Web Archiving in the United States: A 2016 Survey

An NDSA Report

Results of a Survey of Organizations Preserving Web Content
February 2017

AUTHORS
Jefferson Bailey, Internet Archive
Abigail Grotke, Library of Congress | Content Interest Group
Edward McCain, Reynolds Journalism Institute / MU Libraries | Content Interest Group
Christie Moffatt, U.S. National Library of Medicine | Content Interest Group
Nicholas Taylor, Stanford University Libraries

TABLE OF 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, non-profits, 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.

DOI: 10.17605/OSF.IO/R5PQK

INTRODUCTION

From January 20 to February 16, 2016, a team of individuals representing multiple NDSA member institutions and interest groups conducted a survey of organizations in the United States actively involved in, or planning to start, programs to archive content from the Web. This effort built upon a similar survey undertaken by NDSA in late 2011 and published online in June 2012[1] and a second survey in late 2013 published online in September 2014.[2] The goal of these surveys is to better understand the landscape of Web archiving activities in the United States 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 US Web archiving initiatives, comparison with the results of the 2011 and 2013 surveys 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 points to future opportunities and developments.

METHODOLOGY

The survey team self-organized into a working group in late 2015 to begin drafting the survey questions. Two goals were identified early in the process: to enable historical comparisons with the 2011 and 2013 surveys and to inquire about additional program details that were not previously included. Accordingly, the group reviewed and refined the questions from the previous surveys and added new questions to address emerging activities and issues. The updated survey was conducted from January 20 to February 16, 2016, using the SurveyMonkey online survey tool, and was promoted via blogs, mailing lists, social media, and other channels. When the survey concluded, the group reviewed the responses and removed test or mostly incomplete entries. Respondents were not required to answer every question. The percentages reported for individual questions reflect the total number of responses to that question, rather than the total number of respondents participating in the survey.

The Survey Content

The 2016 NDSA Web Archiving Survey consisted of 31 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

[1] "Web Archiving Survey Report," *NDSA Report*, June 19, 2012, accessed December 16, 2016, http://www.digitalpreservation.gov/documents/ndsa_web_archiving_survey_report_2012.pdf.
[2] "Web Archiving in the United States: A 2013 Survey," *NDSA Report*, September 2014, accessed December 16, 2016, http://ndsa.org/documents/NDSA_USWebArchivingSurvey_2013.pdf.

their program; access and discovery systems and approaches; and program policies involving capture, availability, and types of Web content.

The working group shared the survey via Library of Congress's *The Signal*; NDSA-ALL LISTSERV; Society of American Archivists LISTSERV; Archive-It Partner News; GOVDOC-L LISTSERV; Federal Web Archiving Working Group; Legal Information Preservation Alliance membership; International Internet Preservation Consortium Members' LISTSERV; and Twitter. The survey was completed 104 times, an increase of 13% in the number of respondents from the 92 completed responses to the 2013 survey, and an increase of 35% from the 77 completed responses to 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

Survey instrument and anonymized survey data are available on Dataverse at https://dataverse.harvard.edu/dataverse.xhtml?alias=ndsa.

RESPONDENT CHARACTERISTICS

The survey (see appendix A) opened with questions about respondents' organizations, asking the organization name and type, whether they belonged to any of three formal Web archiving-related professional groups, and the status of their program, be it in the planning, pilot, production, or discontinued stage.

Organization Type

As in past years, the question about organization type elicited more responses than any other question. Each of the 104 respondents answered this question in 2016, as did each of the 92 in 2013 and each of the 77 in 2011. Besides a notable increase in one key community, the proportion of represented organization types was largely similar to those of previous survey years. The percentages of archives, commercial organizations, museums, and cultural heritage organizations were mostly within a few percentage points of the prior surveys. The one significant increase was in the number of academic institutions responding. This group represented 47% (36 of 77) of respondents in 2011, 52% (48 of 92) in 2013, and 63% (66 of 104) in 2016. The steady increase in Web archiving at college and university libraries and archives (as well as research institutions and disciplinary programs within universities) demonstrates the popularization of Web archiving as a core collection development and preservation activity within academic institutions.

Other noteworthy findings include that governments (federal, local, and state) doing Web archiving have held steady over the three surveys (17 respondents every year), though

their percentage of representation decreased slightly (decreasing 8% since 2011) due to the overall growth in college and university respondents.

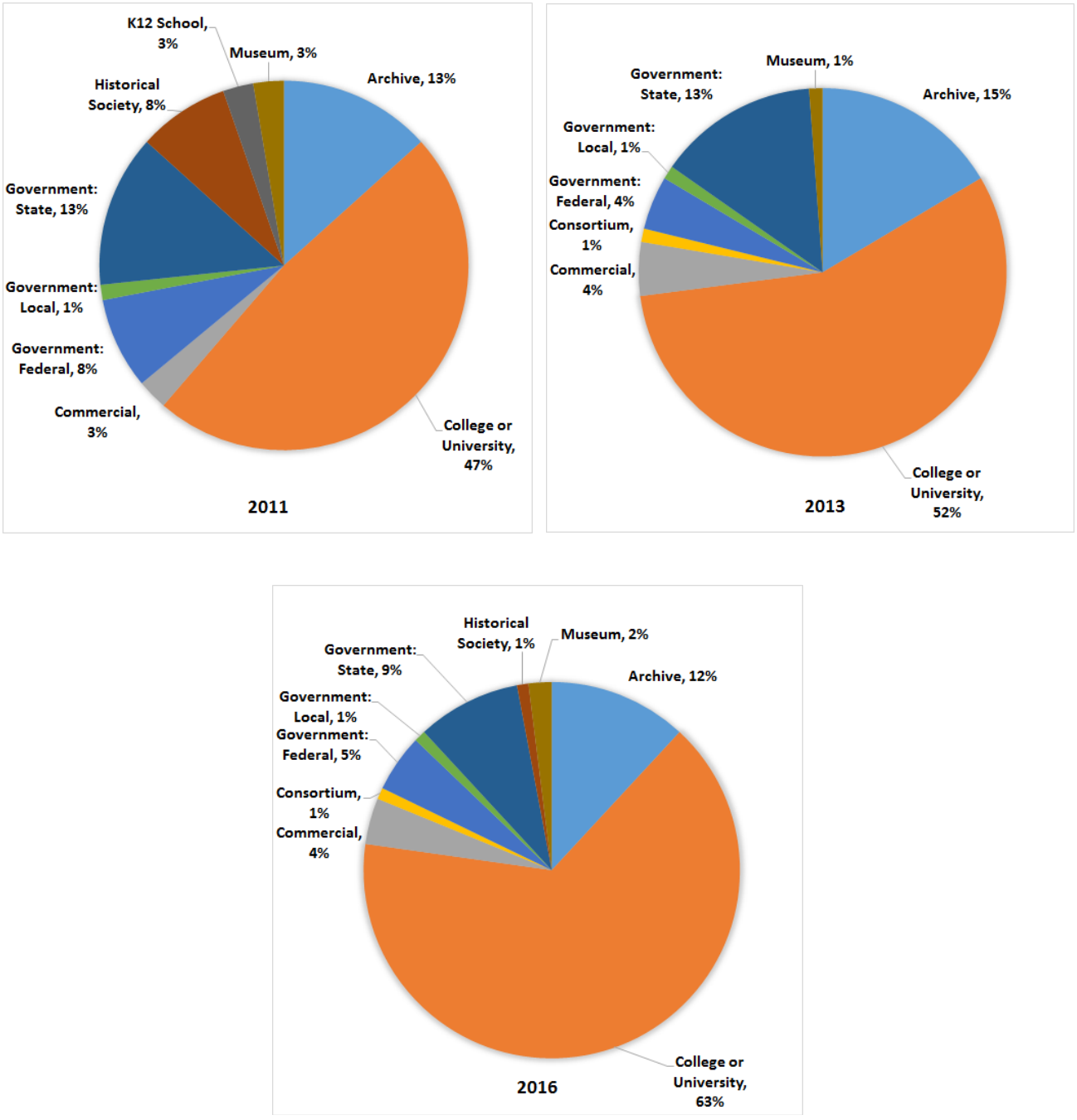

FIGURE 1: RESPONDING ORGANIZATION TYPE

Group Affiliations

Web archiving-related professional group affiliations remained consistent across the surveys. In 2016, 10% (8 of 77) were members of the International Internet Preservation Consortium (IIPC), the same as in 2013 and similar to the 8% (6 of 77) from 2011. Membership in the NDSA also remained consistent: 52% (30 of 58) of survey respondents belonged to the NDSA in 2013 and 53% (41 of 77) were members in 2016. The 2011 survey did not ask about respondents' NDSA membership. The 2013 survey additionally asked if organizations belonged to the Society of American Archivists (SAA) Web Archiving Roundtable, which was established after the 2011 survey. Again, membership percentages stayed consistent, with 71% (41 of 58) of respondents in 2013 and 69% (53 of 77) in 2016 affiliated with this group.

Interestingly, the number of responding organizations that reported multiple affiliations grew in 2016. The 2016 survey had 30% (23 of 77) of respondents choosing affiliation with two of the three possible groups, an increase from 26% (15 of 58) in 2013 and 25% (6 of 24) in 2011. As affiliations often follow occupational boundaries and the diversity of responding institutions features a range of professions, growth in this area is heartening, suggesting the recognition of Web archiving as a professional activity mature enough to support multiple affiliate groups.

Activity Status

The 2016 survey asked about the current status of respondents' Web archiving programs; answer options were planning, pilot, production, and discontinued. The 2016 survey found an ongoing increase in formalized programs with 79% (81 of 103) classifying their status as production, continuing an upward trend from 74% (66 of 89) in 2013 and 64% (49 of 77) in 2011. The current survey found a return to percentages in 2011 for those in the planning stage of program activity, with 15% (15 of 103) identifying as being in the planning stage, similar to the 17% (13 of 77) in 2011, but a slight increase over 9% (8 of 89) of programs in the planning stage in 2013.

Most notable among the program status responses was the sharp drop in respondents classifying their program as in the pilot phase, with only 5% (5 of 103) identifying as such in 2016, as opposed to 15% (13 of 89) in 2013 and 16% (12 of 77) in 2011. This could indicate that, with the continued growth in the community and better awareness of best practices, standards, and policies, organizations are taking less time to pilot a program and are instead moving directly from the planning stage to establishing production-level programs more quickly and easily than in the past. The return to 2011-era rates of respondents in the planning stage, along with the overall growth in respondents to the survey, may indicate the ongoing spread of Web archiving as a formal activity in organizations.[3]

[3] Some caution is appropriate in interpreting these changes, as the group of institutions participating in each year of the survey differ. Of the institutions that have participated in the surveys over time, a significant

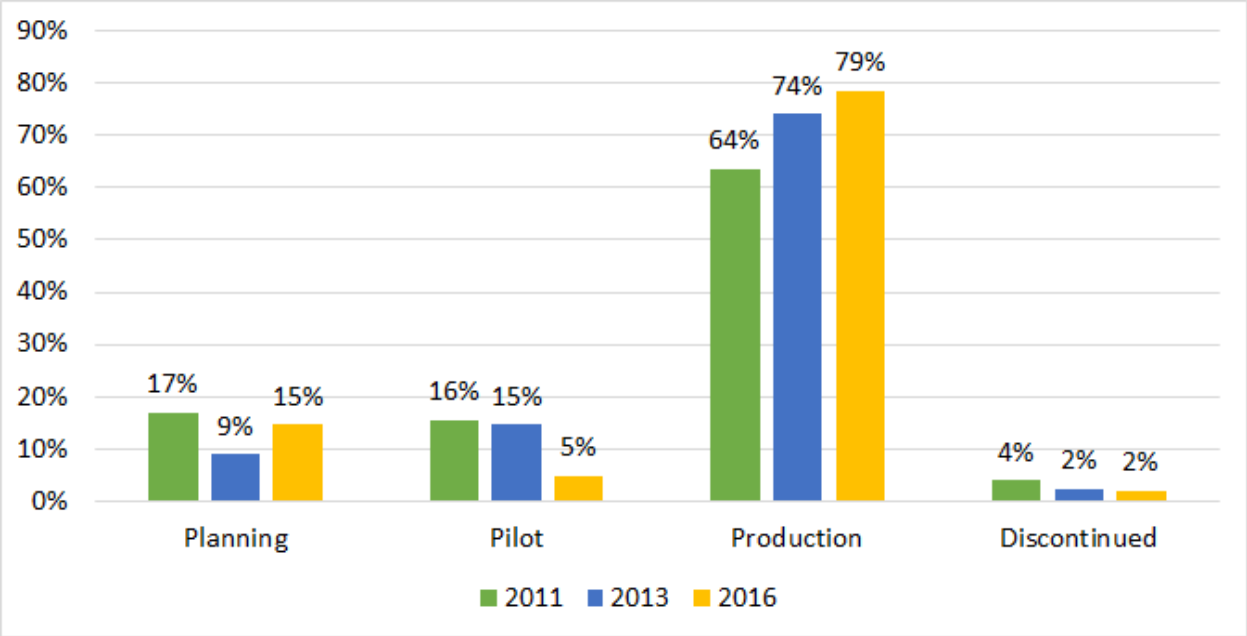

FIGURE 2: STATUS OF WEB ARCHIVING ACTIVITY

PROGRAM INFORMATION

This next section of the survey addressed the staffing and management of Web archiving programs across the United States as well as the goals and perceptions of progress within these programs.

Perceptions of Progress

Respondents to the 2016 survey were asked to compare the current state of their organization's Web archiving program to what it was two years ago. While perceptions of progress may be inherently subjective, the overall trend was remarkably positive: over three-quarters of the respondents (77%, 68 of 88) reported that their program had made either significant or some progress over the past two years. The number of those indicating that their programs made significant progress increased by 16% from the 2013 survey, when the question was first asked (2016: 53% [47 of 88]; 2013: 38% [35 of 93]). Of the organizations reporting significant progress, 47% (22 of 47) began archiving Web content in the past two years. Not one respondent to the 2016 survey indicated that their program was slightly worse off or much worse off. While two organizations in the 2016 survey reported their program status as "Discontinued," one left the question about progress in the last two years blank and the other indicated that progress was about the same.

majority, 71% (139 of 195), have participated one time only. Also worth noting is the possibility that different individuals with varying perspectives completed the survey on behalf of their organizations.

Looking closer at the kinds of progress underway, the 2016 survey included a new question asking respondents to identify the top three areas where their organizations had made the *most* progress. Multiple choice answers were based on Archive-It's Web Archiving Life Cycle Model.[4] Organizations seem to have made the most progress in areas of determining what they should capture and how to do it. The top three areas of progress identified were in data capture (59%, 50 of 85), appraisal and selection (46%, 39 of 85), and vision and objectives (40%, 34 of 85).

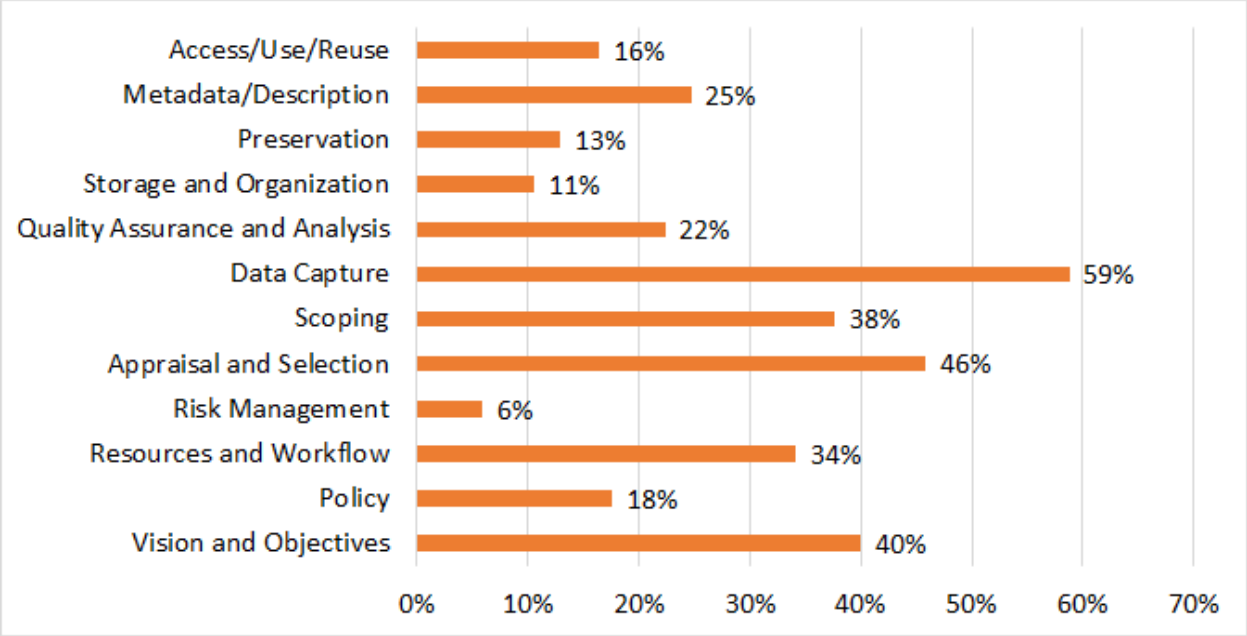

FIGURE 3: PERCEPTIONS OF MOST PROGRESS IN LAST TWO YEARS

When asked in a separate question, also for the first time, to share the top three areas of the Web Archiving Life Cycle Model in which the respondents' organizations had made the *least* progress, the top number of responses point towards aspects of Web archiving that take place later in the life cycle: access/use/reuse (52%, 43 of 82) was the top response, followed by metadata/description (38%, 31 of 82), and quality assurance and analysis (37%, 30 of 82).

[4] The Archive-It Team at the Internet Archive, "The Web Archiving Life Cycle Model," March 2013, accessed October 11, 2016, https://archive-it.org/static/files/archiveit_life_cycle_model.pdf.

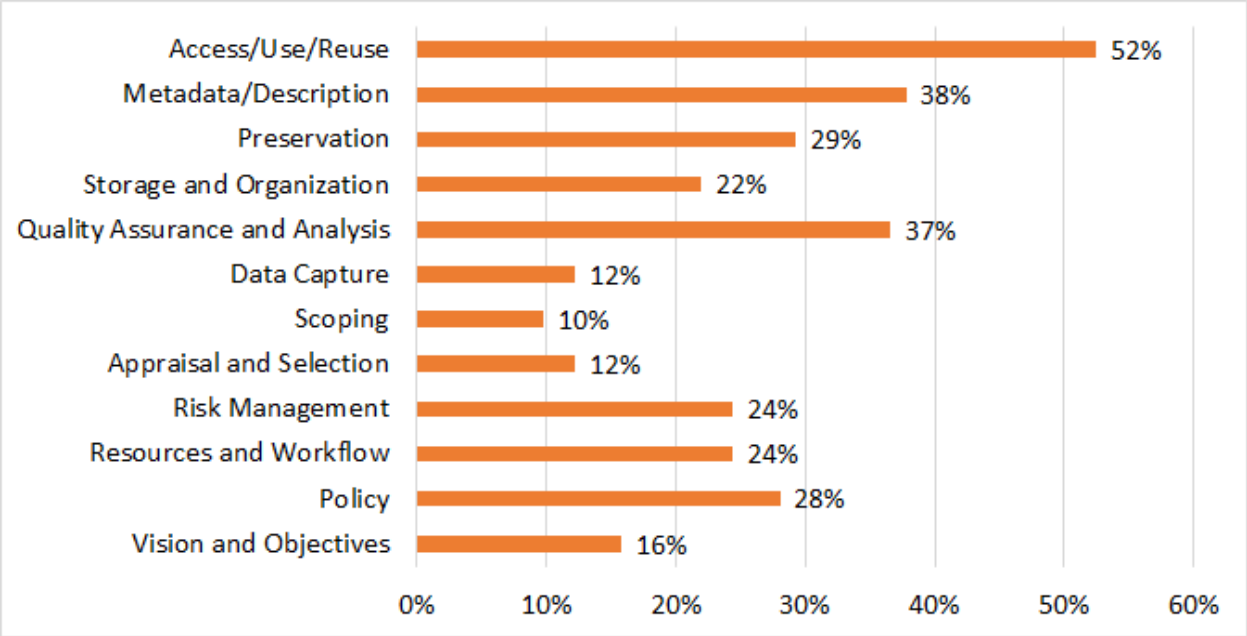

FIGURE 4: PERCEPTIONS OF LEAST PROGRESS IN LAST TWO YEARS

Content Being Archived

The 2011, 2013, and 2016 surveys asked organizations about the goals of their Web archiving activities. Responses indicated a continuing trend towards organizations archiving their own or affiliated Web content as a type of institutional record, with 89% (79 of 89) doing so in 2016, 20% higher than in 2011. The number of those archiving content from other organizations or individuals decreased 24% during the same period, with only 56% (50 of 89) indicating this as a goal in 2016, as compared with 80% (57 of 71) in 2011. The number of organizations that checked both boxes, indicating that they are both archiving one's own content and the content of others, remained fairly consistent, though declining slightly over time: 49% (35 of 71) in 2011; 46% (41 of 89) in 2013; and 45% (40 of 89) in 2016.

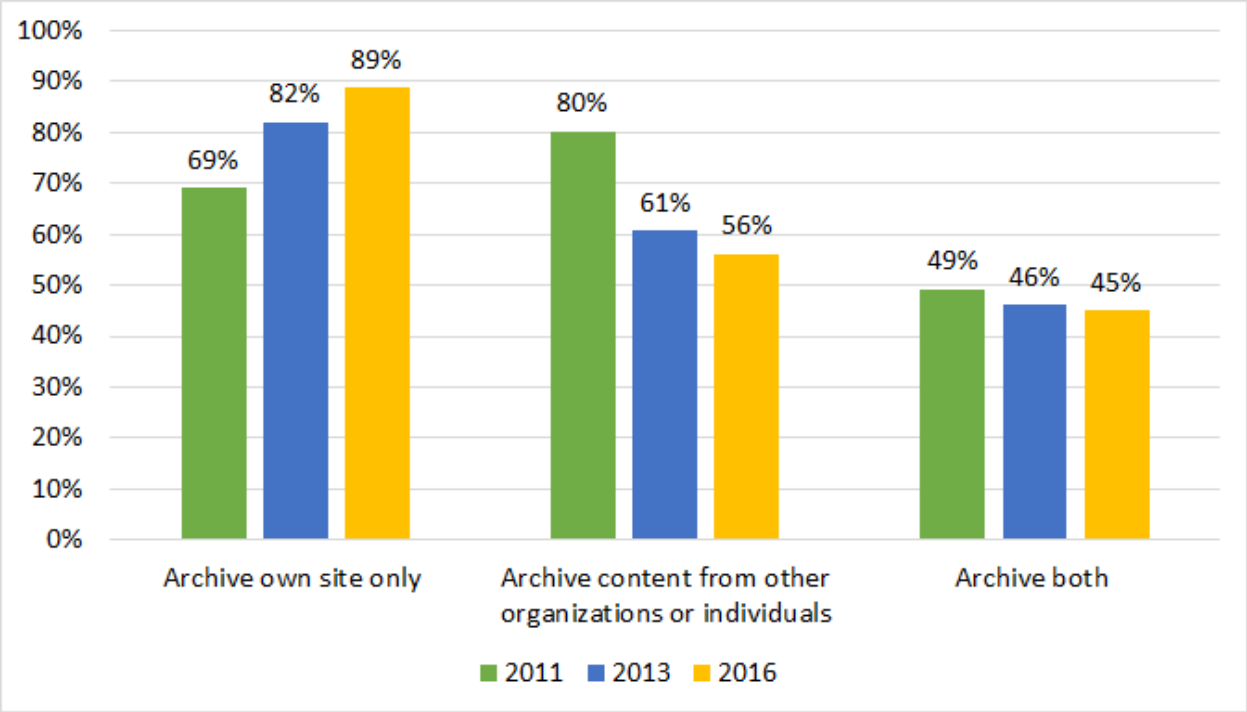

FIGURE 5: ARCHIVING ACTIVITY GOALS

When Programs Started

Each of the three surveys also included a question about the year participating organizations began archiving Web content. Responses ranged from two organizations that have been archiving since 1996 and two that indicated they were making plans to start soon. 65% (53 of 82) of respondents to the 2016 survey have been archiving since 2010. Organizations beginning to archive Web content in 2015 made up the largest individual group, with 18% (15 of 82) of respondents beginning that year.[5] In the previous two surveys, the largest groups of respondents were those beginning to archive Web content during the survey year, too, with 23% (19 of 81) in 2013 and 18% (14 of 77) in 2011.

[5] The 2016 survey took place in early January 2016, so "current year" extends to the two weeks in January.

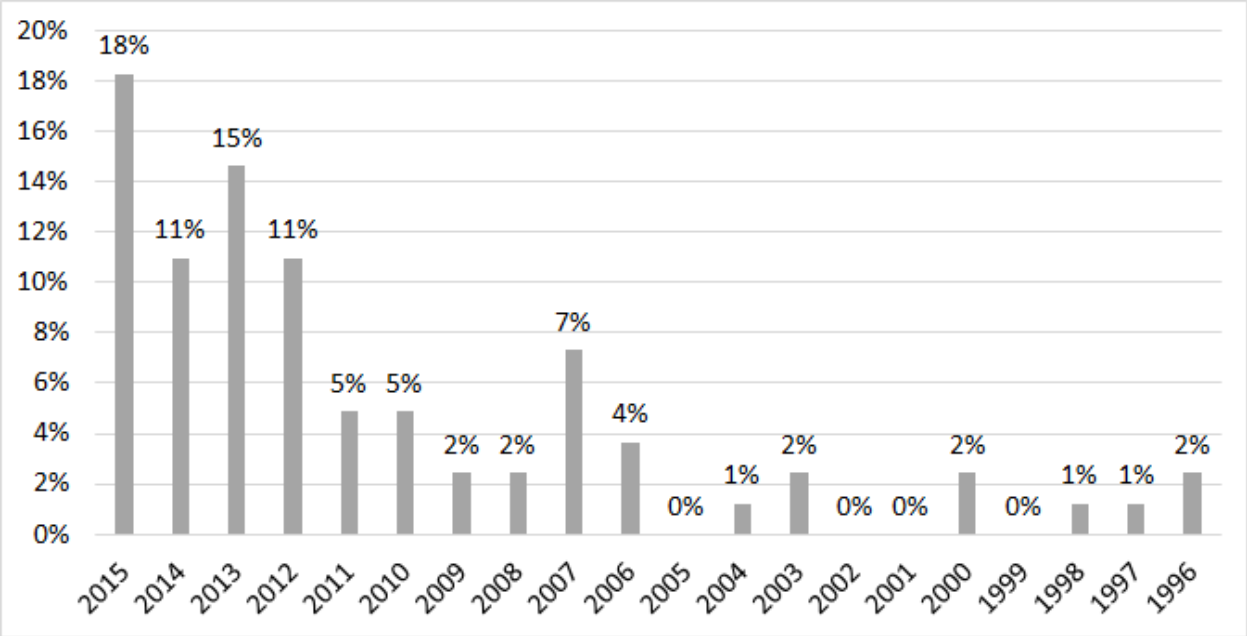

FIGURE 6: YEAR INSTITUTIONS BEGAN ARCHIVING WEB CONTENT

Devoted Staff Time

Responses indicated that organizations continued to devote only fractional time to Web archiving activities, with 76% (64 of 84) devoting less than the equivalent of one full-time employee's (FTE) time. There has been little change since the question was first included in the survey in 2013, with more than half of the organizations, 58% (49 of 84) in 2016 and 57% (49 of 86) in 2013, still only devoting one-quarter FTE to Web archiving activities. These figures indicate that Web archiving continues to be an activity included among the many other duties professionals are tasked to do within in their organizations. Whether these organizations believe current staffing levels to be sufficient to achieve the goals of individual Web archiving programs, or the broader goals of the digital preservation community,[6] might be explored further in future surveys.

There was one area of increase in staffing, with the number of organizations dedicating one to three FTE to Web archiving activities nearly doubling, growing from 7% (6 of 86) in 2013 to 13% (11 of 84) in 2016. Perhaps this is an indication that once Web archiving has achieved organizational support for one FTE, it is easier to advocate for additional resources.

[6] "2015 National Agenda for Digital Stewardship," *NDSA Report*, September 2014, accessed October 11, 2016, http://www.digitalpreservation.gov:8081/ndsa/documents/2015NationalAgendaExecSummary.pdf.

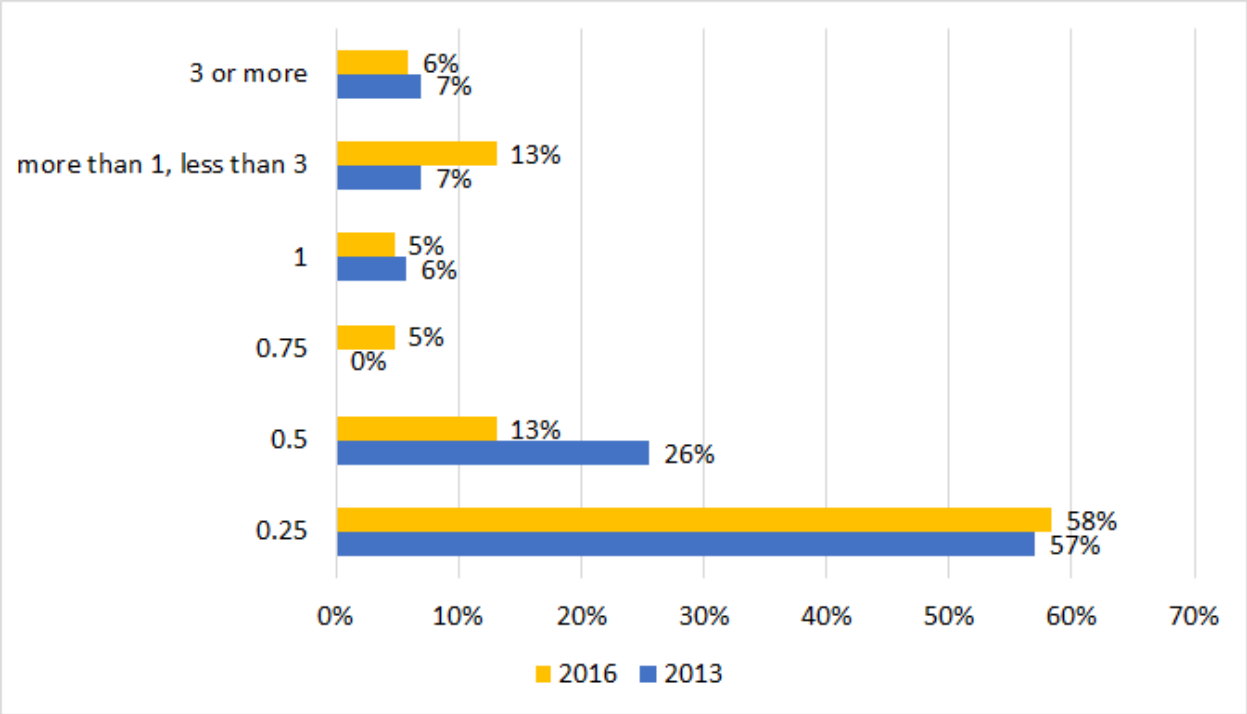

FIGURE 7: FTE STAFF DEDICATED TO WEB ARCHIVING

Considerations in Development of Programs

The options for responses to the question "What are the top 3 considerations for the development of your Web archiving program?" were drawn from open ended responses to a similar question first asked in the 2013 survey. The question was altered slightly in 2016 to try to get a better sense of the most important considerations in program development.

Over 50% of participating organizations identified access and use, cost (which includes staffing level requirements), and quality as their top three considerations for the development of Web archiving programs. Notably, these considerations corresponded fairly closely to the areas where organizations indicated they have made the least progress: access/use/reuse, metadata/description, and quality assurance and analysis (see Perceptions of Progress). Data volume, which topped the metrics of importance to organizations in 2013, fell to sixth place out of seven available choices in the 2016 survey, while open ended responses coded for usage and cost remained in the top three. This trend indicates a growing concern for quality over concerns about volume, which could be a sign of maturity of the field. Other responses provided by respondents in an open ended field broadly addressed fulfilling institutional requirements and collecting goals, staff time, or simply that it was too early to say.

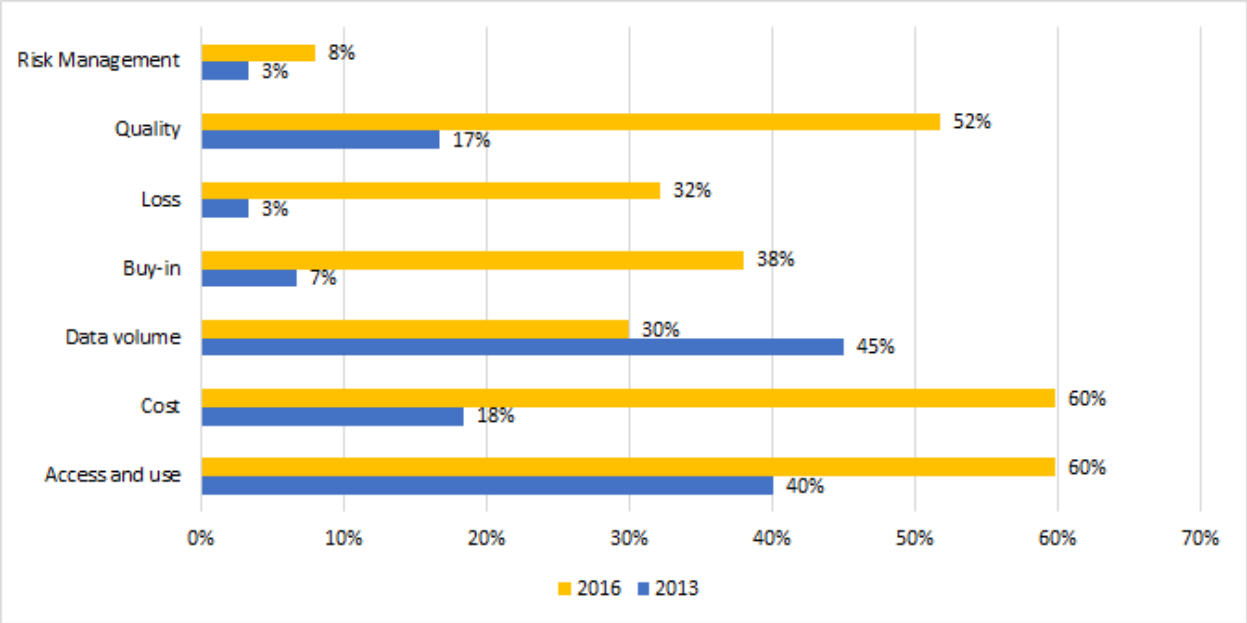

FIGURE 8: TOP CONSIDERATIONS WHEN DEVELOPING WEB ARCHIVING PROGRAM

Skills

The options for responses to the 2016 question on staff skills essential to the development and success of Web archiving within their organization were similarly developed from open ended responses to the same question in 2013. More than 50% of respondents indicated that the top three skills needed are facility with archiving tools (e.g., configuring or operating Web archiving crawler, access, and curatorial tools), indicated by 69% (61 of 88) of respondents, followed by skills for appraisal and selection (determining what content to select), indicated by 61% (54 of 88) of respondents, and for performing quality assurance (analyzing and troubleshooting Web archive quality issues), indicated by 51% (45 of 88) of respondents. Quality, once again, appears to be of growing importance and concern for organizations. This is a significant change from the 2013 survey, when quality assurance was cited least frequently out of all the coded categories, with only 6% (4 of 63) of the open ended responses indicating this as an essential skill.

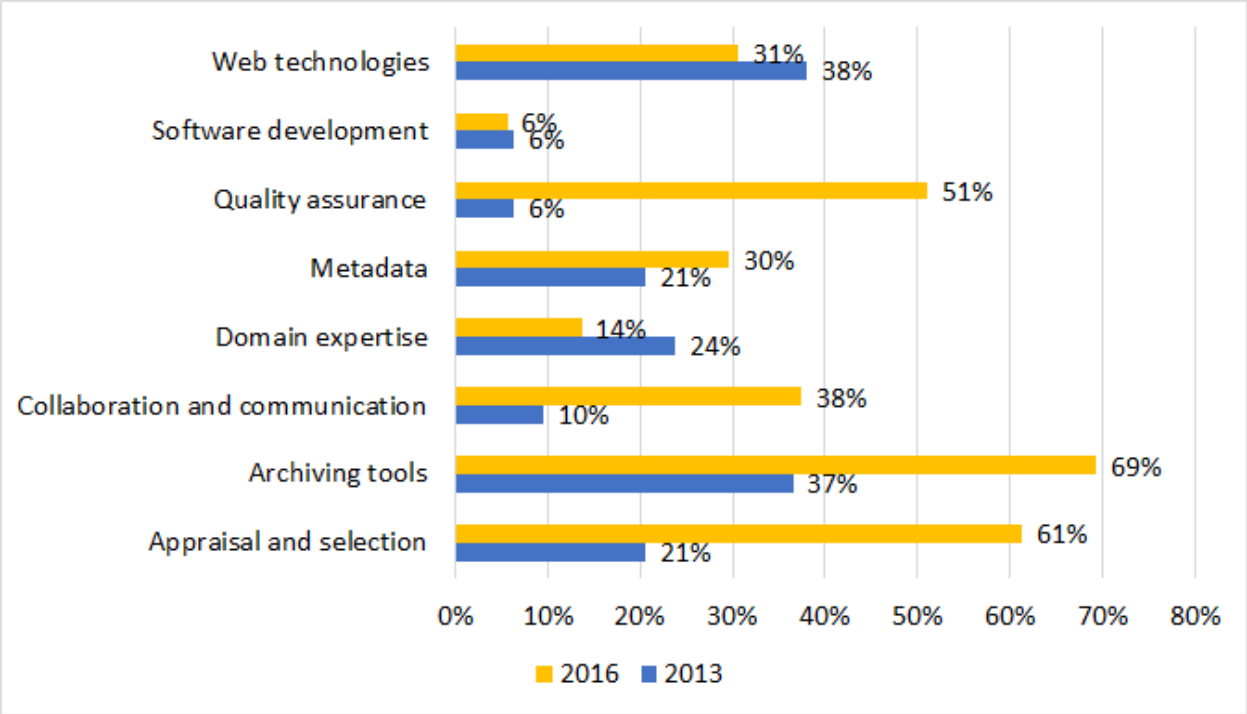

FIGURE 9: SKILLS DEEMED ESSENTIAL FOR STAFF

Content Types of Concern

As in 2013, the 2016 survey asked what types of content organizations have concerns about their capacity to archive. "Art" was removed from possible responses as it is less a *type* of content itself, and more of a subject area. The other options for answers (audio, blogs, databases, interactive media, social media, and video) remained the same, with the option to write in other content of concern. Respondents were asked to select all options that applied. The ability to capture social media remained a top concern in 2016, with 70% (60 of 86) indicating that they are concerned about their capacity to archive this content. Video was a close second, indicated by 69% (59 of 86) of respondents, followed by interactive media, and databases, both indicated by 62% (53 of 86). Across all but interactive media (which perhaps picked up numbers owing to the exclusion of "art" as a type in this survey), concerns about capacity to archive dropped from 2013. This could be an indicator that respondents gained more confidence about their own ability to archive using the tools they have (see Perceptions of Progress section), or possibly a result of increased focus on archiving institutional content, which might be comparatively simpler than third party "in the wild" webpages. On the other hand, decreased concerns about capacity to archive could simply reflect an increased recognition of limitations such as staffing, tools, and time, to archive content.

The responses to other content of concern written in the open ended field included password/log-in/intranet/deep Web content, geographic data, and very large Web sites that were difficult to archive comprehensively.

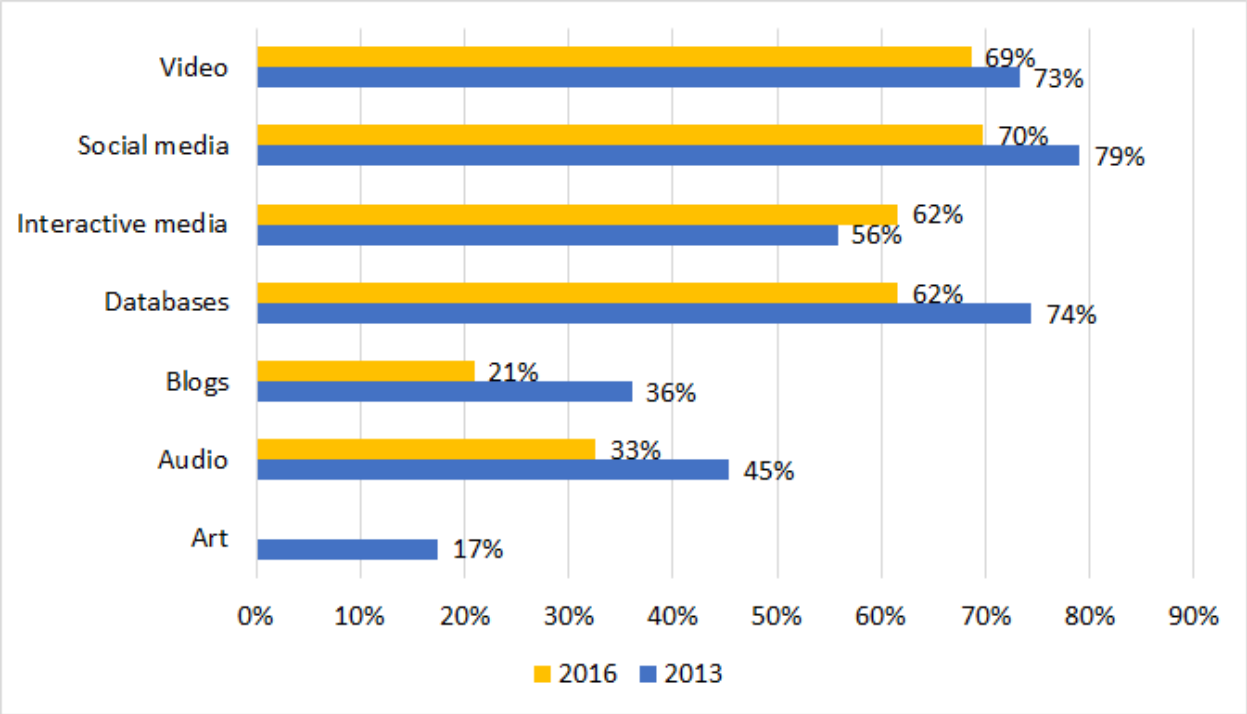

FIGURE 10: TYPE OF CONTENT PROVOKING CONCERN OVER CAPACITY TO ARCHIVE

Collaborative Archiving

The 2011 and 2013 NDSA Web archiving surveys included a question about whether organizations had participated in building collaborative Web archives around a specific event, theme, or domain, such as the United States End of Term government Web Archive or the International Internet Preservation Consortium's Olympic Games Web archives. In 2013, nearly half of responding organizations (51%, 45 of 89) indicated that they had either participated in a collaborative Web archive or were interested in doing so. In 2011, there were two separate questions on the topic: 23% (15 of 66) of respondents indicated that they had participated in a collaborative archive and 51% (34 of 67) indicated that their organization would be interested in future collaborative Web archives if the topic fit within their collecting scope and interests. The 2016 survey expanded the question to focus on identifying the areas in which organizations are most interested in collaborating, in addition to building collaborative collections, and any barriers that they face in doing so. Respondents were asked to check all responses that applied, with an opportunity to provide additional information in response to either or both questions.

Over 50% of responding organizations indicated interest in collaborating on quality assurance techniques and strategies (65%, 57 of 88), capture configuration and optimization (52%, 46 of 88), and metadata standards and application (51%, 45 of 88). Results indicated lower levels of interest in input on APIs and standards as well as collaborative collection development.

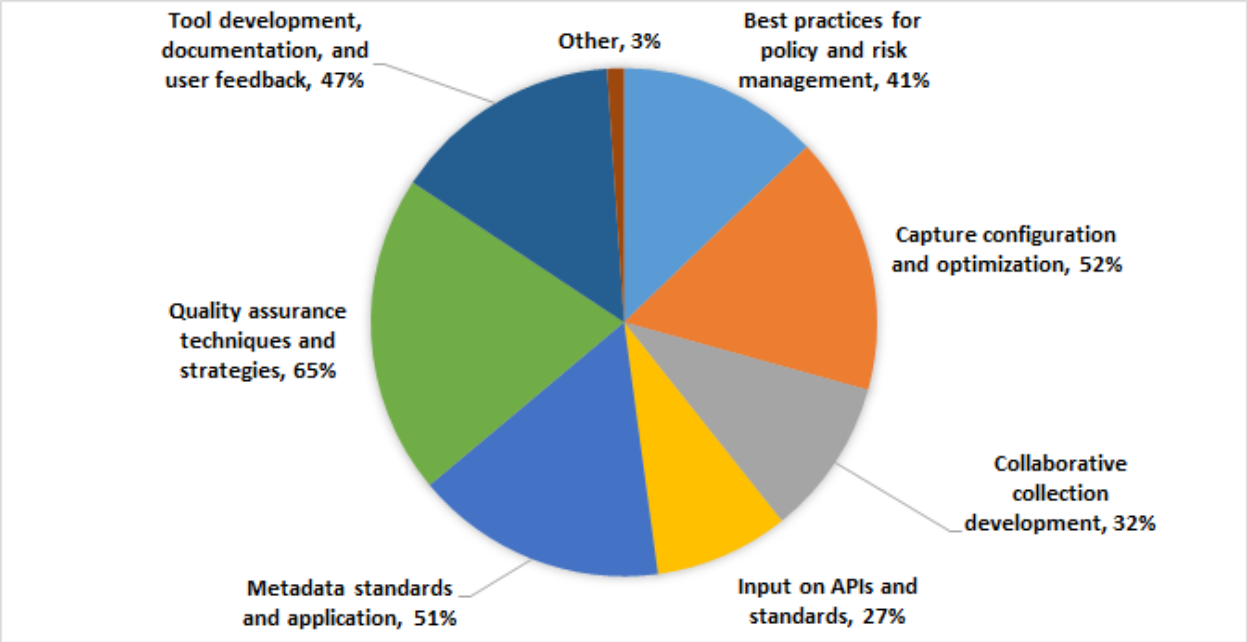

FIGURE 11: AREAS OF INTEREST FOR COLLABORATION

The primary barrier to collaboration is lack of time, identified by 75% (62 of 83) of respondents to the survey. 30% (25 of 83) of respondents reported they were still in a planning phase and did not have much to share is a distant second barrier. Taken together, the following perspective provided in an optional "other" response and anonymized for this report, may be a common experience:

> As a medium/small university, sometimes it seems like we don't fit into a clear niche with collaboration. I don't have the resources (or expertise) to do high level, cutting edge development (like an R1 university might). I also have to take lack of staff time into consideration when coming up with workflows, etc - I'd love to do more QA and more description, but it's just not a top priority in my department's work right now. So while it's helpful to talk with others about workflows/procedures that they've developed, they're often not applicable for me.

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 notification and permission requirements, approaches to robots.txt directives, policy guidelines specific to social media, and copyright and access policy development resources.

Notification and Permission

There was no change in survey questions about notifications and permissions between 2013 and 2016. Respondents indicated significant shifts away from either notifying or seeking permission from content owners when collecting Web content, which may reflect the notable increase in organizations focused solely on crawling their own or affiliated Web content as a type of institutional record. Roughly two-thirds of those surveyed (67%, 46 of 69) indicated that they took an approach of no action to notify or seek permission when capturing Web content, a 9% increase from the 58% (42 of 73) who reported this approach in 2013. The number of respondents notifying website owners before capture decreased slightly, from 23% (17 of 73) in 2013 to 22% (15 of 69) in 2016. The number of organizations seeking permissions when collecting content also decreased, with only 12% (8 of 69) reporting that they do so in the 2016 survey. Of those seeking permission when capturing Web content, one organization was from the federal government, one from local government, one was a museum, and six were programs in colleges or universities. In 2013, 19% (14 of 73) of organizations reported seeking permission, and 13% (8 of 61) did so in 2011. These decreasing numbers may not simply be linked to an increasing number of organizations focused on archiving their own content. Of the organizations indicating that they archive content from other organizations or individuals for future research (see Content Being Archived section), only 14% (7 of 50) seek permission.

The same proportion – 68% (47 of 69) of 2016 respondents – indicate that they take no action to notify or seek permission from content owners when providing either restricted or public access to Web content. This is roughly equivalent to the results in the 2013 survey, with 68% (42 of 62) of respondents indicating that they took no action when providing restricted access to Web content and 63% (45 of 71) indicating that they took no action when providing public access.

Statistics for practices surrounding restricted access indicated small changes in the proportion of respondents notifying and seeking permission from content owners. The percentage of respondents notifying content owners of their intent to provide restricted access to the content they have archived decreased from 11% (7 of 62) in 2013 to 6% (4 of 69) in 2016. The percentage of respondents seeking permission from content owners to be able to provide restricted access to content they have archived also decreased from 21% (13 of 62) in 2013 to 14% (10 of 69) in 2016. According to the 2016 survey, when providing public access, 13% (9 of 69) notify content owners and 16% (11 of 69) seek permission.

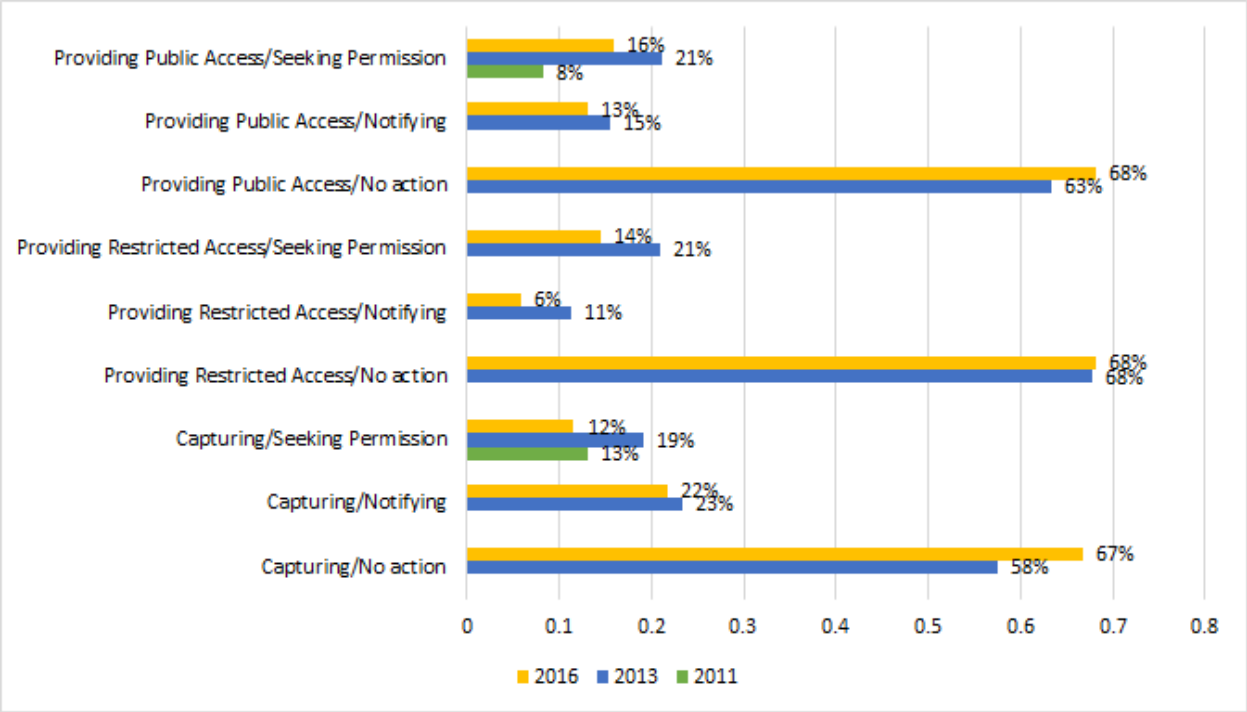

FIGURE 12: APPROACHES TOWARD SEEKING PERMISSIONS

Access Embargo

Access to archived content may be blocked or embargoed for defined time periods to reduce confusion with live websites, to reduce competition (with a news site, for example), or to address any number of other reasons. This section of the survey was launched in the 2013 survey, so all comparisons relate to that year.

The percentage of organizations that indicated the use of embargoes decreased by 15% from 28% in 2013 (21 of 76) to 13% (10 of 78) in 2016. Interestingly, those responding that they did not employ embargoes dipped by only 3% with "No" responses at 65% (51 of 78). About 17% of organizations (13 of 78) responded that they are "Considering" the use of embargoes, a new response option this year. The "Not Applicable" (dark archive) response rate came in at 5% (4 of 78).

Organizations that responded "Yes" to the question of employing embargoes as a matter of policy were asked in a follow up question about the duration of such embargoes. Nearly one-third, 27% (3 of 11), reported six-month embargoes in 2016, in contrast to 45% (10 of 22) in 2013. The percentage of one-year embargoes doubled to 18% (2 of 11) from 9% (2 of 22) in 2013, and the use of other embargo intervals, including several respondents who indicated that they let creators of the content decide, rose by 20%, to 55% (6 of 11).

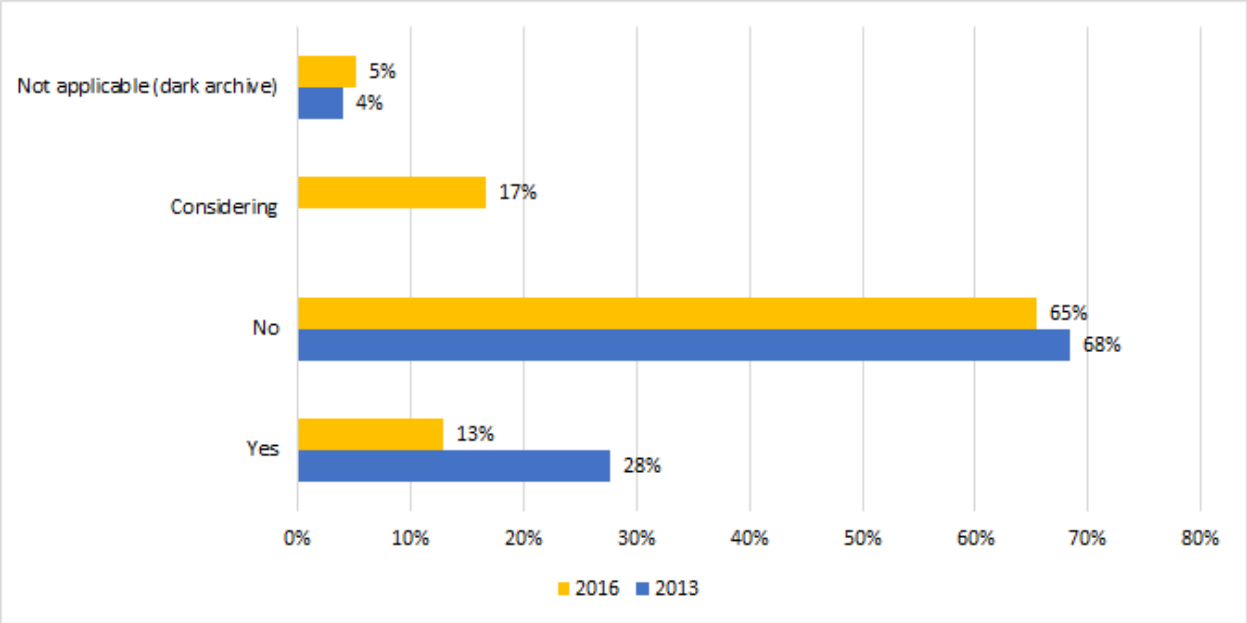

FIGURE 13: USE OF EMBARGOES IN ORGANIZATION POLICIES

Robots.txt Policies

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.

Responses to questions on policies for addressing robots.txt in the 2011, 2013, and 2016 surveys indicated a distinct and continued trend toward a conditional approach to the exclusionary protocol. In the 2016 survey, 9% (7 of 79) of respondents indicated that they always respected robots.txt, decreasing from 22% (17 of 77) in 2013 and 38% (22 of 58) in 2011. Organizations indicating that they never respected robots.txt also fell to a 4% rate (3 of 79), from 8% (6 of 77) in 2013 and 9% (5 of 58) in 2011. Organizations that selected the "Sometimes/it depends" option grew by 21% relative to 2013, with 76% (60 of 79) of respondents indicating this response. For the first time in the survey series, there were no organizations reporting that they "Don't know" how they respond to robots.txt.

Both 2013 and 2016 surveys requested information from those who selected "Sometimes/it depends" about the conditions under which respondents may decide to ignore the robots.txt protocol. As was the case in 2013, most organizations (61%, 37 of 61) indicated that they owned the copyright or had special permissions (e.g., in-house collection or an organization charged with archiving government records). Slightly less than half of respondents, 48% (29 of 61), said that they secured permission before ignoring robots.txt. In 2016, 52% (32 of 61) of responding organizations reported that they have

adopted policies to bypass robots.txt protocols when capturing "essential" content (e.g., stylesheets and images), an increase from 2013, when 43% (19 of 44) of respondents indicated such policies.

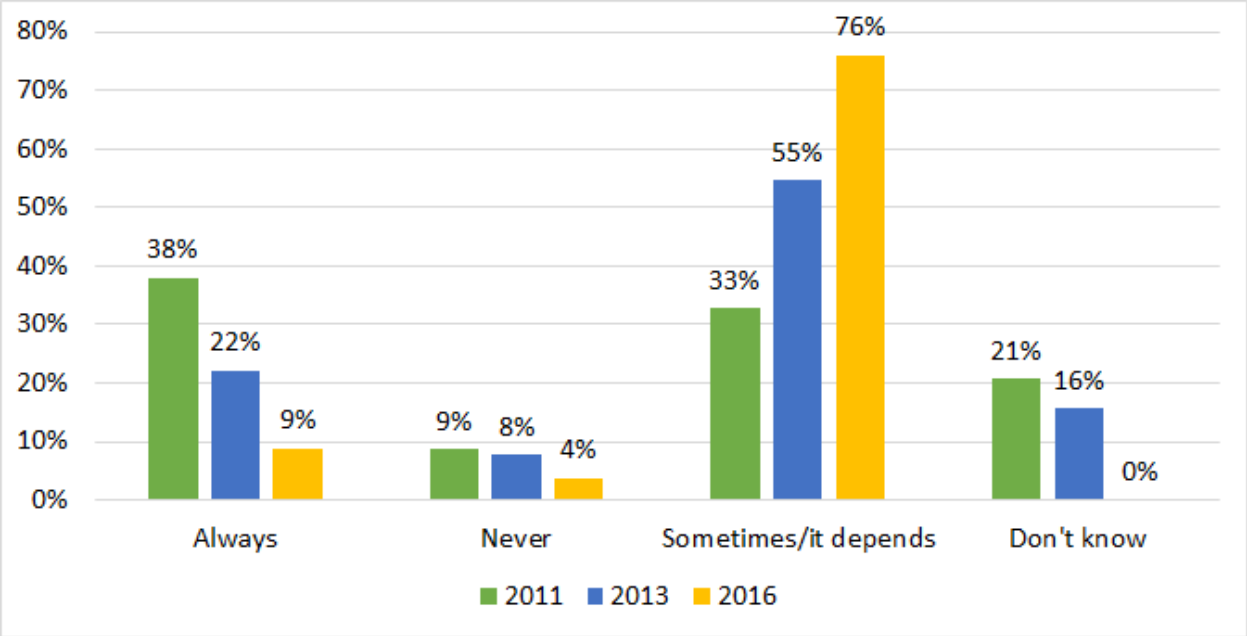

FIGURE 14: POLICIES FOR RESPECTING ROBOTS.TXT

Copyright and Policy Development Resources

More than half of 2016 respondents, 60% (41 of 68), said they rely on the policies of similar organizations for developing their own copyright and access policies, increasing slightly from 55% (30 of 55) in 2013. In a tie for second place, 40% (27 of 68) of organizations indicated a reliance on both the "Association for Resource Libraries Code of Best Practices in Fair Use for Academic and Research Libraries"[7] and "Legal Counsel." The 2016 survey for the first time included "Legal Counsel" as an optional response, which may have influenced the significant increase in response (as opposed to a response provided in an open ended text field in 2013). Only 25% (17 of 68) of organizations cited "The Section 108 Study Group Report"[8] in 2016, similar to responses in 2013. "Statutory authority" and "Other" were indicated by 16% of responding organizations each (11 of 68), previous NDSA surveys indicated by 15% (10 of 68), and "Oakland Archive Policy"[9] by 4% (3 of 68). Other open ended responses mostly indicated that the development of policy is in progress, though

[7] The Association of Research Libraries (ARL), "Code of Best Practices in Fair Use for Academic and Research Libraries," January 2012, accessed December 23, 2016, http://www.arl.org/storage/documents/publications/code-of-best-practices-fair-use.pdf.
[8] "The Section 108 Study Group Report," March 2008, accessed December 23, 2016, http://www.section108.gov/docs/Sec108StudyGroupReport.pdf.
[9] "The Oakland Archive Policy," December 2002, accessed December 23, 2016, http://groups.ischool.berkeley.edu/archive/aps/removal-policy.html.

one indicated SAA standards for guidance, and another referenced leadership in the copyright and archives community.

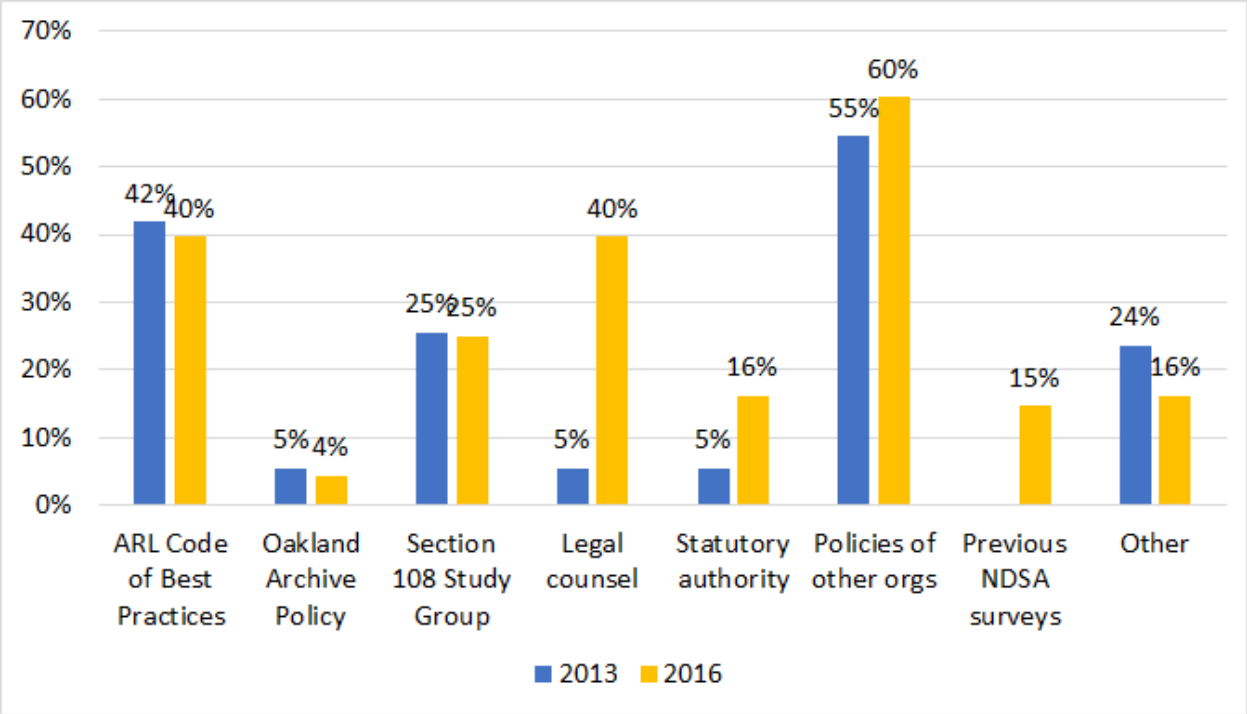

FIGURE 15: RESOURCES USED IN COPYRIGHT AND ACCESS POLICY DEVELOPMENT

Social Media

The rapid growth of social media sites such as Facebook, YouTube, Twitter, and others, present new opportunities and challenges for those collecting Web content. Given the significant impact of social media sites, responses to the survey's query regarding whether organizations had policies that specifically addressed social media were surprising: only 7% (6 of 82) answered affirmatively. This was a 17% drop relative to 2013, the first year that the survey inquired about social media, when the "Yes" response was 24% (19 of 78). One plausible guess from the survey team is that social media is simply not being treated differently than other content platforms.

TOOLS AND SERVICES

The Web, and necessarily the tools and services needed to archive it, change quickly. The goal of this section of the survey was to learn about what technologies organizations are using to archive the Web, comparative reliance on free tools and external services, and interest and intentions for replicating data from external services.

Local and External

The highest observed proportion of respondents were using external service providers to carry out Web archiving in the latest survey: 94% (74 of 79) in 2016, compared with 79% (65 of 82) in 2013 and 65% (47 of 63) in 2011. The growth appeared to be among those that both leverage external services as well as make use of other tools, since the proportion of programs relying exclusively on service providers only ranged 60-63% across the three surveys. The percentage of organizations using both external service providers and local tools increased to 30% (24 of 79) in 2016 from 16% (13 of 82) in 2013 and 14% (9 of 63) in 2011, suggesting increased local experimentation and interest in hybrid approaches.

The array of tools used by respondents remained diverse. Heritrix and HTTrack were again the most popular, with the highest percentage of users in 2016: 31% (9 of 29) and 28% (8 of 29), respectively. Webrecorder, a Web archiving service released since the last survey, was used by 21% (6 of 29) of respondents. Other tools used by at least one respondent included Adobe Web Capture, Brozzler, Grab-a-Site, Snagit, Teleport Pro, Umbra, WAIL, Web Curator Tool, and Wget. Use of tools that do not support the WARC format continued, but it was unclear whether they were used in a related capacity (e.g., link checking) rather than for Web archiving itself.

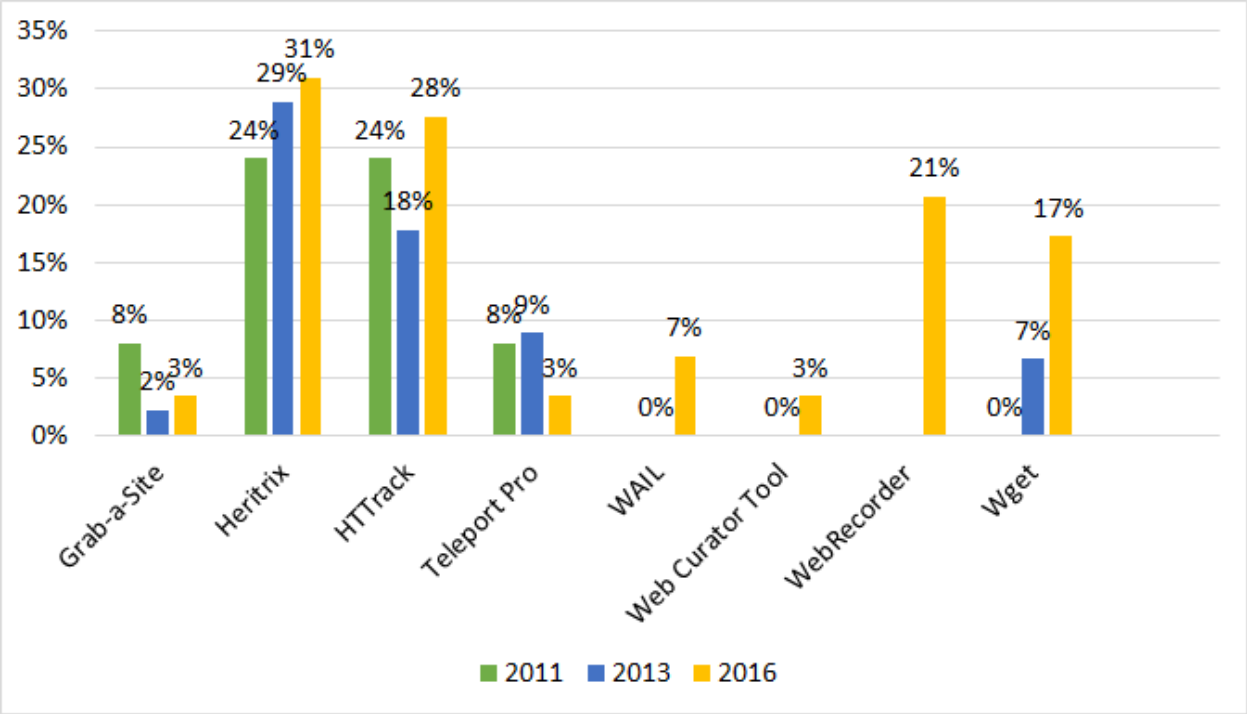

FIGURE 16: TOOLS USED FOR LOCAL CAPTURE

The California Digital Library Web Archiving Service ceased operation and transitioned customers to Archive-It during the time that the survey was active. Given this and the rates of subscription to Archive-It reported in previous surveys 71% (53 of 75) in 2013 and 72%

(36 of 50) in 2011, it was less surprising to see the increase in Archive-It use to 87% (69 of 79). Archive-It was both the most popular external service, as well as the most popular mechanism for US organizations to carry out Web archiving in general.

Data Transfer

The proportion of respondents who were transferring their Web archive data from an external service remained consistently low at 19-20% across all surveys, even as use of external services has grown. The longstanding question about data transfer was genericized in the latest survey to account for local and remote repositories as possible destinations for transferred data. There was also a follow-up question to learn about what kinds of repositories organizations were transferring data to. Most respondents were replicating to local repositories (59%, 10 of 17), with almost half replicating to external repositories (47%, 8 of 17), and a small number (6%, 1 of 17) replicating to both.

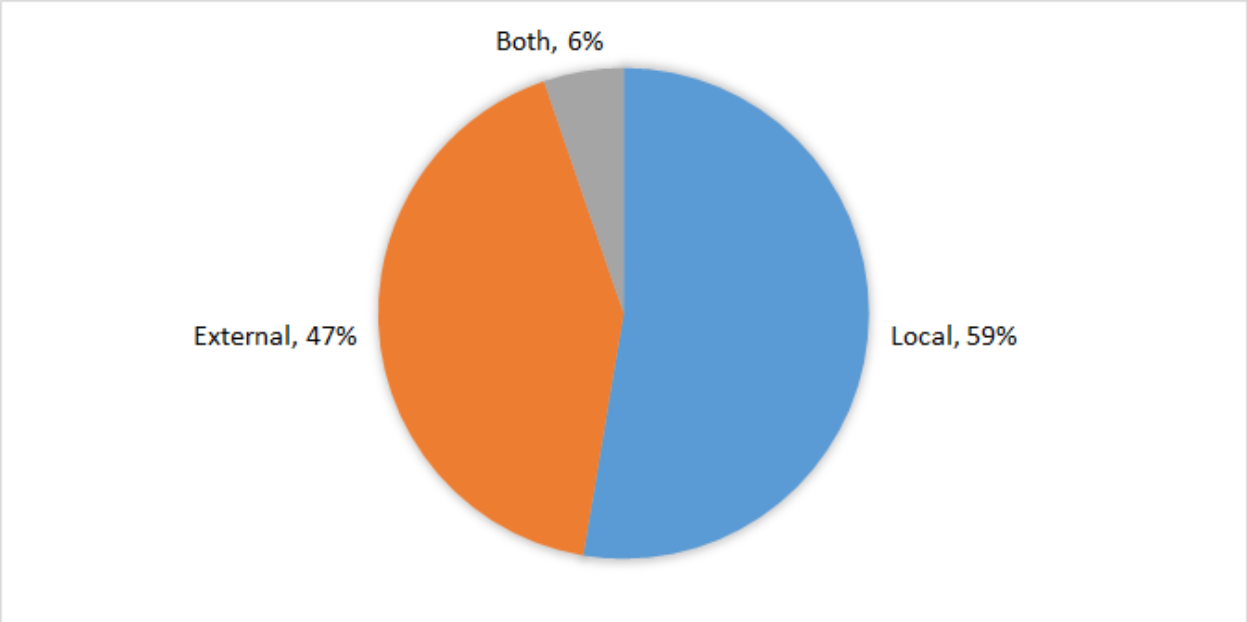

FIGURE 17: REPOSITORY FOR TRANSFER OF DATA

For the first time, trusting an external data capture service provider was the top reason for not replicating data to another repository. One-third or more of the respondents cited building local infrastructure (44%, 27 of 61), no place to store downloaded data (41%, 25 of 61), or being unsure about what to do with downloaded data once they had it (33%, 20 of 61). There were also a range of "Other" comments from 20% (16 of 21) of respondents in the 2016 survey, including concerns about local infrastructure, staffing, and funding; having future intentions to transfer data; not having captured data yet; and object data model mismatches between local systems and content from many websites being co-packaged in individual WARC files.

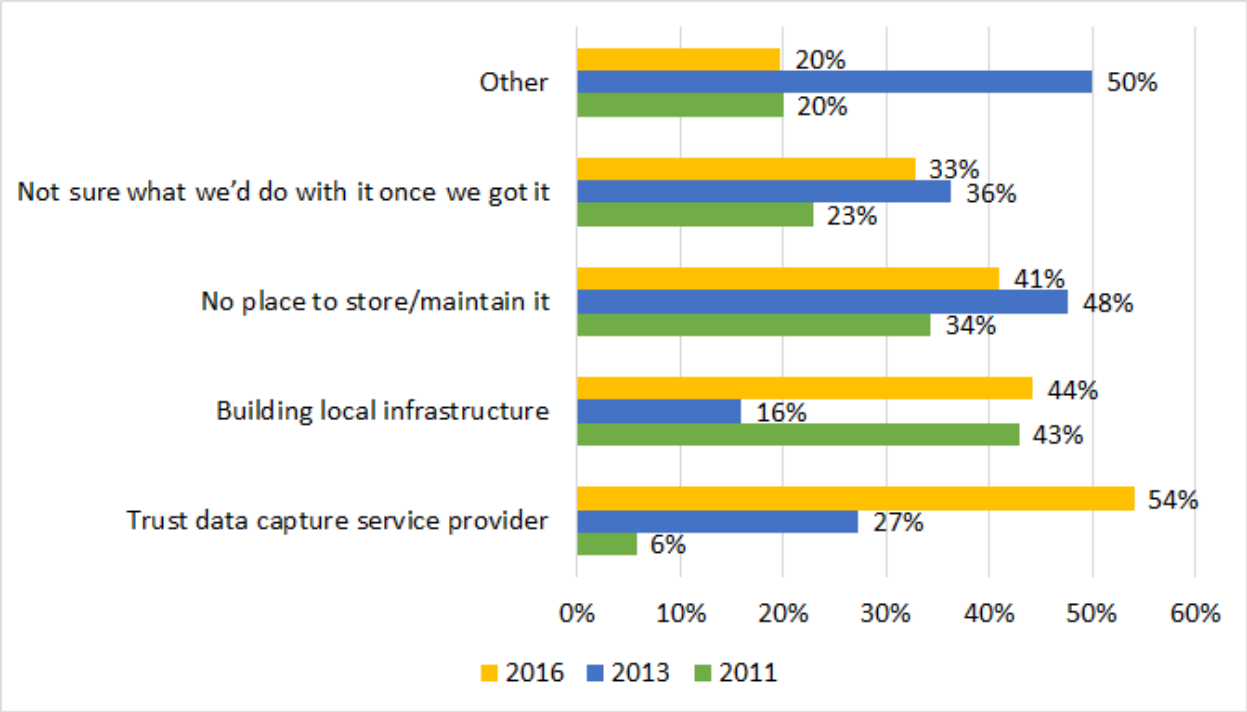

FIGURE 18: REASONS FOR NOT TRANSFERRING DATA FROM AN EXTERNAL SERVICE

ACCESS AND DISCOVERY

This section of the survey aimed to learn about how organizations are facilitating access to their Web archives. The 2016 survey included a new question, asking organizations if researchers were using their Web archives and, if so, to provide a summary of those users and activities.

Access Mechanisms

Organizations continued to provide many forms of access to their Web archives. In the 2016 survey asking respondents to choose from a variety of multiple choice answers, two key trends emerged. First, the number of organizations supporting methods such as search and item-level access points continued to decline. Second, the percentage of organizations creating collection-level catalog records and finding aids continued to grow. Search and browse access features covered four choices: URL search, full-text search, browse list by URL, and browse list by title. The number of organizations supporting search by URL dropped to 41% (26 of 63), down from 55% in 2013 and 62% in 2011. The percentage of those supporting full-text search saw a similar decline to 52% (33 of 63) in 2016, from 67% in 2013 and 66% in 2011. Supporting browsing by URL and title also continued a downward trend from the highest percentage in 2011, with only 30% (19 of 63) supporting URL-browsing, down from 44% in 2013 and 47% in 2011. The number of organizations browsing by title fell to 41% (26 of 63) from a high of 55% in 2011. These trends were offset by the increase in the number of organizations creating collection-level catalog records and

finding aids. Those creating collection-level catalog records increased to 30% (19 of 63), up from 22% of 2013, and those creating finding aids grew to 30% (19 of 63) from 21% in 2013; the survey did not contain this response option in 2011. Percentages for other access method responses remained largely consistent across all surveys.

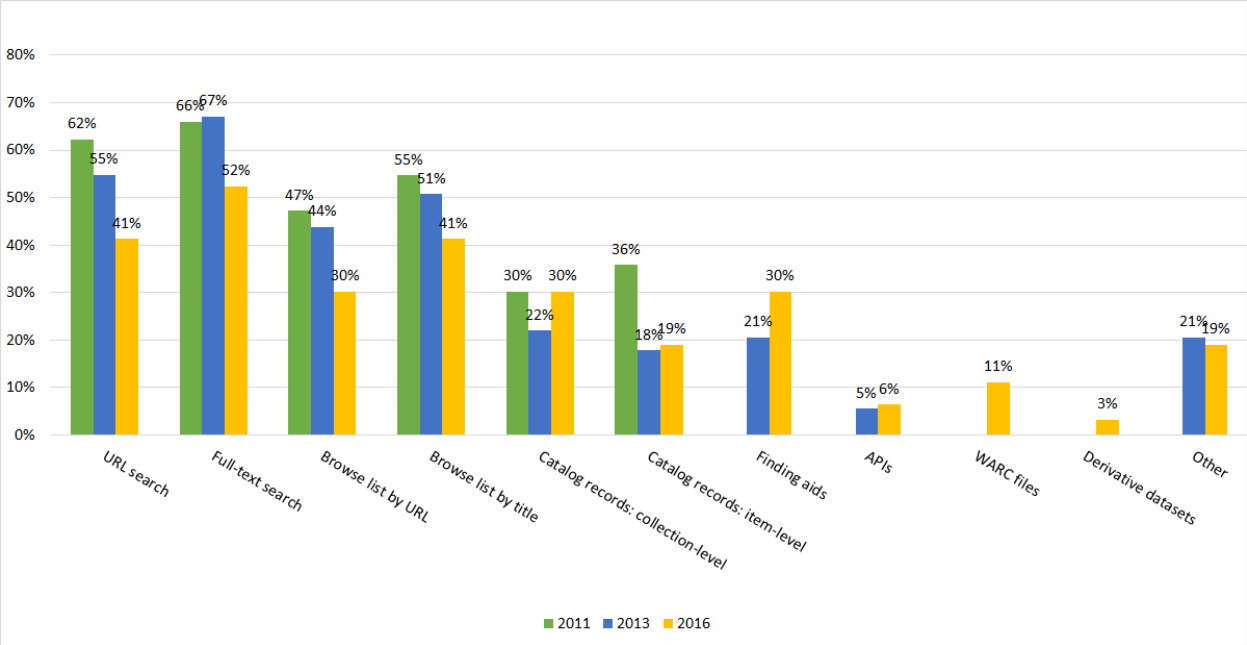

FIGURE 19: KINDS OF ACCESS PROVIDED

In considering the conclusions to be drawn from these trends, a number of possible interpretations emerge. For one, the widespread use of Archive-It, which includes full-text search and browsable lists in its public portal, could play a role in diminishing institutional interest in providing localized full-text or URL-based search internally, especially given that custom portals can link collections hosted in Archive-It to local discovery layers; this explanation is plausible, given a rise in Archive-It subscription from 71% in 2013 to 87% in 2016 (see Tools and Services section). The increase in organizations creating catalog records and finding aids suggests an increasing institutionalization of Web archives within traditional methods of description and discovery. This suggests the ongoing integration of Web archiving into existing library and archival strategies for intellectual control and a further movement of collection-level descriptive metadata into established online catalogs and discovery portals.

The trend away from specialized access methods and towards consolidation with established systems brings with it both benefits and dangers. On the one hand, the trend away from search and browse and towards traditional cataloging and finding aids signals a growing legitimacy of Web archive collections as they are folded into already-supported access systems and descriptive methods. At the same time, the move away from discovery

systems tailored to, or taking advantage of, the unique affordances and characteristics of Web archives, and the subsequent attempts to fit them into traditional methods based on bibliographic categories, standards, or workflows not created with Web archives in mind, has the potential to stifle a notable opportunity for creativity and innovation around access and discovery. This holds especially true for access and discovery methods that may be more familiar to native users of the Web who are less familiar with library OPACs, cataloging, and discovery systems.

Integrated, institutionalized access for Web archives reveals itself as a complex activity area that merits thoughtful consideration of resource allotment and strategic benefits. The survey does, however, reveal additional forms of access in the "Other" category that are not well evidenced by the existing survey responses and point to novel thinking around enhancing access. For instance, one respondent noted: "we have a link to our Web archives on my university's 404 error page." Others noted tools like LibGuides and blogging that served as additional forms of access.

Use by Researchers

The 2016 survey included a new question prompted by the growing maturity of many Web archiving programs and their subsequent turn to focusing on use and user communities. The question was: "Do you have active researchers utilizing your web archives?" Given the relative youth of many programs, as well as the fractional nature of staffing and other resource limitations, lack of knowledge of downstream use is perhaps not surprising. Of the 80 responses, 19% (15) answered "Yes," 30% (24) answered "No," and 51% (41) answered "Don't know." Many programs did, however, track access metrics using tools like Google Analytics, but translating these numbers into on-site or substantive researcher use can be challenging. Thus, the lack of clarity on formal research use is understandable, but it does signal an area of activity that merits community attention, knowledge sharing, and success stories. Organizations that answered "Yes" were asked to provide a summary of how researchers were using their Web archives. Narrative responses identified historians, social and political scientists, and institutional faculty as the primary research user communities.

LONGITUDINAL ANALYSIS

For the first time, analysis on repeat responders to the NDSA survey has been done. Between 2011 and 2016, 195 organizations have participated in the NDSA Web Archiving Survey. The vast majority, 139 (71%), has participated only one time. Forty organizations have participated twice, and 16 have participated in all three surveys. Of these 16 organizations, 10 programs were at universities, four were in the federal government, one was a museum and one represented a program in state government. By looking at the data collected from these 16 organizations over the past six years, one can gain a sense for how Web archiving programs have evolved. One should consider when reviewing the data,

however, the possibility that different staff representing Web archiving activities within an organization responded to the different iterations of the survey. Respondents within the same organization may have had differing perspectives on the questions posed. Indeed, answers to perhaps less subjective questions, like the year that an organization started Web archiving, changed over time; this date changed twice in responses of 8 of the 16 organizations and three times for 1.

In 2011, 12 of the 16 organizations characterized the status of their Web archiving activities as production/actively crawling. Two organizations indicated they were planning/considering archiving, but hadn't started yet, and 2 were in a pilot/testing phase. By 2013, those that were planning/considering and pilot/testing moved to production/actively crawling, and all reported this same status in 2016. When asked about perceptions of progress since the previous survey, in 2013 (since 2011) and 2016 (since 2013), 75% (12 of 16) indicated some or significant progress in 2013 (8 indicated that they had made significant progress) and 69% (11 of 16) in 2016 (1 left this question blank). These results strengthen the very positive perceptions of progress described in this report.

As reflected across organizations in the Content Being Archived section, there was a similar trend among these 16 organizations toward focusing more on their own or affiliated Web content, and less on the content from other organizations or individuals for future research. This may suggest that the data does not simply reflect a common initial goal of newer programs: 81% (13 of 16) indicated archiving their own content as a goal in 2016, compared with 75% (12 of 16) in 2013 and 69% (11 of 16) in 2011. Likewise, the data reflects a decreasing focus on harvesting outside content, with only 68% (11 of 16) in 2016 indicating this as a goal, compared with 81% (13 of 16) in 2011. The number of organizations marking both "Other" and "Own" has remained nearly the same (8 of 16 in 2011; 9 of 16 in 2013; 9 of 16 in 2016).

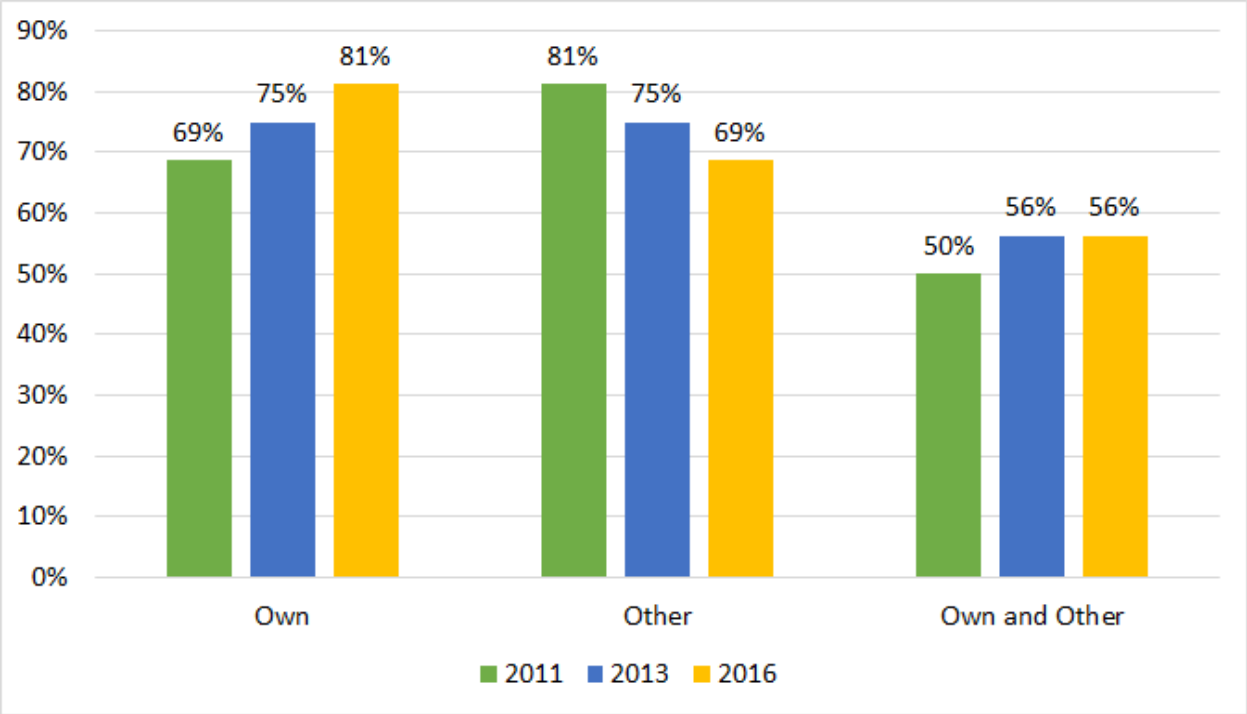

FIGURE 20: GOALS OF WEB ARCHIVING ACTIVITIES

SUMMARY

Overall the 2016 survey reflects positive developments in the Web archiving community in the United States. Respondents noted progress in key activity areas such as capture, integration with discovery systems, appraisal, and professionalization. Other areas, such as metadata, policy, quality assurance, and collaboration continued to introduce challenges to program and community growth. Key themes emerged from the report.

Institutionalization

A number of the survey results pointed to an increase in the institutionalization of Web archiving. There was a noted increase in production-level programs, with fewer respondents identifying their programs in the pilot stage. Additionally, the growth in both the number of overall survey respondents and those participating in professional groups indicates advances in specialization, expertise, and community building—a heartening trend. This growth is perhaps informed by the increase in academic institutions pursuing Web archiving, a trend which likely also informs the further formal integration of Web archive collections into existing discovery and access tools such as online catalogs and finding aids.

Increased institutionalization elucidated other program characteristics. Most notable, perhaps, is the larger focus on preserving internal or institutional content and a concomitant decrease in the perceived mandate or ability to archive materials outside institutional purviews. This trend raises interesting questions about state of perceived

institutional responsibility for archiving content created by others. With limited resources (see Devoted Staff Time section), is it only natural for organizations to focus internally? Are more organizations engaging in Web archiving activities with an intent to fulfill individual records management or institutional history needs rather than to preserve our broader cultural heritage? Do we need greater advocacy for support structures for collecting content created by others? What do these figures indicate about perceptions of collective responsibility for preservation of the Web?

Perceptions of Progress

The activity area which had the highest perception of progress, and the second-highest interest in collaboration, was data capture. Following that, appraisal was the second-highest area in which respondents noted progress and was also the second-highest area of interest for collaborative possibilities. This suggests an overall comfort, confidence, and routinization of choosing what to capture and getting it. This trend could be the result of multiple factors: the proliferation of capture tools, the slight uptick in hybrid Web archiving programs, or a general maturation of programs leading to more targeted scope of collecting policies. Whatever the case, perceived progress in successfully archiving selected content is a positive development for the community.

Areas of Opportunity

The survey results illuminate several areas of opportunity for continued progress. Access and use, especially by researchers, remains a perceived area of need. Likewise, metadata is identified as an area that would benefit from ongoing knowledge-sharing around best practices. Social media and quality assurance continue to be recognized as areas for which better and more accessible tools are needed.

Web archiving continues to be an activity that is fractionally staffed at most institutions. This paints a complex picture regarding perceptions of progress, making it challenging to identify whether this is a community-wide area of need. Clearly, some organizations view advocacy for additional resources (staffing) as a necessary component of program growth. This is generally in line with themes identified in the 2015 NDSA National Agenda for Digital Stewardship, which notes that "despite continued preservation mandates and over ten years of work and progress in building a professional practice around digital preservation, the community still struggles with advocating for resources, adequately staffing to support digital preservation and articulating the shared responsibility for stewardship."[10] For other organizations, however, fractional staffing seems adequate for meeting program requirements. Diversity of institution types and a wide range of extent, volume, and scope of Web collecting goals make it difficult to generalize.

Resource commitments and needs for Web archiving programs, including staffing requirements and time, merit a more detailed investigation. Notable, however, is the

[10] "2015 National Agenda for Digital Stewardship," 5.

significant interest in collaboration on a number of fronts, a trend which has grown across each survey. Respondents elicit a broad desire for collaboration in many areas of Web archiving, though many institutions feel they have neither the time nor resources to participate in collaborative activities. The community and the stakeholders need to invest in research and development efforts to create sustainable frameworks that facilitate meaningful, practical, and effective collaboration.

ACKNOWLEDGEMENTS

We would like to thank all the participants in the 2016 Web Archiving Survey for their time and willingness to share information about their programs with the broader Web archiving community. Many thanks to NDSA, CLIR, and others for their support and feedback throughout the project, especially Micah Altman, Oliver Bendorf, Aly Desrochers, Maureen Harlow, Katherine Kim, Carol Kussmann, Bethany Nowviskie, Mark Phillips, Abbey Potter, Christian W. Skipper, Kathlin Smith, and Lauren Work. The Intramural Research Program of the U.S. National Institutes of Health, National Library of Medicine, supported the contributions of Christie Moffatt to this report.

APPENDIX A

2016 Web Archiving Survey Questions

PDF of survey questions available at http://ndsa.org/publications/.