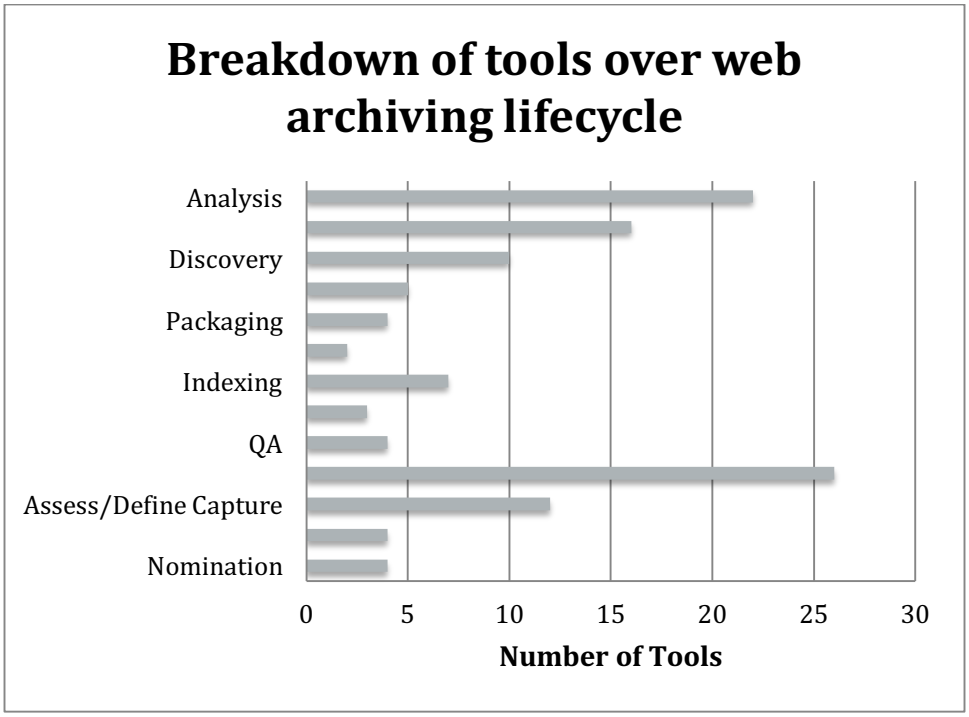

Figure 6: Showing the distribution of tools along the web archiving life cycle from nomination to analysis. Some tools do more than one function and tool-count by function represents this.

[30] http://netpreserve.org/web-archiving/tools-and-software
[31] http://netpreserve.org/general-assembly/2014/presentations
[32] http://netpreserve.org/general-assembly/ga2015-schedule
[33] https://www.youtube.com/playlist?list=PLf1Dab4lwQhBpFRB1dpUnKLglmM2iScjl

The tool categorization by function along the web archiving lifecycle illustrates that tools seemingly serve some areas of functionality very well – such as capture and analysis. However, many of these capture and analysis tools are very specific to narrow types of media (e.g. capturing tweets) or support for particular types of analysis (e.g. link analysis). And, depending on the sites or research of interest, multiple different types of capture and analysis tools may be needed. A further problem is that these tools were not designed to be used together or in modular ways – for example there is not an API framework that could allow these tools to be used together more easily and swapped out as needs or technologies change. As one researcher who expressed the need for better tools put it:

> *Many tools are being developed for very specific collections or use cases which is cost prohibitive to tailor for someone else's use. We need broader tools that a large population can use.*
> *Weber, Matthew*

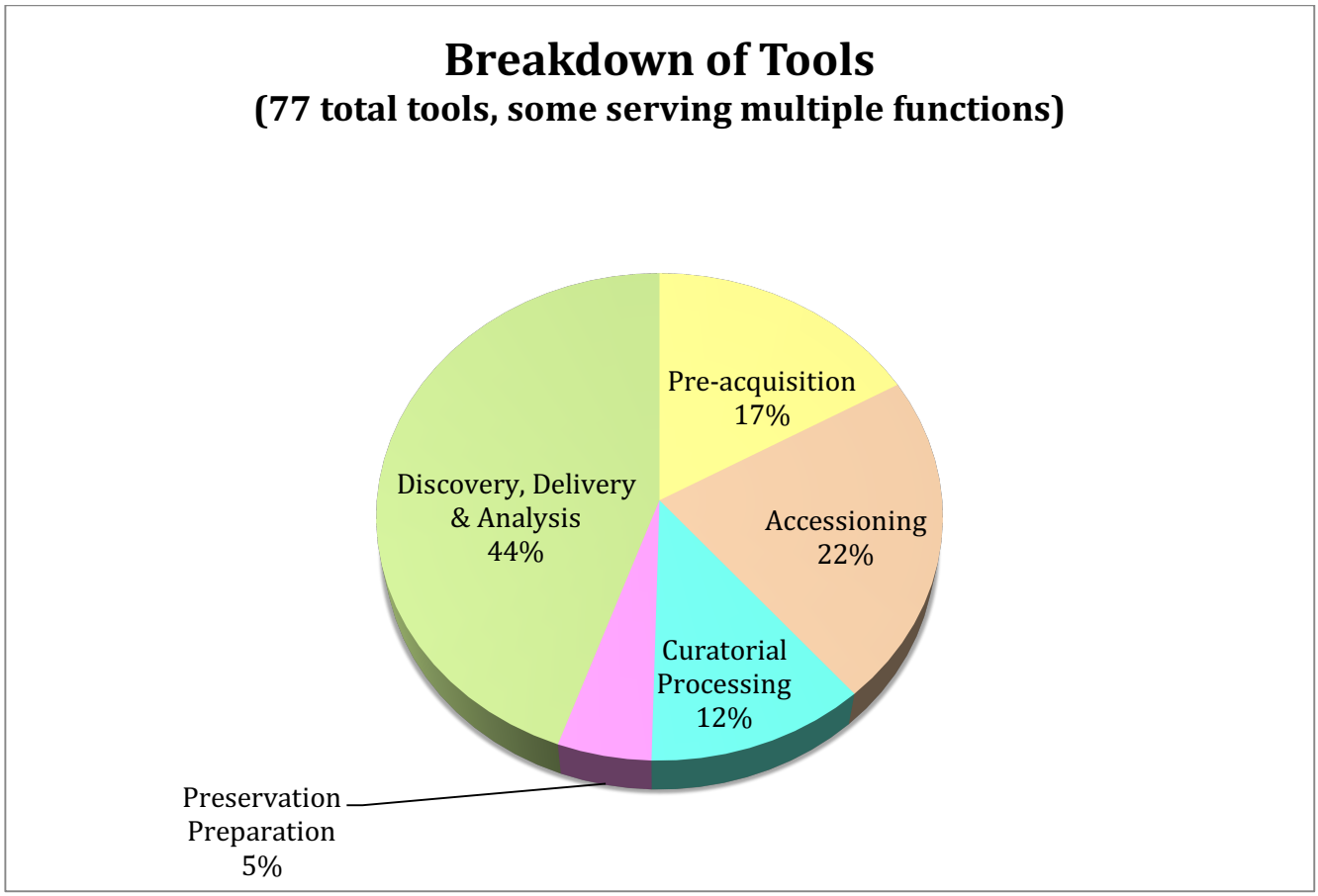

Figure 7: Showing the breakdown of tools by broader categories.

For institutions and users who want to start web archiving, there seems to be a lack of any off-the-shelf tool to help start initial collection definition and collaboration: meaning, deciding what to collect, allowing others to nominate sites, dividing up the workload among users. And before a person starts collecting, it would be useful to be able to easily find out what's already been collected, and by whom. While there are

four nomination tools identified in the tools list (Appendix C) they are currently very specific to the workflow of the tool they are embedded within, or the institution that developed and is using them, making them currently unsuitable for broader-scale out-of-the-box adoption. That said, the UNT Nomination Tool is targeted for open source distribution (Phillips), and the Library of Congress' Digiboard, which was also targeted for open source, has been transferred to a different part of the library so its fate is as yet unknown (Groetke).

Two stakeholders interviewed (Perricci & Knight) spoke directly to the need for gathering user requirements – an important step before any development effort, but also one that often gets short-changed or overlooked. As Anna Perricci put it:

> *I really think we do need more emphasis on software tools that are going to do a better job meeting researchers' variety of needs. I think the Internet Archive has made amazing strides, for example Archive-It is a fantastic tool and associated user services are being rolled out in addition to other researcher support given, but there's plenty of room for growth…. It's known that better software tools are needed and more tools are needed. We will never find a technical solution that will alleviate all the problems surrounding web archives but I think emphasis on software tool development should be of a very high priority. Also I think communicating with users and stakeholders to make sure they're getting what they need in the first place is really important. That said, there's a chicken and egg situation in terms of use of web archives and there are competing demands in an environment of very finite resources.*
> *Perricci*

Jefferson Baily, from Archive-It explained in an email:

> *Archive-It feature/technical development is first and foremost driven by the needs of the Archive-It community, something I noted on the [community] calls.*
> *Bailey*

Interview and profile data indicate that capture tools and application suites developed during the early years of web archiving are outdated and in need of refresh or expansion. These tools include Heritrix (the de facto, most prominent crawler) and other tools developed around it, such as the Web Curator Tool (WCT),[34] and the NetArchive Suite.[35] As Andrew Jackson from the UK Web Archive (UKWA) at the British Library who wrote the profile submission for the UKWA pointed out, these older tools are challenged to keep up with the current web:

> *Heritrix3 is monolithic and, given the staff and skills we have available, difficult to manage and change. We want to use the latest version of H3 in order to get the best crawl results, but*

[34] Developed by UK and New Zealand
[35] Originally developed by the Danish but also used by the French and Spanish national libraries that were interviewed

> *as our deployment does not quite match the way it's used by the Internet Archive, we seem to frequently hit odd bugs and edge cases. At the same time, Heritrix3 is far behind the current web, and although IIPC partners have long known that we require more browser-assisted crawling, very few of them appear to have invested in this area. We have developed improved crawling technology, but we can't integrate it into Heritrix3 as it stands, as the framework is not sufficiently scalable.*
> *Jackson, Appendix B: UKWA Profile*

Steve Knight of New Zealand also talked about how difficult it is to keep up with technology and how he prefers smaller and modular tools and components versus monolithic approaches:

> *"Small and modular means freedom to build what we need."*
> *Knight*

The French National Library in its Profile explains the challenge of upgrading to Heritrix3 due to statistics and preservation workflow being so closely tied to Heritrix1 configurations.

Opportunity #11: Fund a collection development, nomination tool that can enable rapid collection development decisions, possibly building on one or more of the current tools that are targeted for open source deployment.

Opportunity #12: Gather requirements across institutions and among web researchers for next generation of tools that need to be developed.

Opportunity #13: Develop specifications for a web archiving API that would allow web archiving tools and services to be used interchangeably.

Researcher Use

To explore and document researcher use of web archives, we interviewed four researchers who actively work with web archives (Alam, Belovari, Graham, Webber), talked to librarians and archivists who interact with researchers, and attended SAA, Digital Library Federation (DLF), International Conference on Digital Preservation (iPRES) and the University of Michigan's Web Archiving conference, as well as the Internet Archive Researcher Services workshop given by Jefferson Bailey and Vinay Goel, both from the Internet Archive, during the IIPC 2015 General Assembly.

Researcher use of archived websites ranges from using individual archived sites or small collections to large-scale data mining against petabyte-scale collections, such as those of the Internet Archive. Bailey

introduced his workshop by presenting the topology of researchers with whom the Internet Archive works. He breaks the researchers into the following groups, differentiated by current uses:[36]

- Legal - legal discovery, evidentiary, patent, documentary
- Social/political scientists –communications, politics, government, social anthropology
- Web scientists – web technologies, systems, protocols, benchmarking
- Digital humanities – historians and humanities disciplines, network graphing, text mining, topic modeling
- Computer scientists – information retrieval, data enrichment, technical development, technology over time
- Data analysts – data mining and model training, natural language processing, trend analysis, named entity recognition, machine learning

Outside of using the Internet Archive's huge Wayback data archive, or those web archives created by institutional web archive programs, researchers are likely to create their own collections. One might want to ensure the integrity of her datasets, but another might simply not know that his desired collection has already been compiled by a library or archive.

Alternate ways of access

While access to web archives today is predominately accomplished by knowing a particular URL in advance and using the Wayback interface to access a web archive, or by conducting a keyword search, researchers are more interested in tools and programmatic access to analyze and explore the data. But as Pamela Graham observed:

> *There are technical barriers to working with web archives more analytically, very high barriers…. Researchers may need programming skills to work with web data, and there aren't many easy to use, out-of-the-box tools.*
> *Graham*

And in a paper from Cathy Hartman in which she looks at researcher use of the Government 2008 end-of-term archives, a political scientist points to lack of training and expertise in mining web archives as a hurdle:

> *If we can mine 16 terabytes of data with a few lines of code and be able to put that into a spreadsheet format, a tabular format that we can analyze statistically, that's really cool. And some political scientists have the computing skills to do that. Many of us don't. We studied content and how to do the statistics, but not this. It's not our training. It's not what we're trained to do.*
> *Hartman 33*

[36] Bailey, Jefferson. "Research Datasets Workshop."

Opportunity #14: Train researchers with the skills they need to be able to analyze big data found in web archives.

Opportunity #15: Provide tools to make researcher analysis of big data found in web archives easier, leveraging existing tools where possible.

Provenance and decisions made

To make web archives useful and usable by researchers they need to be able to understand the decisions that went into each collection, such as why certain sites or portions of sites were selected or omitted, or the reason a crawl was stopped for a period of time or altogether. During conference sessions this past year (SAA, IIPC, U. Michigan) and via email and discussion list conversations (IIPC members), this issue of provenance was brought up repeatedly.

As Pamela Graham from Columbia University explained:

> *How do you make a claim and ensure that the data is representative? For peer review you need to meet certain standards.*
> *Graham*

And as Mark Phillips from UNT commented:

> *It's a chicken and egg issue. We want them [researchers] to work with our collections but it's hard to go through and describe them in meaningful ways for researchers.*
> *Phillips*

To better understand this, a group of institutions are undertaking an as yet unpublished[37] 2015 survey of researcher use, pointing to the problem that *"researchers need metadata about web archives in order to document/interpret the validity of the results of their web archived data analysis*." The survey seeks to explore "what exactly do they need, and how much does it vary by discipline?"

Michael Neubert of the Library of Congress said:

> *Even though I have only heard from a very small sample of researchers, the message that they want to know about how and why the items in the archive were selected and made part of the archive is a clear one.*
> *Neubert, Michael. Email to IIPC Members.*

[37] Sponsored by Rutgers School of Communications and Information, University of Waterloo Department of History, Columbia University Libraries & Information Service, Web Resources Collection Program, International Internet Preservation Consortium (IIPC), California Digital Library

Indeed, several postulate that the more decisions we can document, the better for the researcher. Niels Brügger of Aarhus University, Denmark, said in an email on the IIPC mailing list:

> *What we need as researchers when using web archives — the Internet Archive and any other web archive — is as much documentation as possible. Basically, we'd like to know why did this web element/page/site end up in the web archive, and how? And this documentation can be anything between documentation on the collection level down to each individual web entity. This documentation will probably have to be collected from a variety of sources, and it will probably also have to be granulated to fit the different phases of the research process, we need documentation about the collection before we start, and we may need other types of documentation as we move along. And the need for establishing documentation becomes more and more imperative as web archives grow older, because those who created the archive will not continue to be around.*
> *Brügger. Message to Neubert reposted to IIPC members list.*

Andrew Jackson from the UK Web Archive says that if we are to capture and record decisions at scale, then we need to automate this:

> *We don't explicitly document precisely why certain URLs were rejected from the crawl, and if we make a mistake and miss a daily crawl, or mis-classify a site, it's hard to tell the difference between accident and intent from the data. Similarly, we don't document every aspect of our curatorial decisions, e.g. precisely why we choose to pursue permissions to crawl specific sites that are not in the UK domain. Capturing every mistake, decision or rationale simply isn't possible, and realistically we're only going to record information when the process of doing so can be largely or completely automated.*
> *Jackson, Andy. "The Provenance of Web Archives."*

But while some argue for as much documentation about decisions made as possible, others like Anna Perricci of Columbia University Libraries, assert that we should not hold web archives to a higher standard than other types of archival material such that it would be counterproductive:

> *Collecting decisions need to be defined. I absolutely appreciate that; I don't, however, want to get into a situation where web archivists are being asked for so much more than other archivists would reasonably be asked for…. How much questioning do archivists of other materials really get about their appraisal decisions? Beyond the collection development policy, I don't think researchers can expect a clear explanation about why things were done a certain way at this point. With web archives it's possible to articulate this to some extent, but I don't think it's yet established what should be said and how should it be stated. I do think there's a role for something which I think is coming up and that is a template for describing these*

decisions. I think that kind of tool is really important because it's hard enough to have a really succinct, clear collection development policy but to have to go too deep into the nuts and bolts of every decision impedes our ability to do other things. Documentation of our decisions and methods needs to be factored in as one of several priorities and should not be such a large use of time that it would hinder other important parts of the web archiving process…. I would hate to see something that I think is a secondary priority at this point taking away from other things that we need to do.
Perricci

Susanna Belovari from University of Illinois sums it up when she explains the information she needs when conducting her research:

Many web archives have a startling mismatch between stated appraisal, collection policy and scope and what is actually included in their archives. When we don't know the reasons behind appraisal and scope, for example, when only sites about a particular event are preserved, the web archive is of limited research use. Researchers need precise metadata about what web archives are doing and NOT doing anymore. Take web archives that have gone inactive for whatever reason. Imagine a physical archive that proclaims to have collected the significant manuscripts of the 14-17th century on a particular topic. As a researcher you then visit that archives to look for a document from the 16th century only to find that they actually stopped collecting anything after the 15th century. That's not good professional practice but we see this in web archives. What they say they have is frequently not what you actually get when you go to their portal … and remember that many national web archives only offer onsite access to preserved websites and online catalogs frequently do not list preserved sites for you to decide whether to travel across the globe for your project.
Belovari

Opportunity #16: Establish a standard for describing the curatorial decisions behind collecting web archives so that there is consistent (and machine-actionable) information for researchers.

Opportunity #17: Establish a feedback loop between researchers and the librarians/archivists – see Memberships and Collaboration where this was identified as an area to explore.

Derived data sets / Reading room only

Lack of access due to copyright concerns can be overcome by providing subsets of the data that hide or remove personal and/or copyrighted information. Jackson describes what the UK Web Archiving program is doing in this area:

Lack of public access to material makes it harder to articulate the value of the collection. We are investigating generating analytics and datasets as a way of helping researchers get something valuable out of the collection even if they can't access the individual items.
Jackson, Appendix B: UKWA Profile

Providing researcher services

This year the Internet Archive announced its Researcher Services.[38] Rosalie Lack of CDL had very positive things to say about the new service, but suggests that more work may be warranted to expand the "services" element (Lack). This is a sentiment echoed by Graham, who adds that she needs to think about how to add local services to support researchers at Columbia:

We probably need them [researcher services] more locally and will probably have to think how to support people. Using WARC derivatives and Archive-It Research Services is a new area for us. We need to look at how to provide user services, and how to provide support to our users. There are not any peers or other universities to ask, which suggests an opportunity for a community around these data services, which may blend in to other data services.
Graham

Opportunity #18: Explore how institutions can augment the Archive-It service and provide local support to researchers, possibly using a collaborative model.

Researcher feedback

Jackson from the UKWA urges that the curators of web archives engage with researchers to learn what they need and how to improve what is collected (such interaction between the two groups has come up several times as an opportunity for future exploration):

No corpus, digital or otherwise, is perfect. Every archival sliver can only come to be understood through use, and we must open up to and engage with researchers in order to discover what provenance we need and how our crawls and curation can be improved.
Jackson, "The Provenance of Web Archives"

Cathy Hartman of UNT conducted a small study to investigate researchers' use or anticipated use of the 2008 EOT archives. In the area of future research and development she concludes:

It is clear that researchers in most disciplines will need assistance to extract the data they need from the [web] archive. Researchers will need to identify the content of interest to their research and to specify the data elements and data formats needed in the extracted content. Collaborations between researchers, librarians, information scientists, and computer scientists

[38] https://archive-it.org/blog/post/launching-archive-it-research-services-part-1/

> *appear necessary to build the tools that will enable researchers to discover and extract content.*
> Hartman 33

As far as gathering researcher feedback, Michael Neubert sums it up nicely:

> *I'm actually OK with the researchers being mad that we didn't fulfill their expectations. I don't want the researcher to grab my necktie and give it a tug to express their annoyance, but to calmly tell me where we are failing them, yeah. How else do we learn?*
> *Neubert, Message to IIPC Members*

Opportunity #19: Increase interaction with users, and develop deep collaborations with computer scientists.

Infrastructure

Any digital program, web archiving no exception, requires technical infrastructure. A little over half (12 of the 23) of the institutions surveyed are outsourcing portions of, or the entire web crawling and hosting, to an external vendor (11 use either Internet Archive as a contractor, or Internet Archive's Archive-It service, and Germany outsources to a German company, oia.[39] The barrier to entry for web archiving is lowered by the Archive-It service, since it alleviates the need to run local infrastructure or have IT support. But what about access to tools that these institutions may want to run against their archives? Without resources, how might broader access to tools and compute be made possible?

Onsite and external

Over half of the respondents (13 out of 23) reported that they use both onsite and offsite infrastructure for web archiving. Of these respondents, all but six (the national programs out of Iceland, Finland, Germany, Netherlands, United Kingdom and the Rhizome) are using Archive-It or a contract with Internet Archive to create and house their collections offsite (Columbia, Stanford and George Washington Universities, Library of Congress, NYARC, Smithsonian, UCLA).

Onsite only

Six institutions (MIT, Harvard University, UNT, and the national programs out of Denmark, France and Spain) reported onsite infrastructure *only*, although Harvard – whose WAX service has been maintained onsite until now – is investigating moving some crawling activities to Archive-It; and MIT – early in its web archiving program – has started evaluating Archive-It.

External only

Four institutions reported they only have external infrastructure (Cornell and Yale Universities, NMWA, and the National Library of New Zealand), with the first three subscribing to Archive-It and the New

[39] http://oia-owa.de/de/home/

Zealand web archiving program using a combination of Internet Archive, which is contracted for its domain crawls, and its own service hosted at a Government-mandated data center (using a private cloud).

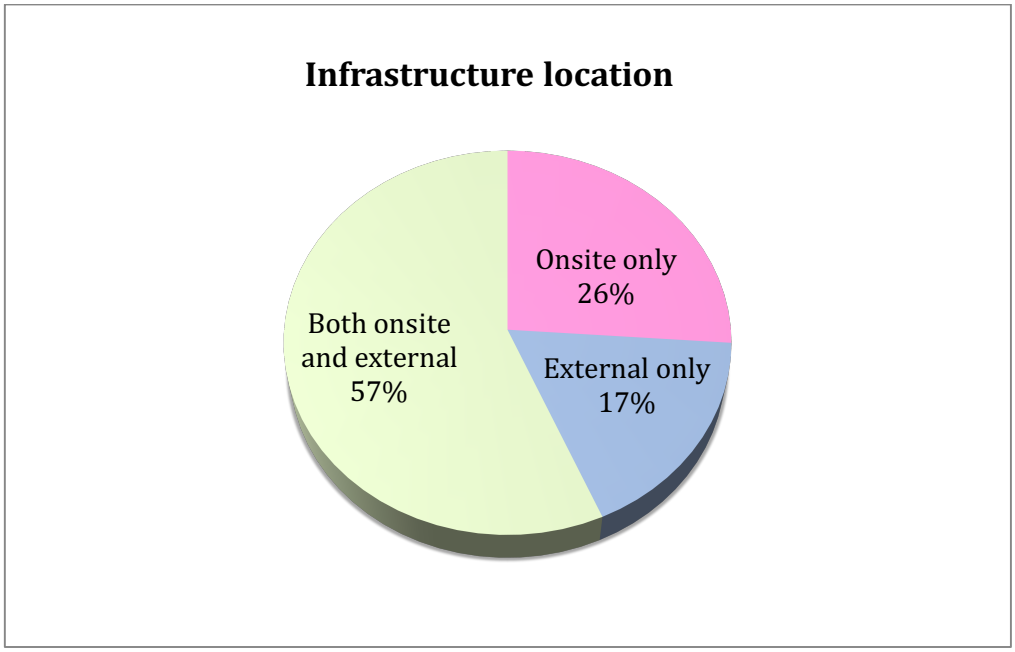

Figure 8: Showing the location of infrastructure for each of the surveyed web archiving programs.

In 2015 the CDL's Web Archiving Service (WAS) collections and all core infrastructure activities, i.e., crawling, indexing, search, display, and storage, were transferred to the Internet Archive's Archive-It.[40]

> *We decided to make the change because we really needed to reduce our operational and developmental costs…we were just keeping up through a lot of time and effort but we were never getting ahead in terms of function. Also, as the archive continued to grow we were continually running up against scaling issues. You have to have a lot of machinery available – a lot of processing power, an awful lot of storage, and that's cumbersome to manage, so we were very much interested in making the move to what could be seen as much more of a commodity solution that's providing a certain baseline function and the idea always was, and continues to be, that by doing that we will be freeing up our limited resources that could then be re-applied to areas that we could uniquely add value to. Seems to be a better allocation of our resources.*
> *Abrams*

Archive-It, with over 400 subscribers, combined with the broader captures and national domain crawls provided out of its parent institution, the Internet Archive, houses a large portion of today's web archives. As a way of opening up these archives to researcher use, Archive-It launched its Researcher Services in 2015. While these services offer access to subsets of data more suitable to run analysis against (derived datasets), those institutions with little or no onsite infrastructure or IT support may be at a disadvantage.

[40] https://was.cdlib.org/

Compute resources are best deployed close to the data they run against. The derived datasets are a fraction of the size of a web archive file, and analysis can be run on a researcher's laptop. But for researchers who desire a large corpus of data to analyze – including that of the Wayback Machine's broader-capture archive – getting compute close to the data often poses a challenge.

Sawood Alam, a researcher from Old Dominion University, has 300TB of web archive data that he has indexed. He says a problem for him is the insufficient compute or storage resources for derivative work (Alam). One idea he offers is to distribute tasks and perform computation on users' distributed machines. This would permit a BOINC-SETI@home-type[1] of implementation to perform Hadoop-style distributed research operations on web archive data on multiple distributed users' machines simultaneously. Alam in email explained that his BOINC inspired idea would be implemented using JavaScript Web Workers to distribute the task to users via the browser as they visit certain participating sites instead of asking them to install a separate software for the purpose.

Another researcher, Matthew Weber of Rutgers University, helped create archivehub.rutgers.edu, which he described as a Hadoop-based service with about 80TB of datasets from a variety of web archives. He has processed the data and it runs on a locally hosted cluster. He currently has 4 institutions working with the data.

As more researchers start to investigate how to use web archives and the tools available to analyze them, the need for access to compute, storage and other infrastructure resources will need to be addressed. Does the Internet Archive need to consider offering Platform as a Service[41] (PaaS) to host a suite of analysis tools that can be run against the data it hosts? If so it would need to add additional infrastructure and support personnel. Or, as Rosenthal, Bailey and Taylor suggest,[42] archives could transfer the data to be mined to the cloud using the cloud provider's servers at the researcher's expense.

Opportunity #20: Explore what, and how, a service might support running computing and software tools and infrastructure for institutions that lack their own onsite infrastructure to do so.

Preservation of Web Archiving Collections

Over half the institutions surveyed report that they have a local preservation copy of their web archives. As Andrea Goethals of Harvard explains:

> *At Harvard we collect and make sure our web archives are preserved, just as we do our other digital library collections. We take care of it; but in a collaboration how can we ensure that it is preserved?*
> *Goethals*

[41] https://en.wikipedia.org/wiki/Platform_as_a_service
42 Rosenthal, D. Taylor, N. Bailey, J. Interoperation Among Web Archiving Technologies. N.d. TS.

Interestingly, 23% (7 institutions) reported Archive-It as providing preservation, while others were quite adamant that Archive-It has neither disclosed its preservation strategies, nor has it been audited by a third party and was not considered a preservation service. Michele Paolillo explained Cornell's position on preservation:

> *Evaluation of a repository for suitability in preservation requires much more transparency with regards to technology, institutional organization and resource commitment (both funding and FTE). There are ways to explore these areas: TRAC Assessment, TDR Certification, and ISO 16363, etc. The lack of disclosure from Internet Archive on many key points has not made any such comprehensive assessment possible. Even the statement "Data integrity and system availability are assured using a combination of internal and external systems and processes," does not assure me of bit fixity. I'd like to better understand what they are actually doing technologically to monitor for and guard against unwanted changes. Much more transparency is in order.*
> *Paolillo*

Jason Kovari of Cornell says he plans to eventually export the WARC files back to preserve them in Cornell's local preservation solution (Kovari). Archive-It includes the ability for any user to download their WARCs at any time, even after they potentially discontinue using the services (Bailey, email).

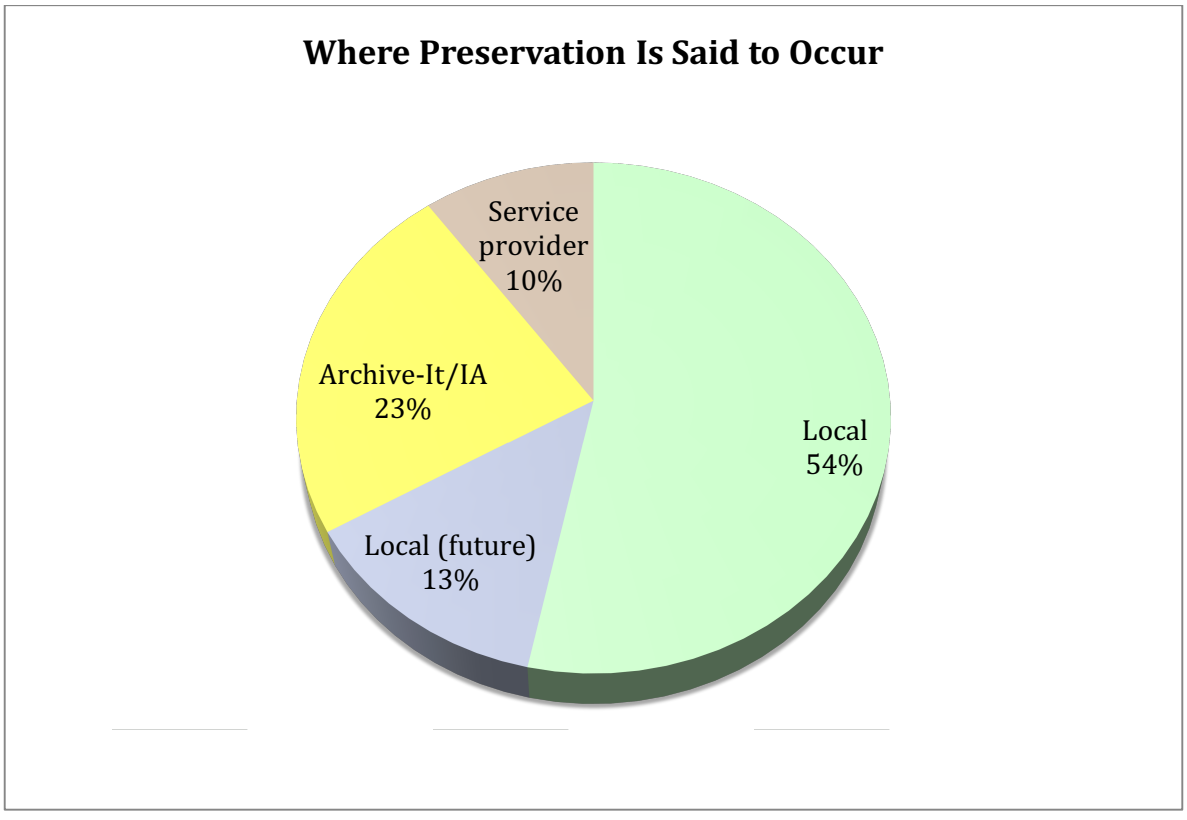

Figure 9: Shows location of each web archiving program's preservation copies (now and planned).

The National Library of New Zealand contracts with the Internet Archive for its whole domain crawls and does selective domain crawling itself using the Web Curator Tool. The library maintains its preservation capability in house, stating it would not entrust a third party to undertake the preservation aspect of content care (Appendix B: National Library of New Zealand Profile).

The Smithsonian retrieves WARC files every week from Archive-It for local preservation. The Smithsonian uses Archive-It but doesn't identify copies stored there as being externally preserved, answering "no" to that question (Appendix B: Smithsonian Profile). The Smithsonian developed the WARC Transfer Tool (Schmitz Fuhrig, Lynda) to more efficiently pull down their weekly Web Archiving File (WARC) retrieval (WARC Transfer Tool uses dates and checksums to look for duplicate content).

The Archive-It service includes a provision to export the WARCs to the DuraCloud service for bit-level preservation.[43] NYARC is the single organization interviewed that subscribes to Archive-It and also has a paid subscription with DuraSpace, the non-profit organization that developed DuraCloud. Columbia has considered using DuraCloud and reports a concern that they lack an assigned dedicated budget for this should they decide to do so (Appendix B: Columbia, Profile).

While bit preservation is important, several interviewees remind us that preservation is about more than ensuring data usability over time, it's about access and renderability over time. And with the browser technologies and user experiences of websites evolving so rapidly, accurate rendering/replay of these sites will likely require preservation strategies to include both emulation and migration. Skip Kendall of Harvard and Kari Smith of MIT spoke of these challenges:

> *Lots of preserved websites use Flash, so they can deliver the archive but having it work depends on the technology in the browser and then the harvest becomes unusable. This seems to point to emulation, which is a challenge since we now have to consider multiple preservation strategies. We're already doing format migration, so perhaps the Flash problem could be resolved by keeping an old browser/station around.*
> *Kendall*

> *I think the preservation of web archives over time is going to be something that will need more and more guarantee as we need to move through versions of technology and as we're changing from formats of attachments and things. How does that continue to be rendered within a web based file structure?*
> *Smith*

Accurate replay is particularly important for art resources as Heather Slania of NMWA points out:

[43] http://www.duracloud.org/archive-it

Art related websites frequently break when being archived due to their high levels of dynamic content and interactivity. Preserving that interactivity is currently not possible – and highly desired.
Slania

Another art resources institution, Rhizome, indicates that emulation plays a key part in their preservation activities. "Preservation" for Rhizome means being able to recall computational performances. (Appendix B: Rhizome Profile). And both Rhizome and NYARC indicate that non-renderable formats from within WARCs, such as Flash files, remain a challenge (Appendix B: NYARC and Rhizome Profiles).

The National Library of Finland locally preserves its web archiving files as well as externally preserving copies at the National Digital Library's Digital Preservation System, maintained by CSC.[44] And at the National Library of New Zealand, its National Library's Preservation Research and Consultancy team conducts local preservation based on the Rosetta digital preservation system[45].

Service Providers

Of the service providers in the web archiving space, arguably the most well known – and subscribed to – is the Archive-It service offered out of the Internet Archive, and in use globally. In 2015 another web archiving service provider, the California Digital Library, announced the transfer of its Web Archiving Service (WAS) collections and all core infrastructure activities, i.e., crawling, indexing, search, display, and storage, to Archive-It.[46] Other service providers cover this space too, such as Hanzo Archives.[47] Hanzo is being used by NYARC to conduct crawls that cannot be captured adequately by Archive-It,[48] but is not broadly used by the demographic covered for this scan. Hanzo's focus (and those of its competitors, such as Iterasi[49]) is on web archiving for eDiscovery, compliance and corporate heritage.[50]

Despite the large number of web archiving tools (Appendix C), there are very few service providers maintaining and supporting all these tools. Users are often required to download the source code from GitHub or other code repositories and have sufficient programming skills and IT knowledge and support to use the tools. MIT is fortunate to have a library fellow, Jessica Venlet, who has tested different tools in the MIT Digital Sustainability Lab (Smith, Kari). But many users and researchers do not have the necessary resources to use the tools on offer, suggesting an opportunity for more service offerings in this space or the addition of a services component to those tools that are already offered today via a web user interface, such as webrecorder.

[44] http://www.csc.fi
[45] http://www.kdk.fi/index.php/en/long-term-preservation
[46] https://was.cdlib.org/
[47] http://www.hanzoarchives.com/
[48] http://ndsr.nycdigital.org/wp-content/uploads/2014/05/nyarc.pdf
[49] http://www.iterasi.com/
[50] http://www.hanzoarchives.com/solutions/

With the reliance on the Internet Archive and its Archive-It service, what are the risks and opportunities this might afford? Steve Knight of the National Library of New Zealand observed that:

> *We need to be careful about resting on one architecture. Informal research from Denmark suggests that multiple methods can deliver differing results. This was in the context of identifying national parts of the internet but the principle still applies. We need to be periodically re-evolving the criteria and methods we use until we can be certain we are getting the results that we need.*
> *Knight*

The Smithsonian presents a more optimistic outlook:

> *I see an advantage to using a service that a large amount of institutions are using. For example we can collectively put pressure on them if there's a problem we need help finding a solution to. And if their business falters it's more likely a good solution will be found to make sure we all can get our content out – there's 'safety in numbers'.*
> *Wright, Jennifer*

Opportunity #21: Service providers develop more offerings around the available tools to lower the barrier to entry and make them accessible to those lacking programming skills and/or IT support.

Opportunity #22: Work with service providers to help reduce any risks of reliance on them (e.g. support for APIs so that service providers could more easily be changed and content exported if needed).

Findings and Opportunities for Future Research and Development

The purpose of conducting this environmental scan is to identify common practices, concerns, needs, and expectations in the collection and provision of web archives to users; the provision and maintenance of web archiving infrastructure and services; and the use of web archives by researchers. Through engagement with 23 institutions with web archiving programs, two service providers and four web archive researchers, along with independent research, it uncovered 22 opportunities for future research and development. At a high level these opportunities fall under four themes: (1) *increase communication and collaboration, (2) focus on "smart" technical development. (3), focus on training and skills development, and (4) build local capacity.*

Theme 1: Increase communication and collaboration

Our investigation into current practices in web archiving reveals the need to radically *increase communication and collaboration*. Of the 22 opportunities identified for future exploration, 13 fall under

this theme, making *increase communication and collaboration* the number one theme (note, some opportunities fall under more than one theme).

More communication is needed both across the librarians and archivists who build and steward collections of archived websites, but also between these stewards and the historians, data scientists and researchers who use them. Collection building for web archives predominantly falls within the libraries and archives, but individuals, departments, or groups of researchers might also build their own web archiving collections to support specific needs. This indicates an opportunity (see opportunity #3) for increasing communication and collaboration across these different types of collectors. Absent sufficient communication, institutions today lack insight into the collection decisions and practices of others, and this can result in either duplication or gaps in coverage. Opportunities #6 and #7 point to the need for more transparency of web archive holdings (#6) and exposing web archive holdings information via a registry or similar method to researchers and other collecting institutions - even if the content is only viewable in on-site reading rooms (#7).

Communication for outreach and education purposes surfaces in opportunities # 2 and # 8 which identify the need to train existing staff working with more traditional collections who may be new to web archiving (#2), and website developers may need training to create more easily archived and described sites (#8).

Opportunities #4, #5, #17, #18, and #19 address researcher use of web archives and the need for *greater communication and collaboration* with this community. These opportunities identify the need to gather researcher feedback on requirements and impediments to the use of web archives (#4), and leveraging membership overlap between RESAW and IIPC membership to facilitate formal researcher, librarian, archivist collaboration projects (#5). Opportunities #17, #18 and #19 call for a feedback loop between researchers and librarians/archivists (#17); a possible collaborative model of providing researcher support to augment the new Archive-It Researcher Services (#18); and increasing interaction with users, and developing deep collaborations with computer scientists (#19).

From a discovery perspective, we see an opportunity (#10) to communicate across web archiving institutions (for example, conducting surveys or user studies) to investigate whether Memento should be adopted more broadly as part of their discovery infrastructure.

The tools section identifies the need to communicate across institutions to gather requirements for the next generation of tools that need to be developed (#12). And opportunity #9 suggests outreach and communication to influence tech company content hosting sites (e.g. Google/YouTube).

Theme 2: Focus on "smart" technical development

Focus on "smart" technical development ranks as the second most popular theme, showing up in eight of the 22 opportunities. Gathering requirements for next generation tools, a *smart* start for any technical development, is identified in opportunity #12. Opportunity #13 suggests developing specifications for a

web archiving API that would allow web archiving tools and services to be used interchangeably and would enable them to be "chained" together. Continuing with APIs, opportunity #22 suggests working with service providers to help reduce the risk of reliance in them, where APIs could help move content as needed. Opportunity #11 suggests funding technical development for a collection development and nomination tool to enable rapid collection development decisions, possibly building on one or more current tools. Development of a collection development tool, such as a registry or directory (#7) would also require some technical development and oversight.

Opportunity #15 identifies the need to provide tools (leveraging existing tools where possible) to make researcher analysis of big data found in web archives easier. Establishing a standard for describing the curatorial decisions behind collecting web archives, so that there is consistent and machine actionable information for researchers, is identified in opportunity #15.

Finally, opportunities #20 and #21 are around service providers offering "as-a-service" tools and infrastructure to those institutions lacking their own onsite infrastructure to do so (#20), and making tools more accessible to those lacking programming skills and/or IT support (#21).

Theme 3: Focus on training and skills development

Training and skills development ranks as the third most prevalent theme - it surfaces in six of the 22 opportunities for future exploration. The need for training and skills development for existing staff new to web archiving is identified in opportunity #2. Training also pertains to opportunity #8, which calls for training for website developers. Opportunities #4 and #5 identify the need to understand researcher use and requirements around web archiving – and might easily include a training and skills development component. Training researchers with skills they need to analyze big data found in web archives is specifically called out in opportunity #14. Finally, opportunity #9 suggests the need to train content hosting sites on the importance of supporting libraries and archives in their efforts to archive their content.

Theme 4: Build local capacity

Build local capacity relates to augmenting an institution's resources and proficiency in the area of web archiving and related services. It ranks fourth in themes with four of the 22 opportunities falling here. *Build local capacity* first appears in opportunity #1, the dedication of full-time staff to work in web archiving so that they can stay abreast of latest developments, best practices and consequently fully engage in the community. Opportunity #18 suggests exploring how institutions can augment the Archive-It service and provide their own, local support to researchers. And for service providers, there's an opportunity (#20) to increase their local capacity in the area of providing computing and software tools and infrastructure for those institutions lacking their own onsite infrastructure, and also for them to develop more offerings around the available tools (#21).

A summary of all the opportunities follows:

Staffing:

Opportunity #1: Dedicate full-time staff to work in web archiving so that institutions can stay abreast of latest developments, best practices and fully engage in the web archiving community.

Opportunity #2: Conduct outreach, training and professional development for existing staff, particularly those working with more traditional collections, such as print, who are being asked to collect web archives.

Location in organization:

Opportunity #3: Increase communication and collaboration across types of collectors since they might collect in different areas or for different reasons. See also Memberships and Collaborations section.

Memberships and collaborations:

Opportunity #4: A funded collaboration program (bursary award, for example) to support researcher use of web archives by gathering feedback on requirements and impediments to the use of web archives.

Opportunity #5: Leverage the membership overlap between RESAW and European IIPC membership to facilitate formal researcher/librarian/archivist collaboration projects. Such collaborations' goals might include to:

- facilitate understanding of how researchers want to use web-based content in their research
- determine how to provide web archives to researchers, including APIs and other means of access
- determine description and metadata that would help validate research samples
- determine how to provide support services to researchers

Collection development:

Opportunity #6: Institutional web archiving programs become transparent about holdings, indicating what material each has, terms of use, preservation commitment, plus curatorial decisions made for each capture.

Opportunity #7: Develop a collection development tool (e.g. registry or directory) to expose holdings information to researchers and other collecting institutions even if the content is viewable only in on-site reading rooms.

Opportunity #8: Conduct outreach and education to website developers to provide guidance on creating sites that can be more easily archived and described by web archiving practitioners.

Opportunity #9: IIPC, or similar large international organization, attempts to educate and influence tech company content hosting sites (e.g. Google/YouTube) on the importance of supporting libraries and archives in their efforts to archive their content (even if the content cannot be made immediately available to researchers).

Discovery:

Opportunity #10: Investigate Memento further, for example conduct user studies, to see if more web archiving institutions should adopt it as part of their discovery infrastructure.

Tools:

Opportunity #11: Fund a collection development, nomination tool that can enable rapid collection development decisions, possibly building on one or more of the current tools that are targeted for open source deployment.

Opportunity #12: Gather requirements across institutions and among web researchers for next generation of tools that need to be developed.

Opportunity #13: Develop specifications for a web archiving API that would allow web archiving tools and services to be used interchangeably.

Researcher use:

Opportunity #14: Train researchers with the skills they need to be able to analyze big data found in web archives.

Opportunity #15: Provide tools to make researcher analysis of big data found in web archives easier, leveraging existing tools where possible.

Opportunity #16: Establish a standard for describing the curatorial decisions behind collecting web archives so that there is consistent (and machine-actionable) information for researchers.

Opportunity #17: Establish a feedback loop between researchers and the librarians/archivists – see Memberships and Collaboration where this was identified as an area to explore.

Opportunity #18: Explore how institutions can augment the Archive-It service and provide local support to researchers, possibly using a collaborative model.

Opportunity #19: Increase interaction with users, and develop deep collaborations with computer scientists.

Infrastructure:

Opportunity #20: Explore what, and how, a service might support running computing and software tools and infrastructure for institutions that lack their own onsite infrastructure to do so.

Service providers:

Opportunity #21: Service providers develop more offerings around the available tools to lower the barrier to entry and make them accessible to those lacking programming skills and/or IT support.

Opportunity #22: Work with service providers to help reduce any risks of reliance on them (e.g. support for APIs so that service providers could more easily be changed and content exported if needed).

Appendices

Appendix A: List of Institutions and Participants Consulted for the Environmental Scan

- California Digital Library (service provider) – Stephen Abrams, Scott Fisher, Rosalie Lack, David Moles
- Columbia University Libraries – Anna Perricci, Pamela Graham, Alex Thurman
- Cornell University – Jason Kovari, Michelle Paolillo
- Danish Royal Library and State and Local Library – Nicholas Clarke
- George Washington University – Daniel Chudnov, Christie Peterson, Rachel Trent, Laura Wrubel
- Harvard Library – Abigail Bordeaux, Andrea Goethals, Skip Kendall
- Internet Archive (service provider) – Jefferson Bailey, Vinay Goel
- Library of Congress – Helen Conkle, Rick Fitzgerald, Abigail Grotke, Gina Jones, Andrew Weber
- Massachusetts Institute of Technology (MIT) –Kari Smith, Jessica Venlet
- National Library of Finland – Lassi Lager
- National Library of France – Sara Aubry
- National Library of Germany – Tobias Steinke
- National Library of Netherlands - Peter de Bode, Barbara Sierman
- National and University Library of Iceland - Kristinn Sigurðsson
- National Library of New Zealand – Jay Gattuso, Gillian Lee, Steve Knight
- National Library of Spain - Dragan Espenschied
- National Museum of Women in the Arts (NMWA) – Heather Slania
- New York Art Resources Consortium (NYARC) – Sumitra Duncan, Deborah Kempe
- Old Dominion University (researcher/user) – Sawood Alam
- Rhizome - Dragan Espenschied
- Rutgers University (researcher/user) – Matthew Weber
- Stanford University Libraries – Nicholas Taylor
- Smithsonian Institution Archives – Lynda Schmitz Fuhrig, Jennifer Wright
- UK Web Archives/British Library – Andy Jackson
- University of California, Los Angeles (UCLA) – Martin Klein
- University of Illinois (researcher/user/archivist) – Susanne Belovari
- University of North Texas – Mark Phillips
- Yale University – Rachel Chatalbash, Gabriela Redwine

Appendix B: Institutional Profiles

<table>
<tr><td>Institution Name: Columbia University Libraries (CUL)</td><td colspan="2">https://www.archive-it.org/organizations/304https://www.archive-it.org/organizations/304
http://hrwa.cul.columbia.edu/</td></tr>
<tr><td colspan="3">Columbia University Libraries (CUL) has 12 TB of content and has been archiving websites since 2008 (with 3 Mellon Foundation grants followed by CUL-funding). CUL uses Archive-It and also downloads WARC files each quarter to offer local access to a single collection, the Human Rights Web Archive. The sustainability of ongoing support for this local portal is under discussion. CUL focuses its web archiving on collection development and collaborations rather than technical development. CUL initiated the Ivy Plus collaborative collection development pilot and has presented a proposal to jointly fund the expansion of the pilot into an ongoing program to the Borrow Direct/Ivy Plus university librarians group.</td></tr>
<tr><td colspan="3">Main Use Cases:
• Thematic or topical web archives (including collaborative collecting) aligning with existing CUL collecting focus
• Websites of organizations or individuals whose records or papers are held at CUL
• Columbia.edu domain and other Columbia-affiliated websites</td></tr>
<tr><td>Collection Development Location: Libraries; Web Resources Collection Coordinator works with subject specialists and university archivist.</td><td colspan="2">Additional Info: Impetus was collection development and extending collections to include missing (web) content. They more recently collaborated with U. Archives for institutional sites, but program still resides within Columbia University Libraries.</td></tr>
<tr><td>Membership and Collaborations:
• IIPC
• NDSA/NDIIP
• SAA
• Ivy Plus
• Archive-It partner meetings, including NYC AIT group
• Mellon Foundation</td><td colspan="2">Details:
• IIPC - CUL Coordinator serving as co-chair of Content Development Group (collaborative collections)
• Ivy Plus - collaborative collection development at Archive-It – Contemporary Composers Web Archive (CCWA) and Collaborative Architecture, Urbanism, and Sustainability web Archive (CAUSEWAY) pilot web collections
• With Ivy Plus are considering options for preservation (possibly DuraCloud)
• Mellon funding of program and tool development – viewed by CUL as a collaborator/partner</td></tr>
<tr><td colspan="3">Funding: Self-funded since 2013 with additional Grant funding from Mellon Foundation (three since 2008).</td></tr>
<tr><td colspan="3">Staffing: One FTE Librarian (Web Resources Collection Coordinator) and one FTE Bibliographic Assistant (vacant). Another grant-funded FTE librarian position recently ended. CUL pays for supervisory staff time dedicated to web archiving (web archiving steering committee meets twice/month), estimated at ¼ person time/month.</td></tr>
<tr><td>Integration of Web Archives with Other Collections:</td><td>Internally: Yes
CLIO online catalog</td><td>Externally: No
All Archive-It partner collections can be jointly searched at archive-it.org.</td></tr>
</table>

Details and Concerns:		MARC records for CUL's archived websites are shared in Worldcat.
External Infrastructure: Yes. (Archive-It)	External Preservation: Yes. (Archive-It)	External Access Portal: Yes. (Archive-It)
Onsite Infrastructure: Yes	Local Preservation: No	Local Access Portal:Yes
Tools Used Onsite: HR Manager (Blacklight/SOLR display of metadata), Search expansion, Search extension, FileMakerPro database for administrative/permissions metadata		
Tools/Infrastructure Challenges: Maintaining quarterly download of Human Rights collection WARC data from Archive-It for indexing in HRWA local portal, and reindexing millions of items, without ongoing dedicated funding for developer time.		
Preservation Challenges: Lack of assigned budget should CUL decide to use DuraCloud .		
Collection Development Challenges: Choosing subjects for thematic web archives of external content, aligning with existing CUL collection strengths in other formats		
Other:		

Institution Name: Cornell University Library (CUL)	https://www.library.cornell.edu/ https://archive-it.org/organizations/529	
Cornell University Library (CUL) has been web archiving since 2011 and has 4 TB of content in Archive-It.		
Main Use Cases: • Institutional archives (websites and social media) • Thematic or topical web archives • Enhancement of manuscript collections (organizational materials collected by CUL repositories) • Faculty member teaching collection		
Collection Development Location: Library	Additional Info: Distributed between main library collection development group and archival repositories, dependent on collection; one collection was selected by a faculty member.	
Membership and Collaborations: No institutional membership related to web archiving	Details: IvyPlus collaborative collection development	
Funding: Internal		
Staffing: CUL devotes approximately .7 FTE to the technical services aspects of web archiving (crawling, QA, non-MARC metadata); this does not include collection development staffing, which is more difficult to quantify.		
Integration of Web Archives with Other Collections: Details and Concerns:	Internally: Nothing systematic beyond the creation of a collection-level MARC record. Better	Externally: None

	integrating these resources into our discovery environment is a topic of conversation but is not yet on the development timeline.	
External Infrastructure: Yes–Archive-It.	External Preservation: None	External Access Portal: Archive-It
Onsite Infrastructure: No	Local Preservation: In-queue	Local Access Portal: None
Tools Used Onsite: Archive-It; we have not pursued tool development beyond what is available in that service		
Tools/Infrastructure Challenges: • Better capturing of dynamic media • Need for automation of quality control / assessment procedures		
Preservation Challenges:		
Collection Development Challenges: We are currently in the process of assessing collection development needs.		
Other:		

Name: Danish Royal Library and State and Local Library (Netarkivet.dk)	URL: http://netarkivet.dk/in-english/
About: Netarkivet.dk (Danish Royal Library and State and Local Library) has been crawling Danish domains since 2005. Besides the frequent crawls of media sites to capture impromptu events, a fixed number of broad crawls are also run every year. The last few years 4 broad crawls have been run. The archive has exceeded 700TB. There are approximately 1.3 million dk domains.	
Main Use Cases: • News sites crawled several times a day • Event harvests when something out of the ordinary needs to be preserved • Preserve as much of the Danish web as possible • Ebooks	
Collection Development Location: Netarkivet.dk is a collaboration between the Royal Library and The State Library.	Additional Info: Storage and crawlers are distributed between these two institutions. Storage is kept in 3 copies. One institution handles broad crawls while the other handles selective and event harvests.
Membership and Collaborations: • IIPC • OPF • PREMIS	Details: • WARC/1.1 – work on the next version • OPF – With particular interest in PDF validation and emulation

• WARC		
Funding: Finance act.		
Staffing: Approx. 20 people distributed over 2 institutions.		
Integration of Web Archives with Other Collections: No Details and Concerns:	Internally: No	Externally: No
External Infrastructure: No	External Preservation: No	External Access Portal: No
Onsite Infrastructure: Yes	Local Preservation: Yes	Local Access Portal: Yes
Tools Used Onsite: Heritrix 1/3, Wayback, SOLR, Archive-IT, JWAT-Tools		
Tools/Infrastrucure Challenges: • Funding always has an impact on how much development can be done • We are switching from an old distributed archive to a newly implemented one (Danish BitRepository) • Switching from Heritrix 1 to Heritrix 3		
Preservation Challenges: Archiving website with advanced AJAX use (We are not using Umbra or similar yet).		
Collection Development Challenges: Making sure we are not missing parts of the Danish web. Research projects are in progress to determine if there are parts of the Danish web we are not currently preserving.		
Other:		

Institution Name: George Washington Libraries (GWL)	https://archive-it.org/home/gwlibraries
George Washington Libraries (GWL) has been web archiving using Archive-It since June 2014 and has 1TB of content. GWL has migrated content from the Internet Archives public Wayback collections to one of the GWL collections, going back to 1996. GWL is experimenting with Social Feed Manager for collection of social media archival data.	
Main Use Cases: • Institutional (gwu.edu) archives (websites and social media) • Thematic or topical web archives for special collections and of interest to Global Resources Center and faculty • Twitter (with Flickr, Tumblr, and Weibo being added) using Social Feed Manager harvesting tool in support of the mission of the University Archives, special collections, and faculty and student researchers on campus	
Collection Development Location: Activities are distributed across several schools and the university archives.	Additional Info: This is in transition, but all activities are within the Libraries. Archive-It administration had initially been done as a one-year cross-divisional project team. It is moving under the management of the Digital Services Manager in the Special Collections and Research Center within GW Libraries. A web archiving team will have participation from a number of people in other areas of the library.
Membership and	Details: NCSU, UNT, Yale, NYU, UVA, UCSD, RRCHNM and others met in

Collaborations: • IIPC • Archive-It partner meetings, including Mid-Atlantic AIT • SAA	2013 to help identify areas and priorities for future development of GWU's Social Feed Manager prototype.

Funding: Archive-It is funded out of the library's budget. Social Feed Manager is partially supported by grants, from the NHPRC and Council on East Asian Libraries (via Mellon). Social Feed Manager also had initial support from IMLS through a Sparks Innovation Grant.		
Staffing: No dedicated personnel today. Combined staff hours of approximately five cross-division curators from various departments in the library, estimated 5-10 hours/week, plus a part-time grad student. The Digital Services Manager coordinates the program.		
Integration of Web Archives with Other Collections: Details and Concerns:	Internally: No	Externally: No
External Infrastructure: Yes. Archive-It	External Preservation: Yes. Archive-It	External Access Portal: Yes. Archive-It
Onsite Infrastructure: Yes	Local Preservation: No. In the future, may also store the data in GW libraries' digital repository.	Local Access Portal: No. In the future, may also store the data in GW libraries' digital repository.
Tools Used Onsite: Archive-It; experimental use of Social Feed Manager for capturing social media archives (in-house tool for capture of Twitter, Tumblr, Weibo, and Flickr)		
Tools/Infrastructure Challenges: • Meaningful integration of Archive-It content with related collections, including social media datasets • Legal uncertainties to capturing, preserving, and making available social media data harvested via platform APIs • Need to develop technical standards for preservation and access to social media data		
Preservation Challenges: Diversity and evolution of social media data formats; vendor storage integration		
Collection Development Challenges: • Identifying GW-managed social media presences (no definitive list exists) • Identifying portions of the gwu.edu domain to prioritize for collection & QA • Identifying and capturing international, unstable content in a timely manner		
Other:		

Institution Name: Harvard Library (HL)	http://wax.lib.harvard.edu/collections/home.do

<table>
<tr><td colspan="3">Harvard Library (HL) has been web archiving since 2006 and has approx. 4TB of content. It currently maintains its own WAX service but is considering outsourcing a portion (to Archive-It). Should Archive-It provide the crawling and QA service, the library would request copies of WARC files for local preservation, plus possibly a subset of the WARCs would be used to continue to offer local access from the HL portal.</td></tr>
<tr><td colspan="3">Main Use Cases:
• Institutional archives (websites and social media)
• Thematic or topical web archives
• Enhancement of manuscript collections (Harvard people/fellows, companies)
• Archive PDF publications no longer published in print form</td></tr>
<tr><td>Collection Development Location: Activities are distributed across several schools and the university archives.</td><td colspan="2">Additional Info: Central system is run out of Library Technology Services and Preservation Services. Considering centralizing a web services manager position to help coordinate distributed activities.</td></tr>
<tr><td>Membership and Collaborations:
• IIPC
• NDSA/NDIIP
• SAA
• Ivy Plus
• Chesapeake Project</td><td colspan="2">Details:
• Ivy Plus– collaborative collection development at Archive-It – Contemporary Composers Web Archive (CCWA) and Collaborative Architecture, Urbanism, and Sustainability web Archive (CAUSEWAY) pilot web collections
• IIPC/NDIIP partners– End of Term Archive, US Government - collaborative collection development Kennedy School of Government, Internet Archive, Library Congress, California Digital Library, Government Printing Office, University North Texas
• Chesapeake Project– (part of the Legal Information Archive)– collaborative collection development at OCLC/CONTENTdm (Harvard Law School Library)</td></tr>
<tr><td colspan="3">Funding: Charge-back to Libraries and Archives using the Harvard WAX service partially covers the cost; internal funding covers the remainder.</td></tr>
<tr><td colspan="3">Staffing: No dedicated personnel today (4 persons 1/4 time), plan to recommend a full-time web service manager and a developer.</td></tr>
<tr><td>Integration of Web Archives with Other Collections:</td><td>Internally: Yes.
HOLLIS online catalog, searchable finding aids</td><td>Externally: No.
Concerns about trusting a collaborating institution to preserve WARC content</td></tr>
<tr><td>External Infrastructure: No. (Until/unless move to AIT)</td><td>External Preservation: No.</td><td>External Access Portal: No. (Until/unless move to AIT)</td></tr>
<tr><td>Onsite Infrastructure: Yes</td><td>Local Preservation: Yes</td><td>Local Access Portal: Yes</td></tr>
<tr><td colspan="3">Tools Used Onsite: Heritrix, Wayback, NutchWAX (in the event Archive-It is used but with local access also via HL WAX portal updates to Wayback and deployment of SOLR are likely necessary).</td></tr>
<tr><td colspan="3">Tools/Infrastructure Challenges:
• Local tools are versions behind. Considering moving core functionality (crawling and QA) to Archive-It, freeing developers to deliver new modes of access to the web archives for Harvard and other</td></tr>
</table>

researchers • Integrating WARC files from other sources (e.g. hard drive, personal web archives)
Preservation Challenges: Off-the-shelf tools don't work well with WARCs (e.g. virus check); replay when the preserved site includes non-renderable content (e.g. newer browsers/flash)
Collection Development Challenges: No coordination/communication to find out what others are collecting. Internally– overlap with Archives and schools; externally– overlap of themed collections.
Other:

Institution Name: Library of Congress (LC)	www.loc.gov/websites/collections/	
The Library of Congress (LC) started web archiving in 2000 and has 763 TB of content. It contracts with the Internet Archive (not Archive-It) to crawl using a seed URL list they provide, and crawl reports are accessed via a password-protected area of archive.org. LC content is not pushed into the Internet Archive's Wayback collection at archive.org/web for copyright and embargo reasons. WARC files are transferred to the LC (using BagIt) for local preservation and for access via a local Wayback installation.		
Main Use Cases: • Selective (versus the domain approach of other national libraries collecting all .fr domains, for e.g.) • US and foreign government, political commentary, religious organizations, media, advocacy groups, etc. • Thematic, event based web archives, based on subject expertise of Library Services and Law Library staff		
Collection Development Location: Library	Additional Info: Staff working on web archiving are primarily in Library Services, but some Law Library staff select content also, and OCIO staff support the IT and infrastructure side of the activity.	
Membership and Collaborations: • IIPC (a founding member) • NDSA/NDIIPP • SAA	Details: • IIPC/NDIIPP partners– End of Term Archive, US Government - collaborative collection development with Internet Archive, California Digital Library, Government Printing Office, University of North Texas, Harvard • University North Carolina automated vocabularies project 2012 • French National Library and Archive-It – Ukraine conflict 2014, North Africa and Middle East 2011, Jasmine revolution 2011 • California Digital Library and Internet Archive – Hurricane Katrina and Rita Web 2005 • Virginia Tech, Archive-It, Diet Library – Japanese earthquake 2011	
Funding: Federal funds		
Staffing: Currently 4 FTEs on the Web Archiving team, 1 FTE developer working currently on Digiboard; 1 cataloger working part-time (2 days/week) on web archiving. Other IT staff involved are estimated equal to 1 FTE.		
Integration of Web Archives with Other Collections: Details and Concerns:	Internally: Yes. ILS points to MARC record at collection level. Each website has MODS with controlled	Externally:

	names, subject headings for indexing/searching. MODSs record data is searchable alongside other Library materials via the loc.gov website.	
External Infrastructure: Yes. (Internet Archive).	External Preservation: Yes. (Internet Archive.)	External Access Portal: No
Onsite Infrastructure: Yes	Local Preservation: Yes	Local Access Portal: Yes
Tools Used Onsite: DigiBoard (in-house tool manages selection, permissions), Heritrix, Wayback, SOLR		
Tools/Infrastructure Challenges: • How to more efficiently index/store and serve up content; large collection/large index (CDX files ~4TB) • Lack of technical resources and time to work with tools (such as IBM BigSheets) and derived datasets (such as WANE, WAT) • DigiBoard:development resources		
Preservation Challenges: • Transfer of WARC files from Internet Archive via Internet2 takes time • Infrastructure maintenance (disk/tape) • WARC files not always readable across platforms (e.g. extra carriage returns). De-duplication introduces complexity.		
Collection Development Challenges: No legal deposit mandate for web. Often several institutions are crawling a site at the same time, but the LC gives written notice ahead of time and so is sometimes "blamed" for crawls when experienced as disruptive.		
Other: LC is interested in tools or collaborations that will increase awareness and use of its web archiving collections (such as broader adoption of Memento Time Travel, or a registry site or central hub).		

Institution Name: Massachusetts Institute of Technology (MIT)	
Massachusetts Institute of Technology (MIT) Archives and Special Collections is currently evaluating web archiving tools and considering an Archive-It account. Its web captures thus far are publicly accessible only in the on-site reading room.	
Main Use Cases: • Institutional archives (the MIT.edu domain) including handbooks and catalogs not available in print form • Grey literature (e.g. conference proceedings for sites hosted at MIT) • Extension of existing special collections • Topical website curation by librarians (CAUSEWAY and CCWA)	
Collection Development Location: Institute Archives and Special Collections (IASC)	Additional Info:
Membership and	Details:

Collaborations: • NDSA/NDIIP • SAA • Ivy Plus • ArchiveSpace • Archivematica	• Ivy Plus - collaborative collection development at Archive-It – Contemporary Composers Web Archive (CCWA) and Collaborative Architecture, Urbanism, and Sustainability web Archive (CAUSEWAY) pilot web collections • Archivematica (processing WARC files) and Archivespace (metadata for describing crawls).	
Funding:		
Staffing: ½ FTE library fellow. Identified as a priority area for funding.		
Integration of Web Archives with Other Collections: Details and Concerns:	Internally: Yes, Currently websites are described with their administrative collections and noted in finding aid.	Externally: No
External Infrastructure: No. (Unless/until move to AIT)	External Preservation: No. (Unless/until move to AIT)	External Access Portal: No. (Unless/until move to AIT)
Onsite Infrastructure: Yes	Local Preservation: Yes	Local Access Portal: No
Tools Used Onsite: MIT library fellow tested a variety of capture tools in the MIT Digital Sustainability Lab. The goal was to identify a tool(s) that would allow the archivist to capture websites without extensive support from IT. No single tool has proven to offer a complete solution for domain capture; most sueful are Webrecorder (targeted captures), Wget 1.14 or later with WARC output (for larger portions of mit.edu). For evaluating websites, they are testing the following tools: Archiveready.com, Builtwith.com and Wappalyzer browser extension. The onsite playback tool in use is Web Archive Player.		
Tools/Infrastructure Challenges: Managing OSS for web archiving locally is a challenge because of the time commitment and skills required.		
Preservation Challenges: Size of WARC files, making decisions about deduplication, changes to browsers and the effect on playback overtime, preserving non-text content for playback overtime		
Collection Development Challenges: The IASC feels confident in current collection development scope because of its relatively narrow focus on the mit.edu domain. If the Libraries' general collections develop local topical website collections, we will need to have discussions regarding collecting scope and responsibilities for description across Libraries.		
Other:		

Institution Name: National and University Library of Iceland	http://vefsafn.is
National and University Library of Iceland maintains its own in-house service. It does 3 domain crawls a year, getting the seedlist from the .is registrar, and has 67 TB (about 3.3 billion URIs). The library also uses a curated list of non .is domains containing relevant material. The library does a weekly crawl of sites of interest (political, news etc.) and runs a constant crawl on the RSS feeds of some websites. Occasional topical crawls are conducted, mostly related to elections, but budgetary concerns usually limit these. Large-scale / data mining access has been limited (2 occasions).	

<table>
<tr><td colspan="3">Main Use Cases:
• National domain
• Iceland related material (using the seedlist from the .is registrar)
• Occasional topical crawls, mostly related to elections. Budgetary concerns usually limit these.</td></tr>
<tr><td colspan="2">Collection Development Location: Legal deposits/ IT</td><td>Additional Info:</td></tr>
<tr><td colspan="2">Membership and Collaborations: IIPC</td><td>Details:</td></tr>
<tr><td colspan="3">Funding: Operating budget</td></tr>
<tr><td colspan="3">Staffing: Less than one FTE in total. IT estimated at 1/3-1/2 FTE and legal deposit 1/4 FTE.</td></tr>
<tr><td>Integration of Web Archives with Other Collections:

Details and Concerns:</td><td>Internally: No.
Not currently.</td><td>Externally: No</td></tr>
<tr><td>External Infrastructure: Yes</td><td>External Preservation: No</td><td>External Access Portal: No</td></tr>
<tr><td>Onsite Infrastructure: Yes</td><td>Local Preservation: Yes.
WARC files are stored on mirrored storage arrays.
Tertiary copies are stored offline on HDDs.</td><td>Local Access Portal: Yes</td></tr>
<tr><td colspan="3">Tools Used Onsite: Heritrix, OpenWayback</td></tr>
<tr><td colspan="3">Tools/Infrastructure Challenges: Local tools are the latest versions. Any challenges are due to lack of funding.</td></tr>
<tr><td colspan="3">Preservation Challenges: We are acutely aware that our current setup is a bit “low tech” and would benefit from a higher-level management system. But that basically gets us right back to funding.</td></tr>
<tr><td colspan="3">Collection Development Challenges: For curated collections, we find that we lack the resources to properly engage in them. But, again, this is mostly a funding issue.</td></tr>
<tr><td colspan="3">Other:</td></tr>
</table>

<table>
<tr><td>Institution Name: National Library of Finland</td><td>http://webarchive.nationallibrary.fi/
(only index)</td></tr>
<tr><td colspan="2">National Library of Finland has been web archiving since 2006 and has now over 80 TB of content. Because of the copyright law, access to the archive is only in a few workstations. From November 2015 contents of the archive (WARC files in METS packages) are being sent to the national preservation system (http://www.kdk.fi/index.php/en/long-term-preservation). Access to Finnish Web Archive is available only via local access legal deposit terminals (due to the Copyright Law).</td></tr>
<tr><td colspan="2">Main Use Cases:
• Annual web harvest. Large Finnish domain harvesting is conducted at least once a year with an automatic web crawler. The goal is to harvest as much online material as possible using Internet</td></tr>
</table>

<table>
<tr><td colspan="3">domains such as 'fi' and 'ax'. Also other domestic webpages are archived extensively.
• Daily harvests if about 40 Finnish online news sites and weekly harvests of over 200 online journal sites etc.
• Thematic harvests of some particular subjects or topical issues: important national and state affairs, events that are in danger of disappearing from the internet soon after the event, unexpected events with global importance, harvests that are conducted in cooperation with other organizations.
• Institutional repositories etc. are mainly harvested using OAI-PMH (to get metadata and the whole collections).
• The National Library may also request an online publisher to deposit materials, if automatic harvesting is not possible.
• The last two cases are not included in Web Archive, but a separate Electronic Legal Deposit Archive (a D-Space repository).</td></tr>
<tr><td>Collection Development Location: cooperation with academic researchers</td><td colspan="2">Additional Info: Web harvesting and the Web Archive are maintained by the National Library of Finland. National Digital Library's Preservation Service is maintained by CSC, IT Center for Science Ltd, which is a non-profit, state-owned company administered by the Ministry of Education and Culture.</td></tr>
<tr><td>Membership and Collaborations: IIPC</td><td colspan="2">Details:</td></tr>
<tr><td colspan="3">Funding: Funded by the Ministry of Education and Culture</td></tr>
<tr><td colspan="3">Staffing: Approximately 3 persons/week in web archiving & related issues</td></tr>
<tr><td>Integration of Web Archives with Other Collections:

Details and Concerns:</td><td>Internally: Yes.
Other collections available via legal deposit terminals:
• Electronic Legal Deposit Archive
• Finnish Radio and Television Archive
• Digitized collections with copyright restrictions</td><td>Externally: No</td></tr>
<tr><td>External Infrastructure:</td><td>External Preservation: National Digital Library's Digital Preservation System, maintained by CSC</td><td>External Access Portal:</td></tr>
<tr><td>Onsite Infrastructure: Yes</td><td>Local Preservation: Short/Mid time preservation (tape secured)</td><td>Local Access Portal: Yes (only through legal deposit terminals)</td></tr>
<tr><td colspan="3">Tools Used Onsite: Heritrix, Wayback</td></tr>
<tr><td colspan="3">Tools/Infrastructure Challenges:</td></tr>
<tr><td colspan="3">Preservation Challenges:</td></tr>
<tr><td colspan="3">Collection Development Challenges: To increase the cooperation with academic researchers and make data and text mining of Finnish Web Archive possible.</td></tr>
<tr><td colspan="3">Other:</td></tr>
</table>

<table>
<tr><td>Institution Name: National Library of France (Bibliothèque nationale de France– BnF)</td><td colspan="2">http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html
http://archivesinternet.bnf.fr (only on premises)</td></tr>
<tr><td colspan="3">After a series of experimentations, the Bibliothèque nationale de France (BnF) started archiving the web in 2006 under the terms of the French Heritage Law which was extended to the Internet. As of the end of 2014, BnF holds 23,6 billions URL and 567 TB of data.
Our services are all run internally from selection and harvest to access and preservation.</td></tr>
<tr><td colspan="3">Main Use Cases:
• Legal deposit (yearly broad harvest of French websites)
• Thematic and event (regular harvests of selected websites)
• PDFs of local newspapers</td></tr>
<tr><td>Collection Development Location:
Websites for thematic harvests are selected inline with the Library main collection development policy.</td><td colspan="2">Additional Info: BnF also gets contributions from regional libraries, researchers, associations and public institutions.</td></tr>
<tr><td>Membership and Collaborations:
• IIPC
• NetarchiveSuite
• .fr registry</td><td colspan="2">Details:
• IIPC: to share the use and development of common tools, techniques and standards; to share collection policies and build international collections.
• NetarchiveSuite: to share the use and development of NetarchiveSuite (a tool to plan, schedule and run web harvests with Heritrix)</td></tr>
<tr><td colspan="3">Funding: Public funding.</td></tr>
<tr><td colspan="3">Staffing: 11 dedicated personnel (7 digital librarians – 2 are also in charge ebooks legal deposit, 4 IT), lots of contributors on selection.</td></tr>
<tr><td>Integration of Web Archives with Other Collections:

Details and Concerns:</td><td>Internally: Yes/No
Some selected websites are referenced in the Library Catalog.</td><td>Externally: No</td></tr>
<tr><td>External Infrastructure: No</td><td>External Preservation: No</td><td>External Access Portal: Yes.
We give remote access to the web archives in regional libraries sharing legal deposit with BnF.</td></tr>
<tr><td>Onsite Infrastructure: Yes</td><td>Local Preservation: Yes</td><td>Local Access Portal: Yes</td></tr>
<tr><td colspan="3">Tools Used Onsite: BCWeb (curator tool), NetarchiveSuite, Heritrix, OpenWayback, Solr, nas-preload (prepare broad harvest), nas-qual (generate stats on crawls)</td></tr>
<tr><td colspan="3">Tools/Infrastructure Challenges:
• Build and scale access and data mining tools.
• Update to Heritrix 3 (stats and preservation workflows are tightly linked to H1 configurations, logs and reports.</td></tr>
</table>

Preservation Challenges: Keep up with data structures
Collection Development Challenges:
Other:

Institution Name: National Library of Germany (Deutsche Nationalbibliothek–DNB)	http://www.dnb.de/EN/Netzpublikationen/Webarchiv/webarchiv_node.html
The German National Library (DNB) has been web archiving since 2012. The actual crawling and hosting is outsourced to the German company oia in Düsseldorf. Access is given exclusively in the reading rooms in Frankfurt and Leipzig. WARC files are additionally stored in the preservation system of DNB. One crawl of the German top level domain .de was done in 2014 with the Internet Memory Foundation. Its holdings are estimated at 10 TB of the selective crawls and 120 TB of the one .de domain crawl	
Main Use Cases: • Collection according to the legal deposit (all German publications) • Topic collections • Event crawls • .de domain crawl (only one)	
Collection Development Location: Library	Additional Info: Selection is done by librarians in Frankfurt and Leipzig with a tool by oia. The crawling is then done in Düsseldorf.
Membership and Collaborations: IIPC	Details: Several collaborations with other German institutions doing web archiving, e. g. Bavarian State Library, archives of the German political parties.
Funding: Internal (the national library is a federal institution)	
Staffing: No dedicated personnel, about 2 persons full in summary (plus people of the company oia).	

Integration of Web Archives with Other Collections: Details and Concerns:	Internally: Yes. Every website has a catalogue entry.	Externally: Yes. The catalogue of DNB is fully accessible on the web (all metadata). But the content of the web archive is only accessible onsite.
External Infrastructure: Yes	External Preservation: No. (just hosting for giving access)	External Access Portal: No
Onsite Infrastructure: Yes	Local Preservation: Yes	Local Access Portal: Yes

Tools Used Onsite: Tool OWA by oia
Tools/Infrastructure Challenges: • OWA is a specific Windows tool to input data for oia • WARC files for preservation are just stored without access
Preservation Challenges: The integration of the WARC files in the preservation is still in discussion.
Collection Development Challenges: New pages
Other:

<table>
<tr><td>Name: National Library of the Netherlands (Koninklijke Bibliotheek–KB)</td><td colspan="2">URL: https://www.kb.nl/bronnen-zoekwijzers/databanken-mede-gemaakt-door-de-kb/webarchief-kb
https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving</td></tr>
<tr><td colspan="3">About: The Koninklijke Bibliotheek (KB) has been web archiving since 2007 and has 17 TB of content. As per November 2015, over 10,000 websites have been selected.</td></tr>
<tr><td colspan="3">Main Use Cases: The KB has decided on a selective approach, because this is more in line with the remit of the KB, the available resources and the chosen legal approach. The KB's selection is based on the KB's collection policy. Within this framework, a cross section of the Dutch domain is selected for archiving. Primarily, we select websites with cultural and academic content, but we do include websites which exemplify present trends on the Dutch domain. Finally, we take into account relevance for Dutch society and popularity. Those (Dutch) sites which are consulted most frequently by Dutch internet users and/or have a high ranking, are prioritized for archiving.</td></tr>
<tr><td>Collection Development Location: Division Collections, collection specialists</td><td colspan="2">Additional Info:</td></tr>
<tr><td>Membership and Collaborations:
• IIPC
• A number of national institutions
• A Research Infrastructure for the Study of Archived Web Materials (RESAW)</td><td colspan="2">Details:</td></tr>
<tr><td colspan="3">Funding: Internal</td></tr>
<tr><td colspan="3">Staffing: All the “part-time” people add up to two persons/week.</td></tr>
<tr><td>Integration of Web Archives with Other Collections:

Details and Concerns:</td><td>Internally: No</td><td>Externally: No.
The KB web archive is available onsite (see: https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources/web-archiving/legal-issues)</td></tr>
<tr><td>External Infrastructure: Yes</td><td>External Preservation: Yes</td><td>External Access Portal: Yes</td></tr>
<tr><td>Onsite Infrastructure: Yes</td><td>Local Preservation: No</td><td>Local Access Portal: Yes</td></tr>
<tr><td colspan="3">Tools Used Onsite: Heritrix ; WayBackMachine ; custom-made e-mail software</td></tr>
<tr><td colspan="3">Tools/Infrastructure Challenges: upgrading Heretrix and local tools</td></tr>
<tr><td colspan="3">Preservation Challenges: Top priority for quality assessment</td></tr>
<tr><td colspan="3">Collection Development Challenges: National coordination of collection development</td></tr>
</table>

Other:

Institution Name: National Library of New Zealand (NLNZ)	http://www.natlib.govt.nz
National Library of New Zealand (NLNZ) has been web archiving in earnest since 2007 (actually 1999 via HTTrack) and has 45 TB of content. The Web collections are split into two areas, selective and whole-of-domain. The selective harvest is presently undertaken via Web Curator Tool (WCT), a joint development between National Library of New Zealand and The British Library. The selective harvest is approximately 4.5 TB and spans 22,000 website crawls. The whole of domain is a contracted work, undertaken by the Internet Archive. We have completed 4 harvests, 2008, 2010, 2013 and 2015. The next harvest is being planned for 2016. Each harvest returns between 8 to 15tb of data.	
Main Use Cases: • Institutional archives, e.g. government sites (websites and blogs) • Periodic crawling of the .nz domain • Periodic crawling of the .com, .net, and .org domains where New Zealand hosting is clear • Thematic or topical web archives • HTML serials	
Collection Development Location: National Library of New Zealand	Additional Info: Technical stack is run by the Department of Internal Affairs (the Library's host department) centralised IT service, technical oversight is with Preservation Research & Consultancy, and content decisions are with the Digital Collection Strategy and Collection Development (Legal Deposit) teams.
Membership and Collaborations: • IIPC • PREMIS • PARBICA	Details: • Collaborative crawls with IIPC (e.g. Winter Olympics, WWI) • Discussions with other Pacific institutions regarding web archiving and possible collaboration • NLNZ is currently Chair of PREMIS
Funding: Internal	
Staffing: Collectors: "4 persons/week" for selection, harvesting and other library tasks; technical: 0.5 persons/ month for technical parts including whole of domain harvests.	

Integration of Web Archives with Other Collections: Details and Concerns: Current approach to access is bibliographic. We need to move towards more flexible indexing and access mechanisms including item, group and collection level access.	Internally: Selective – yes Whole of domain - no	Externally: Selective - no Whole of domain – no Legal issues related to re-publishing, liability, hate and porn material.

<table>
<tr><td>External Infrastructure: Yes.
Government mandated data centre (private cloud)</td><td>External Preservation: No.
Managed by National Library's Preservation Research & Consultancy team.</td><td>External Access Portal:
Selective – yes
• Catalogued in externally accessible ILS
Whole of domain – no
• As above</td></tr>
<tr><td>Onsite Infrastructure: No.
Government mandates data centre (private cloud)</td><td>Local Preservation: Yes.
Managed by National Library's Preservation Research & Consultancy team.</td><td>Local Access Portal:
Selective– yes
Whole of domain– no
• Working on implementing Open WayBack</td></tr>
<tr><td colspan="3">Tools Used Onsite: Wayback, Openwayback, Web Curator Tool, Rosetta, ARC and WARC viewers</td></tr>
<tr><td colspan="3">Tools/Infrastructure Challenges:
• WCT is 10 years old now. While development has continued under both The British Library and National Library of New Zealand the point has come where technical deficiency suggests that the tool is at end-of-life as web technologies evolve.
• We will need to assess the cost/benefit of upgrading or migrating to a new tool or suite of tools.
• Being clear that we are actually getting what we want when we crawl the web. It is disconcerting to hear that different institutions using different tools might be picking up different content. In that context how do we make choices about the tools/stack we want to use? Maybe some work on what is the ideal output from crawling would be useful, against which tools could be assessed.
• We maintain our preservation capability in house, and would not entrust a third party to undertake the preservation aspect of content care.
• Key challenge for us is social media for both technical and legal reasons.
• We lack specialist resource to advance our web harvesting and preservation activities.</td></tr>
<tr><td colspan="3">Preservation Challenges: Web preservation is at a very immature stage – we are not ingesting the whole of domain content into the Rosetta preservation system, and the selective items are (currently) maintained in their parent WARC/ARC containers (i.e. we do not explode the WARC/ARC and manage the items at the otherwise typical file level). We are most the way through undertaking an ARC to WARC migration for all the ARC containers we have. This will result in some 3rd generation migrations in the container space, HTTrack to ARC to WARC.</td></tr>
<tr><td colspan="3">Collection Development Challenges:
• Legal issues around Legal Deposit scope vs international law
• Social media - not collecting social media is creating gaps and an imbalance in the web collections, because we're not reflecting what's on the web.</td></tr>
<tr><td colspan="3">Other:</td></tr>
</table>

<table>
<tr><td>Institution Name: National Library of Spain (Biblioteca Nacional de España– BNE)</td><td>The BNE doesn't give access to its web archive collections yet.</td></tr>
<tr><td colspan="2">The National Library of Spain (BNE) has been web archiving since 2009 and has 117 TB of content. From 2009 to 2013, the Library did 8 domain crawls and 2 selective crawls of the Spanish web with Internet Archive. This collection is 111 Tb. It was delivered to the Library servers by the beginning of 2015. Since 2014 the BNE has been testing and doing some selective crawls with NetarchiveSuite. This collection</td></tr>
</table>

(that includes emergency –of websites at risk-, events and thematic crawls) is now 6 Tb. Thanks to a collaboration agreement with the public entity Red.es, we are improving our infrastructure to widen our selective crawls and launch our first domain crawl (with our own resources) in the first months of 2016.		
Main Use Cases: We don't have use cases yet, as we don't give access so far.		
Collection Development Location: Library	Additional Info:	
Membership and Collaborations: • IIPC • ISO TC 46/SC 8 • Regional Libraries in Spain	Details:	
Funding: Internal. Agreement with the public entity Red.es.		
Staffing: 5 people/week		
Integration of Web Archives with Other Collections: Details and Concerns:	Internally: No	Externally: Yes. If we consider the collaboration with Archive-It an "integration"…
External Infrastructure: No	External Preservation:	External Access Portal:
Onsite Infrastructure: Yes	Local Preservation: We are planning it.	Local Access Portal: We are planning it.
Tools Used Onsite:		
Tools/Infrastructure Challenges: • Access • Full text search • Preservation • Deposit of online publications not freely available on internet		
Preservation Challenges: To build the whole environment, which is only planned in its main lines.		
Collection Development Challenges: We are working with the depository libraries (in the framework of the legal deposit) for them to manage their own web archiving collections.		
Other:		

Institution Name: National Museum of Women in the Arts	https://archive-it.org/organizations/587
National Museum of Women in the Arts Betty Boyd Dettre Library and Research Center has been web archiving since 2011 and has 798 GB of content. NMWA has only used Archive-It as a tool.	
Main Use Cases: • Institutional archives (websites and social media) • Archiving websites of new media women artists • Archiving websites of important groups related to women in the arts	
Collection Development Location: Library and Archives	Additional Info:
Membership and Collaborations:	Details: Have worked with ARLIS/NA Artist Files' Special

• ARLIS/NA • Archive-It	Interest Group on looking at what web archiving means for art libraries.

Funding: Internally funded		
Staffing: No dedicated personnel. Part of the library director's duties, estimated at 1/16th of a person (2-5 hours).		
Integration of Web Archives with Other Collections: Details and Concerns:	Internally: No	Externally: No
External Infrastructure: Yes	External Preservation: Yes	External Access Portal: Yes
Onsite Infrastructure: No	Local Preservation: No	Local Access Portal: No
Tools Used Onsite: Archive-It		
Tools/Infrastructure Challenges: Lack of time and money to do more with collection and add other collections		
Preservation Challenges:		
Collection Development Challenges: Few art libraries doing this—there would be a great deal of duplication if they did. We need to figure out a collaborative collecting effort that can utilize those with robust budgets and staffs as well as those with fewer resources. There is enough out there that everyone should and could be doing something.		
Other: It needs to become standard for research institutions to archive their own web activities and then do collaborative collection development since these are not physical collections. Also need better ways for researchers to access information across collections.		

Institution Name: New York Art Resources Consortium	www.nyarc.org/webarchive
The New York Art Resources Consortium (NYARC), consisting of the research libraries of the Brooklyn Museum, The Frick Collection, and The Museum of Modern Art, has been web archiving since 2014 and has approximately 1.2 TB of content. NYARC primarily utilizes the Archive-It subscription service for web archiving (with supplemental captures via Hanzo Archives).	
Main Use Cases: • Institutional web archive collections, consisting of the websites of NYARC, The Brooklyn Museum, The Frick Collection, The Museum of Modern Art, MoMA PS1, and institutional blog content • Topical/thematic collections: 1. Born-digital catalogues raisonnés 2. Artists' websites 3. New York City gallery and art dealer websites 4. Auction house websites and embedded auction catalogs 5. Websites related to the restitution of lost or looted art *6.* Born-digital art resources in danger of impermanence, such as PDF publications no longer published in print	
Collection Development Location: Libraries	Additional Info: Our collection development policy for websites was crafted for use by the three consortial libraries.

<table>
<tr><td>Membership and Collaborations:
• Columbia University Libraries (CUL) Web Resources Collection Program
• Rhizome (Webrecorder.io tool)
• Old Dominion University (ODU)
• Archive-It
• ARLIS/NA
• NDSA
• OCLC Research Libraries Partner
• DLf
• NDSR-NY
• METRO
• SAA</td><td colspan="2">Details:
• Collaborations with CUL have been ongoing and are informal. Meetings take place 3-4 times per year and involve information sharing and collaborative collection development discussions, tool and workflow development, etc. Similarly, the NYARC group collaborates with Rhizome and ODU colleagues for information sharing and tool testing to further development of our respective programs.
• Archive-It membership began in early 2014 (an initial pilot study in partnership with Archive-It was conducted in 2010).
• Membership in ARLIS/NA, NDSA (includes involvement in various working groups), OCLC Research Libraries Partner Group, and DLF – all lend themselves to collaborative discussions with these overlapping communities.
• NYARC hosted an NDSR-NY resident for the 2014-2015 term and participates on the advisory board.
• SAA: participation is not formalized through membership (archives staff are SAA members; web archiving staff is not); participation primarily via SAA Web Archiving RT.</td></tr>
<tr><td colspan="3">Funding: Mellon grant for initial two-year period; institutional funding post-grant</td></tr>
<tr><td colspan="3">Staffing: 1 FTE for project management; 4 part-time staff (each works one day per week on QA); staff in other departments actively collaborate on program elements.</td></tr>
<tr><td>Integration of Web Archives with Other Collections:

Details and Concerns:</td><td>Internally: Yes.
Arcade (NYARC's consortial OPAC); NYARC Discovery (Archive-It integration with Primo)</td><td>Externally: Yes.
Collections publicly accessible via Archive-It and Wayback Machine; Arcade bibliographic records available via Arcade and WorldCat</td></tr>
<tr><td>External Infrastructure: Yes</td><td>External Preservation: Yes</td><td>External Access Portal: Yes</td></tr>
<tr><td>Onsite Infrastructure: Yes (local database)</td><td>Local Preservation: No</td><td>Local Access Portal: No (although provided locally via web access portals)</td></tr>
<tr><td colspan="3">Tools Used Onsite: Local FileMaker Pro database for administrative data tracking; web-based access to Archive-It</td></tr>
<tr><td colspan="3">Tools/Infrastructure Challenges:
• Great need for tools to automate QA to improve efficiencies
• Current tools and infrastructure remain expensive, esp. given the needed scale
• We have achieved the integration of our WARC files from other sources into our Archive-It account, but playback of those files with Wayback Machine is often not possible</td></tr>
<tr><td colspan="3">Preservation Challenges: While we have integrated an automated DuraCloud backup of our Archive-It collections, the under-development of tools, processes, and standards to reliably create and package standard preservation metadata for web archives remains a challenge. Non-renderable formats from WARCs, such as Flash files, will remain a challenge.</td></tr>
<tr><td colspan="3">Collection Development Challenges: Presently few tools exist to improve awareness of what others are collecting and overlap of collecting scopes is likely. Duplication is not ideal, but it remains difficult to</td></tr>
</table>

determine to what extent other institutions are committed to effectively capturing websites, conducting a high level of quality assurance, providing sufficient access points, and ensuring ongoing preservation of WARCs. The universe of content is vast and collaboration is needed to web archive even a small portion of the content we'd like to preserve.
Other:

Institution Name: Rhizome	http://rhizome.org/art/ http://rhizome.org/ http://webenact.rhizome.org/
Rhizome has been web archiving since 1998 and has 95 GB of content. Rhizome's ArtBase contains 2000+ pieces of internet art. We are focused on fidelity rather than collection size. Rhizome is a born-digital arts organization founded 1996. Rhizome's preservation program is using the highly dynamic field of internet art to develop new conservation tools and practices.	
Main Use Cases: • Providing access to an archive of historical and contemporary internet art and digital culture • Researching and developing new digital preservation methods	
Collection Development Location: ln-house, curatorial and research-driven	Additional Info: We are dealing with a web that is very much not document-based or even URL-based. We need to regard the web as a software delivery mechanism and records of performances. Emulation plays a key part in our preservation activities. Preservation for Rhizome means being able to recall computational performances.
Membership and Collaborations: • University of Freiburg, Germany • University of Yale • Vilem Flusser Archive • Archive for German Literature Marbach • Individual artists, like Cory Arcangel	Details: • Webrecorder.io and oldweb.today are projects created in partnership with Rhizome • Emulation + Web-Archiving research with University of Freiburg and Yale • Rapid Response Archiving, highly integrated into curatorial direction and current discourses
Funding: Continuous and stable funding from arts organizations and NEH.	
Staffing: 1 p 80% conservation program lead, 1 p 100% NDSR resident	

Integration of Web Archives with Other Collections: Via http://oldweb.today other archives are seamlessly integrated (Memento aggregator) Details and Concerns:	Internally: Rhizome doesn't have "internally" :)	Externally: Full list at http://oldweb.today
External Infrastructure: Yes. Storage at partner	External Preservation: NA	External Access Portal: NA

institution New Musem in New York, cloud storage		
Onsite Infrastructure:	Local Preservation: Via own staff	Local Access Portal: pywb-based
Tools Used Onsite: • webrecorder (own tool) • pywb (own tool) • EaaS (own tool) • oldweb.today (own tool)		
Tools/Infrastructure Challenges: • Developing our own tools is challenging resource-wise. • Esp. Developing meaningful automation is challenging		
Preservation Challenges: Full integration of Emulation and Web-Archiving to for example be able to play sound from legacy Flash embeds.		
Collection Development Challenges: Workflows		
Other:		

Institution Name: Smithsonian Institution Archives	
Smithsonian Institution Archives (SIA) has crawled 1.7 TB of content since September 2012. The Smithsonian has more than 400 websites and blogs and more than 600 social media accounts (Twitter, Facebook, Instagram, and YouTube) across the Institution. SIA uses Archive-It for most captures but is testing newer tools (such as WebRecorder) for capturing problematic social media sites. Prior to 2012 they had an in-house instance of Heritrix and also received sets of files from webmasters' servers (which they still receive on occasion), although sites not crawled with Archive-It are not readily available for access. They retrieve WARC files from Archive-It every week for preservation.	
Main Use Cases: Institutional archives. The Smithsonian Institution Archives collects, preserves and makes available the official records of the Smithsonian's nineteen museums, nine research center, and the National Zoo that document Smithsonian staff, artifacts, benefactors, events, exhibits, buildings, and research.	
Collection Development Location: Institutional Archives	Additional Info:
Membership and Collaborations: • Federal web archiving working group • Archive-It • NDSA • SAA	Details: • IIPC considered too expensive • Is sharing their "WARC-grabber" tool with National Library of Medicine and Government Printing Office for evaluation purposes • Collaborations not a priority – cited insufficient resources/time for projects outside their core missions, and procedures that differ from other institutions
Funding: Smithsonian year-end funding	

Staffing: No dedicated personnel. Combined, staff hours ~ 1/3rd person/week, plus seasonal interns and volunteers.		
Integration of Web Archives with Other Collections: Details and Concerns:	Internally: Yes. SIA creates MARC records for catalogs plus online finding aids in EAD format with <dao> tag to take users directly to the Archive-It crawl. Processes for accessioning and description of archived websites are designed to be as similar as possible to those processes for non-web accessions.	Externally: No
External Infrastructure: Yes	External Preservation: No	External Access Portal: Archive-It
Onsite Infrastructure: Yes	Local Preservation: Yes. (weekly WARCs from Archive-It)	Local Access Portal: No
Tools Used Onsite: SharePoint-based registry of all websites and social media accounts maintained by the Smithsonian, including when they were captured and tool used; SIA-developed tool to download WARC files from Archive-It, WebRecorder; and TAGS (Twitter Archiving Google Sheet).		
Tools/Infrastructure Challenges: Dynamic content and social media sites can require specialized tools (outside of Archive-It) as their set-up tends to change often resulting in incomplete capture and playback.		
Preservation Challenges: Maintaining the WARC files over time (both storage and format); providing access to content that was not captured using Archive-It (considering local copy of Wayback, Web Archive Player or other tool); created their own tool to retrieve WARC files from Archive-It (checks dates and checksums to prevent downloading duplicate content) in order to streamline the download process.		
Collection Development Challenges: Attempting to capture all of the websites and social media accounts as often as they ideally should be, especially as offices across the Smithsonian continue to add new websites and social media accounts; finding tools to appropriately capture all social media types, including those that are just emerging. Smithsonian policy requires that all social media accounts must link to the Institution's privacy statement before they can be captured by the Archives, but not all social media coordinators follow through.		
Other: Anything not crawled by Archive-It is not easily accessible to the public at this point in time. Ultimately, the Archives would like a solution for making all captured websites and other web-based material more accessible.		

Institution Name: Stanford University Libraries (SUL)	Stanford Web Archive Portal (https://swap.stanford.edu/) Archive-It (https://archiveit.org/explore?q=stanford)

<table>
<tr><td colspan="3">SUL has been web archiving since 2007 and has 40+ TB of content. We are pursuing a hybrid architecture, using Archive-It for capture and building infrastructure for local preservation, discovery, and access. We are interested in fostering a renewed collaborative, community-source approach to development of web archiving tools and APIs.</td></tr>
<tr><td colspan="3">Main Use Cases:
• To document Stanford University and affiliated events, previously print published material as well as new publications with no print analogue
• To facilitate research and teaching, especially by capturing at-risk web materials of scholarly value
• To collect complementary (and otherwise absent) materials for special collections
• To make Federal Depository Library responsibilities easier to complete
• Data management services for student/faculty projects, and preservation of license-free web content cited in documents submitted through Stanford Digital Repository Online Deposit Form
• Transparent documentation of legal policies and affairs as they change over time</td></tr>
<tr><td>Collection Development Location: Distributed; current web content collecting selectors are located in University Archives, Humanities and Social Sciences, East Asia Library, Hoover Institution, and Graduate School of Business.</td><td colspan="2">Additional Info: Service is coordinated out of the Digital Library Systems and Services (DLSS) group. Stakeholders are broader than SUL; also includes campus webmasters, IT security, communications, general counsel, and researchers. These groups also inform collecting.</td></tr>
<tr><td>Membership and Collaborations:
• Archive-It Partners
• IIPC
• LOCKSS Alliance
• NDSA
• SAA</td><td colspan="2">Details:
• Starting in 2016, we'll be working with IA, UNT, Rutgers, and other interested parties on community building for web archiving tool and API development.
• We're exploring merging a couple of legacy collections and prospective collaborative collection development with UCLA.</td></tr>
<tr><td colspan="3">Funding: DLSS staff on four-year term funding 2013-2017. Archive-It accounts and commensurate local storage funded through collecting budgets. Two-year IMLS grant 2016-2018 will fund a few months of DLSS staff time.</td></tr>
<tr><td colspan="3">Staffing: About 2.5 FTE, most of that represented by 1 FTE service manager and 1 FTE developer and the remainder represented by fractional time committed by curators and metadata staff.</td></tr>
<tr><td>Integration of Web Archives with Other Collections:

Details and Concerns: We plan to treat web archives as first-class digital objects in our integrated discovery</td><td>Internally: Yes.
We expect to have support for web archive records in our integrated discovery environment in early 2016. Individual records will be at the website-level, feature</td><td>Externally: Yes.
We have legacy finding aids on OAC describing the SU websites collection, because we don't have a local hosting environment for finding aids. The granularity at which we will continue to maintain this external finding aid, or rely</td></tr>
</table>

environment, SearchWorks.	a thumbnail filmstrip of major versions of the seed URL, and link off to an Open Wayback access point on SWAP.	on finding aids for discovery of web archives in general, is TBD.
External Infrastructure: Yes	External Preservation: Yes. (i.e., as provided by Archive-It)	External Access Portal: Yes. (Archive-It)
Onsite Infrastructure: Yes	Local Preservation: Yes	Local Access Portal: Yes
Tools Used Onsite: We use Heritrix and wget for (limited) local capture and Open Wayback for our local access portal. We're interested in eventually using Umbra in-line with Heritrix to improve capture efficacy; WebRecorder as a repository-integrated self-service web archiving tool; and Shine to improve the access capabilities of SWAP.		
Tools/Infrastructure Challenges: • The smaller the archiving job, the less efficient it is to curate and process in terms of staff time. • The most important websites to archive are often the most challenging to archive.		
Preservation Challenges: Haven't yet figured out division of roles and best-effort quality assurance practices.		
Collection Development Challenges: Web content collecting is new for many curators, let alone institutions. There's a vast amount of material that could be archived, and largely unsystematic methods to assess what's been or is being archived. In light of these challenges, training, education, and clarifying service processes and responsibilities are essential.		
Other:		

Institution Name: UK Web Archive at the British Library	http://www.webarchive.org.uk/ukwa/
The British Library has led the UK Web Archive effort since 2003. From 2003 to 2013 this was as an *ad-hoc* consortium and only by permission, and we collected multiple instances of thousands of sites. However, since Legal Deposit regulations were enacted in 2013, we have been working with the Legal Deposit libraries to archive the entire UK web domain at least once a year, and to archive notable UK sites much more frequently than that (e.g. news sites are currently crawled daily). We gain about 2.5 billion URLs from about 5 million sites each year. This means we are growing at around 65TB of compressed WARCs per annum. Outsourcing is not currently considered an option due to the specialism and scale, and the legal constraints. This may be revised in time, but as the Legal Deposit regulations are due for review in 2017 we are highly unlikely to make any major changes before then, unless governmental cuts force our hand. We are considering investing some of our resources to make significant changes and improvements to our tools, particularly the crawl process, during the next calendar year. Total capacity is estimated at 280TB stored.	
Main Use Cases: • Implementation of national Legal Deposit legislation • Thematic/topical collections	

<table>
<tr><td colspan="3">• Document harvesting, e.g. grey literature or where publishers have gone electronic-only
• Understanding researcher needs and developing services to increase usage and drive collection-care/preservation</td></tr>
<tr><td>Collection Development Location: Archivist/curators across the UK Legal Deposit Libraries, and in partner institutions.</td><td colspan="2">Additional Info: Services are run out of Boston Spa (North of England) but content is cloned out across four different sites, and access to Legal Deposit material is permitted from all the Legal Deposit libraries. Suitably licensed material is made publically available via the website.</td></tr>
<tr><td>Membership and Collaborations:
• UK Legal Deposit Libraries
• IIPC
• DPC
• OPF</td><td colspan="2">Details:
• The Legal Deposit libraries co-fund the web archive. The British Library performs the specialized technical work, whereas all partners contribute curatorial effort.
• We work with the IIPC and it's members for many reasons. Crucially, we work with the IIPC to improve and maintain the tools we need.
• The Digital Preservation Coalition and the Open Preservation Foundation provide links to the broader preservation communities.</td></tr>
<tr><td colspan="3">Funding: Government 'grant in aid' funding, but this is reducing over time.</td></tr>
<tr><td colspan="3">Staffing: About four FTE dedicated technical staff (currently 1FT unfilled), and about three FTE of archival/curatorial/management staff.</td></tr>
<tr><td>Integration of Web Archives with Other Collections:

Details and Concerns:</td><td>Internally: Yes.
Currently 'title level' records are generated from the crawl and combined with the curated annotations, and injected into the main catalogue. Using machine-generated catalogue records is acceptable, but could be improved greatly. Would also like to allow full text search of Legal Deposit material over the public web.</td><td>Externally: No.
We link out to other web archives wherever possible (e.g. via Memento). The content itself is largely harvested under Legal Deposit and as such cannot be re-distributed unless additional licensing is sought.</td></tr>
<tr><td>External Infrastructure: No</td><td>External Preservation: No</td><td>External Access Portal: No</td></tr>
<tr><td>Onsite Infrastructure: Yes (multi-site)</td><td>Local Preservation: Yes</td><td>Local Access Portal: Yes</td></tr>
<tr><td colspan="3">Tools Used Onsite: Heritrix3, OpenWayback, W3ACT (replacing our pre-Legal Deposit WCT and Selection & Permission Tool), webarchive-discovery and Apache Solr (full-text indexing and data mining), various orchestration tools.</td></tr>
<tr><td colspan="3">Tools/Infrastructure Challenges:
• Heritrix3 is monolithic and, given the staff and skills we have available, difficult to manage and</td></tr>
</table>

change. We want to use the latest version of H3 in order to get the best crawl results, but as our deployment does not quite match the way it's used by the Internet Archive, we seem to frequently hit odd bugs and edge cases

- At the same time, Heritrix3 is far behind the current web, and although IIPC partners have long known that we require more browser-assisted crawling, very few of them appear to have invested in this area. We have developed improved crawling technology, but we can't integrate it into Heritrix3 as it stands, as the framework is not sufficiently scalable.
- Manual QA can't scale and must be more automated.
- Automated are also needed to improve collection management given the limited effort available to catalogue web material.
    - Playback also lags behind the current state of the web. This is somewhat less pressing than the failure to collect material, but still concerning. Basic tooling issues, like the lack of easy access to some kind of proxy-based playback, make the debugging of quality issues unnecessarily difficult.

Preservation Challenges:

- As the crawler is not scalable the quality of the crawl is at risk – if we're not capturing the transcluded resources that modern pages depend upon, no later preservation action can fill the gap. This is our biggest worry right now.
- During the crawl, we collect screenshots and 'final' HTML from the original hosts via browser-based rendering and capture. We believe this will be valuable asset to future preservation work, but have not exploited it so far.
- We perform format and preservation risk scans during full-text indexing, but have not had enough time to really analyze the results of this in detail.
- We have not been made aware of any access issues relating to format obsolescence, but have discovered a few through our own explorations. Although we are preparing to deal with these issues later on, they are not a major concern right now. The basic crawling and playback of modern content is a significantly more serious issue.

Collection Development Challenges:

- Clear communication and integration of collection development effort between the LDLs.
- Better integration of our collected material with other classes of content, e.g. holistic delivery of historical news material across TV, radio, web, newspapers, etc.
- Scale – we can't manually QA enough, and we can't manually catalogue enough, and need computer-assisted tactics.
- Legal restrictions – lack of public access to material makes it harder to articulate the value of the collection. We are investigating generating analytics and datasets as a way of helping researchers get something valuable out of the collection even if they can't access the individual items.

Other:

Name: UCLA Library	URL: https://archive-it.org/organizations/877

About: UCLA Library has been web archiving since 1998 and currently has more than 7 TB of content. The library was utilizing CDL's WAS until its discontinuation earlier in 2015. Since then, the library has been using Archive-It. We are in the process of building our own web crawling and archiving infrastructure based on IIPC-featured software with the intention to do more experimental and research

work in the realm of web archiving, for example with social media content.		
Main Use Cases: • Institutional archives (websites and social media) • Thematic or topical web archives • Linking of "traditional" web archival content to other related collections such as TV news, social media, and other news content		
Collection Development Location: Library	Additional Info:	
Membership and Collaborations: • IIPC • NDSA/NDIIP • SAA	Details: UCLA Lib just recently joined the IIPC and we are in the process of establishing collaborations with partners in CA and beyond.	
Funding: Internal		
Staffing: No dedicated FTEs, 5 persons ¼ time each		
Integration of Web Archives with Other Collections: Details and Concerns:	Internally: Yes. Planned, not yet implemented.	Externally: Yes. Planned, not yet implemented.
External Infrastructure: Yes	External Preservation: AIT	External Access Portal: AIT
Onsite Infrastructure: Yes	Local Preservation: Yes	Local Access Portal: Yes
Tools Used Onsite: Open Wayback, Heritrix, Solr, Browsertrix, Social Feed Manager, Twarc		
Tools/Infrastructure Challenges: • Increased interest in social media capture but no convenient tools in place to capture (Facebook and Flickr, for example) • AIT interface not suitable for staff with little to no experience, and needs to be transferred to 2-3 admins; this is counter productive to our goal to "quickly capture this before it's gone" • Integration with existing collections and implementation of linking is subject to current research		
Preservation Challenges: Lack of convenient WARC replay tools for quality control		
Collection Development Challenges: We are in need of coherent collection development policy to address multiple questions. • What to collect, when, how often, who is responsible, what is the sustainability plan, who oversees the AIT account quota? • How do we know if other orgs have started a similar (enough) collection already? • How can we engage in collective collection development?		
Other:		

Institution Name: University of North Texas	http://digital.library.unt.edu/explore/collections/GDCC/ http://webarchive.library.unt.edu/eot2008/ http://digital.library.unt.edu/explore/collections/UNTWEB/
University of North Texas (UNT) commenced web archiving in 1997 with the CyberCemetery collection of government websites that have ceased operation. In 2005 it started collecting the UNT domain and in	

2008 collaborated on the end-of-term (EOT) election project. It now has ~200 TB of content. UNT is considering outsourcing (to Archive-It) to allow for curators to conduct their own crawls, in which case UNT would request copies of WARC files for local preservation and local access. Access to the CyberCemetery and EOT archives is provided from the UNT web portal; UNT domain crawls are not publicly available (access requires UNT IP address or VPN)		
Main Use Cases: • Institutional archives • Thematic or topical web archives • Extension of existing special collections		
Collection Development Location: The digital collections group of the Library	Additional Info:	
Membership and Collaborations:	Details: More involvement with the Special Collections group and collections management is anticipated.	
Funding:		
Staffing: No dedicated personnel; 2/3 of a developer, part time of 2 library staff (1/4 person/week), until recently ½ a grad student's time.		
Integration of Web Archives with Other Collections: Details and Concerns:	Internally: Yes. Site-level DC metadata record for online catalog, searchable finding aids. Users can also restrict their search to just websites	Externally: No
External Infrastructure: No. (Until/unless offer AIT)	External Preservation: No. (Until/unless offer AIT)	External Access Portal: No. (Until/unless offer AIT)
Onsite Infrastructure: Yes	Local Preservation: Yes	Local Access Portal: Yes
Tools Used Onsite: Heritrix, Wayback, Memento Time Travel, URL Nomination Tool (UNT developed), considering adding SOLR for full text search against smaller collections		
Tools/Infrastructure Challenges: • Looking to newer crawler technologies (e.g. Umbra) to enhance crawls • Preference for Python (versus Java) tools – not all tools are Python • WARC files are difficult to work with (level of knowledge required is too high for most people)		
Preservation Challenges:		
Collection Development Challenges: See a need for better ways to communicate where collections exist and how to get access to them (but no more registries!)		
Other:		

Institution Name: Yale University	https://archive-it.org/organizations/976 https://archive-it.org/organizations/977 https://archive-it.org/organizations/978 https://archive-it.org/organizations/1048

<table>
<tr><td colspan="3">Yale University has been web archiving since 2015. Four Yale units --Beinecke, ITS, the Yale Center for British Art, and Manuscripts and Archives--entered into a one-year contract with Archive-It this year in order to begin capturing content from websites and social media sites. A University-wide web archiving group has been formed to develop a web archiving strategy for Yale University, including website harvesting, description of the archived web content, development of access methods, and investigation and management of rights issues. The group will consult with other staff members as needed about digital preservation, description, and discovery; it will also maintain necessary contracts.</td></tr>
<tr><td colspan="3">Main Use Cases:
• University and institutional archives (websites and social media)
• Enhancement of manuscript collections in repositories' collecting areas
• Thematic collecting of web archives</td></tr>
<tr><td colspan="2">Collection Development Location: Individual units throughout the University, includes library, museum, and administrative units</td><td>Additional Info:</td></tr>
<tr><td colspan="2">Membership and Collaborations:
• CCWA
• SAA
• Ivy Plus
• NDSA
• OCLC</td><td>• Details: Ivy Plus– collaborative collection development at Archive-It– Contemporary Composers Web Archive (CCWA) and Collaborative Architecture, Urbanism, and Sustainability web Archive (CAUSEWAY) pilot web collections</td></tr>
<tr><td colspan="3">Funding: Internal</td></tr>
<tr><td colspan="3">Staffing: No dedicated staffing, approximately 3FTE across all University units; crawls are run by individual units by existing personnel.</td></tr>
<tr><td>Integration of Web Archives with Other Collections: NA, since this is our first year of web archiving, we have yet to integrate collections.

Details and Concerns:</td><td>Internally:</td><td>Externally:</td></tr>
<tr><td>External Infrastructure: Yes</td><td>External Preservation:</td><td>External Access Portal: No</td></tr>
<tr><td>Onsite Infrastructure: No</td><td>Local Preservation: Some units may store their WARC files in the YUL digital preservation system in the future.</td><td>Local Access Portal: NA</td></tr>
<tr><td colspan="3">Tools Used Onsite: Archive-IT service and HTTrack (https://www.httrack.com/)</td></tr>
<tr><td colspan="3">Tools/Infrastructure Challenges: Archive-It can't meet all of our needs in terms of capturing all desired content, and the interface will require future development to maximize potential.</td></tr>
<tr><td colspan="3">Preservation Challenges: Format, file size</td></tr>
</table>

Collection Development Challenges: rights issues, cost of web archiving, and building researcher interest in archived websites
Other:

Appendix C: Tools Lifecycle Matrix

The living document for the tools lifecycle can be found at at http://bit.ly/1Zok3WB

Key to Activities in Tools Lifecycle Matrix:

	Activity	Activity Description
Activity 1	Nomination	select sites targeted for web archiving
Activity 2	Rights	manage permissions to archive web sites
Activity 3	Assess/Define Capture	assess sites, define scope, create seed lists
Activity 4	Capture	capture web-based content
Activity 5	QA	enables quality assurance
Activity 6	Description	add descriptive metadata
Activity 7	Indexing	index for searching
Activity 8	Characterization	format characterization
Activity 9	Packaging	put into container file
Activity 10	Processing	processing ARC and WARC files
Activity 11	Discovery	search archived web pages
Activity 12	Delivery	fetch and display archived web pages
Activity 13	Analysis	visualization and analysis

Tools Lifecycle Matrix:

	Pre-acquisition			Accessioning	Curatorial Processing			Preservation Preparation		Discovery, Access & Analysis			
Activities --->	Activity 1	Activity 2	Activity 3	Activity 4	Activity 5	Activity 6	Activity 7	Activity 8	Activity 9	Activity 10	Activity 11	Activity 12	Activity 13
Tools ↓													
UNT nomination tool	1												
DigiBoard	1	1			1								
Building Collections on the Web (BCWeb)	1	1	1										
W3ACT	1	1	1			1							
Archive-It			1	1	1	1	1		1	1	1	1	
Web Curator Tool (WCT)		1	1	1	1	1							
NetarchiveSuite			1	1	1								
Compare Lists (of URLs)			1										
Extract URLs			1										
Expand Tiny URLs			1										
Harvester (list of URLs)			1										
Archiveready.com			1										
Builtwith.com			1										

	Pre-acquisition			Accessioning	Curatorial Processing			Preservation Preparation		Discovery, Access & Analysis			
Wappalyzer			1										
CINCH				1				1	1				
Heritrix				1									
WGet				1									
cURL				1									
ArchiveFacebook				1									
DeepArc				1									
HTTrack				1									
PreserveMe! (and PreserveMe! Viz)				1									1
Discus Comment Scraper				1									
Image Scraper				1									
PageFreezer				1									
SiteSucker fo Mac OS X				1									
Juriscraper				1									
webrecorder.io				1									
amberlink.org				1								1	
perma.cc				1								1	
Outwit				1									
Browsertrix				1									
PhantomJS				1									
Social Feed Manager				1									
WARCreate				1									
WAIL				1								1	
SiteStory				1								1	
Synchronicity												1	
Warrick												1	
Reverse Archive-It											1		
HRWA Manager application							1				1		
Terminology evolution, TeVo							1			1			1
SOLR							1						
NutchWAX (Nutch with Web Archive eXtensions)							1				1		
WarcManager							1				1	1	
Kibana							1				1		1
JHOVE2								1					
BagIt									1				
WARC-Grabber									1	1			
HTTrack2ARC										1			
Web Archive Transformation (WAT) Utilities										1			

	Pre-acquisition			Accessioning	Curatorial Processing			Preservation Preparation		Discovery, Access & Analysis			
Shine											1		1
WERA											1	1	
Mink											1	1	
Memento Time Travel											1		
Wayback Machine												1	
OpenWayback												1	
PyWB												1	
MediaWiki Memento Extension												1	
oldweb.today												1	
WarcBase (with Hbase and Spark)												1	1
ArchiveThumbnails													1
Natural Language Toolkit (NLTK)													1
Kimono													1
Leximancer													1
WCopyfind													1
Mallet													1
CarbonDate													1
Bubble Lines													1
Censorship Explorer													1
Colors for Data Scientists													1
Compare Networks Over Time													1
Deduplicate (for tags)													1
Dorling Map Generator													1
Raw Text to Tag Cloud Engine													1
Geo Extraction													1
GeoIP													1
Gephi													1
	4	4	12	26	4	3	7	2	4	5	10	16	22

Note: DigiBoard and UNT Nomination tool, not yet available, are targets/candidates for open source (email, UNT and LC, December 2015)

Appendix D: List of Questions for Web Archiving Institutions

Interviews were semi-structured with the following set of questions used to guide the conversation:

- How are institutions providing and maintaining web archiving infrastructure and services? What are the main challenges?
- What are the trends in web archiving programs?
- What are the trends and challenges in the collecting of web archives?
- What are the trends in the usage of web archives?
- How are people currently integrating their web archives with their library collections? How is that working? What are the challenges? Opportunities?
- Where does web archiving live? In the library? Archives?
- Which staff/how many are responsible for it? Does anyone have staff solely dedicated to web archiving? Or is it seen as just another collecting area?
- Is there an increase in usage/awareness of web archives?
- Which disciplines are currently using web archives? E.g., communications, history, computer science, who else? Is there a way to get even a rough sense of how many researchers are actively using web archives?
- What might be driving these trends? What are the emerging issues? What are future threats and opportunities?
- What are your challenges in integrating your web archives with your other collections or with the collections of others?
- What are the current challenges related to collecting /harvesting web content?
- What are the current challenges related to preserving your web archives?
- What do you consider your institution's current largest challenges related to web archiving?
- Do you have any changes planned in the next couple of years in your web archiving activities or program, and if so what?
- Do you have any long-term plans to change your web archiving activities or program, and if so what?
- What do you think web archiving institutions should be doing collaboratively?
- For components provided in-house, are you able to easily switch out components?
- For components provided by an external party, would it be difficult for you to switch to a different external organization or company? Why or why not?
- Are you able to keep your infrastructure up-to-date with the newest version of tools?
- What web archiving APIs does your institution use?
- Are there any document types you're focusing on collecting? What are the challenges?
- Who uses your web archives? Please provide as much detail as possible, e.g. is access provided only on-site or also remotely, which domains use it, do you have researchers doing large-scale, e.g. data mining research on it?
- What tools or APIs do you think web archiving institutions should be developing and maintaining collaboratively?

Appendix E: Works Cited / Resources

Abrams, Stephen. Personal Interview. 17 Sep 2015.

Alam Sawood. Message to Truman Technologies, LLC. 24 Dec 2015.

Alam, Sawood. Personal Interview. 2 Oct 2015.

Bailey, Jefferson, Abigail Grotke, Kristine Hanna, Cathy Hartman, Edward McCain, Christie Moffat, and Nicholas Taylor. Web Archiving in the United States: A 2013 Survey. Rep. NDSA, n.d. Web. <http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_USWebArchivingSurvey_2013.pdf>.

Bailey, Jefferson. Message to Truman Technologies, LLC. 12 Jan 2016. Email.

Bailey, Jefferson. Proc. of Web Archives 2015: Capture, Curate, Analyze, University of Michigan, Ann Arbor. 12-13 Nov 2015.

Bailey, Jefferson. "Research Datasets Workshop." Web Archives as Research Datasets. Proc. of 2015 General Assembly, Stanford University, Li Ka Shing Conference Center, Silicon Valley, CA. IIPC, 28 Apr. 2015. Web. <http://netpreserve.org/general-assembly/ga2015-schedule>.

Brügger, Niels. Message to Neubert reposted to IIPC members list. Nov 2015. Email.

Belovari, Susanne. Personal Interview. 10 Dec 2015.

Chudnov, Dan. Personal Interview. 25 Sep 2015.

Davis, Gina. Personal Interview. 5 Oct 2015.

Dougherty, M., Meyer, E.T., Madsen, C., van den Heuvel, C., Thomas, A., Wyatt, S. (2010). Researcher Engagement with Web Archives: State of the Art. London: JISC.

Duncan, Sumitra. Personal Interview. 5 Oct 2015.

Fuhrig, Lynda Schmitz. Personal Interview. 13 Oct 2015.

Goethals, Andrea. Personal Interview. 17 Sep 2015.

Graham, Pamela. Personal Interview. 24 Sep 2015.

Grotke, Abigail. Personal Interview. 5 Oct 2015

Hartman, Cathy Nelson; Murray, Kathleen R. & Phillips, Mark Edward. *Classification Of The End-Of-Term Archive: Extending Collection Development Practices To Web Archives*. UNT Digital Library. <http://digital.library.unt.edu/ark:/67531/metadc152437/>.

Jackson, Andy. "The Provenance of Web Archives." Web log post. UK Web Archive Blog. Ed. Jason Weber. The British Library, 20 Nov. 2015. Web. <http://britishlibrary.typepad.co.uk/webarchive/2015/11/>.

Kendall, Skip. Personal Interview. 17 Sep 2015.

Knight, Steve. Personal Interview. 10 Nov 2015.

Kovari, Jason. Personal Interview. 28 Oct 2015.

Lack, Rosalie. Personal Interview. 22 Sep 2015.

Neubert, Michael. Message to IIPC Members. Nov. 2015. E-mail.

Neubert, Michael. Message to Truman Technologies, LLC. 12 Dec 2015.

Neubert, Michael. Personal Interview. 5 Oct 2015.

Paollilo, Michele. Message to Truman Technologies, LLC. Nov. 2015. E-mail.

Perricci, Anna. Message to Truman Technologies, LLC. 21 Dec 2015. Email.

Perricci, Anna. Personal Interview. 25 Sep 2015
Phillps, Mark. Personal Interview. 25 Sep 2015.
"Rhizome (organization)." Wikipedia. Wikimedia Foundation, 7 May 2003. Web. <https://en.wikipedia.org/wiki/Rhizome_(organization)>.
Slania, Heather. Personal Interview. Nov 10, 2015.
Smith, Kari. Personal Interview. 13 Oct 2015.
Thurman, Alex. Personal Interview. 24 Sep 2015.
Weber, Matthew. Personal Interview. 2 Oct 2015.
Wright, Jennifer. Personal Interview. 13 Oct 2015.

HARVARD
LIBRARY