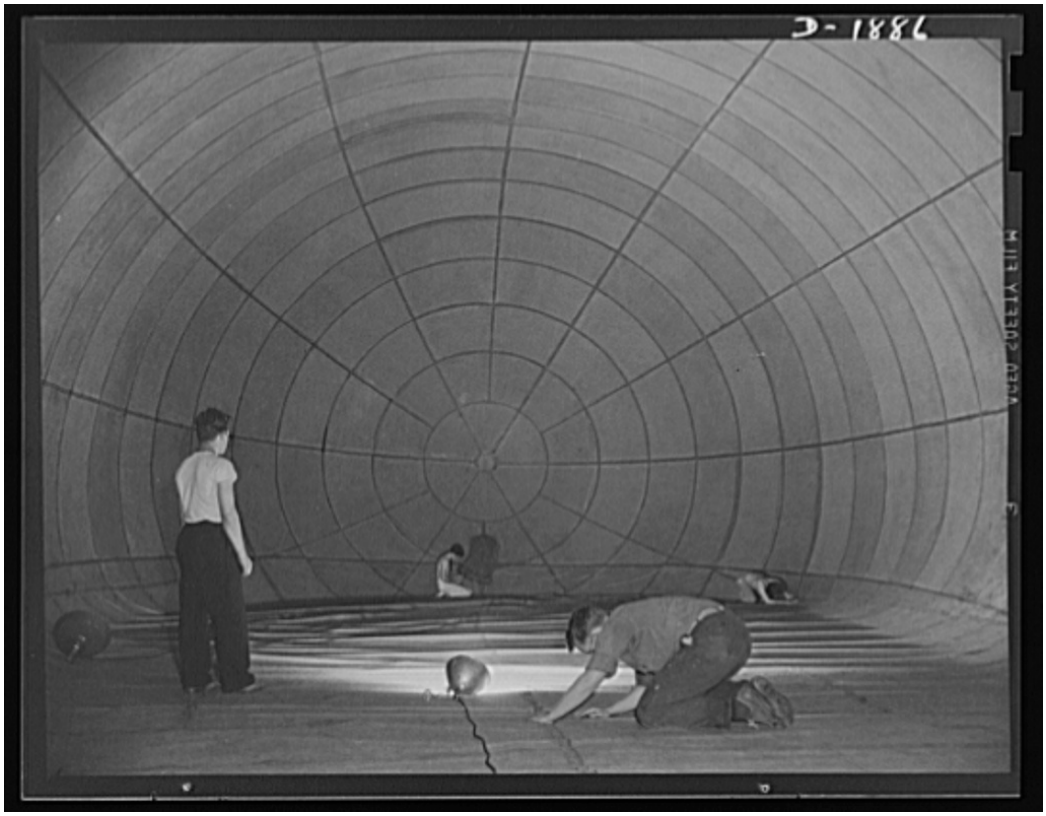
# WEB ARCHIVING IN THE UNITED STATES: A 2013 SURVEY

## AN NDSA REPORT



September 2014

Results of a Survey of Organizations Preserving Web Content


**Authors**
**Jefferson Bailey, Internet Archive**
**Abigail Grotke, Library of Congress**
**Kristine Hanna, Internet Archive**
**Cathy Hartman, University of North Texas**
**Edward McCain, Missouri School of Journalism**
**Christie Moffatt, National Library of Medicine**
**Nicholas Taylor, Stanford University Libraries**

## Contents

## ABOUT THE NATIONAL DIGITAL STEWARDSHIP ALLIANCE

Founded in 2010, the National Digital Stewardship Alliance (NDSA) is a consortium of institutions that are committed to the long-term preservation of digital information. NDSA's mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. NDSA member institutions represent all sectors, and include universities, consortia, professional associations, commercial enterprises, and government agencies at the federal, state, and local levels.

More information about the NDSA is available at http://www.ndsa.org.



Persistent URL: http://hdl.loc.gov/loc.gdc/lcpub.2013655118.1

Cover: Palmer, Alfred T. "Barrage balloon manufacture. Spider and the flies. Putting last touches on a completed barrage balloon are these workers and their supervisor (standing). To inspect the bag for leaks, the light inside is turned off, leaving the workers in a midnight gloom. Then a co-worker on the outside takes a bright light, and works along the length of the entire balloon, back and forth, until he has covered every inch. The tiniest pinhole shines forth like a bright star in the sky, and these alert inspectors spot it and repair it. Cloth moccasins are worn inside the bag, or else stocking feet--to avoid chafing the fabric. General Tire and Rubber Company, Akron, Ohio, " December 1941. Part of: Farm Security Administration - Office of War Information Photograph Collection (Library of Congress). Available online at http://www.loc.gov/pictures/item/oem2002010545/PP/

# INTRODUCTION

From October through November of 2013, a team of individuals representing multiple NDSA member institutions and Working Groups 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. This effort builds upon a similar survey undertaken by the NDSA in late 2011 and published online in June of 2012.[1] The goal of the survey was to better understand the landscape of web archiving activities in the U.S. by investigating the organizations involved, the history and scope of their web archiving programs, the types of web content being preserved, the tools and services being used, access and discovery services being provided, and overall policies related to web archiving programs. While this survey documents the current state of U.S. web archiving initiatives, comparison with the results of the 2011-2012 survey enables an analysis of emerging trends. This report therefore describes the current state of the field, tracks the evolution of the field over the last few years, and forecasts future activities and developments.

# METHODOLOGY

The team that oversaw the survey self-organized in August 2013 to begin drafting the survey questions. Two goals identified early in the process were to enable historical comparisons with the 2011 survey and to learn about program details that were not inquired about in the previous survey. Accordingly, the group reviewed and refined the existing questions from the 2011 survey and added new questions to address emerging activities and issues. The updated survey took place from October 22, 2013 to November 30, 2013 using the SurveyMonkey online survey tool, and was promoted via blogs, listservs, social media, and other channels. When the survey concluded, the group reviewed the responses and removed test or mostly-incomplete entries.

## The Survey Content

The survey consisted of twenty-seven questions organized around five distinct topic areas: background information about the respondent's organization; details regarding the current state of their web archiving program; tools and services used by their program; access and discovery systems and approaches; and program policies involving capture, availability, and types of web content.

The 2013 NDSA Web Archiving Survey was started 109 times, and completed 92 times for an 84% completion rate. The 92 completed responses represented an increase of 19% in the number of respondents compared with the 77 completed responses for the 2011 survey. The survey consisted primarily of multiple-choice questions, with some questions also containing free-text response fields for clarification or elaboration of answers.

## The Survey Data

Anonymized survey data is available online and can also be provided upon request.[2]

[1] "NDSA Web Archiving Survey Report," June 19, 2012. Available at http://www.digitalpreservation.gov/ndsa/working_groups/documents/ndsa_web_archiving_survey_report_2012.pdf

[2] Content & Innovation Working Groups, 2014, "NDSA Web Archiving Surveys", http://dx.doi.org/10.7910/DVN/27593 National Digital Stewardship Alliance [Distributor]. A link to the PDF of survey questions can be found in Appendix B.

# RESPONDENT CHARACTERISTICS

## Organization Type

The survey started by asking respondents for their organization name[3] and organization type and whether they belonged to any of three web archiving-related professional groups. Even with nearly 20% more survey respondents in 2013 than in 2011, the proportion of represented organization types was remarkably similar. In the 2013 survey, just over half, 52% (48 of 92) of the respondents characterized their organizations as Colleges or Universities, a slight increase from the 47% (36 of 77) of the 2011 survey. Percentages for all other possible responses changed by less than five percent between the two surveys. Examples of organizations that self-identified as "other" included non-profits, academic consortia, local governments, and independent research libraries.

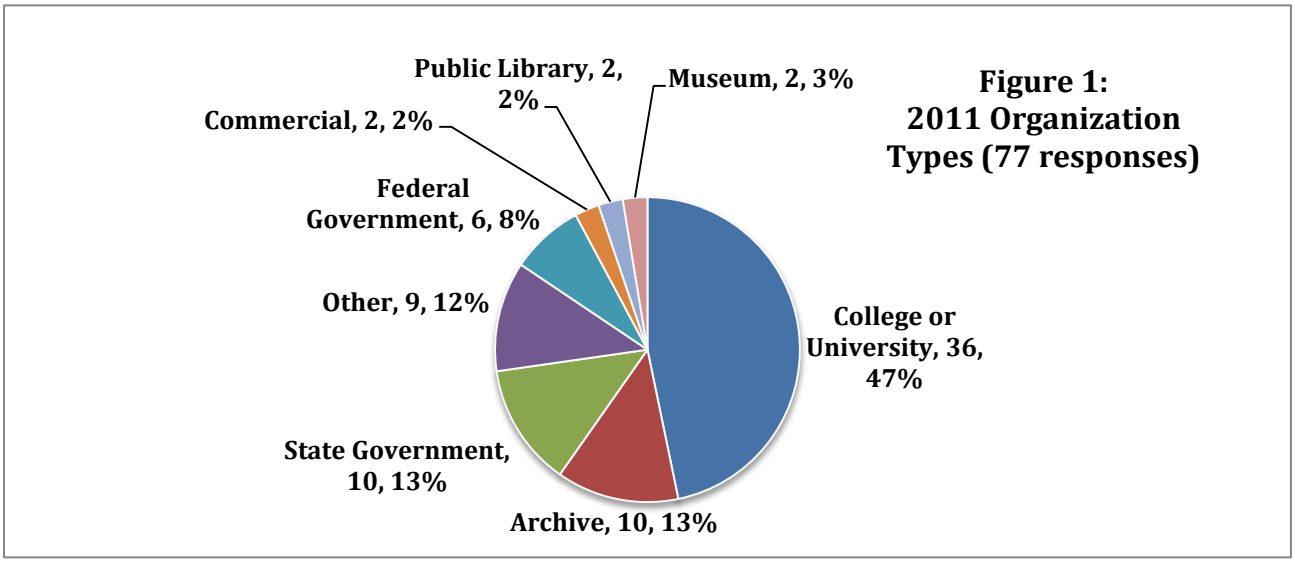



**Figure 1: 2011 Organization Types (77 responses)**

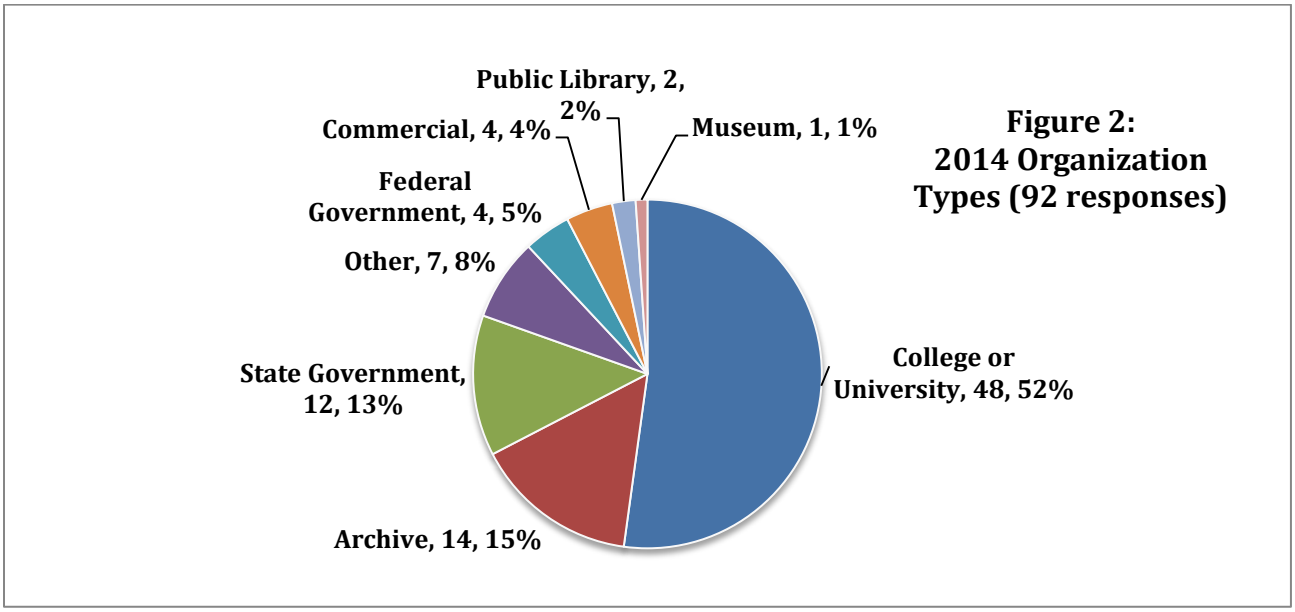



**Figure 2: 2014 Organization Types (92 responses)**

[3] Note: the names of responding organizations have been removed from the publicly-released data.

### Group Affiliations

Web archiving-related professional group affiliations also remained consistent across the 2011 to the 2013 surveys, with 7% (6 of 92) membership in the International Internet Preservation Consortium (IIPC),[4] as opposed to 8% in 2011, and 33% (30 of 92) membership in the NDSA, in comparison to 31% in 2011. The 2013 survey additionally asked if organizations belonged to the Society of American Archivists (SAA) Web Archiving Roundtable,[5] which was established since the 2011 survey with membership open to any interested individual. The SAA Roundtable was the most popular of the three groups, with 45% (41 of 92) of respondents citing affiliation.

## ARCHIVING PROGRAM INFORMATION

The goal of this section of the survey was to learn more about the state of web archiving programs, perceptions of progress, broad areas of collecting interest, metrics, devoted staff time, content types of concern for archiving, and collaborative archiving. More so than the other sections of the survey, the questions about web archiving programs were qualitative and subjective. Though the survey featured only one respondent per organization, and that individual's perspective may diverge from that of others within the same enterprise, in aggregate the responses illuminate a number of specific trends across programs.

### Program Status

Like the 2011 survey, the 2013 survey asked about the status of organizations' web archiving efforts. The available answers to this question remained the same in both instances:  active, testing, planning, and no longer collecting. At least six organizations that listed their status as testing or planning in 2011 report active programs in the 2013 survey. More broadly, testing and planning together account for 23% (21 of 89) of the responses in 2013, compared to 33% of the responses in 2011. This near 10% decrease in testing and planning over the past two years paralleled an 8% increase in the number of active programs, suggesting the growing maturity of existing efforts.

[4] More information on the IIPC can be found at http://netpreserve.org/.

[5] More information on the SAA Web Archiving Round Table can be found at http://webarchivingrt.wordpress.com/.

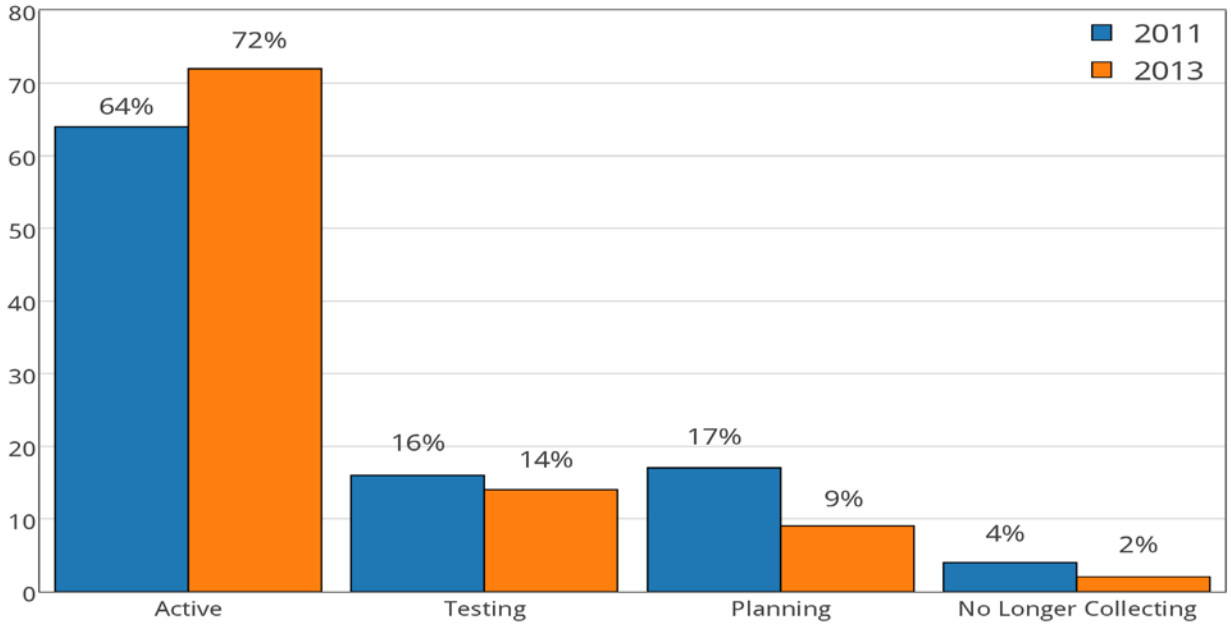



**FIGURE 1: PROGRAM STATUS**

## Perceptions of progress since 2011

Survey respondents are also overwhelmingly positive about the progress of their programs, with 74% (67 of 90) reporting some or significant progress compared to two years ago. From the comments and other responses, it is clear that many of these programs indicating progress are recent initiatives, which may engender perceptions of significant progress (going from having not having a program to having one at all may understandably be described as significant progress). However with only 24% (22 of 90) of respondents indicating a similar or worse level of progress, it is clear that the growth in the number of web archiving programs since the last survey is matched by widespread perceptions of overall progress.

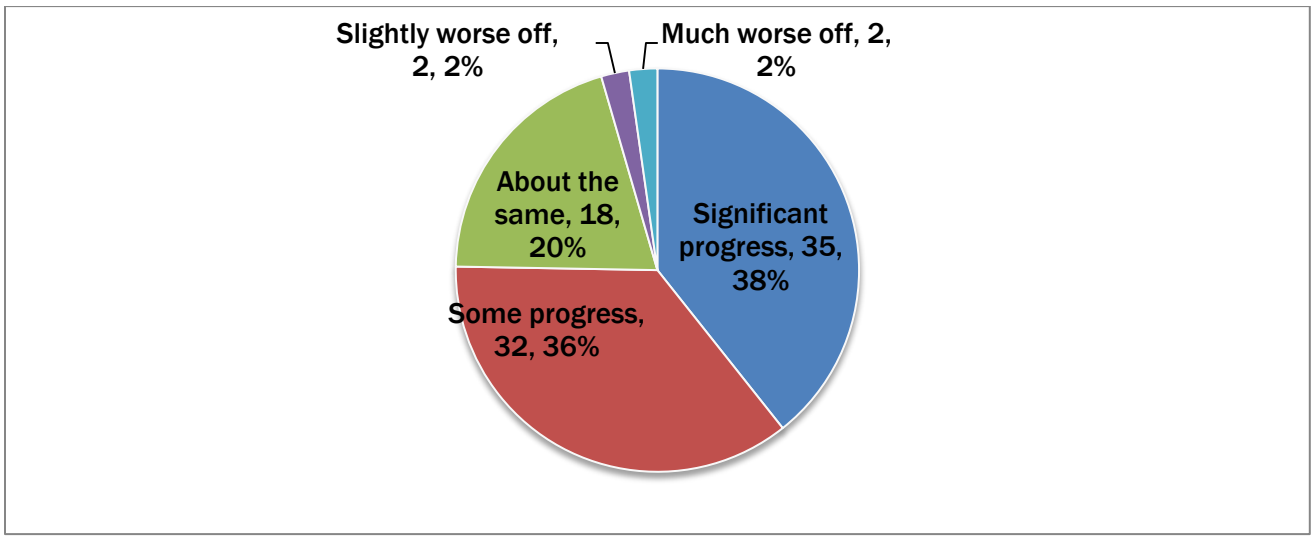



**FIGURE 2: COMPARISON OF ORGANIZATION'S WEB ARCHIVING PROGRAM TO 2011**

## Ownership of Content Being Archived

Respondents are more focused on archiving their own or affiliated content in 2013 than they were in 2011, with those reporting that they are only collecting such content nearly doubling from 20% (14 of 71) to 37% (32 of 86). There was a near-identical drop in the proportion of programs focusing only on third-party content, from 31% (22 of 71) to 15% (13 of 86). Respondents archiving both their own or affiliated and third-party content remained steady across the 2011 and 2013 surveys, with about half of the organizations taking this approach.

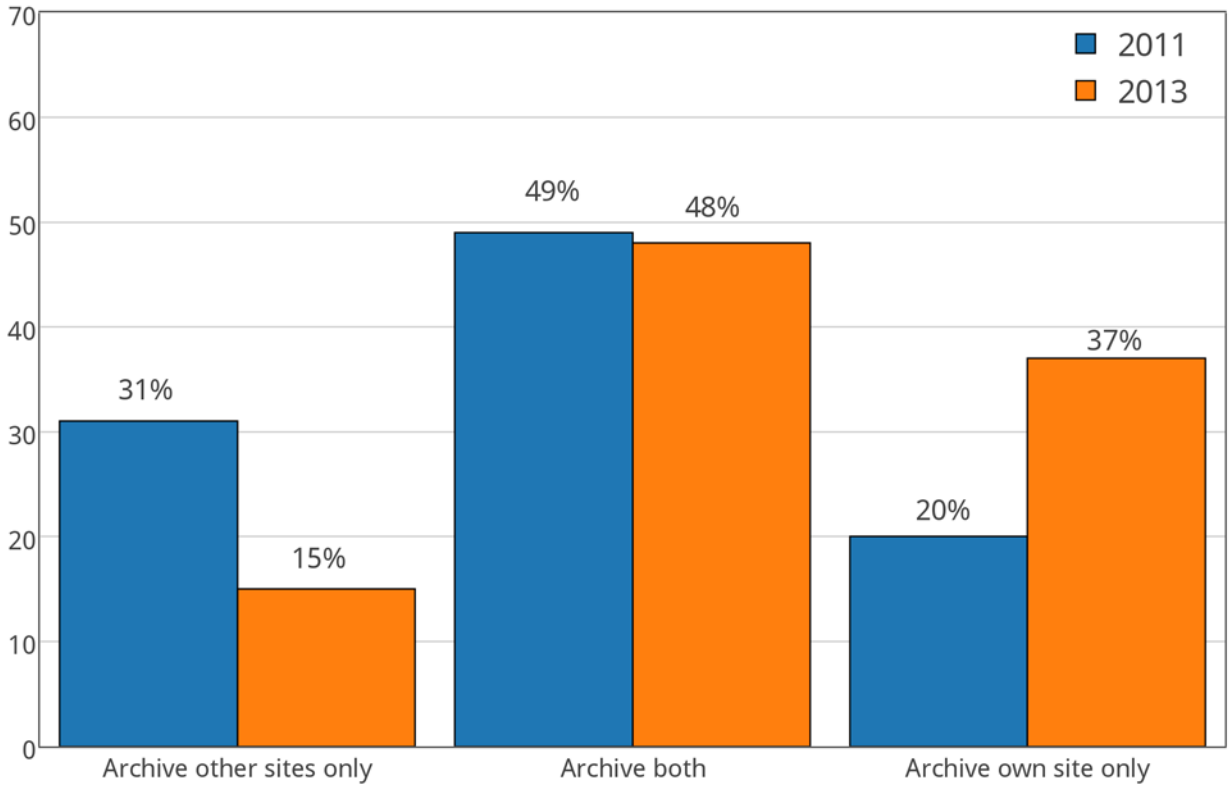



**FIGURE 3: COMPARISON OF ARCHIVING ACTIVITY GOALS**

## When Programs Started

The last two years saw a major increase in the number of organizations starting web archiving programs; as in 2011, 38% (31 of 81) of respondents indicated that their organizations began within the past two years. The distribution of start dates reported in the 2011 and 2013 surveys did not otherwise closely match. This is attributed to variation in which organizations responded to the survey and the fuzziness of program start dates when programs typically go through planning and pilot phases before becoming active.

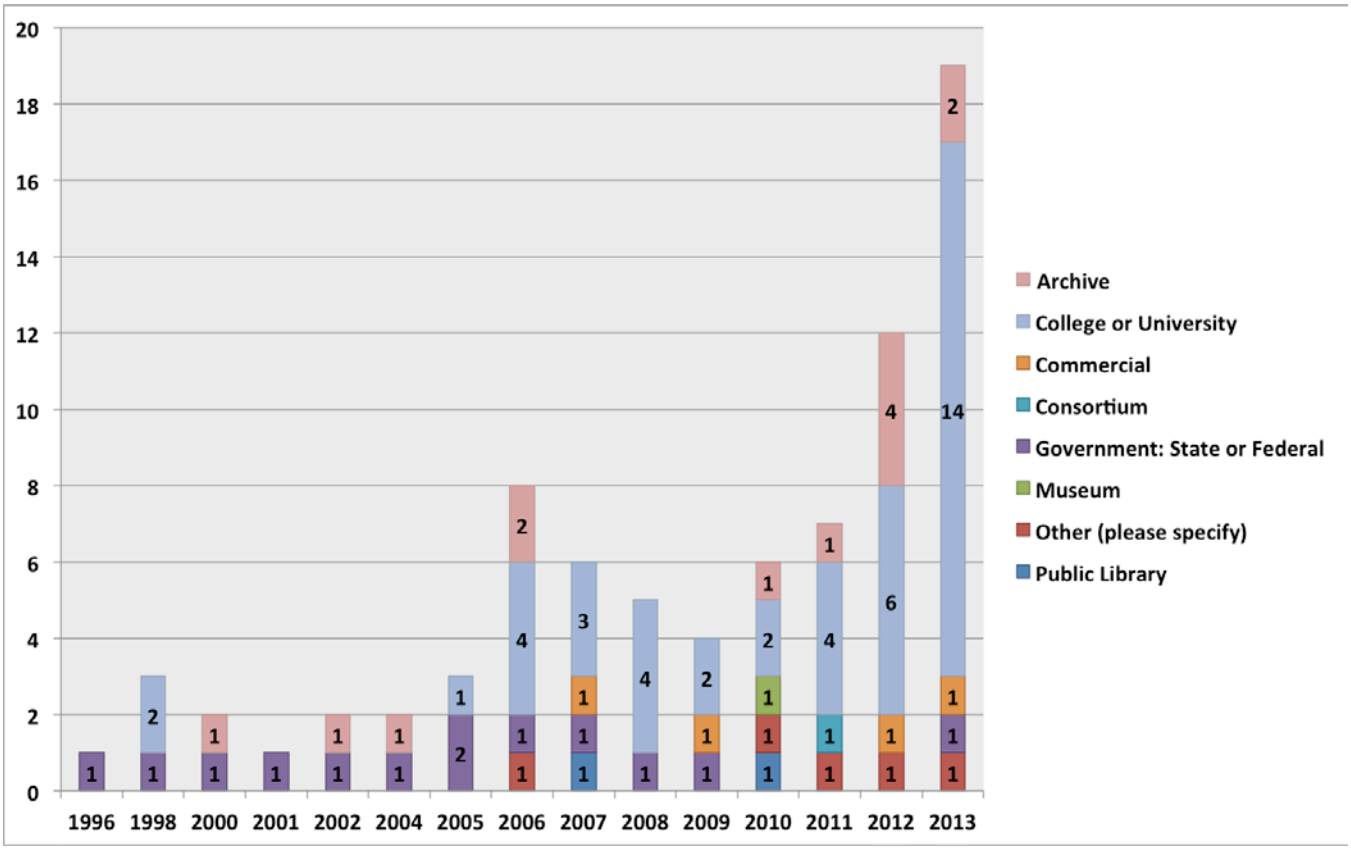



**FIGURE 4: YEAR INSTITUTIONS BEGAN ARCHIVING WEB CONTENT**

## Devoted Staff Time

Organizations most commonly devote fractional staff time to web archiving activities; in the 2013 survey, 81% (71 of 88) devote half or less of the equivalent of one full-time (FTE) staff person's time. Only 19% (17 of 88) of organizations devote at least one FTE staff person's time. The median value of devoted staff time is one-quarter FTE. While devoted staff time was not a question in 2011, and these numbers may reflect the newness of many programs, or the growing use of external services to harvest content (see report section "External or In House?" ), the very high rate of institutions devoting only fractional staff time to web archiving merits further scrutiny. Whether this level of staffing was seen as adequate by respondents was not evaluated; other studies, such as the Staffing for Effective Digital Preservation: An NDSA Report [6], suggest program staff believe digital preservation is generally understaffed. The same may hold true for web archiving programs specifically.

[6] "Staffing for Effective Digital Preservation: An NDSA Report," December 2013, is available at http://lcweb2.loc.gov/master/gdc/lcpubs/2013655113.pdf

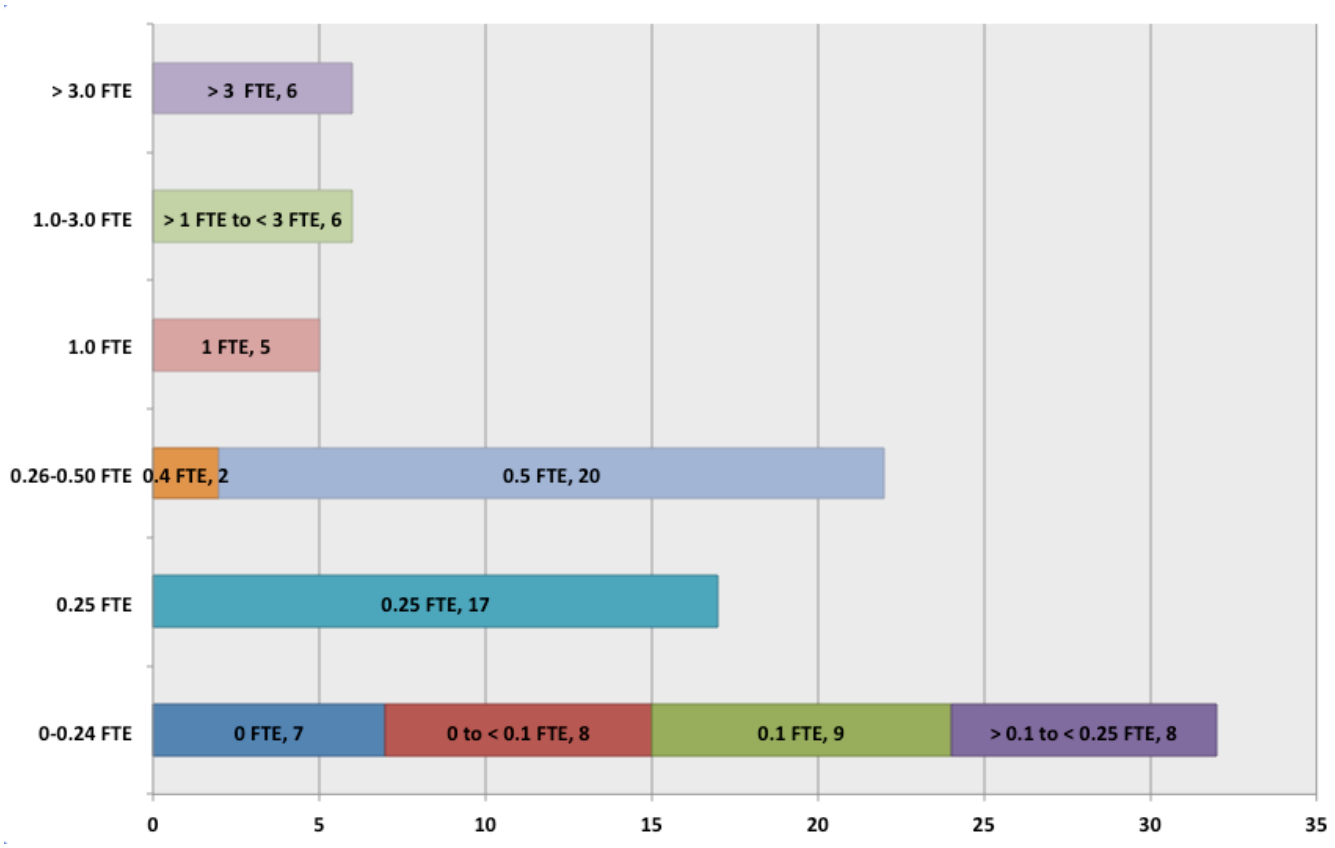



**FIGURE 5: FTE STAFF DEDICATED TO WEB ARCHIVING**

## Skills

One of two new, open-ended questions on the 2013 survey concerned the staff skills necessary to the development and success of organizations' web archiving programs. Responses were coded into one or more of eight categories: web technologies, archiving tools, domain expertise, appraisal, metadata, collaboration and communication, software development, and quality assurance.[7]

Respondents consider technical skills to be the most necessary to the development and success of their programs; almost 40% indicated that knowledge of web technologies (24 of 62) or archiving tools (23 of 62) is essential. The emphasis on technical skills in the responses was followed closely by curatorial skills, with 24% (15 of 62) indicating domain expertise as important and 21% (13 of 62) indicating appraisal. Quality assurance was cited least frequently out of all the coded categories at 6% (4 of 62), suggesting that respondents may be either satisfied with or resigned to the performance of their tools.

[7] A description of how these responses were coded can be found in Appendix A.

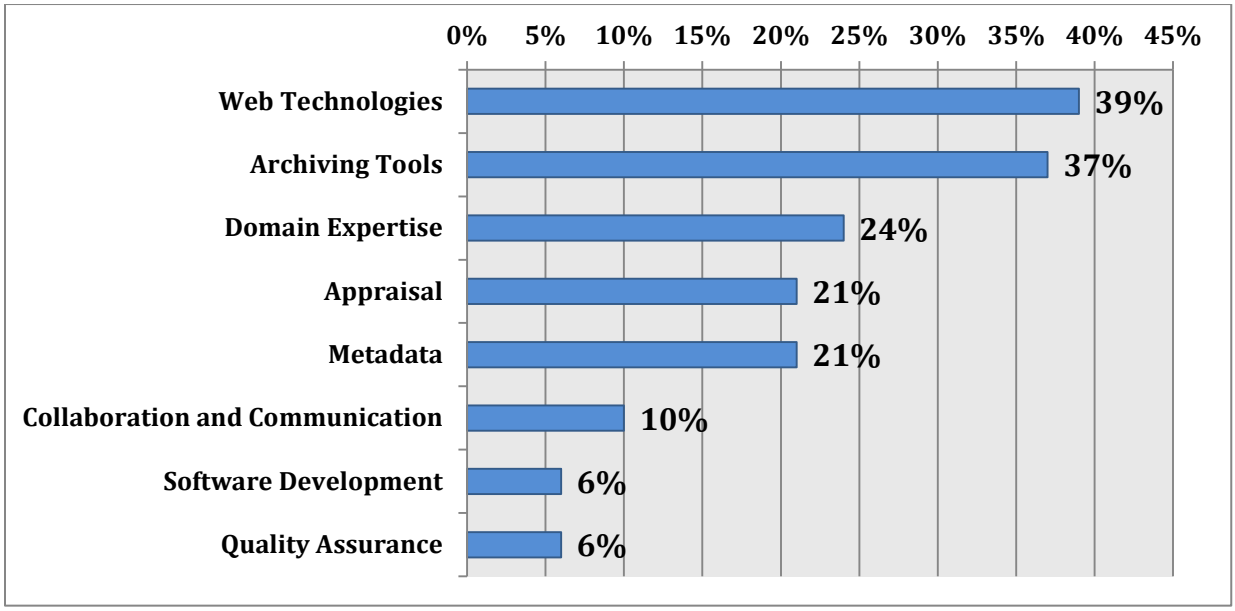



**FIGURE 6: SKILLS DEEMED ESSENTIAL FOR STAFF**

Other skills highlighted by respondents as being important included attention to detail, analysis, organization, project and program management, appropriate application of policy, patience, persistence, flexibility, adaptability, and being a fast learner.

## Metrics

The other new, open-ended question on the 2013 survey concerned the metrics that were important to the development of organizations' web archiving programs. Responses were coded into one or more of seven categories: volume, usage, cost, quality, buy-in, loss, and policy.[8]

Respondents are most interested in metrics relating to volume and usage, with around 50% citing one or the other. Though only 22% (11 of 51)of those who responded to this question specified cost as important to their web archiving program development, volume metrics may be seen as a close proxy for cost, given that third-party service models are typically based upon storage and number of captured object limits. The 47% (24 of 51) of respondents who mentioned usage-related metrics show that there is an interest in the community in access and use. The even split between tracking volume (an administrative, internal-facing concern), and usage (an external, patron-facing concern) is a logical balance, operations-wise. Also interesting is the lower number of responses citing metrics about "loss" or "quality" as a concern. Though cost is somewhat implied in the importance of metrics regarding volume, the disparity in these two responses suggest a dedicated and predetermined financial commitment to web archiving but an ongoing concern regarding matters of managing capture, scoping collections, and the extent of content acquisition.

[8] A description of how these responses were coded can be found in Appendix A.

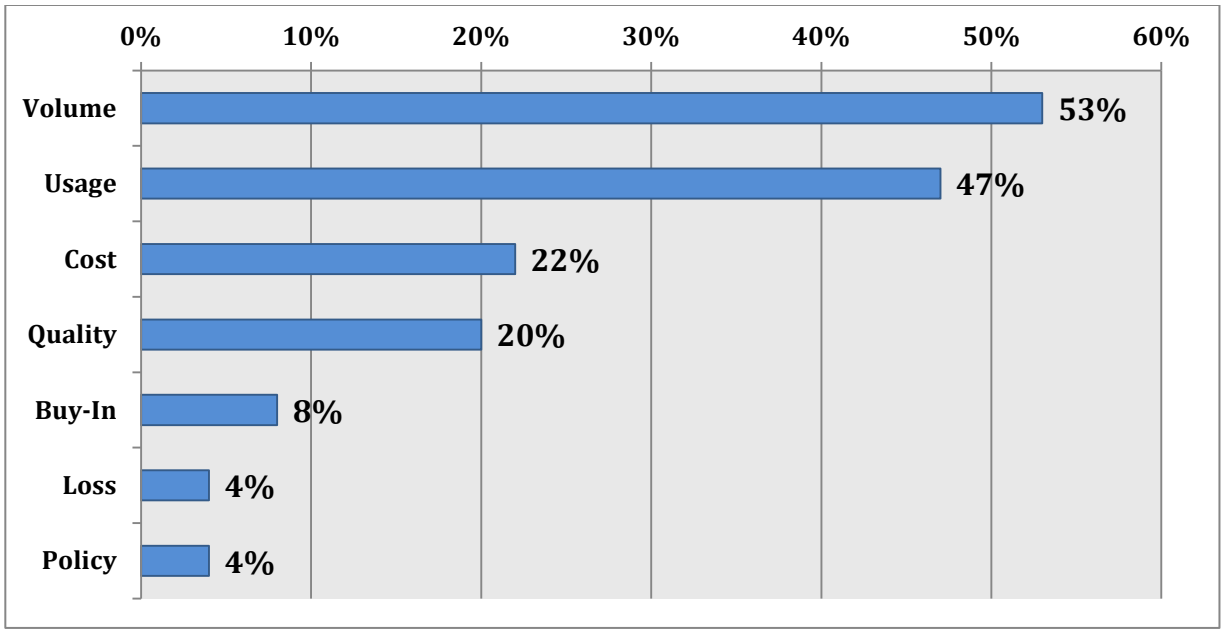



**FIGURE 7: METRICS IMPORTANT TO PROGRAM DEVELOPMENT**

## Content Types of Concern

The 2013 survey inquired for the first time about content types that organizations were concerned about their capacity to archive. The question permitted choosing multiple content types. Respondents expressed the most concern about their ability to archive social media 79% (68 of 86), databases 74% (64 of 86), and video 73% (63 of 86). Art was the content type of concern to the fewest number of respondents 17% (15 of 86); this could reflect the relatively small proportion of responding organizations focused on collecting web-based art and not a diminished recognition that archiving this material is any less problematic.

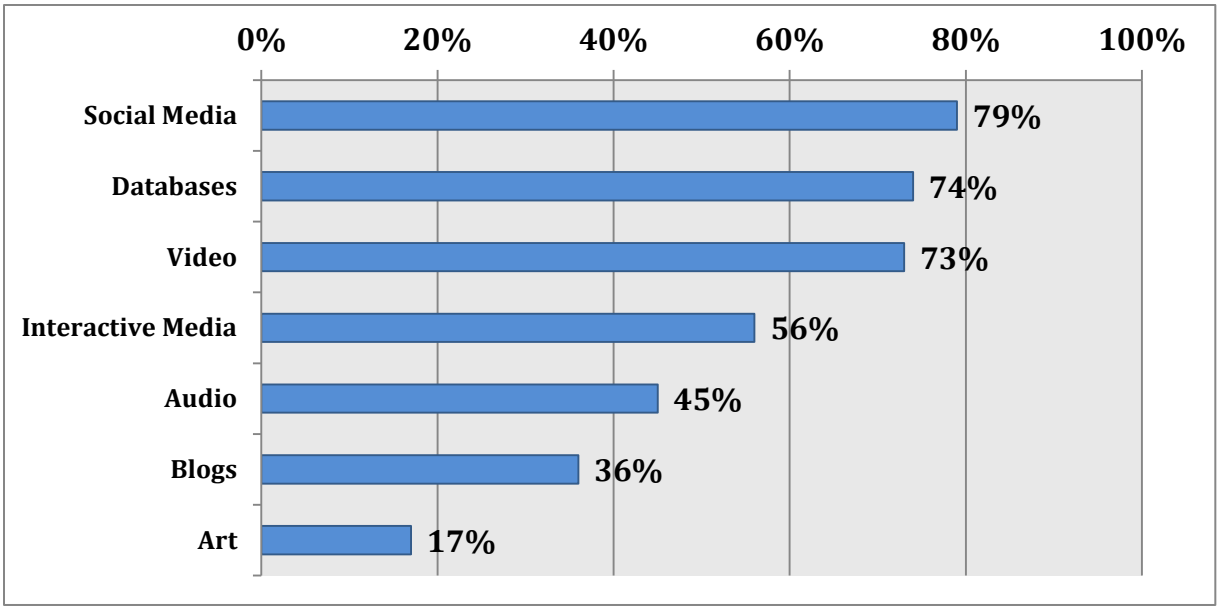



**FIGURE 8: TYPE OF CONTENT PROVOKING CONCERN OVER CAPACITY TO ARCHIVE**

### Collaborative Archiving

Collaborative web archives are those in which multiple organizations provide curatorial expertise by nominating websites to be preserved, but the collection is made available through a single point of access and housed on one institution's infrastructure. Examples of collaborative collections are the Ukraine Conflict Web Archive and the 2010 Winter Olympics Web Archive.[9] The two questions regarding collaborative collecting in the 2011 survey were combined into a single question in the 2013 survey. While a comparable proportion of respondents indicated in 2011 and 2013 that their organizations had or had not participated in a collaborative web archive, the 2013 survey additionally allowed respondents whose organizations had not participated in a collaborative web archive to indicate their interest in doing so. There is substantial interest, indicating an opportunity for greater collaboration.

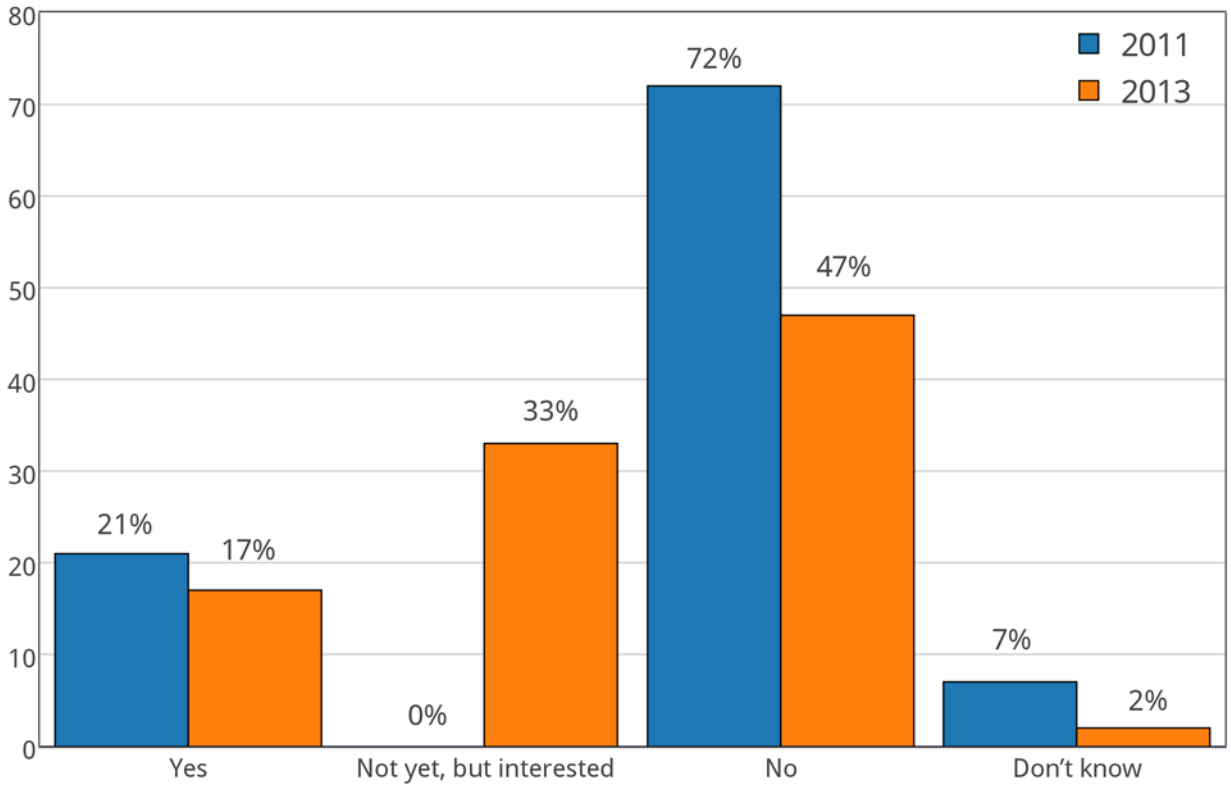



**FIGURE 9: PARTICIPATION IN A COLLABORATIVE WEB ARCHIVE**

## ARCHIVING POLICIES

The goal of this section of the survey was to learn about the policies that govern organizations' web archiving activities. Specific areas of inquiry included permission and notification requirements, approaches to robots.txt directives, policy guidelines specific to social media, and copyright and access policy development resources. The survey also included, for the first time in the 2013 survey, a question on access embargo periods

[9] The collections can be found at https://archive-it.org/collections/4399 and http://webarchives.cdlib.org/a/2010olympics/about

## Notification and Permission

A slight majority of respondents, 58% (42 of 73) capture web content without either notifying or seeking permission from content owners. Accounting for the remainder, 23% (17 of 73) send notice of their intent to capture content and 19% (14 of 73) seek permission from content owners. Direct comparison between the relevant 2011 and 2013 survey questions is difficult due to the former's focus on permissions and omission of notification as a possible option. In the 2011 survey, 25% (15 of 61) indicated that they never sought permission to capture from content owners. The proportion of organizations seeking permission is up slightly in 2013 to 19% (14 of 73) from 13% (8 of 61) in 2011.

Respondents have slightly more liberal approaches to providing public access to archived web content than capturing it, with 63% (45 of 71) neither notifying nor seeking permission, 15% (11 of 71) notifying content owners, and 21% (15 of 71) seeking permission to provide public access. This is contrary to the expectation that organizations are more sensitive about access than capture. Once again, as notification was not provided as a response in the 2011 survey, making a comparison of the 2011 and 2013 survey results a challenge. In the 2011 survey, 36% (21 of 59) of respondents never sought permission and 8% (5 of 59) always sought permission to provide public access.

The 2013 survey also featured a new question regarding notification and permission approaches for providing restricted access, such as through onsite terminals at an organization or narrowly providing access to specific researchers in response to qualified requests. The proportion of respondents neither notifying nor seeking permission approximately mirrors that for providing public access, with 68% (42 of 62) of respondents.

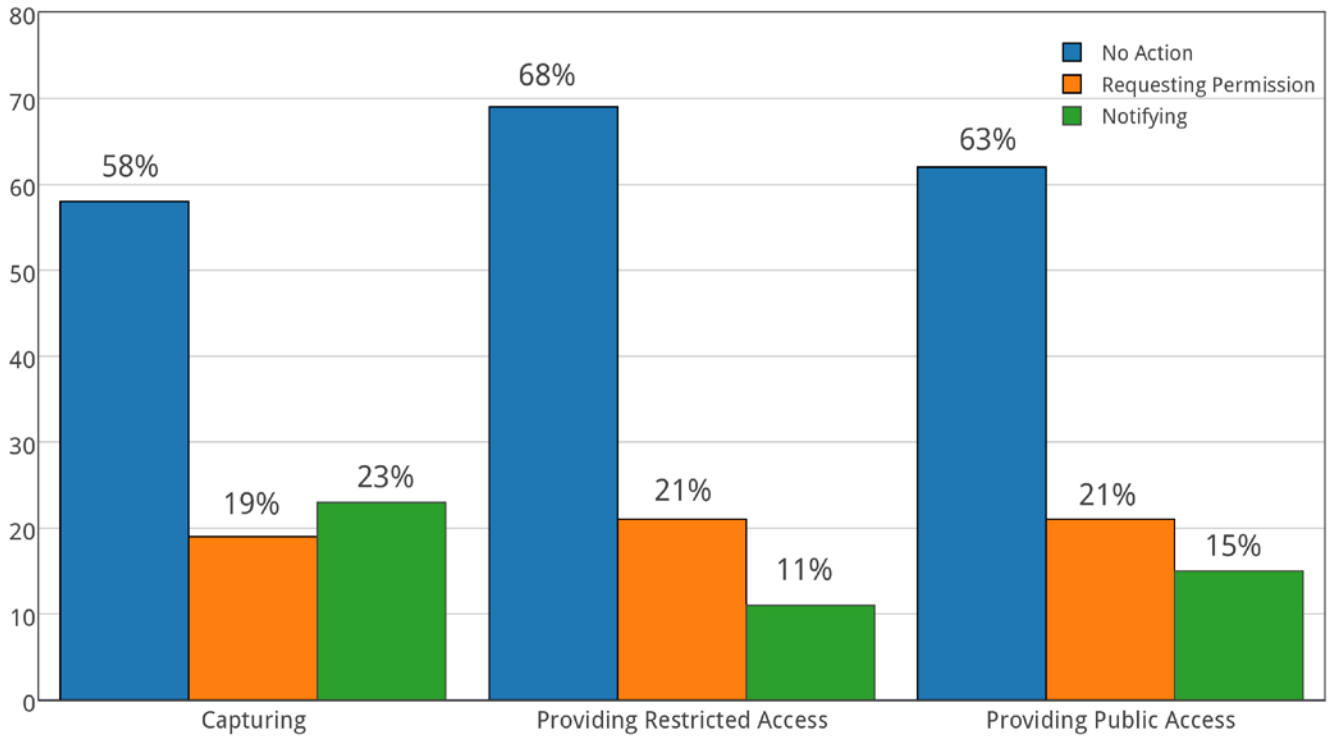



**FIGURE 10: APPROACHES TOWARDS SEEKING PERMISSIONS (2013 RESULTS)**

A comment field provided respondents the opportunity to explain the rationale for particular notification or permission requirements or the lack thereof. Responses provided in this field indicated that the multiple-choice question structure by itself missed the nuance that many organizations have a conditional approach to notifications or permissions. Some respondents seek permission, even multiple times, and proceed if there is no response after a defined period of time. Additionally, some organizations typically send notifications but escalate to seeking permission if robots.txt directives would adversely affect capture.

Organizations that are responsible for archiving government web content cite statutory frameworks at the state level and the public domain status of U.S. Government works at the federal level that exempt them from permission requirements. Some work with government agencies directly to ensure that appropriate content is archived.

Universities similarly eschew notification and permission requirements when archiving web content hosted on their own domains. Several respondents noted that university content hosted on third-party domains did not necessarily qualify for this exception, and other university respondents indicated that some sort of permission requirement was ultimately likely but had not yet been formalized.

## Approaches to robots.txt

Robots.txt is a machine-readable protocol used typically by website owners to request that search engine crawlers ignore certain content so that it does not appear in search results. Robots.txt is often used to mitigate traffic overload from crawling, or for other editorial reasons. Robots.txt directives similarly impact archival crawlers' ability to access web content but are more problematic in the case of web archives, which depend on a greater range of content than is typically useful for a search index.

More organizations have adopted conditional approaches to robots.txt in the last two years, perhaps suggesting the maturation of policies. The 2013 survey found that 55% (42 of 77) of respondents conditionally respect robots.txt compared to 33% (19 of 58) in 2011, a 21% increase. The proportion of respondents that always respect robots.txt meanwhile dropped from 38% (22 of 58) in 2011 to 22% (17 of 77) in 2013. The proportion of respondents always ignoring robots.txt remained the same, at 8% (6 of 77) in 2013 and 9% (5 of 58) in 2011.

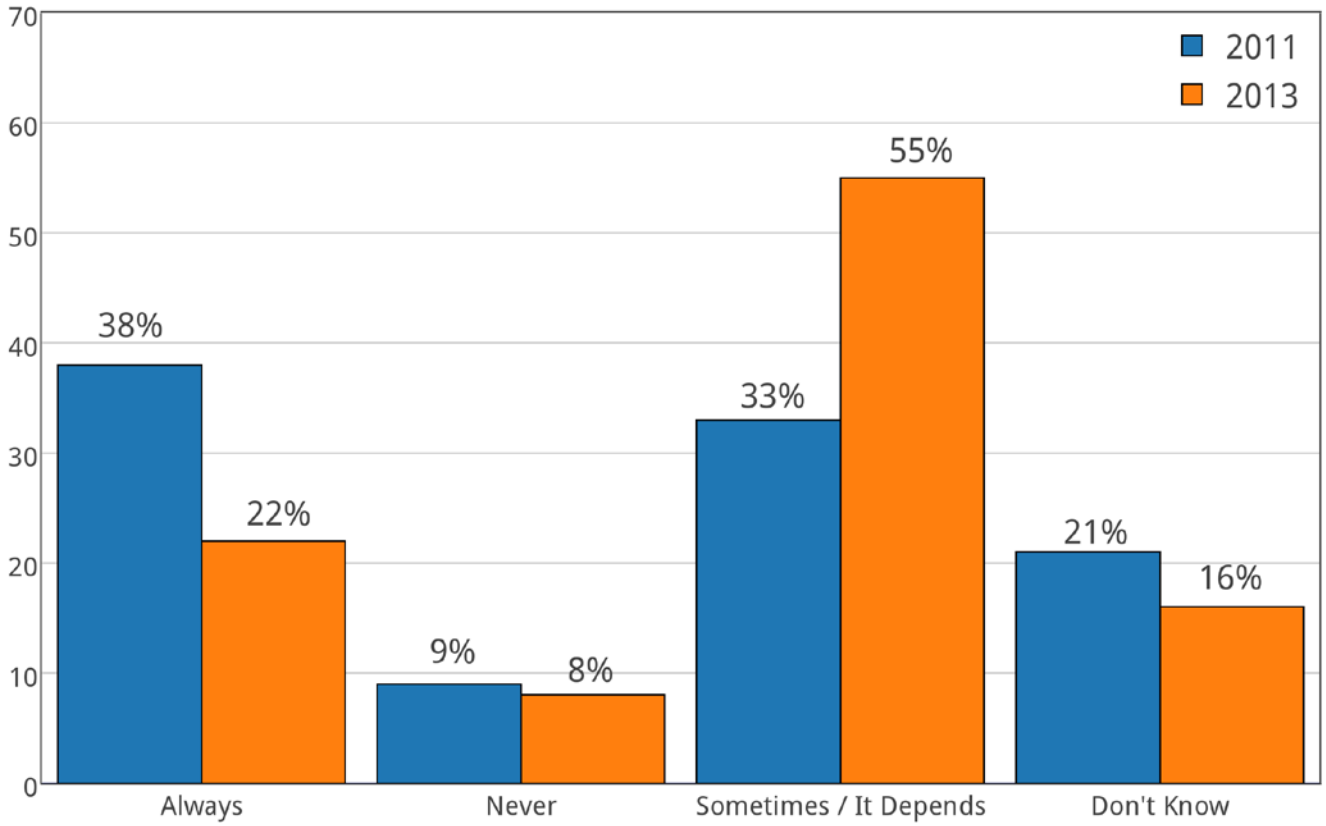



**FIGURE 11: POLICIES FOR RESPECTING ROBOTS.TXT**

The 2013 survey asked for the conditions under which respondents may ignore robots.txt. Most, 64% (27 of 42), indicated that they either owned the copyright or had some other special access right (e.g., as would be the case for an organization's own archive or an organization responsible for government records). A slight majority, 52% (22 of 42), indicated that they ignore robots.txt pursuant to either securing permissions from or sending notifications to content owners. Slightly less than half of respondents, 45% (19 of 42), ignore robots.txt when it is deemed necessary to capture essential content.

## Access Embargoes

Archived content may be embargoed from public access for a defined length of time for a number of reasons, including to minimize competition or confusion with the "live" web content. This is evidently a feature of many policies, as 69% (53 of 77) of respondents indicated that they employ embargoes. For those with embargoes, the survey asked also about embargo lengths. Less than half of those with embargoes responded to this question (22 of the 53), but of those, six months was the most common response, with 45% (10 of 22), followed by a small number, 9% (2 of 22), with one-year embargoes. The remaining 45% (10 of 22) responded "other," with comments indicating largely that embargo policy is still being determined. One respondent noted that their embargo lengths are subject to the preferences of content owners. Not all respondents employ embargoes, however; 27% (21 of 77) make content available without policy-imposed delays. This may be a byproduct of the default absence of embargoes for popular third-party service providers, such as Archive-It and CDL's Web Archiving Service.

### Copyright and Access Policy Development Resources

The survey asked participants to indicate one or more resources as informing the development of their own copyright and access policies. Responses suggest that organizations rely on what their peers and the community is doing in determining their own policies, with many citing either the web archiving policies of other organizations as resources or the community-informed Association of Research Libraries Code of Best Practices in Fair Use for Academic and Research Libraries.[10] The Section 108 Study Group Report[11] was the next-most popular, cited by 25% (14 of 55) of respondents. The survey also inquired about social media archiving policies. Social media, such as Facebook, Twitter, and YouTube, represents a relatively new and growing area of collecting interest and concern for web archiving organizations. Web archiving copyright and access policies are mature, by comparison. The survey did not dig deep into social media archiving policies but merely asked whether organizations had them. The results indicate that few do; 76% (59 of 78) lack social media archiving policies.

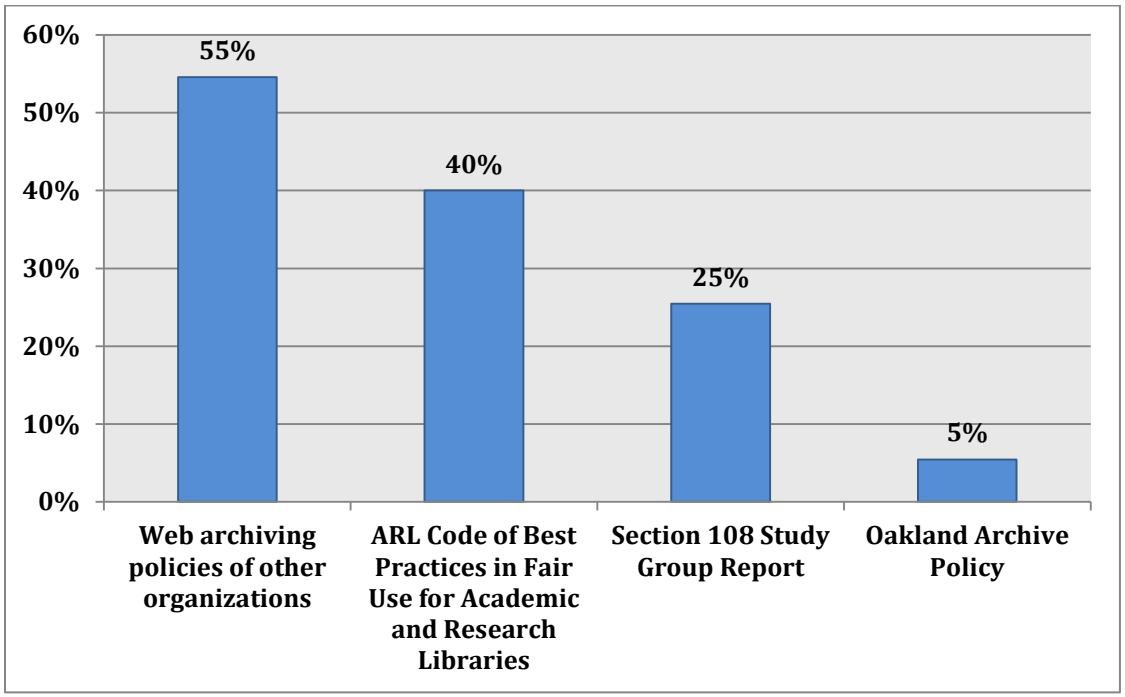



**FIGURE 12: RESOURCES USED IN COPYRIGHT AND ACCESS POLICY DEVELOPMENT**

## TOOLS AND SERVICES

Web archiving is dependent on specialized tools and services. The goal of this section of the survey was to learn more about what decisions web archiving programs are making in this area, understand preferences and rationales for using either external services or in-house tools (or both), and to identify convergent adoption of particular software, data formats, and/or vendors. The 2011 and 2013 surveys highlight trends regarding the use

[10] The Association of Research Libraries (ARL) "Code of Best Practices in Fair Use for Academic and Research Libraries," January 2012 is available at http://www.arl.org/focus-areas/copyright-ip/fair-use/code-of-best-practices

[11] The "Section 108 Study Group Report", March 2008 is available at http://www.section108.gov/

of external web archiving services versus in-house tools, the range of services and tools being used, and the number of institutions transferring their web archive data from external services for local storage.

Some results remained roughly the same: the proportion of organizations using external services versus in-house tools, the profile of external services used, and the proportion of organizations that had transferred web archive data from external services all were very similar in the 2013 survey results to the numbers from the 2011 survey. Notable results that indicated changes between 2011 and 2013: organizations are using more tools that support the WARC and ARC formats; fewer organizations using external services are planning to transfer their web archive data to local storage; and, concomitantly, external services are increasingly viewed as adequate for preservation purposes.

## External or In-House?

Between 2011 and 2013, there has been a slight shift toward external services, with 5% fewer institutions using in-house capabilities exclusively, approximately 3% more using external services exclusively, and about the same proportion using both.

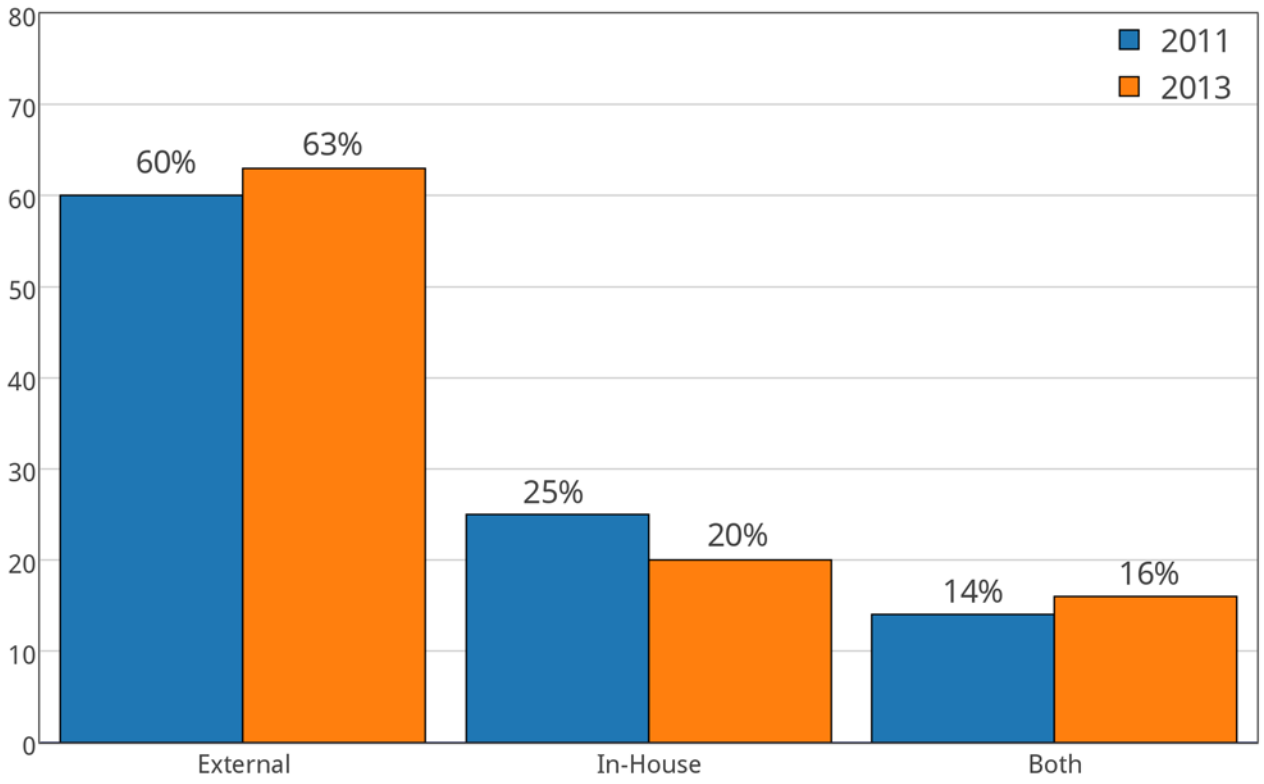



**FIGURE 13: USE OF EXTERNAL SERVICE VERSUS ARCHIVING IN-HOUSE**

Web archiving organizations continue to use a wide variety of software. The chart below reflects the tools and software that two or more organizations reported using in either of the two surveys. Between 2011 and 2013, there was an approximate 5% increase in use of Heritrix and an approximate 7% increase in the use of "other"

software used by no more than a single organization. Examples of "Other" in-house tools reported include: custom-built tools based on content management systems, a modified Heritrix crawler, manual download of individual web files, screenshots, KEN,[12] Social Feed Manager,[13] UXTR,[14] and WAIL.[15]

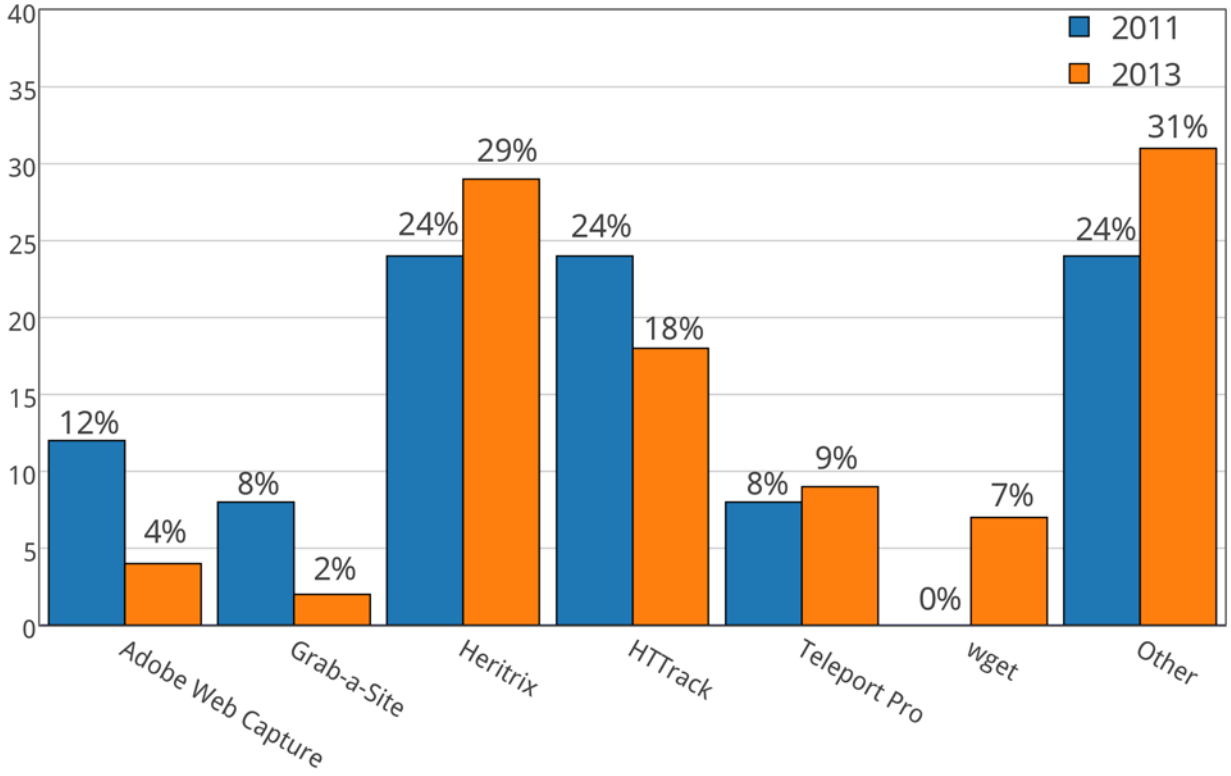



**FIGURE 14: TOOLS USED IF CAPTURING IN-HOUSE**

Most notably, however, there was a 14% increase in the use of archiving software that supports the WARC or ARC formats, suggesting that institutions are paying increasing attention to standards for web archive data formats.

## External Services

Archive-It remains the dominant external service among survey respondents, with approximately 70% (53 of 75) using the service in both 2011 and 2013. California Digital Library's Web Archiving Service is second, with a slight increase in use among respondents from 16% (8 of 50) in 2011 to 17% (13 of 75) in 2013. The main development

[12] More information about KEN Web Archiving Platform is available at https://ken-webarchiving.com/
[13] More information about Social Feed Manager is available at http://gwu-libraries.github.io/social-feed-manager/
[14] More information about UXTR is available at http://webarchivingbucket.com/uxtr/doc/
[15] More information about Web Archiving Integration Layer (WAIL) is available at http://matkelly.com/wail/

since 2011 is the entrance of new commercial providers into the roster of external services reported: Aleph,[16] Hanzo,[17] and Reed,[18] though all together these services accounted for around 5% of the total responses.

## Data Transfer

The percentage of organizations that have or have not transferred web archive data is about the same as in 2011. Notably, 27% fewer organizations than in 2011 report that data transfer is waiting on the establishment of in-house infrastructure. That 27% is split between the two remaining responses: 14% more organizations report not having a place to store transferred data and 13% more organizations report not knowing what they would do with web archive data, were they to transfer it.

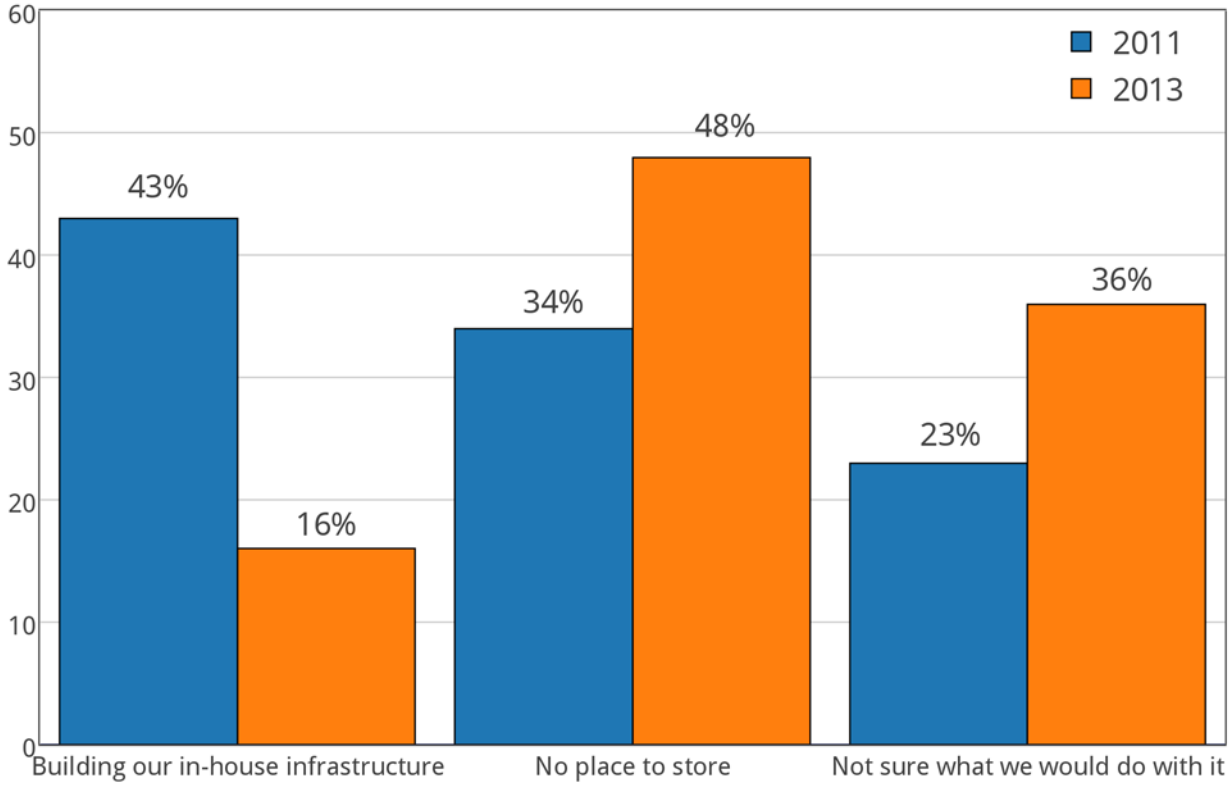



**FIGURE 15: REASONS FOR NOT TRANSFERRING DATA FROM AN EXTERNAL SERVICE**

A slightly higher percentage of organizations report that they are planning to transfer their data - approximately 23% (5 of 22) in 2013 compared to approximately 14% (1 of 7) in 2011. Fewer organizations report that storage costs are the prohibitive reason. The shift, however, appears to be that organizations that have moved out of the pilot phase of their programs - where the question of whether they would transfer their data or not was more open - increasingly trust external services to preserve their data. Extrapolating from the findings in this section is difficult, given the paucity of responses compared to other questions in the survey.

[16] More information about Aleph Archives is available at http://aleph-archives.com/
[17] More information about Hanzo is available at http://www.hanzoarchives.com/
[18] More information about Reed Archives is available at http://www.reedarchives.com/

# ACCESS AND DISCOVERY

The goal of this section of the survey was to learn more about how organizations are facilitating access to and discovery of their web archives. Questions focused on the viewing platform through which web archives are made accessible, the mechanisms provided for discovery, and the degree of integration of web archives into discovery environments for other resources.

## Web Archive Viewer

The Wayback Machine is the most popular access platform, used by 89% (67 of 75) of respondents. This is inclusive of those using external services providers such as Archive-It and California Digital Library's Web Archiving Service, which themselves provide access through Wayback instances. This is an increase from the 76% (41 of 56) using Wayback to provide access in the 2011 survey. Of the 11% of remaining respondents to the 2013 survey, 8% (6 of 75) indicated that they were using "other" viewers, including the ArchiveSocial access portal, the Reed Archives Console, and an internally-developed viewer. The remaining 3% (2 of 75) are not yet providing access to their web archive data.

## Discovery Mechanisms

Organizations are engaged in a wide range of efforts to support the discovery of archived web content. The most common discovery mechanisms are full-text search, URL search, and title browse lists, which are all provided by half or more of respondents. Fewer than a quarter of respondents provide catalog records at either the collection or item level. For those discovery mechanisms presented as possible responses in both the 2011 and 2013 surveys, there was a drop in usage ranging from 4% to 18%. The largest drops were in item-level catalog records, from 36% in the 2011 survey to 18% in the 2013 survey, and collection-level catalog records, from 30% in the 2011 survey to 22% in the 2013 survey.

It is unclear why the numbers for all provided web archive discovery mechanisms are down, though may be explained by the addition of finding aids as a possible response. Respondents representing archives may have previously entered collection or item-level catalog records as proximate responses. In the 2013 survey, 20% (15 of 76) respondents indicated that they are providing discovery of web archives through finding aids. Another new possible response for the 2013 survey was application programming interfaces (APIs), which are provided by 5% of respondents. The 20% (15 of 76) respondents who indicated "other" mostly described their future plans or the status of their current activities.

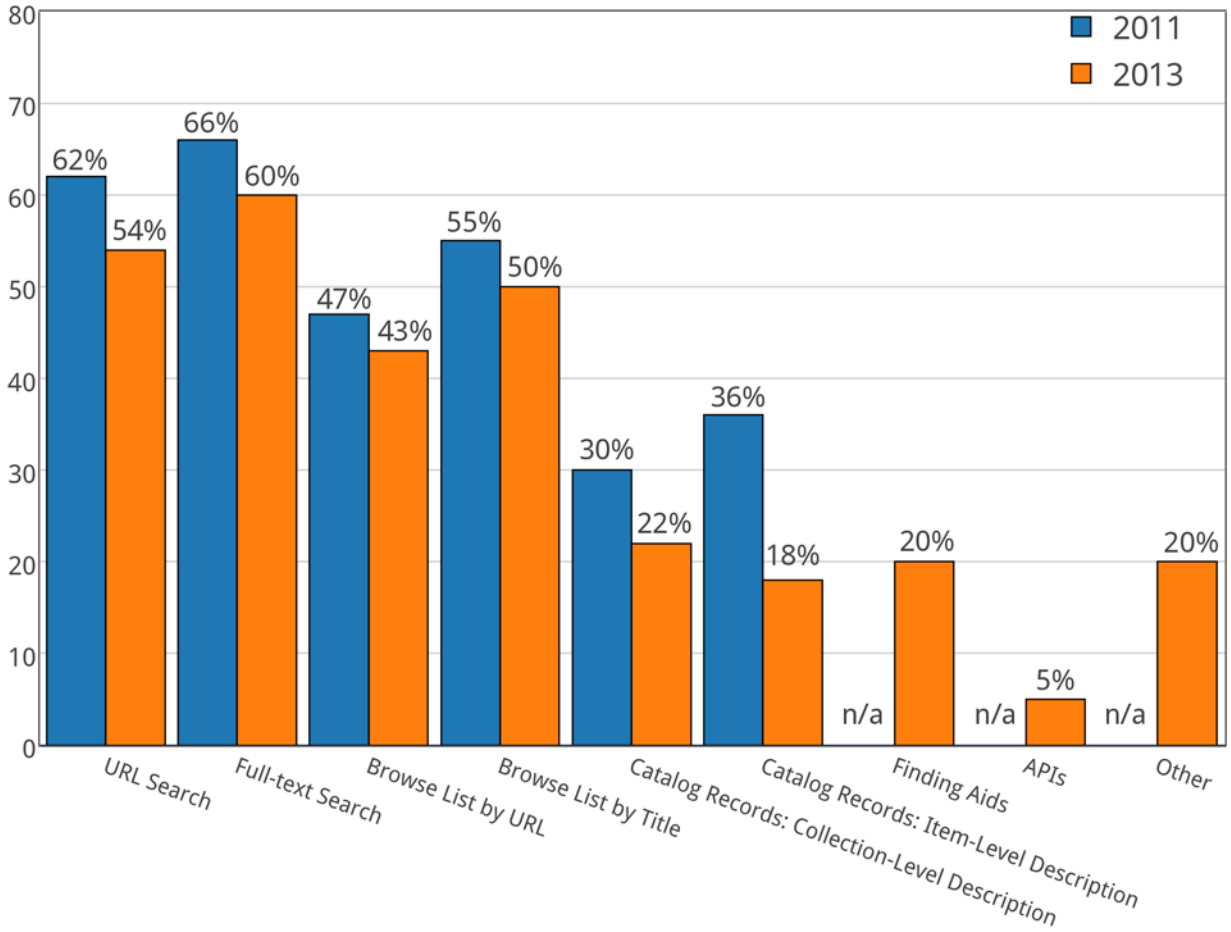



**FIGURE 16: KINDS OF ACCESS PROVIDED**

### Discovery Interface

The majority of respondents, 71% (41 of 58), make their web archives discoverable solely through a dedicated interface (i.e., that is not used for the discovery of other types of resources). A roughly equal proportion of respondents, 14% (8 of 58) versus 15% (9 of 58), respectively make their web archives discoverable through a common discovery environment or both a dedicated interface and a common discovery environment. Thirty four survey participants skipped the question, suggesting perhaps that the question may have needed further explanation.

## SUMMARY

### Maturity and Convergence

Overall, the survey results suggest that web archiving programs nationally are both maturing and converging on common sets of practices. Advancement is demonstrated by 75% of respondents reporting some or significant progress from two years ago, 38% of respondents having started programs in the last two years, and 8% more respondents indicating that their programs are in active status. Adoption of common practices is supported by a 14% increase in the use of tools that support the WARC data format, a 13% increase in the use of Wayback as a

web archive viewer, 81% of organizations devoting one half FTE or less to web archiving activity, and 67% of respondents relying upon other organizations' or community-generated policies in the creation of their own.

## Challenges and Opportunities

The survey results highlight challenges and opportunities that are or could be important areas of focus for the web archiving community. A third of respondents are interested in taking part in a collaborative web archiving project but have not yet done so. It is unclear what, exactly, is needed, but possibilities include better outreach for the periodic collaborations that do take place and a greater variety of projects. The data clearly indicate untapped interest in working across organizational boundaries.

Respondents are highly focused on the data volume associated with their web archiving activity and its implications on cost and the usage of their web archives. Cost modeling, more efficient data capture, storage de-duplication, and anything that promotes web archive usage and/or measurement would be worthwhile investments by the community. Unsurprisingly, respondents continue to be most concerned about their ability to archive social media (79%), databases (74%), and video (73%). The research, development, and technical experimentation necessary to advance the archiving tools on these fronts will not come from the majority of web archiving organizations with their fractional staff time commitments; this seems like a key area of investment for external service providers. Policy development for archiving social media is also needed, as 76% of respondents currently lack such policies.

## Questions to Revisit

There were a number of areas where the data was ambiguous or hinted toward possible trends that would be worthwhile to revisit in the next survey. For example, the drop in organizations building in-house infrastructure to store web archive data, the combination of the widespread perception of the difficulty in archiving particular content types, and the relatively low staff time devoted to web archiving all have implications for the distribution and roles of external service providers and in-house operations. Perceptions of progress, while overall positive and admittedly difficult to quantify, could benefit from great specificity as well as a more granular breakdown by areas of program activity.

The 2013 survey saw a drop in the proportion of organizations providing either item or collection-level catalog records for discovery of web archives but a comparable increase in the proportion of organizations providing finding aids, which were not presented as an option on the 2011 survey. The next survey could examine whether description of web archives is, in fact, decreasing, is instead appearing in new or different systems, or simply suffers from a lack of disambiguation among terminology. Each of these options may have different implications for how web archives are made discoverable.

On the policy front, the 2013 survey asked only whether respondents had any guidelines relating to social media. Considering what a major area of concern social media is, the importance of peer organizations' policies in policy development, and the relative dearth of policies specifically addressing it, it makes sense for the next survey to probe this area more deeply and surface the relevant dimensions of such policies.

# APPENDIX A

In the Archiving Program Information section of this report, two questions on Skills and Metrics allowed for free-form responses. These responses were categorized as follows:

## Skills Categorization:

Responses were coded as relating to **web technologies** if they mentioned knowledge of web architecture, design, formats (e.g., HTML, CSS, JavaScript, common), or website structure.

Responses were coded as relating to **archiving tools** if they mentioned being familiar with, configuring, and/or operating web archiving tools.

Responses were coded as relating to **domain expertise** if they mentioned knowledge of the organizations or subject areas that are the focus of web content collecting.

Responses were coded as relating to **appraisal** if they mentioned appraisal, collection development, or making judgments about what content should be collected.

Responses were coded as relating to **metadata** if they mentioned description or metadata.

Responses were coded as relating to **collaboration and communication** if they mentioned outreach to or coordination with internal or external stakeholders.

Responses were coded as relating to **software development** if they mentioned software or web development.

Responses were coded as relating to **quality assurance** if they mentioned analyzing or troubleshooting web archive quality issues.

## Metrics Categorization:

Responses were coded as relating to **volume** if they mentioned capacity planning, tracking data volume or the number of objects collected, or maintaining acquisitions statistics.

Responses were coded as relating to **usage** if they mentioned measuring access use, assessing researcher value, web analytics, or enabling access.

Responses were coded as relating to **cost** if they mentioned budgeting, object quantity or storage limits, or calculating devoted staff time or allocated resources.

Responses were coded as relating to **quality** if they mentioned accuracy, completeness, or quality of the archived web content.

Responses were coded as relating to **buy-in** if they mentioned securing management approval or buy-in from institutional stakeholders.

Responses were coded as relating to **loss** if they mentioned tracking the disposition of the live versions of resources that had since been archived.

Responses were coded as relating to **policy** if they mentioned tallying permission responses or monitoring conformance with their web archiving policy.

# APPENDIX B

## 2013 Web Archiving Survey Questions

PDF of survey questions available at:
http://www.digitalpreservation.gov/ndsa/documents/ndsa_web_archiving_survey_2013.pdf