National Digital Stewardship Alliance
Web Archiving Survey Report
Produced by the NDSA Content Working Group
June 19, 2012

The National Digital Stewardship Alliance is a member organization whose mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. Comprised of members who span a range of diverse communities, from cultural heritage organizations to educational institutions to commercial entities, the NDSA promotes a shared interest in fostering and supporting digital stewardship. The NDSA established five Working Groups focusing on the following areas: Content; Standards and Practices; Infrastructure; Innovation; and Outreach. The Content Working Group focuses on identifying content already preserved, investigating guidelines for the selection of significant content, discovery of at-risk digital content or collections, and matching orphan content with NDSA partners who will acquire, preserve, and provide access to it.

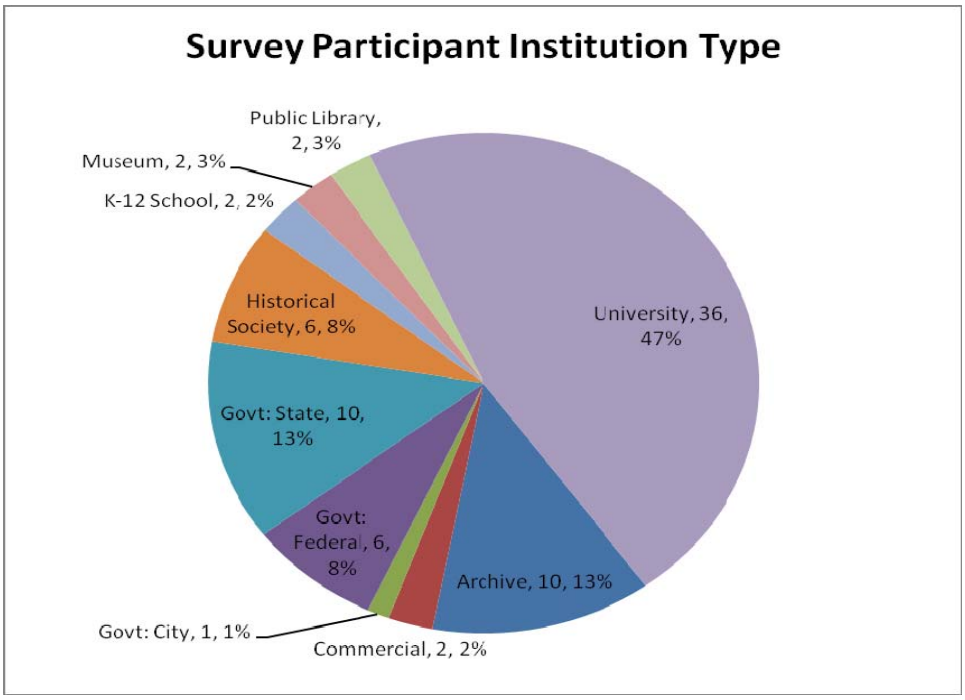
From October 3 through October 31, 2011, the Content Working Group conducted a survey of organizations in the United States that are actively involved in, or planning to start, programs to archive content from the web. The goal of the survey was to better understand the landscape of web archiving activities in the United States, including identifying the organizations or individuals involved, the types of web content being preserved, the tools and services being used, and the types of access being provided. This summary report examines participant responses for the purposes of discerning trends, themes, and emerging practices and challenges in web-based content acquisition and preservation.

Report Contents

1) **Activities & Policies [pg. 2]**: Examines the types of institutions conducting or planning web archiving activities, the range of their operations, and their specific policies towards acquisition, access, and preservation.
2) **Tools [pg. 11]**: Details current software and strategies for acquiring and displaying web archives.
3) **Content [pg. 15]**: Provides an overview of the types of online content currently being harvested.
4) **Conclusion [pg.17]**: Offers a summary of the survey results, highlights areas of potential advancement of the field, and suggests issues or topics that merit further study.
5) **Appendix [pg.19]**: Provides information on the tools referenced in the report, sample survey responses, and other resources.

Survey Participants

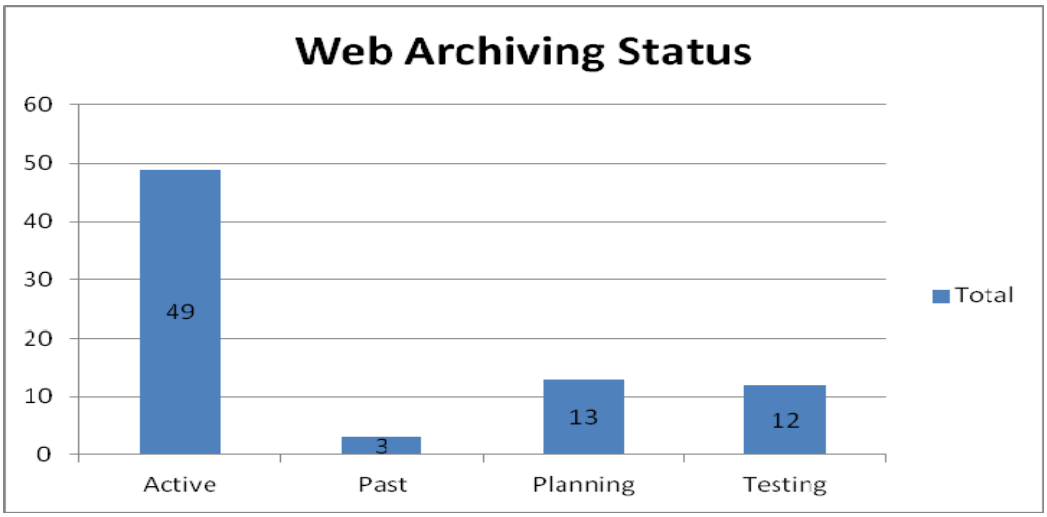
The survey garnered 77 unique responses from a range of institutions, with survey participants primarily representing the cultural heritage (29%, 22 of 77), government (22%, 17 of 77), and university communities (46%, 36 of 77). Of the survey respondents, 31% (24 of 77) were members of the NDSA and 8% (6 of 77) were members of the International Internet Preservation Consortium (IIPC).



1) Activities & Policies

Web Archiving Activity

An active web archiving program is currently being administered by 63% (49 of 77) of the survey respondent institutions. Additionally, 16% (12 of 77) are actively testing such a program and another 17% (13 of 77) are planning on pursuing a web archiving program in the near future, meaning a full 96% (74 of 77) of respondents are actively or planning on archiving web content (the 3 other institutions had formerly managed web archiving programs, but no longer do so).



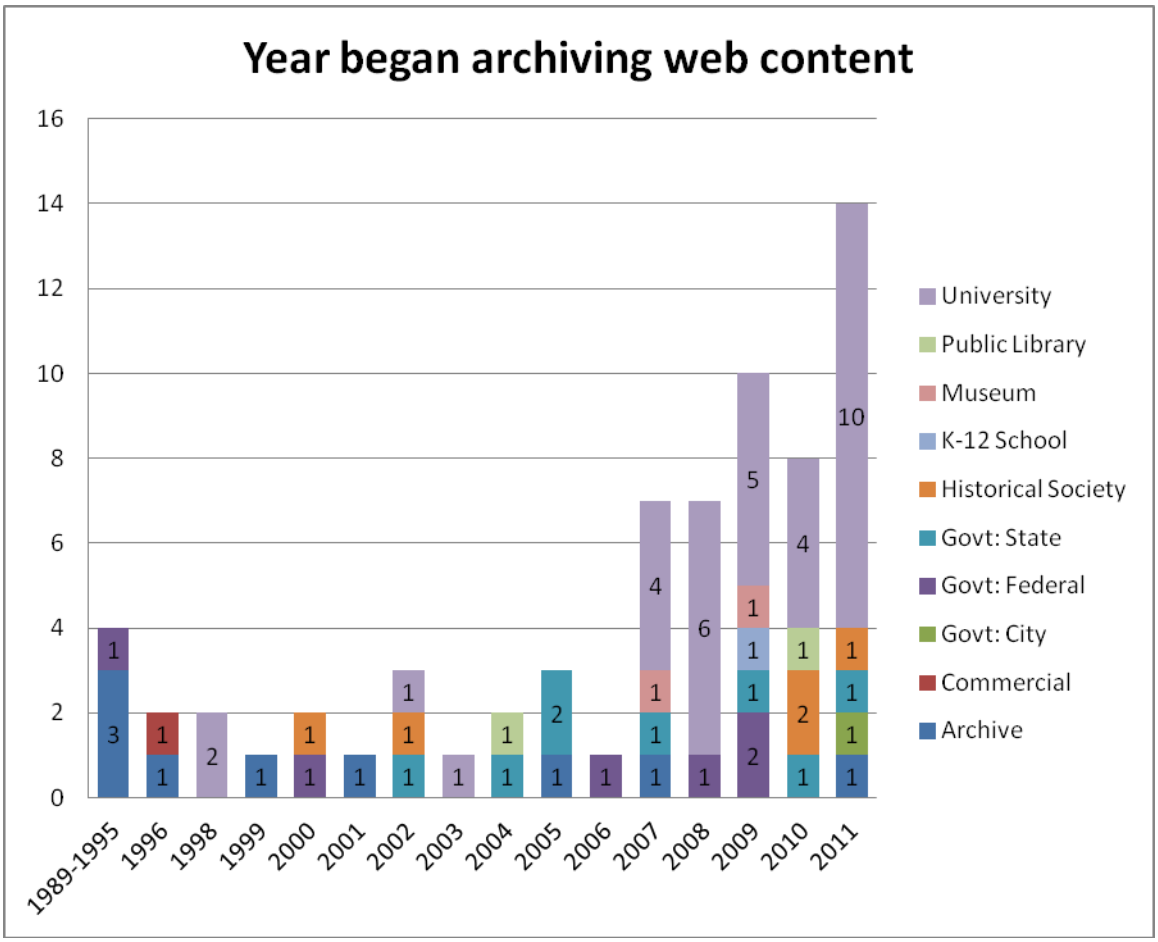
Of the 71 respondents that identified their web archiving goals, 49% (35 of 71) were preserving their own institutional web content as well as archiving content "from other organizations or individuals for future research." Of the remaining, 20% (14 of 71) identified the first goal and 31% (22 of 71) the latter.

What are the goals of your web archiving activity? Select as many as apply.		
Answer Options	**Response Percent**	**Response Count**
Archive your own web site as a type of institutional record.	69.0%	49
Archive content from other organizations or individuals for future research.	80.2%	57
Both of the above	49.2%	35

Comments on the question about goals focused largely on two areas. The first was legal mandate, especially among state governments; this was illustrated by responses noting "the statutory obligation to deposit all state publications, regardless of format" or the need to "stay in compliance with records retention laws." The other common response was centered on web archiving as an enhancement of existing collections, with institutions planning on archiving "websites of organizations and individuals for whom we hold paper/print archives" and desires for "capturing digital faculty projects as well as sites associated with our manuscript collections."

The survey also illustrated the relatively recent emergence of web archiving as an institutional activity. Of the 68 respondents that identified the specific year their web archiving began, nearly a third, 32% (22 of 68) began their programs within the last two years, the exact same number of institutions (22, 32%) that began archiving web content in the 17 years between 1989 and 2006. The recent surge in web archiving within the last

5 years – 68% (46 of 68) of those surveyed – is primarily due to universities starting web archiving programs. Twenty-nine of those 46 institutions undertaking web archiving in the last five years were universities, a rate of 63%. Self-identified archival institutions, on the other hand, accounted for 6 of the 15 institutions that began web archiving prior to 2002 (a rate of 40%), yet only 3 of the 39 (7%) that began programs from 2002 to the current day.

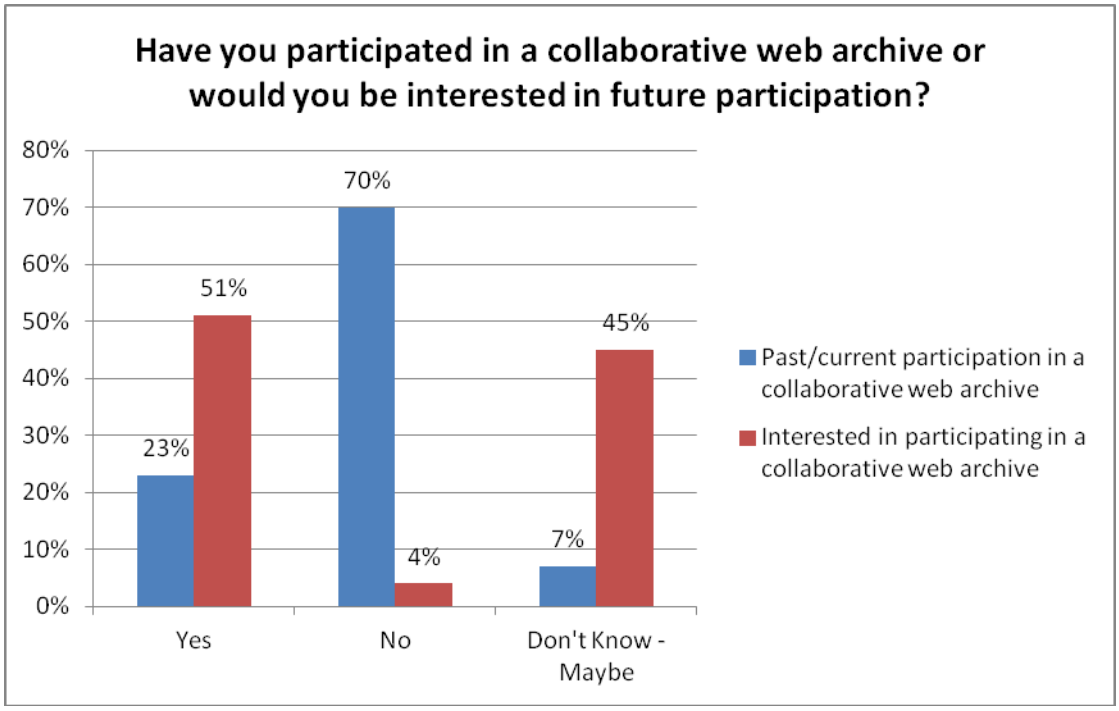


A topic the survey highlighted as ripe for continued exploration and discussion is that of collaborative web archiving. Collaborating on print acquisitions around event or subject areas has long been a goal for institutions sharing a similar collecting focus. That type of collaboration has only begun to include acquiring and preserving born-digital web-based materials. Existing collaborative web archiving projects are often focused on quickly-unfolding events or events with an international scope. In these projects, participants will contribute URLs or orient individual acquisition activities in concert with partner institutions.[1] The survey responses document a wide interest in the potential benefits of

[1] For an elaboration on collective web archiving projects, see Abbie Grotke's blog post, "It Takes a Village… to Archive the Internet." http://blogs.loc.gov/digitalpreservation/2011/07/it-takes-a-village%E2%80%A6to-archive-the-internet/

shared efforts in this area. When asked "has your organization ever participated in a collaborative web archive" only 23% (15 of 66) answered "yes," while 77% (51 of 61) said "no" or "don't know." While those answering "no" sometimes explained how their specific legal requirements or retention policies limited their ability to participate in collaborative archiving (for example, the legal requirements for collecting online state government publications obviously disallow collaboration), a number of institutions noted they are "are working to build collaborations" or "are aware of each other's collections and have discussed our mutual efforts, but have no formal collaborative collection development policy."

These comments foreshadow the responses to the next question, which asked "would your organization be interested in future collaborative web archives (if they fit within your collecting scope and interests)?" Responses to this question evinced a broad desire for greater shared harvesting. Though only 23% of organizations were currently collaborating on web archiving, 96% (64 of 67) answered either "yes" (34, 51%) or "maybe" (30, 45%) when asked if they were interested in participating in future collaborative collecting activities. As these numbers demonstrate, there is a sincere interest in the collaborative opportunities around joint web archiving, but little current action in this area. Research and promotion of current collaborative efforts could provide the information and impetus needed to support the creation of more collaborative web archiving initiatives.

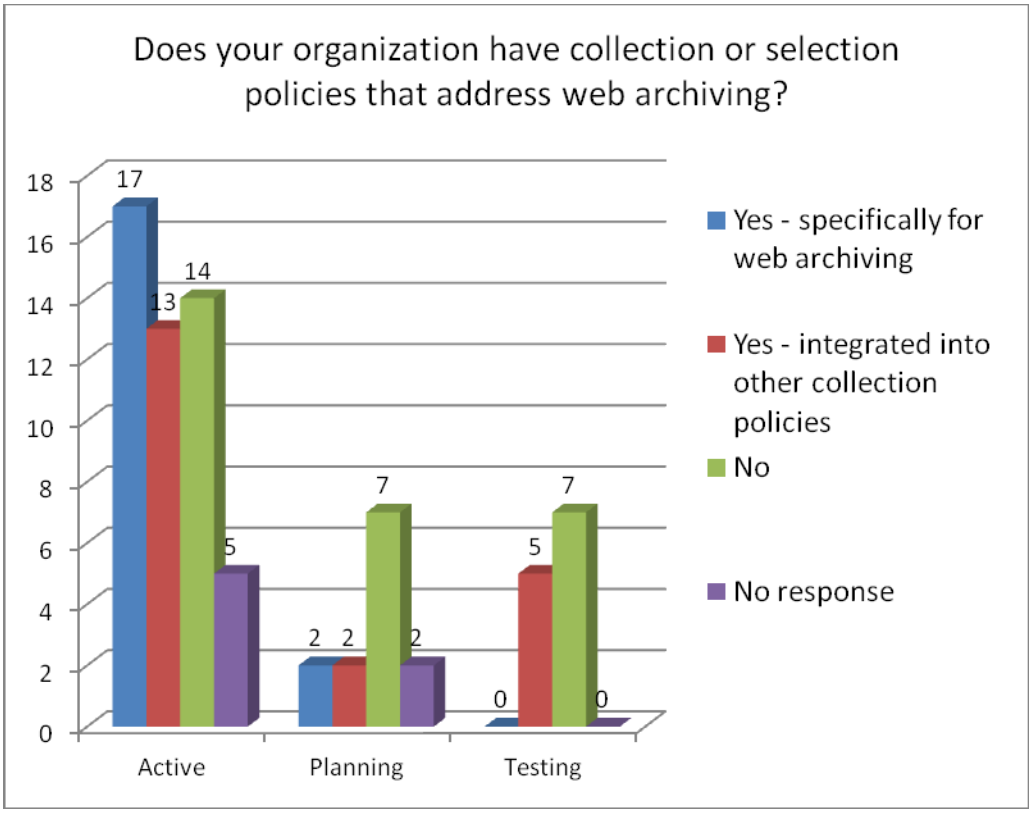


Acquisition & Access Policies

Though collaboration is still an emerging concept, institutional policies for web archiving were in place at many institutions.

- 29% (14 of 49) have "collection or selection policies that specifically address web archiving."
- 35% (17 of 49) have web archiving policies "integrated into other collection policies."
- 27% (13 of 49) had no policy and 5 did not respond.

Of the 25 institutions testing or planning a web archiving program, 36% (9 of 25) had either specifically-written or integrated collection policies covering these materials and 56% (14 of 25) had no documented policies (2 did not respond). Active programs are, by nature, more likely to have documented policies, but the lack of policy documentation for active, planning, and testing web archiving programs, as seen in the following chart, signals another area for potential collaboration and knowledge-sharing.

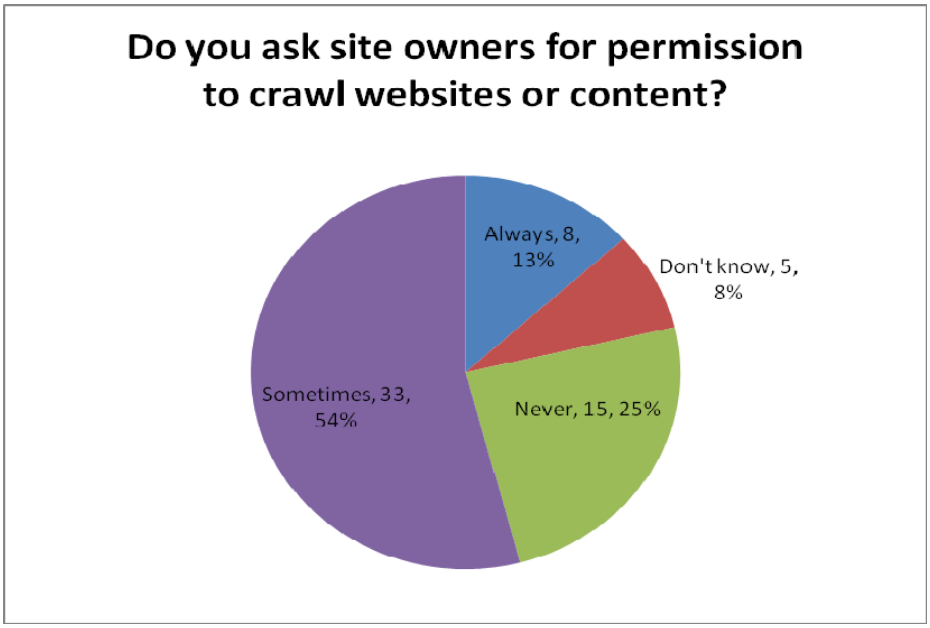


The specific policies covered in the survey examined a number of activities related to acquisition and access practices, soliciting site owner permissions for harvesting and display, policies towards robots.txt files, and how acquired content is accessed and used by researchers. Policies around seeking permission from content creators, both permission to collect and permission to make accessible, evoked a variety of responses.

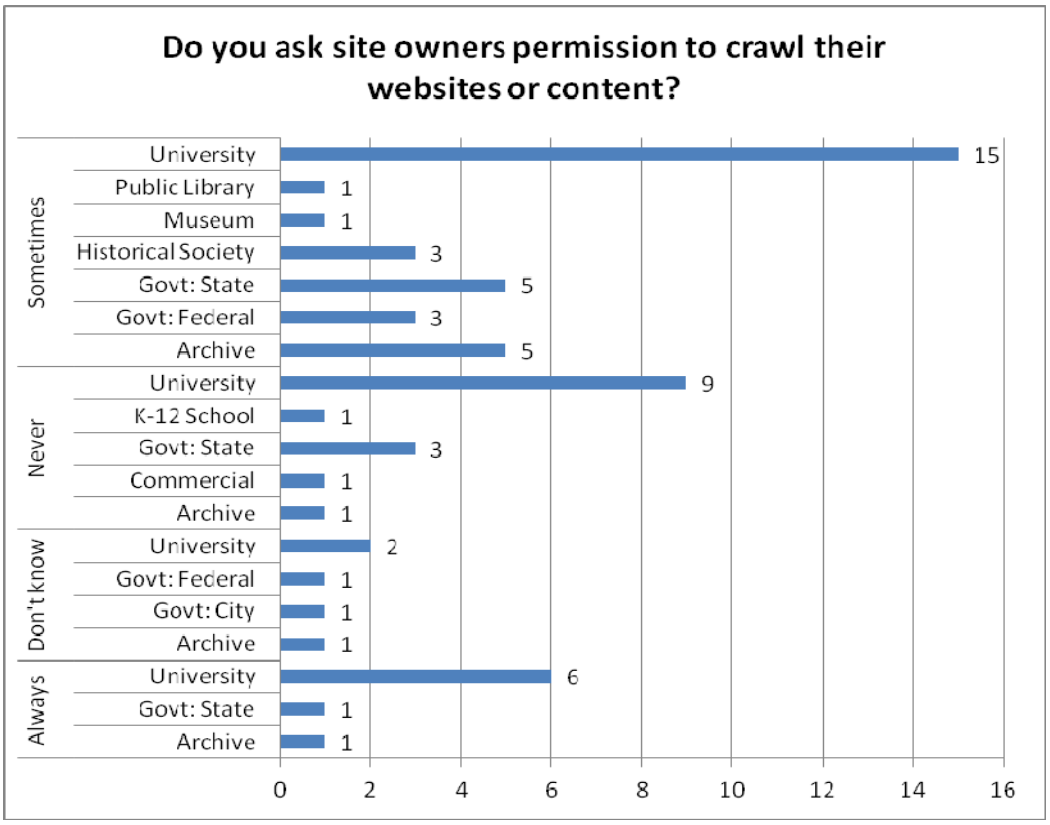
- 54% (33 of 61) answered they "sometimes" ask permission to harvest content
- 25% (15 of 61) never ask for permission to harvest content
- 13% (8 of 61) always ask for permission to harvest content

- 8% (5 of 61) said they "don't know" if they are requesting permission to harvest content

Most striking here is the high number of institutions seeking permission on a case-by-case basis. This question did not include comments, so it is difficult to discern the content or acquisition conditions which prompt permission seeking. Collection type, the type of site (whether a news organization or independent blog or government site, for instance), site extent or structure, presumed ephemerality, or creator accessibility, could all be factors impacting whether a harvesting organization asks for permission to crawl. Idiosyncratic approaches to risk-management, legal due diligence, and institutional culture also influence acquisition, access, and permission policies. At a minimum, greater transparency, either in policies or in scope of collection documentation, could help better illuminate current practices this area and assist other organizations in developing their own policies.

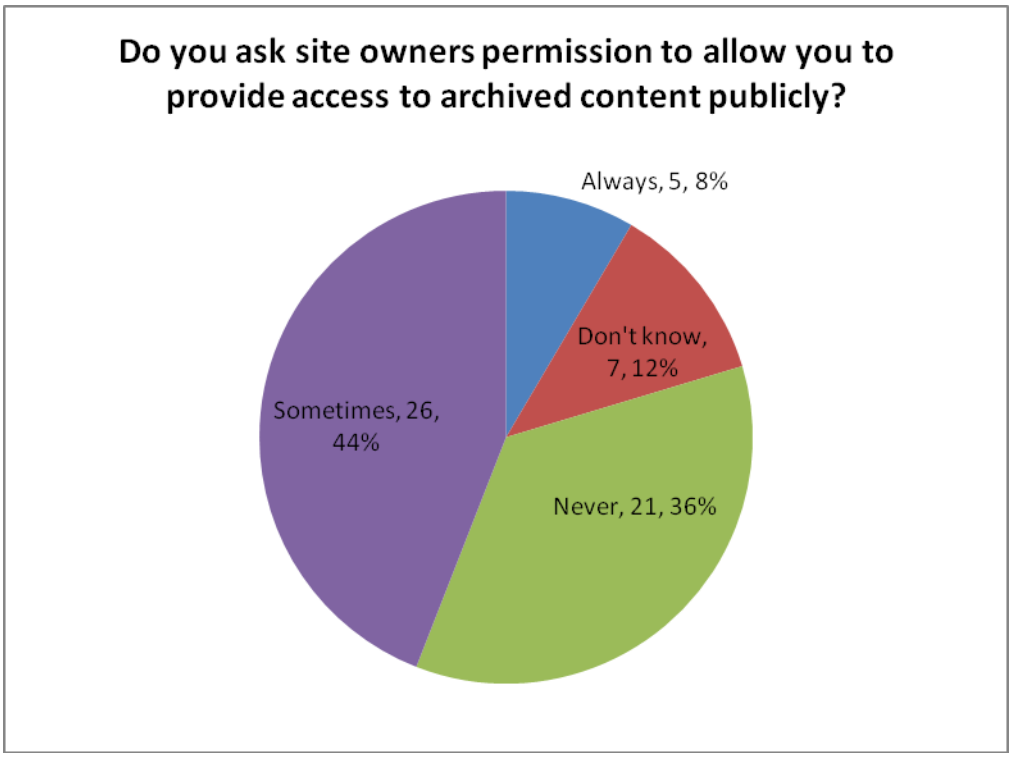


The results can also be broken down by institution type:



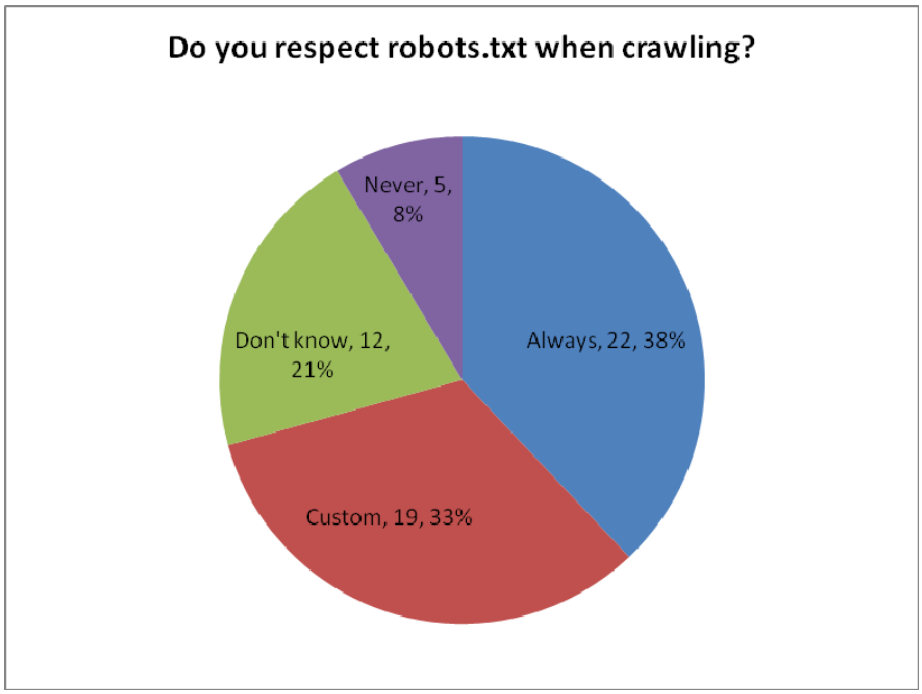
Asking site owners for permission to provide access to their content also elicited a mixed response, though survey respondents were somewhat less inclined to seek permission for access than they were for acquisition. As with acquisition, seeking permission for access was largely conditional:

- 44% (26 of 59) “sometimes” ask permission to provide access to content
- 36% (21 of 59) “never” seek site owner permission to provide public access
- 12% (7 of 59) say they “don’t know” if they ask permission to provide access
- 8% (5 of 59) “always” ask permission to provide access to content

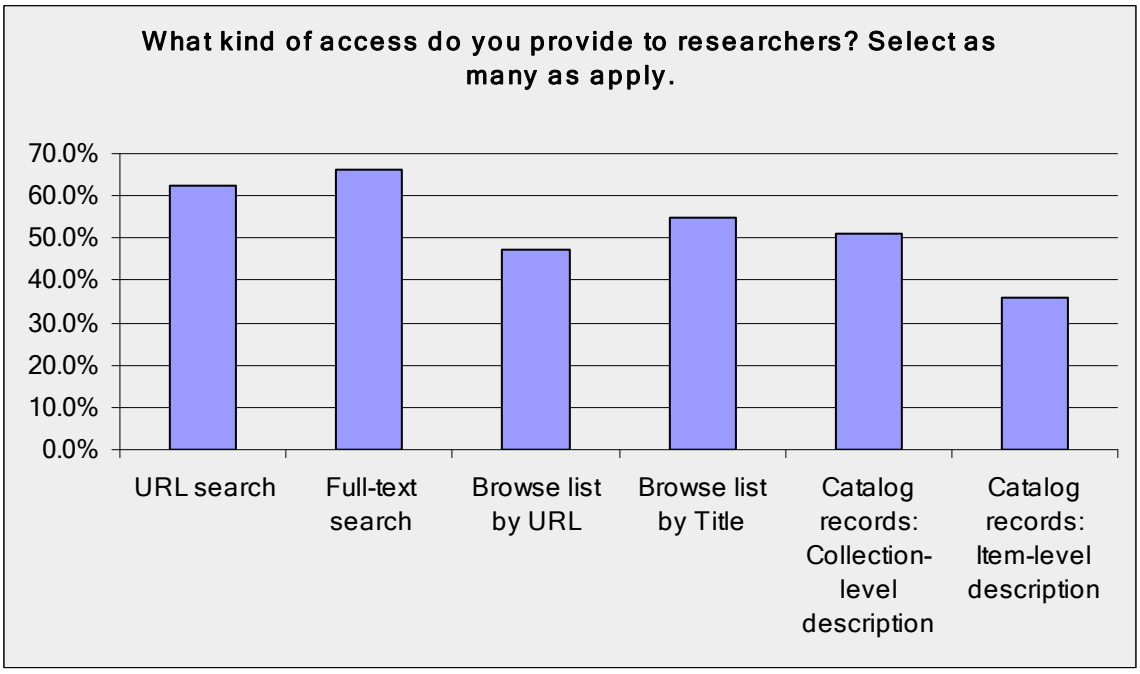


A survey question about respecting robots.txt[2] received a mixed response, 38% (22 of 58) always respect a robots.txt file, 8% (5 of 58) never respect it, and 54% (31 of 58) either conditionally respect it or are not sure of their policies (33% and 21%, respectively). Some web archiving programs respect robots.txt but do not seek permissions or notify site owners that they are archiving. Other archiving programs believe that respecting robots.txt would create a void of valuable content in their archive due to the restrictions of robots.txt. This void includes not just web content, but also things that make up the design and aesthetics of a website, such as images and stylesheets. Some institutions generally do not follow robots.txt once permissions are secured or notice of harvesting given while others, in certain special conditions, will work with site owners to respect robots.txt on a conditional basis.

[2] "Robots.txt" is a file put on a web server which tells web-crawling robots not to visit or harvest that particular website. More information can be found at http://www.robotstxt.org/.



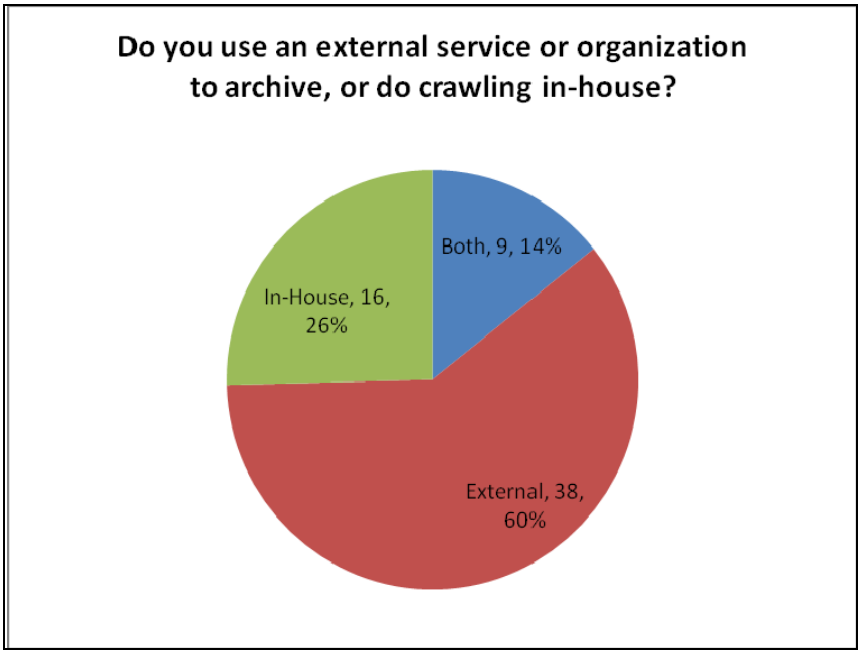
An area of uncertainty exposed by the survey involved institutions' knowledge of *how* their collections were being used. Institutions are providing a variety of means of access to harvested content (see below chart). Even beyond the categories available in the survey, respondents commented on additional means of accessing collections, including through links in EAD finding aids, by geographic coverage, and by tagging.



At the same time, the question "how are researchers using your archives" solicited 52 responses, a significant majority of which were a variation on "unknown," "too soon to tell," or "good question." A number of responses did note a specific use or user community – such as local historians, genealogists, educators, and government officials – but the lack of knowledge about web archive usage and users is clearly a topic that merits further investigation.[3]

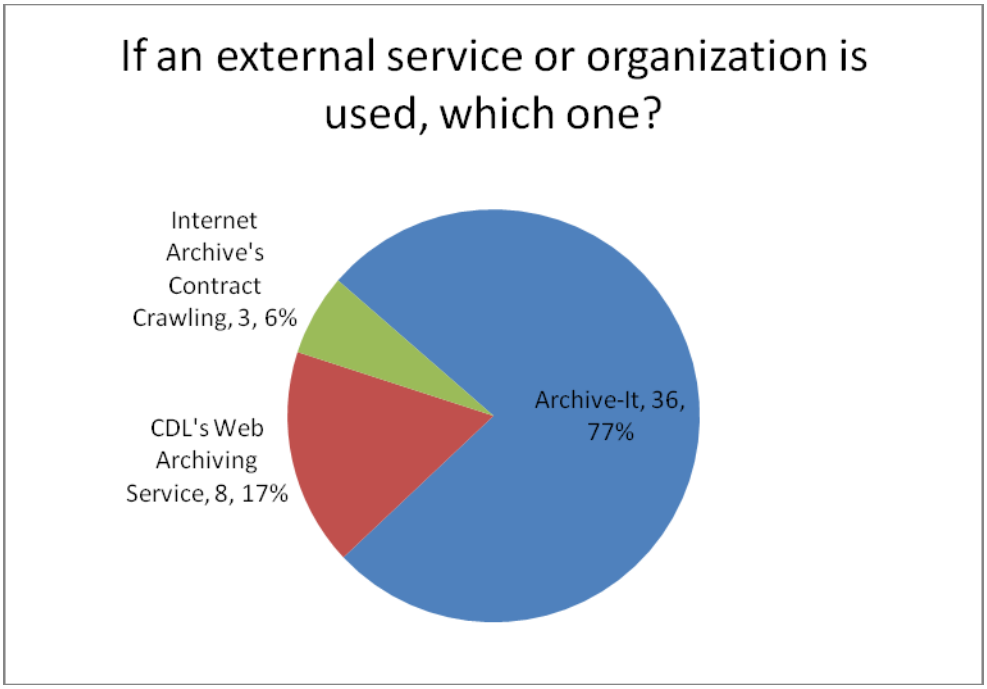
2) Tools[4]

The web archiving survey also sought to gain a better understanding of the specific tools being used both to collect content and to display archival collection. Of the 63 respondents indicating their tools for harvesting web materials, 60% (38) were using an external service for acquisition, 26% (16) were using an in-house method, and 14% (9) were using both in-house and external services.

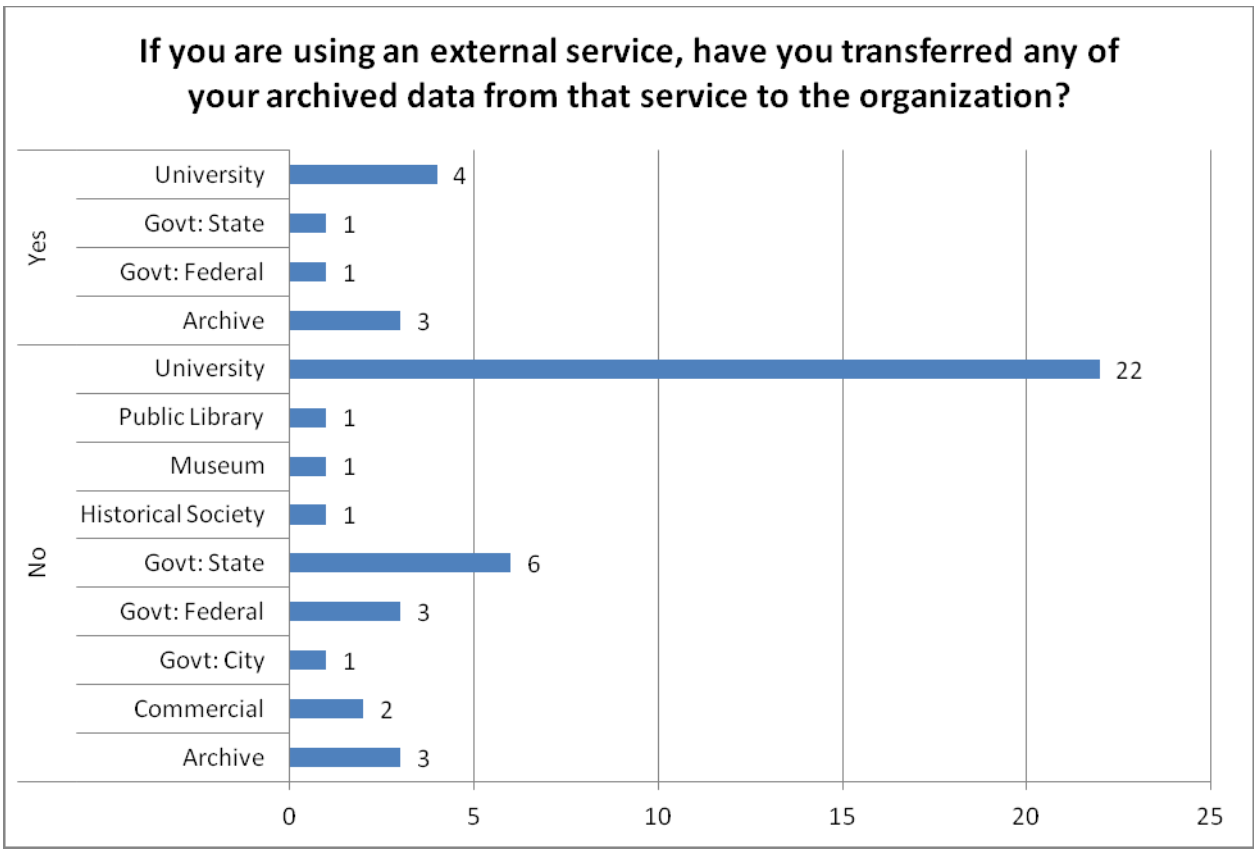


[3] Of interest here is a recent effort by the UK Web Archive to solicit user feedback. The form can be found at http://www.irn-research.com/surveys/webarchive.htm.
[4] A full list of the tools discussed in this section, and links to supporting documentation, can be found in the Appendix.

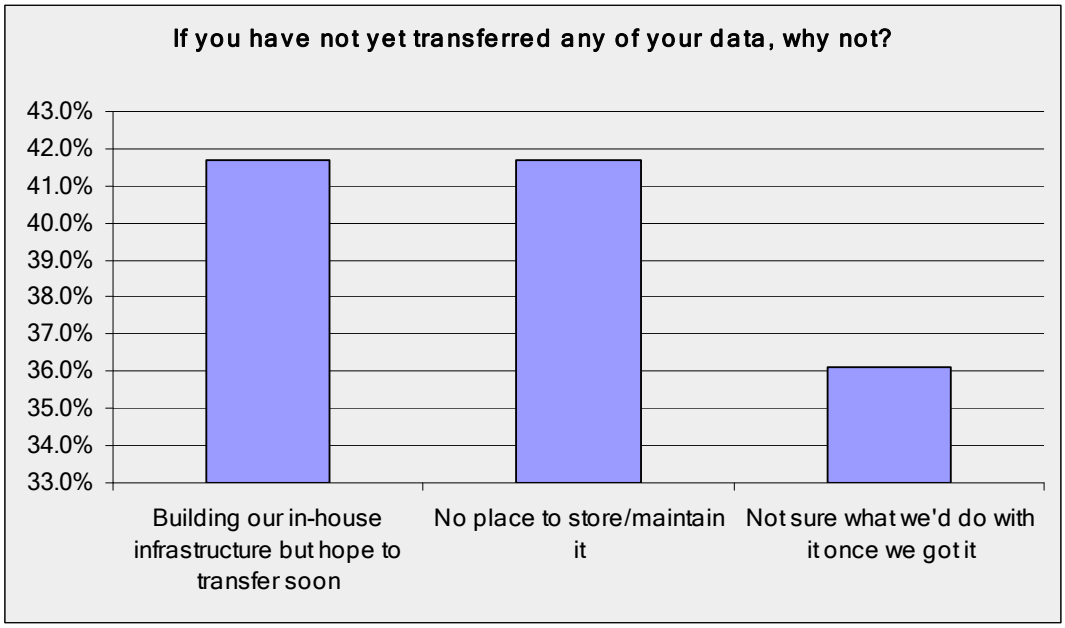
Of those 47 institutions using an external service (either exclusively or in combination with in-house tools), an overwhelming majority, 77% (36 of 47) use Archive-It; the remaining use California Digital Library's Web Archiving Service, 17% (8 of 47) or contract for crawling with the Internet Archive, 6% (3 of 47).



One discovery of the survey was the low percentage of respondents that have transferred their archived data from their external service to their institution. Only 18% (9 of 49) have transferred their data in-house, including only 2 of the 12 government respondents and only 4 of the 25 university respondents. A total of 82% of those using an external service have not transferred data to their institution.

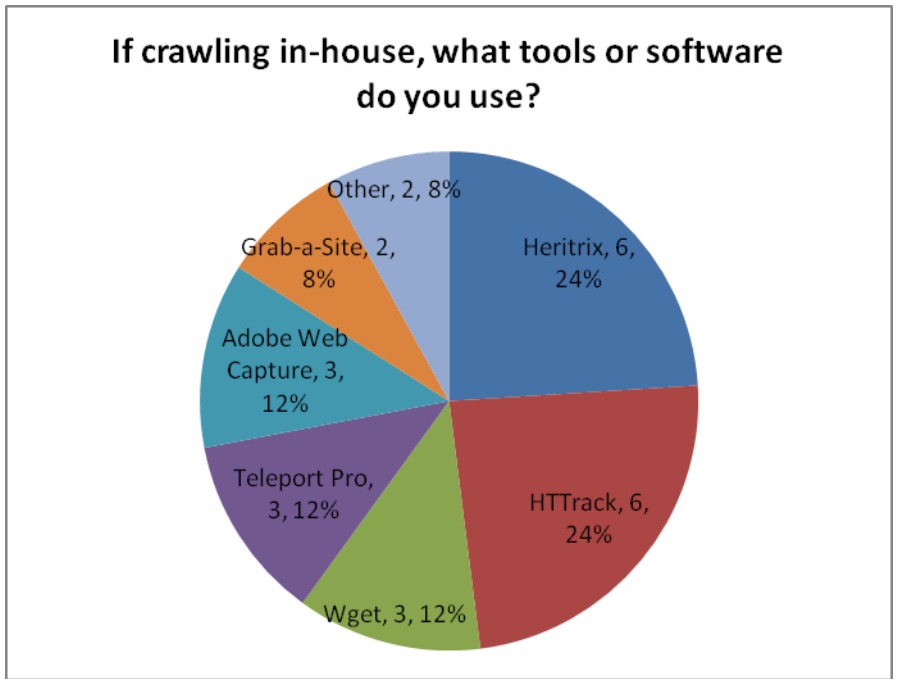


When asked their reasons for not having transferred data, survey participants were split between being in the process of building the infrastructure to support the content once transferred and institutions that have no place to store the data. In addition, a notable percentage (36%, 13 of the 36 respondents) answered that they were "not sure what we'd do with it once we got it." This response hints at the inability to provide storage or adequate infrastructure, but it also suggests the organizational challenges to providing access to web archive collections.



Free text comments for this question pointed to some other concerns for transferring externally harvested data to in-house systems including "duplicate costs," confidentiality concerns, and cataloging and accessibility challenges.

Of the 25 institutions doing their crawling either in-house or in conjunction with an external service, Heritrix and HTTrack were the most popular, each used by roughly a quarter of respondents (24%, 6 of 24 for each), with a variety of other tools in use, including Wget, Teleport Pro, Grab-a-Site, and Adobe Web Capture.

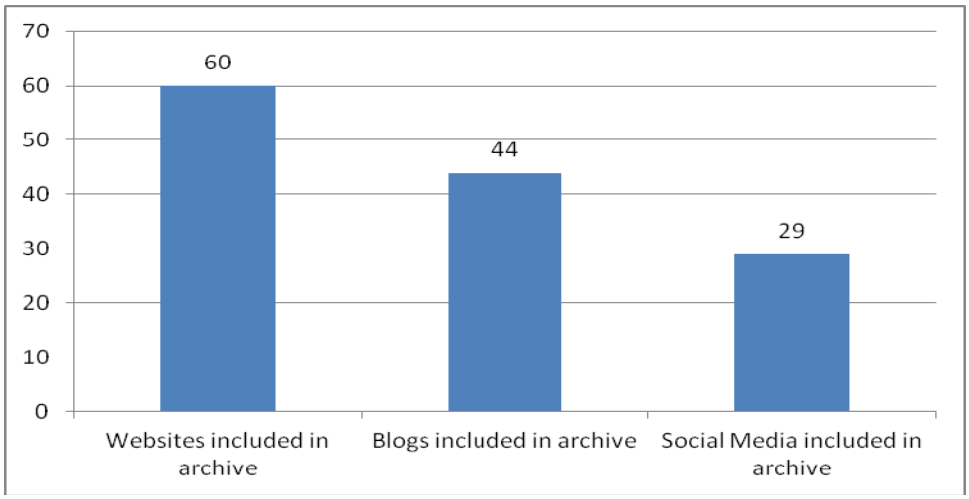


When asked what "viewer or software" institutions were using to provide access to their harvested web content, 76% (41 of 56) were using Internet Archive's Wayback Machine (many of them through their Archive-It accounts, most likely). The "other" category, 25% (13 of 54 of respondents), were using a variety of tools, including OCLC and CDL services, Memento, and custom or in-house HTML displays.

3) Content

Responses revealed a diversity of content of interest to active web archiving programs. The types of content being acquired included websites, blogs, and social media:

- 78% (60 of 77) included or plan on including **websites** in their archive
- 57% (44 of 77) included or plan on including **blogs** in their archive
- 38% (29 of 77) included or plan on including **social media** in their archive

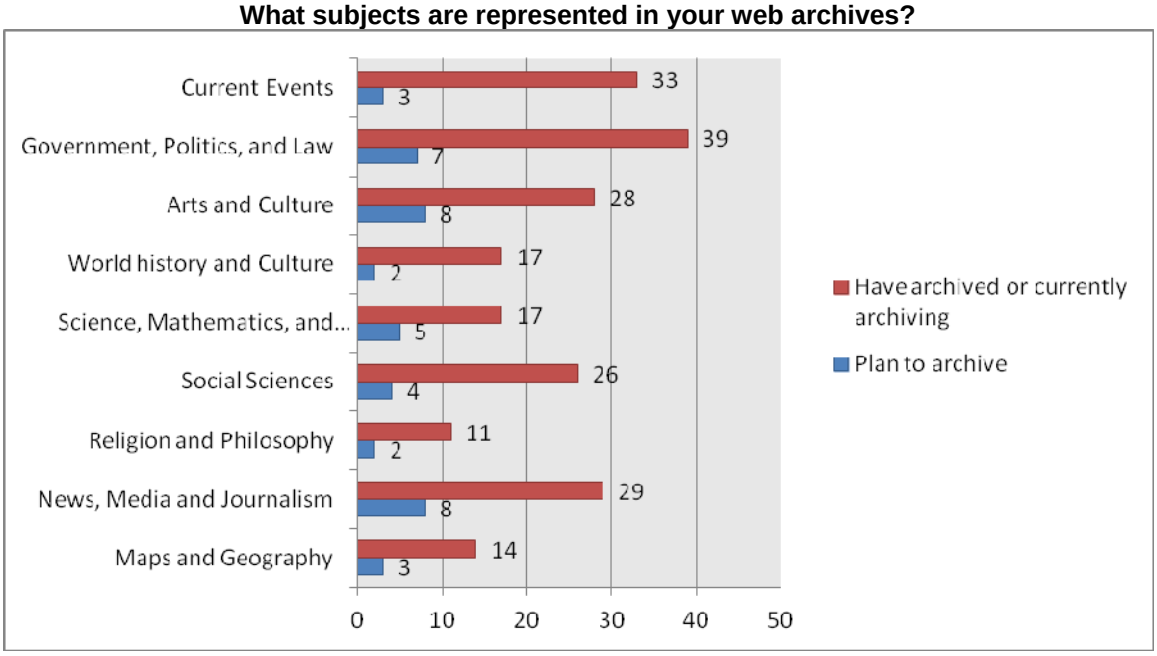


A free-text survey question asked for respondents to "briefly describe the scope of your web archive collections: what type of events, topics, themes, or approaches you take in archiving content from the web." Broadly stated, these responses fell into one of three categories: institutional self-documentation, collection enhancement, and thematic. A number of respondents noted that web archiving was an attempt to capture online promotional and outreach efforts (especially through social media). Unique thematic collections focused on human rights websites, web-based digital art, online evidence of "the left and labor movements," subject-related blogs (law blogs, tobacco-related content), and political events and natural/environmental disasters.[5]

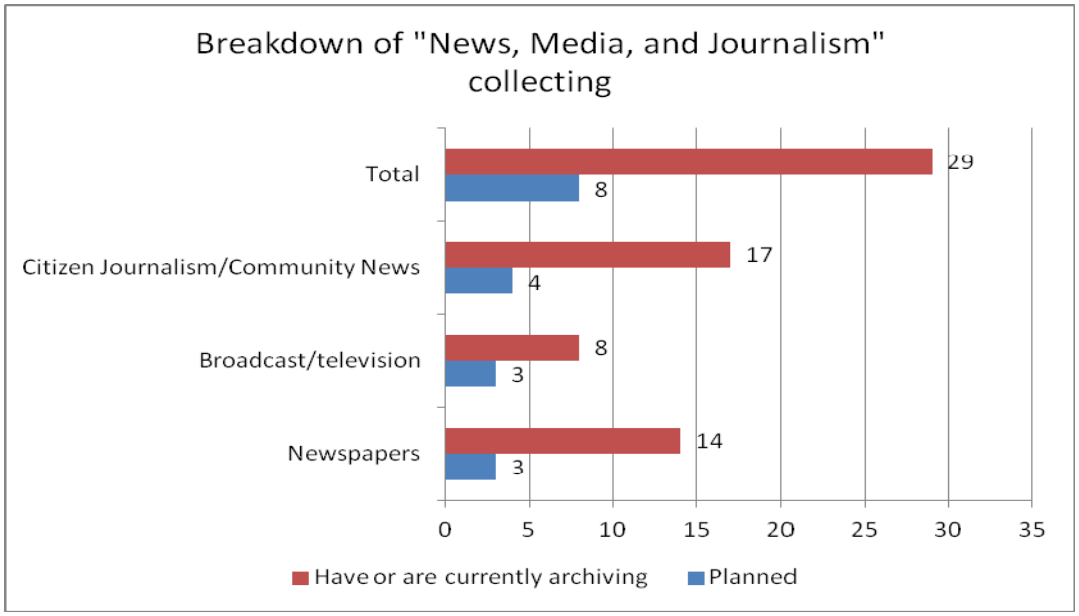
The below chart shows the types of subjects associated with past, current, and planned web archiving.

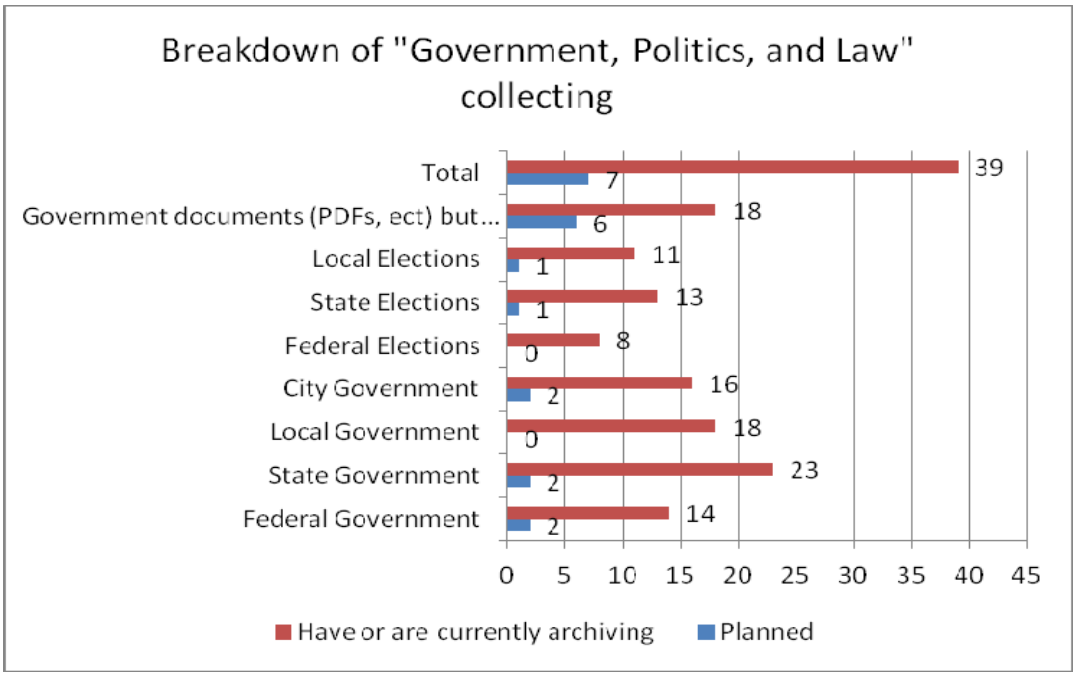
[5] A sample of some of the responses to this question can be found in the appendix.

What subjects are represented in your web archives?



The survey also sought additional detail about two of the above-listed categories: Government, Politics, and Law and News, Media, and Journalism. The following responses provide more granularities about the specific subject areas within these scopes of collecting. The "total" in the following charts is the number of respondents, but many institutions are collecting multiple types of content even within these specific categories, which accounts for the individual selections exceeding the totals.





4) Conclusion

The survey responses from institutions archiving the web and planning on building web collections reveal a number of themes underscoring current activities. These themes demonstrate both the success and collaborative potential of web archiving as well as continuing uncertainty around specific policies and processes.

The recent emergence of web archiving, especially at academic institutions
One finding revealed by the survey was the preponderance of universities that have initiated web archiving programs in the last 5 years. At the same time, self-identified archival and government institutions have initiated programs within the last decade at a similar rate as the previous decade. The preponderance of academic institutions as recent initiates of web archiving programs holds promise for encouraging knowledge-sharing across the community and an institutional alignment on related research into standards and best practices.

Collecting trends and collaborative potential
Collecting among the survey respondents fell into one of three categories: institutional self-documentation, existing collection enhancement, or thematic and event-based collections. The potential for collaboration was a notable aspect of these results. This includes not just knowledge-sharing, as mentioned above, but also shared acquisition projects. Of the survey respondents, 96% would consider participating in a joint collecting project. Web archiving clearly has a unique potential to bring institutions together around collaborative collection development.

Lack of policies and unclear guidance on permissions
Internal policy documentation appeared to be an area of continued improvement for many institutions. While some programs had incorporated web-materials into existing policies and procedures, others had not and some seemed unsure of their institution's current policy status for web content. The survey also brought to light a lack of clarity around seeking permission from content creators, both for harvesting and for providing access to collections – no doubt due to the difficulty of working with web content creators and the accompanying legal and intellectual property challenges. Since the survey, the Association of Research Libraries (ARL) has released both the Code of Best Practices in Fair Use for Academic and Research Libraries. This code, coupled with ARL's prior analysis of the legal issues surrounding web archiving, may provide additional guidance to institutions.[6]

Inconsistent custodianship
Another surprising result of the survey was the small number, only 18%, of institutions holding their own copy of their harvested collections. Storage and infrastructure challenges were the most frequently cited impediment to custodianship. Better documentation or case studies could help clarify the benefits (or perils) of transferring harvested content in-house.

[6] Adler, Prudence S., Patricia Aufderheide, Brandon Butler, and Peter Jaszi, *Code of Best Practices in Fair Use for Academic and Research Libraries* (Washington, D.C.: Association of Research Libraries, 2012), available at http://www.arl.org/bm~doc/code-of-best-practices-fair-use.pdf; for ARL's web archiving legal analysis, see Band, Jonathan, *A New Day for Website Archiving 2.0*, (Washington, D.C.: Association of Research Libraries, 2006), available at http://www.arl.org/bm~doc/band_webarchive2012.pdf.

5) Appendix

Sample Responses to the Scope of Collection Survey Question:

(Question: "Please briefly describe the scope of your web archive collections: what type of events, topics, themes, or approaches you take in archiving content from the web.")

"We are currently archiving our organization's own web page, the web pages of a number of organizations for which we serve as the archival repository (generally Native American missions, Catholic Social Action, etc.), and sites that feature topics related to collections in our holdings (e.g. J.R.R. Tolkien)."

"We archive human rights content, primarily published by grassroots human rights organizations who are publishing primary source material - video, photos, reports"

"Biannual snapshot of entire university-controlled domain and associated official domains. Exclusions are ephemeral, student, personal, restricted pages"

"4 types of collections: 1) subject based, related to other collecting strengths (e.g., human rights); 2) university website, related to University Archives; 3) organizations & individuals for whom we hold paper archives; 4) "rescue" of individual "at risk" websites"

"We archive websites related to the left and labor movement. We have multiple different collections, ranging from communism, labor union websites, alternative mass media, the progressive movement, economic and social justice, and other left activism. For example, we have archived documentation relating to Guantanamo Bay and right now we are focusing on the Occupy Wall Street movement."

"Public archives: 2003 California Recall Election, 2007 Southern California Wildfire Collection, 2010 Winter Olympics, Myanmar Cyclone Archive, Web at Risk Wiki Archive. Forthcoming: Deepwater Horizon Oil Spill. Dark Archives: Google Book Settlement, Nature Publishing Group Controversy, Retired Websites of the CDL (versions prior to significant redesigns)."

"We have gathered web content pertaining to local area government websites, the H1N1 Influenza Outbreak, Mexican Elections, campus racial tensions, and the 2010 Northern Mexico / Easter Sunday Earthquake"

"We are archiving the following:
-University websites
-State of Maryland websites that support our research/teaching mission
-websites related to historic preservation
-sites of organizations/individuals that support our existing archival collections"

Software Tools and Components Used by Survey Respondents:

Adobe Acrobat Pro Web Capture
Websites can be converted into PDFs using Adobe Acrobat's web capture function.

- Developer: Adobe
- Application: http://www.adobe.com/products/acrobat.html
- Documentation: http://help.adobe.com/en_US/Acrobat/9.0/Professional/WS58a04a822e3e50102bd615109794195ff-7f67.w.html
- Licensing: Fee-based

Archive-It
Archive-It, a subscription service from the Internet Archive, allows institutions to build and preserve collections of born digital content. Through a web application, Archive-It partners can harvest, catalog, manage and browse their archived collections. Collections are hosted at the Internet Archive data center and are accessible to the public with full-text search.

- Developer: Internet Archive
- Application: http://www.archive-it.org/
- Documentation: http://webteam.archive.org/confluence/display/ARIH/Welcome
- Licensing: Fee based

Grab-a-Site
"Grab-a-Site 5.0 is a file-based Offline Browser that combines speed, stability, and powerful filtering capabilities… that can that can download an entire web site while retaining the original filenames and directory structure."

- Developer: Blue Squirrel
- Application: http://www.bluesquirrel.com/products/grabasite/
- Documentation: http://www.bluesquirrel.com/products/grabasite/htmlmanual/index.html?Product=Grab-a-Site
- Licensing: Fee-based

Heritrix
Heritrix is a flexible, extensible, robust, and scalable Web crawler capable of fetching, archiving, and analyzing Internet-accessible content.

- Developer: Internet Archive
- Application: http://crawler.archive.org
- Documentation: http://crawler.archive.org/articles/user_manual and http://webteam.archive.org/confluence/display/Heritrix/Home
- License: GNU Lesser General Public License 2.1 (http://crawler.archive.org/license.html); migrating to Apache License 2.0 in future

HTTrack
HTTrack "allows you to download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer. HTTrack arranges the original site's relative link-structure. HTTrack can also update an existing mirrored site, and resume interrupted downloads. HTTrack is fully configurable, and has an integrated help system."

- Developer: Xavier Roche, Yann Philippot , and others
- Application: http://www.httrack.com/page/2/en/index.html
- Documentation: http://www.httrack.com/html/index.html
- Licensing: GNU General Public License, version 3

Teleport Pro
"Teleport Pro is an all-purpose high-speed tool for getting data from the Internet… Capable of reading HTML 4.0, CSS 2.0, and DHTML." Teleport Pro has "server-side image map exploration, automatic dial-up connecting, Java applet support, variable exploration depths, project scheduling, and relinking abilities."

- Developer: Tennyson Maxwell Information Systems, Inc
- Application: http://www.tenmax.com/teleport/pro/download.htm
- Documentation: http://www.tenmax.com/teleport/support.htm
- Licensing: Fee-based

Wayback Machine
The Wayback Machine is a powerful search and discovery tool for use with collections of Web site "snapshots" collected through Web harvesting, usually with Heritrix (ARC or WARC files).

- Developer: Internet Archive
- Application: http://archive-access.sourceforge.net/projects/wayback/
- Documentation: http://archive-access.sourceforge.net/projects/wayback/administrator_manual.html
- Licensing: GNU Lesser General Public License 2.1 (http://archive-access.sourceforge.net/projects/wayback/license.html); migrating to Apache License 2.0 in future

Web Archives Workbench
The Web Archives Workbench is a suite of Web capture tools based on principles of managing archived content in aggregates rather than as individual objects. The suite is comprised of:

Discovery Tool, which helps identify potentially relevant Web sites by crawling relevant "seed" Entry Points to generate a list of domains that they link to.

Properties Tool, which enables you to maintain information about content creators, associate them with the Web sites they are responsible for, and enter high-level metadata.

Analysis Tool, enables you to look at the structure of the Web site to see what kind of content is represented by the file directory.

Harvest Tool, which is used to monitor crawl status, to review and modify harvest settings, and to package harvests for transfer to a repository. The Harvest Tool also offers a separate Quick Harvest feature, which schedules one-time harvests of content. Harvest packages are encoded in METS with Dublin Core metadata embedded.

- Developer: OCLC
- Application: Download from SourceForge, http://sourceforge.net/projects/webarchivwkbnch
- Documentation: Available on SourceForge
- Licensing: Available on SourceForge

Web Archiving Service

The Web Archiving Service (WAS) is a Web-based curatorial tool that enables libraries and archivists to capture, curate, analyze, and preserve Web-based government and political information. The WAS allows users to set parameters of Web crawls, capture sites, provide metadata for archived sites, and build collections of archived Web sites.

- Developer: California Digital Library
- Application: Web based, http://was.cdlib.org
- Documentation: http://was.cdlib.org
- Licensing: Fee-based

Web Harvester

A service that enables users to harvest content from the Web, review it and add the harvested items to their CONTENTdm® collections during the Connexion cataloging process. By integrating digital collection development and capture with standard cataloging workflows, the Web Harvester provides an additional option for expanding participation in growing and maintaining digital collections. Harvested items added to CONTENTdm Digital Collection Management Software using the Web Harvester are discoverable from the CONTENTdm Web interface, as well as WorldCat.org, WorldCat Local and OCLC FirstSearch. Each harvested item added to CONTENTdm using the Web Harvester is associated with its WorldCat record via a persistent URL based on the OCLC number of the WorldCat record.

- Developer: OCLC
- Application: http://oclc.org/webharvester
- Documentation: http://www.oclc.org/webharvester/support/default.htm
- Licensing: Fee-based

Wget

GNU Wget is a free software package for retrieving files using HTTP, HTTPS and FTP, the most widely-used Internet protocols. It is a non-interactive commandline tool, so it may easily be called from scripts, cron jobs, terminals without X-Windows support, etc.

GNU Wget has many features to make retrieving large files or mirroring entire web or FTP sites easy.

- Developer: Hrvoje Nikšić & Giuseppe Scrivano
- Application: http://www.gnu.org/software/wget/
- Documentation: http://www.gnu.org/software/wget/manual/
- Licensing: GNU General Public License