2017 Fixity Survey Report

An NDSA Report

Results of the Fixity Survey

AUTHORS
Sarah Barsness, Minnesota Historical Society
Aaron Collie, Federal Reserve Bank of St. Louis
Michelle Gallinger, Gallinger Consulting
Carol Kussmann, University of Minnesota Libraries
Sibyl Schaefer, University of California, San Diego
Gail Truman, Truman Technologies
Representing the NDSA Fixity Survey Working Group

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/SNJBV

EXECUTIVE SUMMARY

Fixity checking, or the practice of algorithmically reviewing digital content to insure that it has not changed over time, is a complex but essential aspect in digital preservation management. To date, there have been no broadly established best practices surrounding fixity checking, perhaps largely due to the wide variety of digital preservation systems and solutions employed by cultural heritage organizations. In an attempt to understand the common practices that exist for fixity checking, as well as the challenges institutions face when implementing a fixity check routine, the National Digital Stewardship Alliance (NDSA) Fixity Working Group developed and published a survey on fixity practices in fall of 2017. A total of 164 survey responses were recorded, of which 89 completed surveys were used in results analysis. Several themes emerged from the survey results:

- The vast majority of respondents are fixity checking their content or are planning to do so.
- Fixity checking at specific times, such as to establish an initial baseline fixity value after acquiring content and rechecking fixity values after data transfers, is recognized as good practice.
- Fixity checking practices varied widely based on the organization, the amount of resources (human and technical) available, the types of software and hardware implemented, and the size of the data corpus to be preserved.

The most common challenges associated with implementing a fixity check routine appear to be related to a lack of resources available to set up and maintain the infrastructure necessary to run fixity checks.

INTRODUCTION

Preserving physical and digital assets is an important part of the public mission of many institutions. Such assets hold informational, artifactual, cultural, and research value for current and future generations. Over time, degradation of materials in our collections may diminish their value or utility, which in turn could damage the reputation of galleries, libraries, archives, and museums (GLAMs) as trustworthy stewards of the world's accumulated knowledge. While this report focuses specifically on fixity practices in GLAMs and government entities, it's important to note that the responsible stewardship of vital information is important in many related fields, such as finance and law.

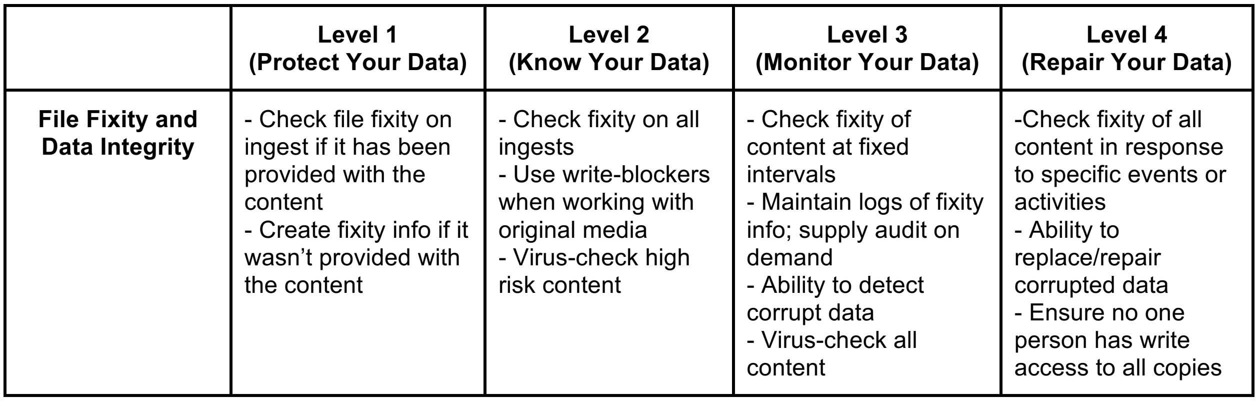
To address this need, many cultural heritage organizations undertake preservation and conservation measures such as condition surveys in order to monitor and attempt to minimize the degradation of physical cultural heritage materials. This is a labor- and skill-intensive process and is oftentimes challenging due to the size and scale of an organization's holdings. In these cases, it may be possible to survey only a representative sample of the organization’s physical collection.

Digital assets must be similarly monitored as there are many factors that may impact their condition such as failing infrastructure, faulty hardware, unstable or interrupted network transactions, human error and more. While there are many different methods that can be employed by institutions to monitor and preserve the integrity of digital assets, the most commonly used is fixity. Most fixity procedures involve a computational method that takes a digital file as input and outputs an alphanumeric value; this output value is used as a baseline comparison each time the fixity operation is rerun. If the values match, the file has not changed. This works because the fixity operation reads every bit (the ones and zeros that make up a file) in a file and mathematically computes a hash or fixity value (a type of digital fingerprint) that can be used to determine any bit-level differences in a file over time. Fixity values can also be used to monitor a file system for duplicate files or missing files.

Checksums (like Cyclic Redundancy Checks, or CRCs) and cryptographic hashes (like MD5 and various SHA algorithms) are two popular methods for fixity checking because they are a relatively straightforward and reliable way of determining that a file has not changed at the bit-level over time. Though somewhat resource intensive due to reading every byte in a file, these computational methods can scale to very high volume data collections.

In 2013, members of the NDSA published a tiered set of recommendations for how organizations should begin to build or enhance their digital preservation activities: the NDSA “Levels of Digital Preservation.”[1] One significant area of focus in “Levels of Digital Preservation" is file fixity and data integrity, which provides guidance on when to check and/or create fixity values. This publication is one of the first that provided guidance on when to create and check fixity as a main digital preservation practice.

[1] NDSA Levels of Preservation Working Group, “Levels of Digital Preservation,” 2013, http://ndsa.org/activities/levels-of-digital-preservation/.



	Level 1 (Protect Your Data)	Level 2 (Know Your Data)	Level 3 (Monitor Your Data)	Level 4 (Repair Your Data)
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	-Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies

FIGURE 1: File Fixity and Data Integrity Row from the NDSA Levels of Digital Preservation Matrix

Following the "Levels of Digital Preservation," members of the NDSA published an eight-page recommendation of best practices for utilizing fixity information in 2014 called "Checking Your Digital Content: What is Fixity, and When Should I be Checking It."[2] This document was based on interviews with experts in the field of digital curation and preservation, as well as a number of popular blog posts regarding fixity information on *The Signal*, a digital preservation blog published by the Library of Congress.[3] In addition to issuing the "Checking Your Digital Content" report, members of the NDSA community raised fixity compliance as an area needing further investigation and support in both the 2014[4] and 2015[5] "National Agenda for Digital Stewardship."

> Fixity checking is of particular concern in ensuring content integrity. Abstract requirements for fixity checking can be useful as principals [sic], but when applied universally can actually be detrimental to some digital preservation system architectures. The digital preservation community needs to establish best practices for fixity strategies for different system configurations. For example, if an organization were keeping multiple copies of material on magnetic tape and wanted to check fixity of content on a monthly basis, they might end up continuously reading their tape and thereby very rapidly push their tape systems to the limit of reads for the lifetime of the medium. There

[2] NDSA Standards and Practices Working Group and Infrastructure Working Group, "Checking Your Digital Content: What is Fixity, and When Should I be Checking it?," 2014, http://ndsa.org/documents/NDSA-Fixity-Guidance-Report-final100214.pdf.

[3] Bailey, Jefferson, "File Fixity and Digital Preservation Storage: More Results from the NDSA Storage Survey," March 6, 2012, https://blogs.loc.gov/thesignal/2012/03/file-fixity-and-digital-preservation-storage-more-results-from-the-ndsa-storage-survey/.

[4] NDSA Coordinating Committee and Working Group Co-Chairs, "2014 National Agenda for Digital Stewardship," June 2013, http://www.digitalpreservation.gov/documents/2014NationalAgenda.pdf.

[5] NDSA Coordinating Committee and Working Group Co-Chairs, "2015 National Agenda for Digital Stewardship," September 2014, http://www.digitalpreservation.gov/docments/2015NationalAgenda.pdf.

> is a clear need for use-case driven examples of best practices for fixity in particular system designs and configurations established to meet particular preservation requirements. This would likely include description of fixity strategies for all spinning disk systems, largely tape-based systems, as well as hierarchical storage management systems.[6]

Complex variables may impact an organization's ability to meet published fixity checking best practices such as their maturity, budget, industry, knowledge, ability, resources, and organizational priorities—among many others. Digital preservation environments vary widely in design and implementation; adapting to recommendations made in “Checking Your Digital Content” may be difficult. For example, some organizations employ a single piece of software while others configure many interoperating pieces, some utilize cloud storage, others local storage, and still others use a combination of local and cloud storage options. File integrity and quality control workflows within digital preservation systems can also vary drastically based on the system’s architecture. For example, fixity checks may occur at different times depending on the institution's environment: during initial deposit only; during any file transmission; during scheduled backup routines; or periodically at specified times or when manually triggered. In addition, digital preservation systems are often interconnected with other systems that may be outside of the organization's control, such as those managed by a consortial partnership or a central IT organization.

For these reasons, members of the NDSA’s Infrastructure Interest Group and Standards and Practices Interest Group held a joint call in 2016 to discuss possible tactics to better understand the challenges and uncertainties facing cultural heritage institutions as they attempt to meet best practices for preserving the integrity of their digital collections. During this conversation, it was determined that a survey may be the best tool to produce a current snapshot of implementation and adoption of fixity practices in the GLAM community. This survey on fixity was then developed and released in 2017 by the newly created NDSA Fixity Working Group.

In crafting the survey, the group worked to identify fixity best practices beyond those laid out in “Checking Your Digital Content,” to learn which of those best practices respondents were implementing and which were challenging to implement, and to perhaps identify possible reasons that real-world practice falls short of best practices.

[6] Ibid.

With these aims in mind, the Fixity Working Group survey team wrote the survey to answer two research questions:

1. What common practices exist for fixity checking?
2. What are the challenges institutions face when implementing a fixity check routine?

This report provides a summary of the survey results and provides insight into the current practices of respondents. Survey respondents were also asked if they would be interested in participating in a follow-up interview in order to gather a number of use cases that may lend more detail into organizational practices. These use cases will be solicited and compiled in the future and are not part of this report.

METHODOLOGY

Survey questions were based on the topics addressed in "Checking Your Digital Content," and consisted of 30 questions organized into four thematic sections: The Basics (addressed if and why respondents' institutions use fixity information); Where, When, and How (solicited information about when, where, and how the respondents' institutions use fixity); Cloud Services (addressed fixity issues specific to using cloud storage services); and About Your Institution (gathered basic demographic information about respondents' institutions). The survey included several optional open-ended questions that aimed to capture the nuance of local practice and the reasons for that practice, as well as follow-up questions that were only displayed to respondents based on previous answers.

The survey was administered in Qualtrics from 8 August 2017 to 22 September 2017. Participation was voluntary, but the authors sought participation from the global digital preservation community (including and reaching beyond NDSA member institutions). The survey announcement and reminders were sent to many professional Listservs and groups to solicit participation; a list is provided in the codebook.

DATA FILES

The survey data was deposited alongside this report on the NDSA's Open Science Framework (OSF) page.[7] While the survey responses were captured using Qualtrics, the results were exported and deposited as a CSV file to allow for interactions without the need for specialized software or login credentials.

[7] NDSA's Open Science Framework (OSF) page can be found at osf.io/4d567/.

A total of 164 responses are recorded in the deposited data. Of these, 89 respondents completed the survey, while 75 did not. The “true” or “false” entry in the “Finished” column of the deposited CSV file will assist in data reuse. Only responses with “true” in the “Finished” column are used in this report.

To keep all responses anonymous, the data from several fields were removed:

- IP addresses that were automatically collected by Qualtrics software
- Location latitude and longitude information that was automatically collected by Qualtrics software
- Responses to Question 30: “Would you be willing to have us work with you to document your fixity practices in the form of a use case?”, as responses included contact information.
- Institution names or other identifying information provided by participants in response to Question 31: “Is there anything else you would like to tell us about your practices around fixity?”

In addition, free-text fields were reviewed and any identifying information was removed to limit the information that could be used to identify the respondent.

Further details about the data itself can be found in the codebook that was deposited alongside this report on the NDSA’s OSF page.

CODEBOOK

The codebook available on the NDSA’s OSF page establishes the context for the survey and its responses. It assumes that future users will have no prior knowledge of the survey or its data, so much effort went into ensuring that the Scope of Study and the Survey Overview sections in the codebook conveyed the goals and provided an overview of the survey. The codebook documents the full text of each question, the possible selections or responses, and the formats in which a respondent could enter their answer.

Specific information about how data was used or interpreted for this report can be found in Appendix 1. In general:

- This report only includes information from completed survey responses.
- Percentages for each question are calculated based on the number of responses to that question. For example, not every respondent may have answered a free-text

question, so percentages are calculated based on the number of comments received, not the total number of respondents who completed the survey.

FINDINGS

This portion of the report is organized by survey section and includes a summary of the survey section followed by the text of the original questions and a basic interpretation of the results.

Sections discussed below include:

- Basic information about respondents fixity practices; if and why they are using fixity information
- When, where, and how institutions are using fixity information
- Cloud services and fixity practices
- Demographic information about responding institutions

Section 1: Basic Information about Fixity Practices

Ninety-two percent of respondents are either using fixity in some way or indicated that they plan to soon. Those that are currently not using fixity information most frequently cited a lack of capacity for fixity checking in their programs and/or policies. The primary reason for utilizing fixity checking is to determine if data has been corrupted or altered over time, though the benefits of collecting, creating, maintaining, and verifying fixity information are many, as discussed below.

Question 1: Do your organizational practices include utilizing fixity information at any point in time?

Of the 88 responses to this question, 84.1% said “yes” (74 responses), while 15.9% said “no” (14 responses). Of those 14 “no” responses, eight cited the immaturity of their program as part of the reason they do not utilize fixity, and seven indicated they wanted to or would begin to use fixity information in the near future. Other reasons for not utilizing fixity were a lack of time or staff (three responses), lack of appropriate/functional tools (two responses), and a lack of knowledge/training (one response).

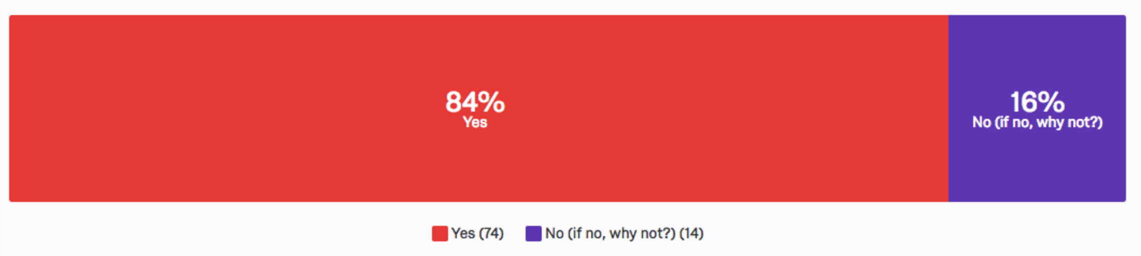


FIGURE 2: Responses to "Does your organization utilize fixity information at any point in time?"

Question 6: What are the reasons your organization collects, checks, maintains, and verifies fixity information? Please rate the importance of each of these items (not important, somewhat important, moderately important, extremely important):

The 74 respondents to this question rated eight provided reasons as extremely important, very important, moderately important, slightly important, or not at all important. The table below shows the responses provided on the importance for each reason (listed in survey order) for collecting, checking, maintaining, and verifying fixity information.

Reason	Extremely important		Very important		Moderately important		Slightly important		Not at all important		Total
Determine if the data has been corrupted or altered over time	86.5%	64	9.5%	7	2.7%	2	0.0%	0	1.4%	1	74
Determine if the data has been corrupted or altered during transmission	66.2%	49	16.2%	12	9.5%	7	5.4%	4	2.7%	2	74
To support the authenticity or trustworthiness of the digital objects	58.1%	43	23.0%	17	14.9%	11	1.4%	1	2.7%	2	74
To monitor hardware degradation	25.7%	19	24.3%	18	25.7%	19	14.9%	11	9.5%	7	74
For authenticity: To prove you are providing the digital object that has been requested	46.0%	34	23.0%	17	10.8%	8	13.5%	10	6.8%	5	74
To permit an update to a portion of a content file while proving the other portions remain unchanged (ex: split video files)	11.0%	8	16.4%	12	17.8%	13	16.4%	12	38.4%	28	73
Meet requirements or best practice guidelines	47.3%	35	33.8%	25	13.5%	10	4.1%	3	1.4%	1	74
Help identify systemic or human error in the management of digital content	51.4%	38	31.1%	23	14.9%	11	2.7%	2	0.0%	0	74
Other	55.6%	5	11.1%	1	33.3%	3	0.0%	0	0.0%	0	9

FIGURE 3: Reasons organizations collect, check, maintain, and verify fixity information

Determining if data has been corrupted/altered either over time or during transmission rank as the two most important reasons for checking, maintaining, and verifying fixity, while supporting authenticity/trustworthiness and identifying systemic or human errors were also popular reasons.

Nine respondents added responses to the “other” category, which offered several different reasons for using fixity, the most common of which were using fixity values to handle multiple versions or copies of files (e.g., de-duplication, comparing backups to master files, or updating specific files in the collection) and using fixity information as a form of inventory control.

Question 2: Does your organization collect fixity information (created by another institution or a separate entity within your institution) along with digital content at the time of acquisition if it is available?

Most respondents are collecting fixity information created by others if it is made available to them. In general, respondents indicated that fixity information is rarely provided, and when it is provided, it is generally from an external vendor as part of the terms of a digitization contract. More donor education and/or better tools to help non-vendors generate fixity information closer to the time of creation (i.e., before ingest) would increase the opportunity for respondents to collect pre-generated fixity information, thereby better verifying successful initial transfer and integrity/authenticity.

Of the 74 respondents who answered this question, 73% said “yes (always, frequently, sometimes, or very rarely),” while 27% said “no.” Of the 73% who collected fixity information in some manner at the time of acquisition, 28.4% said that they always collect fixity information if it is available.

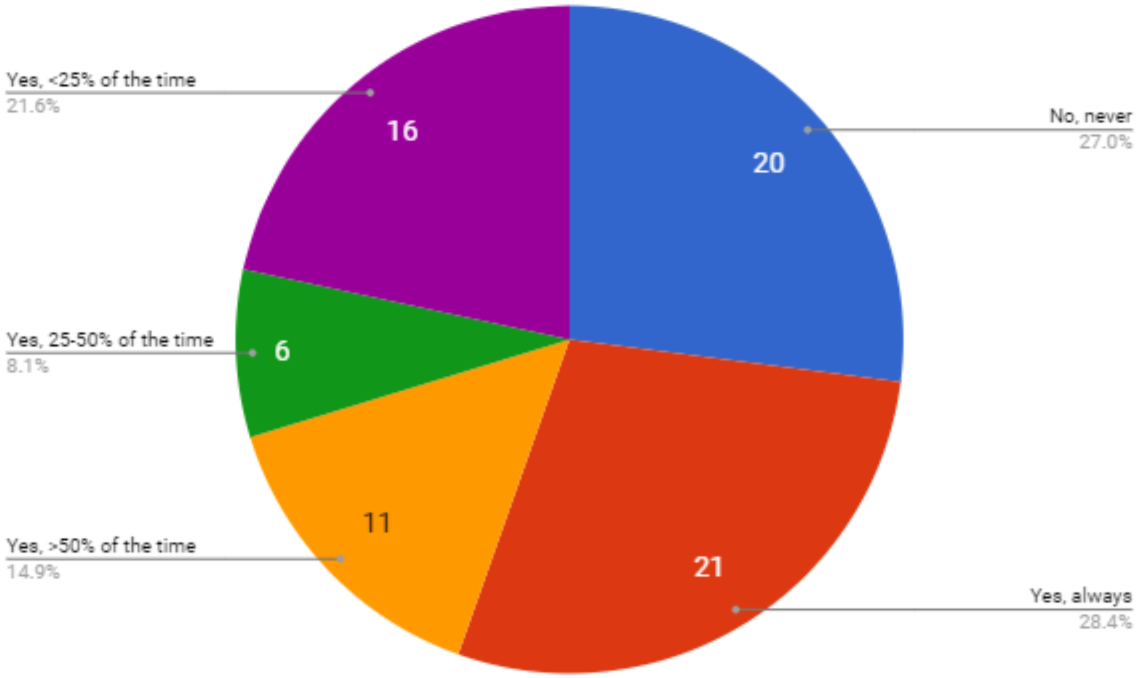


FIGURE 4: Frequency with which respondents collect fixity information created by another institution or a separate entity within their institution at the time of acquisition (if it is available)

Question 3: Please provide any relevant details about why you collect fixity information as frequently as you do.

Fifty-five respondents commented on this free-text question. Many provided details about how they collect externally-generated fixity values. Fourteen (25.5%) discussed receiving fixity information from vendors, and five (9.1%) mentioned receiving them from donors. Eighteen (32.7%) explicitly cited a consistent lack of provided fixity information, specifically from donors as opposed to vendors. Several respondents mentioned that they generated fixity values at ingest even if they were provided, often because of internal system requirements or policies about accepted/preferred checksum algorithms. Seventeen (30.9%) talked about using provided fixity values to ensure the complete and accurate transfer of data from the donor/vendor to the institution. Four (7.3%) mentioned using provided fixity information to ensure the authenticity and/or integrity of files.

Question 4: Does your organization create fixity checks for digital content if they are not provided at the time of acquisition? Please indicate how often you collect fixity information:

The vast majority (87.8%) of respondents are creating fixity information at least half of the time if it is not provided at the time of acquisition. Few respondents create fixity values only sometimes (5.4%) or not at all (6.8%). Many respondents indicate that they create fixity information during the acquisition and/or ingest process in order to capture fixity information as soon as possible as this can identify errors that may occur during future transfers or when held in storage.

Of the 74 responses to this question, 93.2% of respondents said they create fixity checks. Breaking that down further, 74.3% always create fixity values, 13.5% create fixity values frequently (more than 50% of the time), and 5.4% sometimes create fixity checks (between 25%-50% of the time). About 7% reported that they never create fixity values.

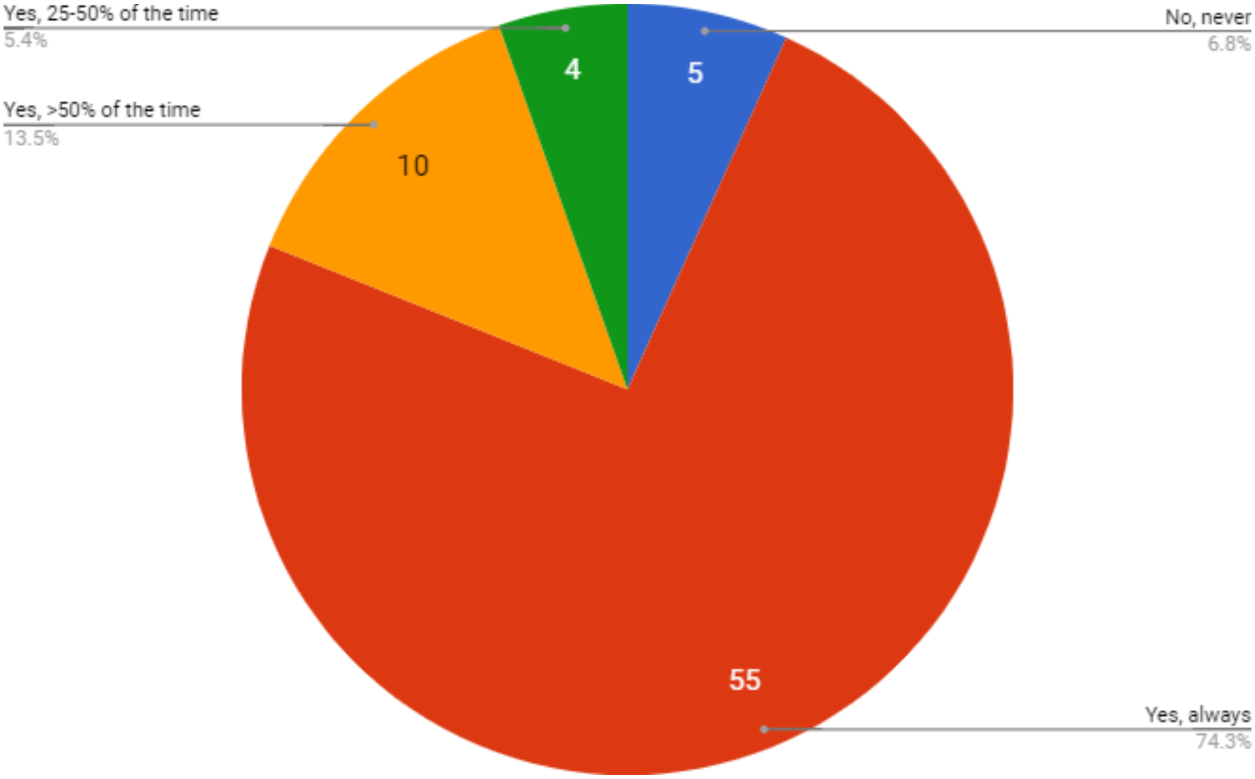


FIGURE 5: Frequency with which respondents create fixity checks for digital content if they are not provided at the time of acquisition

Question 5: Please provide any relevant details about why you create fixity information as frequently as you do.

Respondents provided 61 comments, 41.0% of which explicitly indicated they create fixity information early in the process: at acquisition (four comments), at ingest into a system (15 comments), or generally as soon as possible (six comments). Several respondents indicated the importance of file integrity, authenticity, and trustworthiness (22 comments), and others talked about technical issues such as ensuring complete/accurate transfers (11 comments) or ensuring storage was working properly (three comments). Some comments are provided below.

> *"Creating fixity information (i.e. file checksums) at the point of acquisition helps us to ensure the authenticity and integrity of the content from that point forward."*
>
> *"If fixity information is not provided by a vendor, for example, we create it upon receipt. This is to ensure that we can identify what content was transferred at a given point in time, and so we can ensure that files can be backed up and restored to this original state."*

> *"Calculating and recording the checksums of processed digital material allows us to better understand how our storage is working."*

Section 2: Fixity Practices: Answering Where, When, and How

The second section of the survey was designed to uncover more detailed information about fixity practices at the respondents' institutions.

Question 7: Do you check fixity information after transferring data?

Of the 74 respondents to Question 7, the majority (51 respondents, or 68.9%) stated that they did check fixity information after transferring data. This suggests that it is recognized that the data transfer process is a period where the risk of data alteration is high enough to warrant a fixity review.

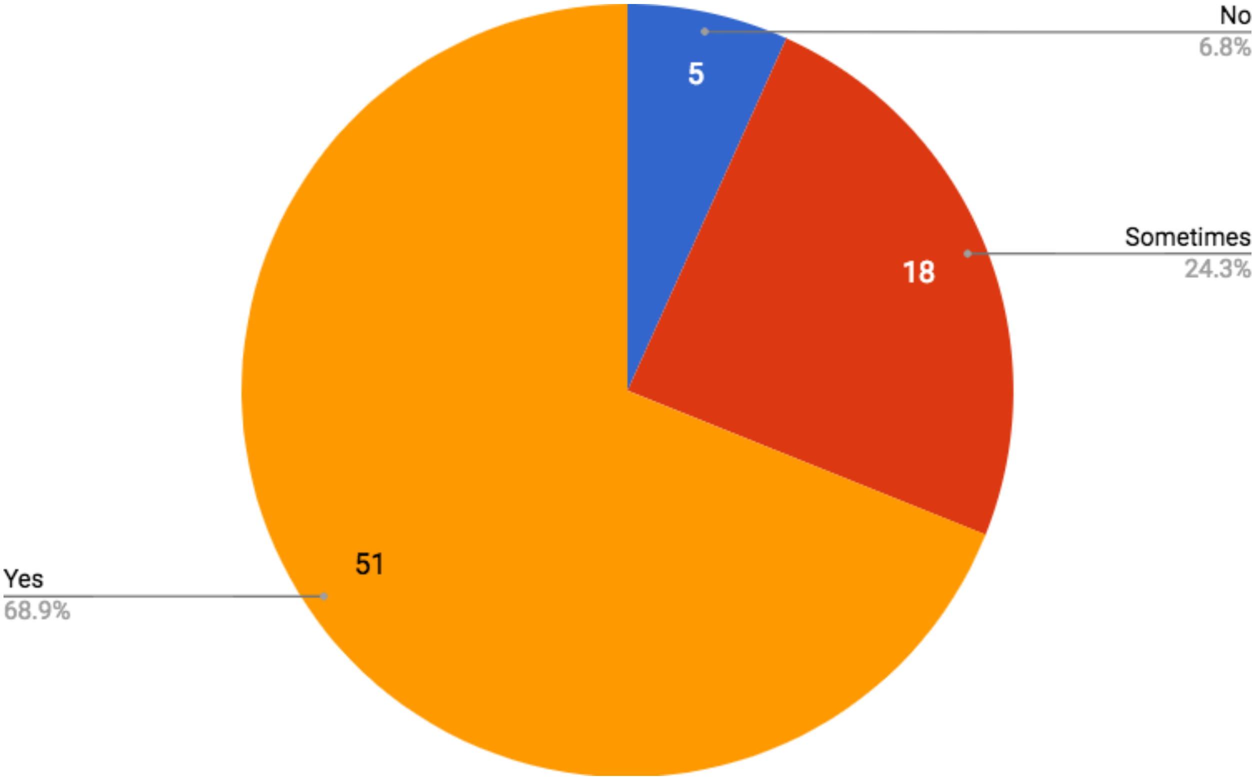


FIGURE 6: Responses to "Do you check fixity information after transferring data?"

Question 8: Do you check fixity information at regular intervals - please specify the intervals that your organization uses. *Select all that apply.*

There were a total of 72 respondents who selected 107 different responses. Nineteen respondents selected multiple fixity checking intervals. Although the survey attempted to be exhaustive in the options listed, over half of the respondents selected the "other" category. The most common responses in the "other" category indicated that respondents checked fixity at "variable intervals" (six respondents), "on migration" (six respondents), or "on a continuous/rolling basis" (five respondents).

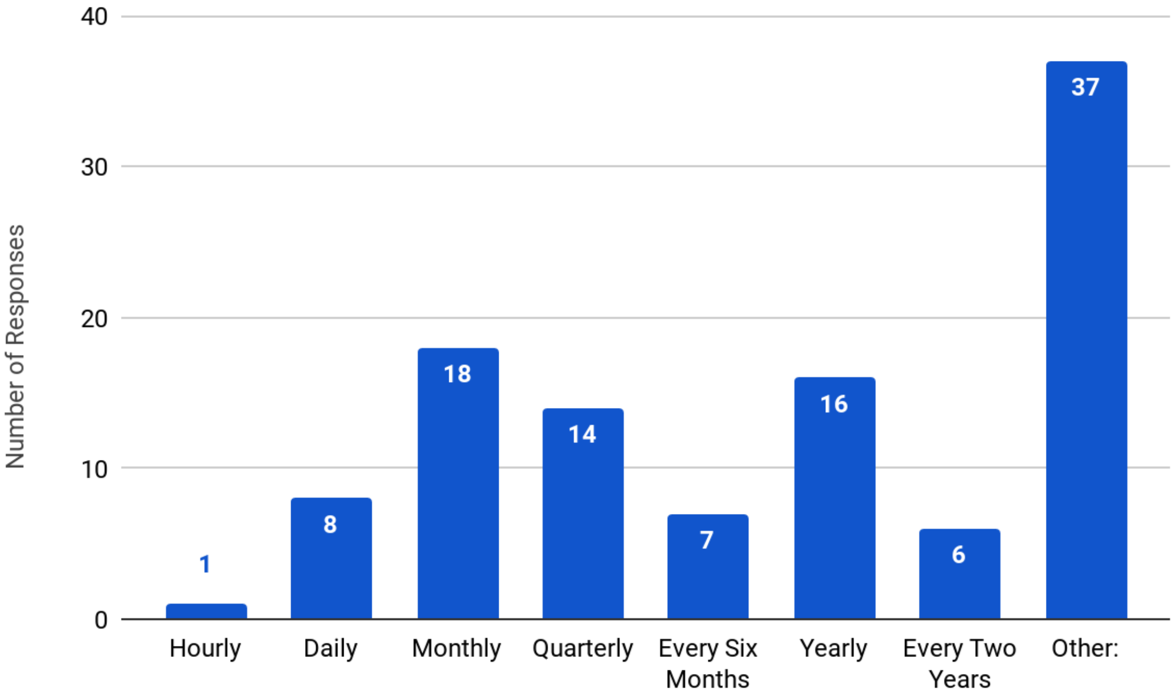


FIGURE 7: The intervals at which respondents said they performed regular fixity checks.

Question 9: Please provide any relevant details about how often you check fixity (e.g., differing frequencies based on storage location).

This open ended free-text question allowed survey participants to elaborate more about the frequency of their fixity checks. Most of the 44 responses clarified their fixity check frequency and also specified which system or storage media/location the checks occurred on. Responses indicated that different storage systems often were subject to different fixity check procedures, and that these procedures varied by the location of the storage, the management (responsibility) of the storage, and the type of system or storage itself. For example:

> *"Fixity checks vary by storage location and responsibility. Born-digital records are monitored on a quarterly basis by running AV Preserve's Fixity tool on a set of dedicated LAN drives—which are run out of servers managed and physically housed by libraries IT. Digital collections are stored on tape via campus IT and with Iron Mountain, and we (libraries) don't know whether/how fixity is monitored there."*
>
> *"Some locations are checked monthly. These locations are accessible to more staff, accessed more frequently, and do not have error-checking built into the filesystem. The digital repository is running on ZFS, is rarely accessed except to ingest new data, is highly restricted, and is large enough that a full fixity check takes more than one week. We have been checking it quarterly (i.e. verifying every bag in addition to the checking built into ZFS), but may reduce that frequency."*
>
> *"Varies primarily based on the storage used, and the affordances of the various systems that store our content."*

Question 10: How much total content (preservation copies that are managed for long-term preservation only) are you running fixity on? *Please provide your answer in total number of TB.*

Of the 74 respondents to Question 10, the largest group of respondents (18 or 24.3%) indicated that they were running fixity checks on 11-50 terabytes (TB) of content. Respondents who selected "More than 500 TB" were asked to indicate the amount. Of the 14 respondents who indicated that they were running fixity on more than 500 TB, six responded that they had less than one petabyte (PB) and five responded that they had between 1 and 5 PB. The largest total amount of content being that fixity was being run on by an institution was listed as 40 PB.

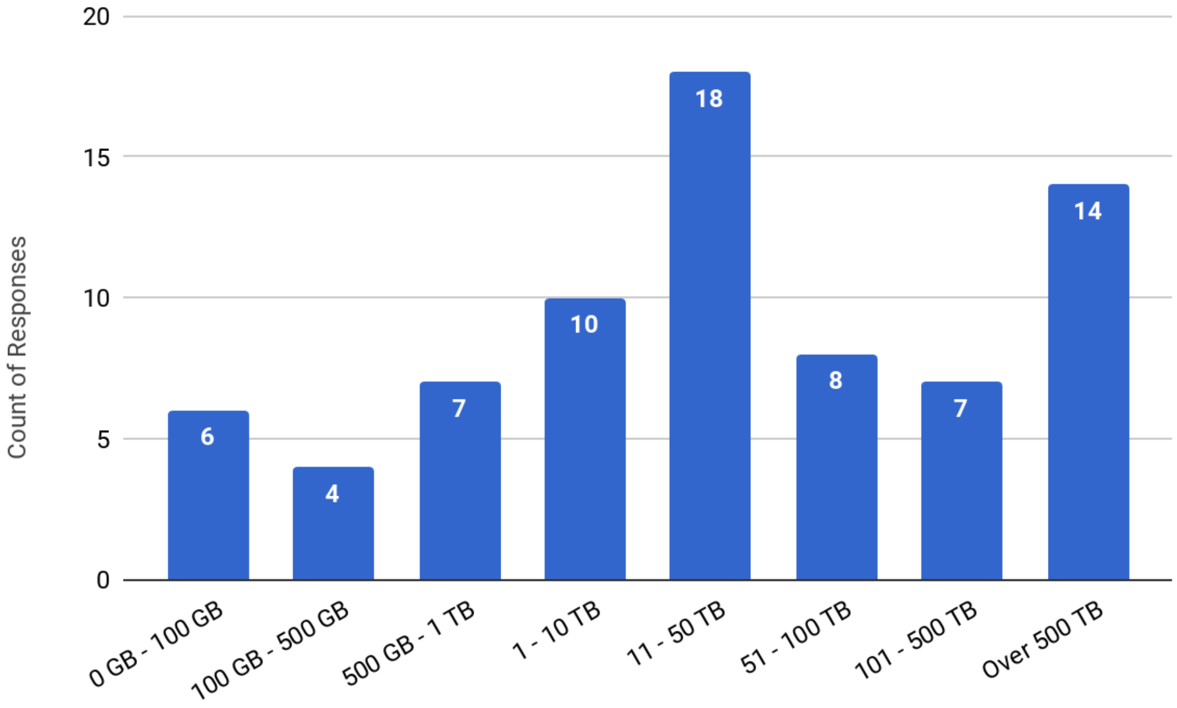


FIGURE 8: The amount of content that fixity is being run on provided by respondents

Question 11: What factors does your organization consider when determining fixity check frequency? *Please rate the importance of each of these items (not important, somewhat important, moderately important, extremely important):*

Respondents were asked to rank the importance of seven factors to consider when determining fixity check frequency. "Number and size of files or objects that require fixity checks" was ranked as an 'extremely important' factor when determining the frequency of fixity checking by 25 of the 72 respondents (34.25%). "Available staff time/resources," was a close second with 24 respondents (32.43%) selecting the ranking of 'extremely important'. Considerations that were most deemed 'not at all important' when determining fixity check frequently were "Reliance on checksums generated by storage providers (i.e. cloud providers and others)" (28 respondents, or 38.89%) and "Regular checks done at the block level via a system" (25 respondents, or 34.72%).

Field	Extremely important		Very important		Moderately important		Slightly important		Not at all important		Total
Concern of media failure due to increased use of storage media (e.g. tape)	20.83%	15	15.28%	11	29.17%	21	13.89%	10	20.83%	15	72
Storage media reaching end of expected lifespan	24.66%	18	21.92%	16	27.40%	20	9.59%	7	16.44%	12	73
Throughput limitations (e.g., network bandwidth)	17.81%	13	34.25%	25	24.66%	18	8.22%	6	15.07%	11	73
Number and size of files or objects that require fixity checks	34.25%	25	28.77%	21	20.55%	15	8.22%	6	8.22%	6	73
Reliance on checksums generated by storage providers (i.e. cloud providers and others)	20.83%	15	11.11%	8	15.28%	11	13.89%	10	38.89%	28	72
Regular checks done at the block level via a system	23.61%	17	11.11%	8	23.61%	17	6.94%	5	34.72%	25	72
Available staff time/resources	32.43%	24	22.97%	17	20.27%	15	13.51%	10	10.81%	8	74

FIGURE 9: Factors organizations consider important when determining fixity checking frequency

Question 12: Do you check fixity at regular intervals of a sampling of your digital content?

Checking fixity over only a sample of content does not appear to be a common practice; an overwhelming 65 out of 74 respondents to Question 12 answered "no." Respondents who answered "yes" were asked what their sample size was. While there was no notable consensus on sampling sizes, responses ranged from 1% to 50% of the total preserved content.

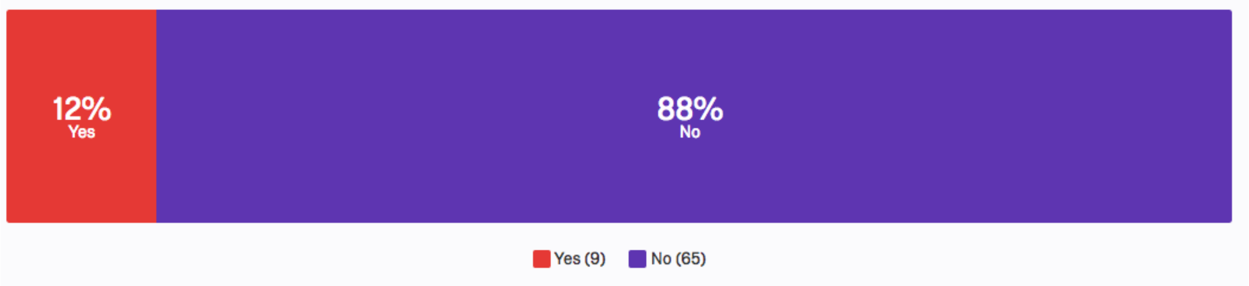


FIGURE 10: Responses to the question "Do you check fixity at regular intervals of a **sampling** of your digital content?"

Question 13: Do you check fixity at regular intervals of all of your digital content?

Given that the majority of respondents indicated in Question 12 that they did not run fixity on a sample of content, it is not surprising that the majority of respondents (55 of 74 respondents, or 74.3%) to Question 13 indicated that they ran fixity checking on *all* of their preserved content.

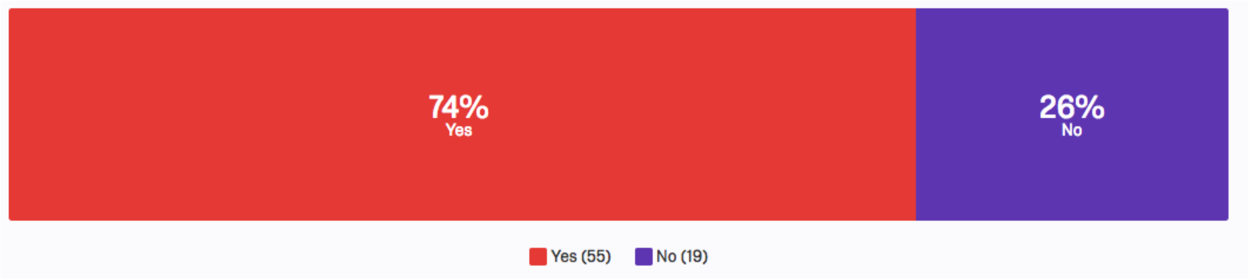


FIGURE 11: Responses to the question "Do you check fixity at regular intervals of **all** of your digital content?"

Question 14: Is your fixity checking done utilizing built in hardware or is it software based?

The survey team was interested in uncovering whether it was common to rely on built-in hardware fixity checks or if using software was more prevalent. The majority (49 respondents, or 68.1%) used software for fixity checking while the remainder (23 respondents, or 31.9%) used both a combination of hardware and software. It is of note that none of the 72 respondents indicated they used only hardware for fixity checking.

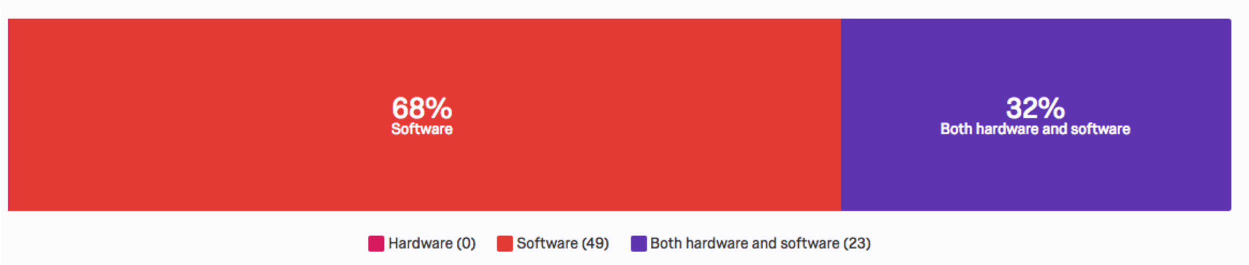


FIGURE 12: Break down of how respondents check fixity: utilizing hardware, software, or a combination of hardware and software.

If people responded "software" or "both hardware and software" in Question 14, Question 15 was displayed to help get a further understanding about what people are using to capture/verify fixity information.

Question 15: What software, tools, or services are you using to capture/verify fixity information? Select all that apply: [displayed based on Question 14 response]

There were 130 selections made by 73 respondents for this question. The highest number of respondents use automated or scheduled software (39 respondents, or 30%). Scripts (or custom code) and manually run software are each used by 26.92% or 35 respondents each. Third party services are used by 6.92% or 12 respondents, while 6.9% or nine respondents selected "other."

For those providing additional information about what tools or services they are using, three indicated Preservica's scheduled fixity checking with Amazon S3. Other respondents listed services built into products such as EMu software, ACE and MetaArchive, OCLO, DuraCloud, and Rosetta. AVP's Exactly and Fixity, AWS Etags, NetBackup, StorageTek, Qstar data archiving solution, AWS, tpverify by Oracle were also mentioned. The "other" responses listed: file system size (ZFS scrubbing), Archivematica's Fixity application, DROID, Archive software - Atempo Digital Archive. In addition, two respondents selected "other" and indicated that they were either not doing this yet, or they were in the process of moving to a more automated approach.

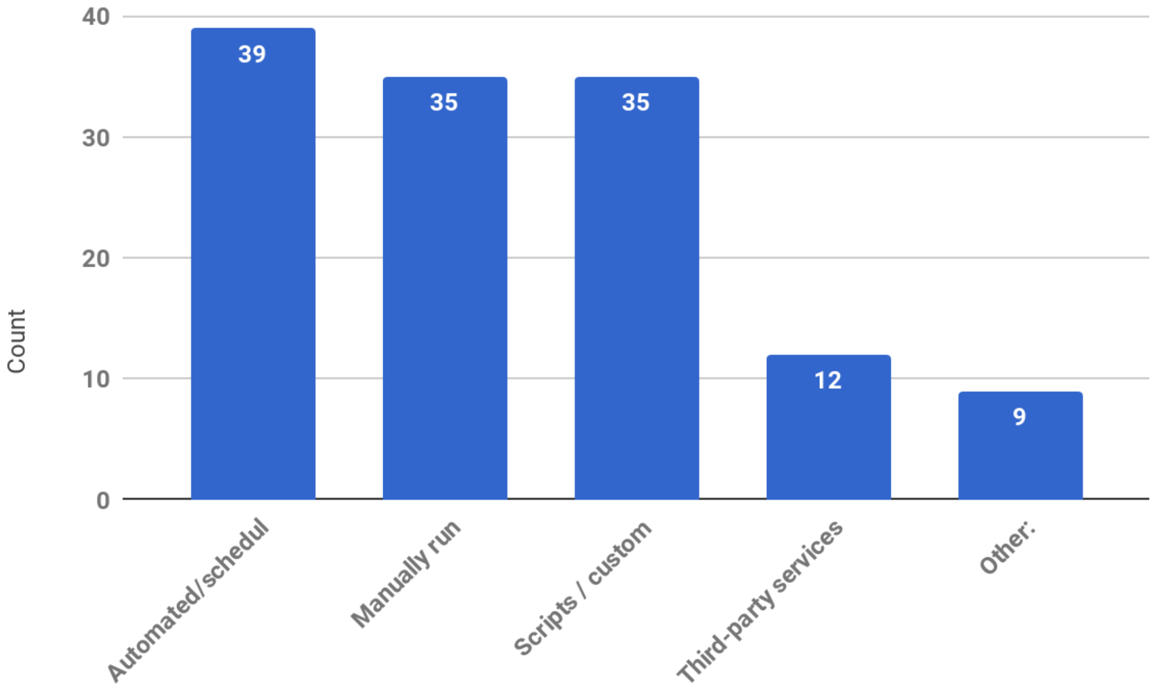


FIGURE 13: Type of software, tools, or services used to capture/verify fixity information

Question 16: What type of fixity checking algorithm does your preservation software use? *Select all that apply: [Displayed if people responded 'software' or 'both hardware and software' in Question 14.]*

When asked what fixity algorithm respondents used, 135 selections were made by 71 respondents. The largest number of respondents, 58 or 42.96%, use the MD5 algorithm; 34 or 25.19% use SHA256, followed by 28 or 20.74% who use SHA1. CRC checksums are used the least, by ten respondents or 7.41%. The five respondents who chose "other" listed other variations of SHA including SHA-224, -348, and -512. SHA512 was listed a total of four times.

Another respondent indicated that their software "creates tokens which are also validated to insure that the checksums have not been altered."

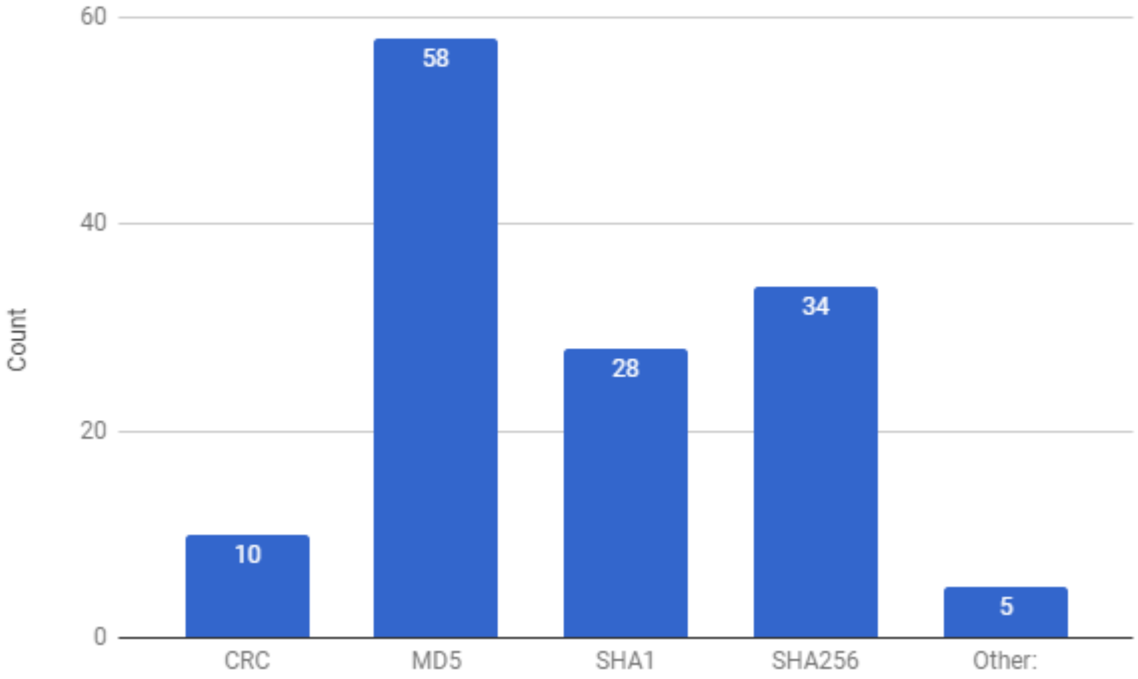


FIGURE 14: Type of fixity checking algorithm utilized within software

Question 17: Who is responsible for fixity checking (e.g., running manual scans, scheduling automated scanning, etc.) *Select all that apply:*

Thinking about the organization of staffing and roles, the survey asked who was responsible for fixity checking; 74 organizations responded and 160 options were selected in response to this question. The results show that system administrators are most often responsible for fixity practices with a 21.25% (34 respondents) response. Digital preservation managers and digital archivists follow with 16.25% (26 respondents) and 13.13% (21 respondents). However, overall the responses show that the responsibilities for fixity checking are spread throughout organizations, and include the roles of administrators, various IT positions, curators, and others.

Comments provided under the "other" option indicated that individuals within the organization often serve in more than one of the listed capacities, or respondents were not sure which categories most closely matched the positions in their institution. For example, three listed "archivist," but did not want to include themselves in the provided category of digital archivist. In situations where fixity is done by third party services, the role responsible for fixity checking also fell to the third party itself.

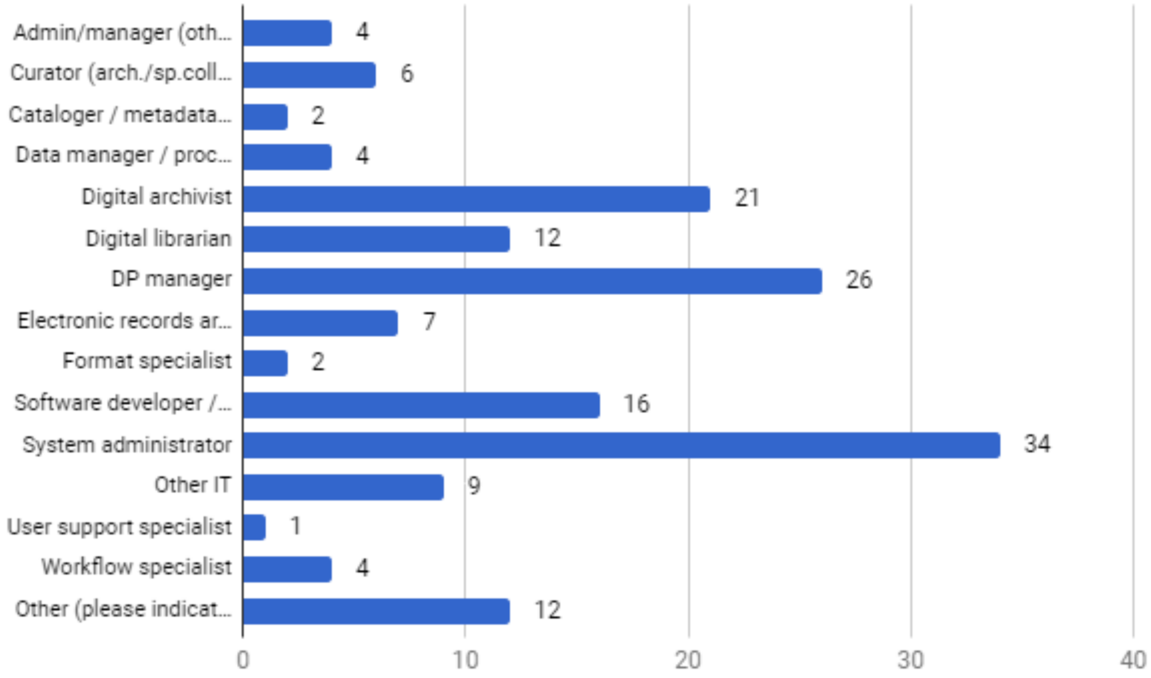


FIGURE 15: Responses to "Who is responsible for fixity checking (e.g., running manual scans, scheduling automated scanning, etc.) within your organization?"

Question 18: Where are the preservation copies stored, upon which the fixity checking occurs? *Select all that apply:*

To better understand the environment in which fixty activities are taking place, this question asked about the storage location of the files upon which fixity checking was done; 74 respondents made 117 selections in response to this question. The most common location on which fixity is being run against preservation copies was on in-house online storage at 42.74% (50 respondents), followed by offsite storage (including cloud vendors) at 29.06% (34 respondents), and then on non-networked in-house materials at 22.22% (26 respondents).

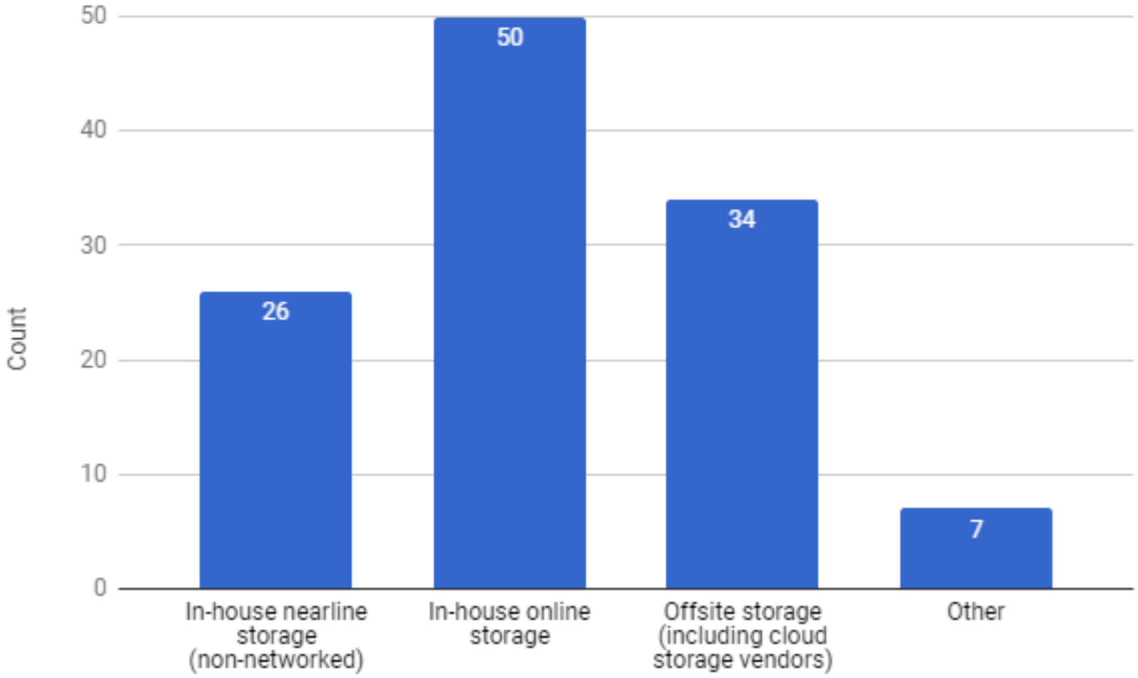


FIGURE 16: Location in which preservation copies are stored and fixity checking occurs

Other responses included LTO tapes of various versions (three respondents), and one each for external hard drives, contracted server farm storage, offline storage in-house, and offsite online storage.

Question 19: Where does your organization record fixity information? Select all that apply:

Seventy-four respondents made 134 selections when answering this question. Recording fixity information in databases or logs was the most common response with 54 respondents (40.30%) taking this action. Storing fixity as part of the metadata record was selected by 39 respondents (29.10%) and storing the information alongside the content was selected by 32 respondents (23.88%). Recording the fixity within the file itself was only selected by nine respondents (6.72%).

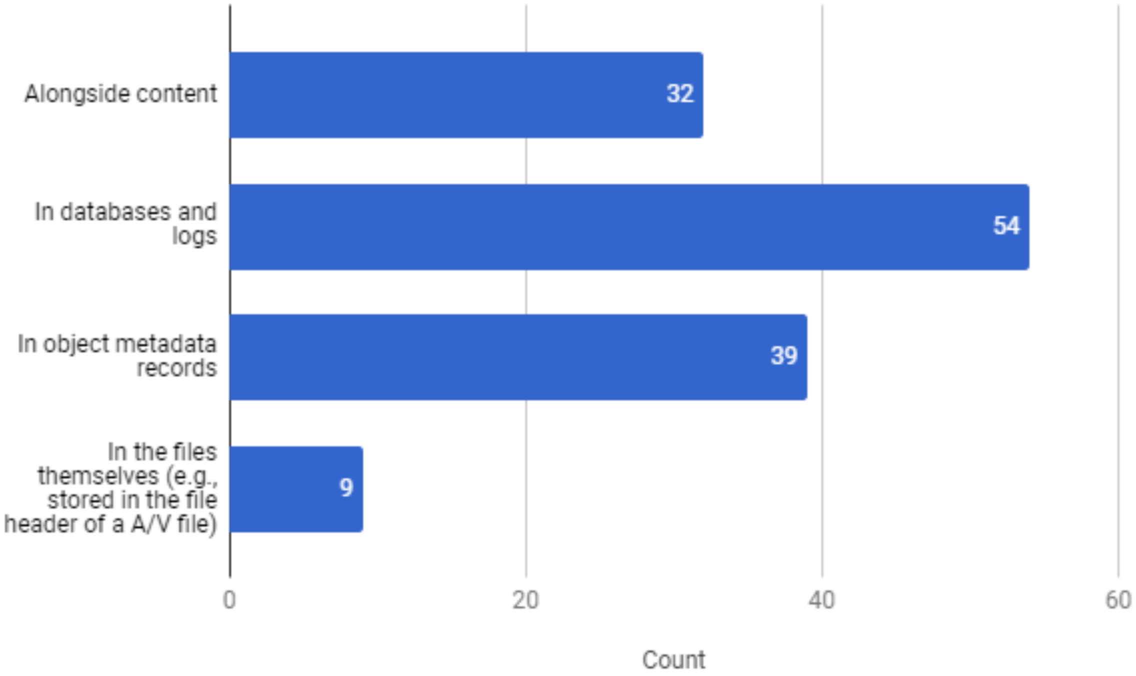


FIGURE 17: Location(s) in which respondents store fixity information

Question 20: What level of granularity do you utilize when running checksums? *Select all that apply:*

When asked about the level of granularity of used when running checksums, 94 selections were made by 72 respondents. The majority of respondents, 70.21% (66 respondents), are using fixity values created for each file, while 21.28% (20 respondents) utilize a single fixity value for a group of files—at the block level, folder level, bag level, etc. Another 8.51% (eight respondents) are using partial-file fixity information, which is to say they are creating multiple fixity values per file.

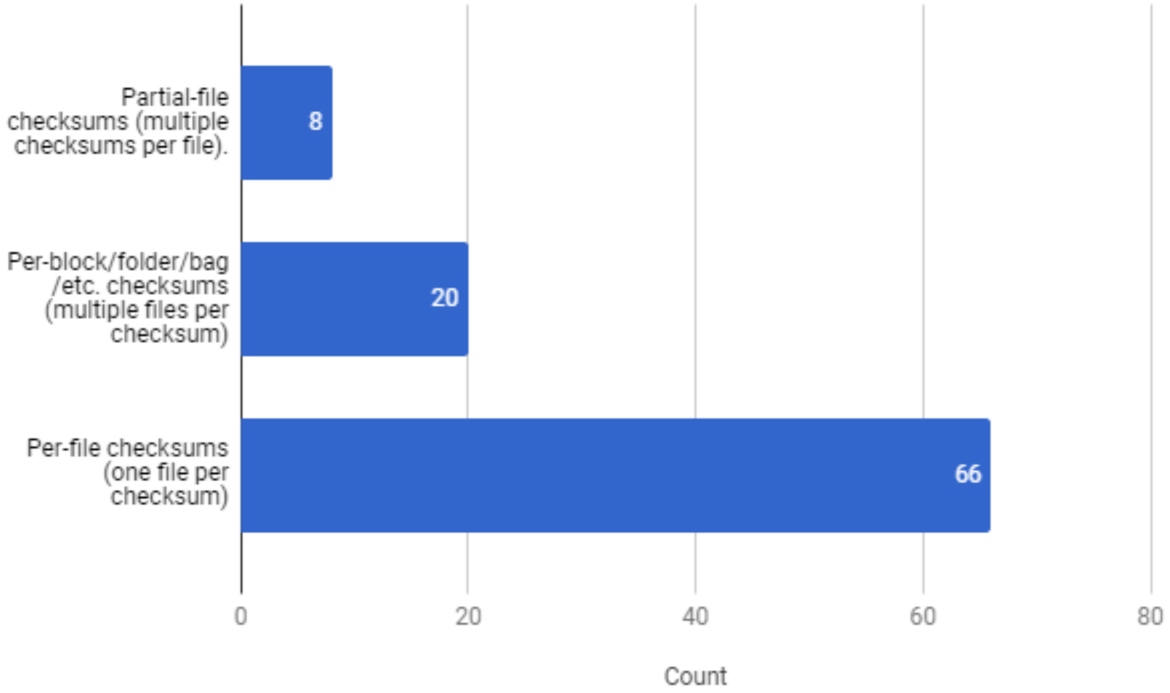


FIGURE 18: The level of granularity utilized when running checksums by respondents

Question 21: If you run partial-file checksums (multiple checksums per file), what is your use case?

Twelve organizations responded to this question about calculating partial-file fixity values, however only eight indicated that they used partial-file fixity values in Question 20. Reviewing the comments provided by these twelve respondents, it is clear not all institutions are truly creating fixity information for parts of files.

The use cases for the eight respondents (as indicated in Question 20) that do create multiple fixity values per file were all related to audio or video files. One respondent said:

> *"I don't consider that the practice of generating one checksum per file is fair for audiovisual archives, as checksums per file don't scale well. With one checksum per file, an av archive might have one checksum for a vast amount of data whereas an archive of smaller files would have more granular fixity. Since one of our use-cases is to use the hash to revert any damage this becomes exponentially less possible when the checksum is documenting a larger data size. Thus we use fixity features in Matroska and FFV1 for more granular levels of fixity."*

The additional four respondents who are not actually calculating partial-file fixity values indicated:

- That this questions was not applicable.
- That they run multiple checksums (different algorithms) for the entire file.
- That they calculate individual checksum values for individual files as well as calculate a fixity value for a group or package of files.
- That fixity information is stored in disk images.

Section 3: Cloud Services and Fixity Practices

This section of the survey was included to better understand if and how respondents were using cloud services, and how that affects their fixity practices.

Question 23: Are you using cloud services that offer fixity services?

Out of 74 responses, only 23 respondents (31.1%) are using cloud services that offer fixity services while over ⅔ of respondents are not (51 respondents, 68.9%). The chart below shows how utilization of cloud-based fixity services is spread out based on amount of content held by the institution; most of the respondents that use cloud fixity services hold between 1 - 50 TB of content.

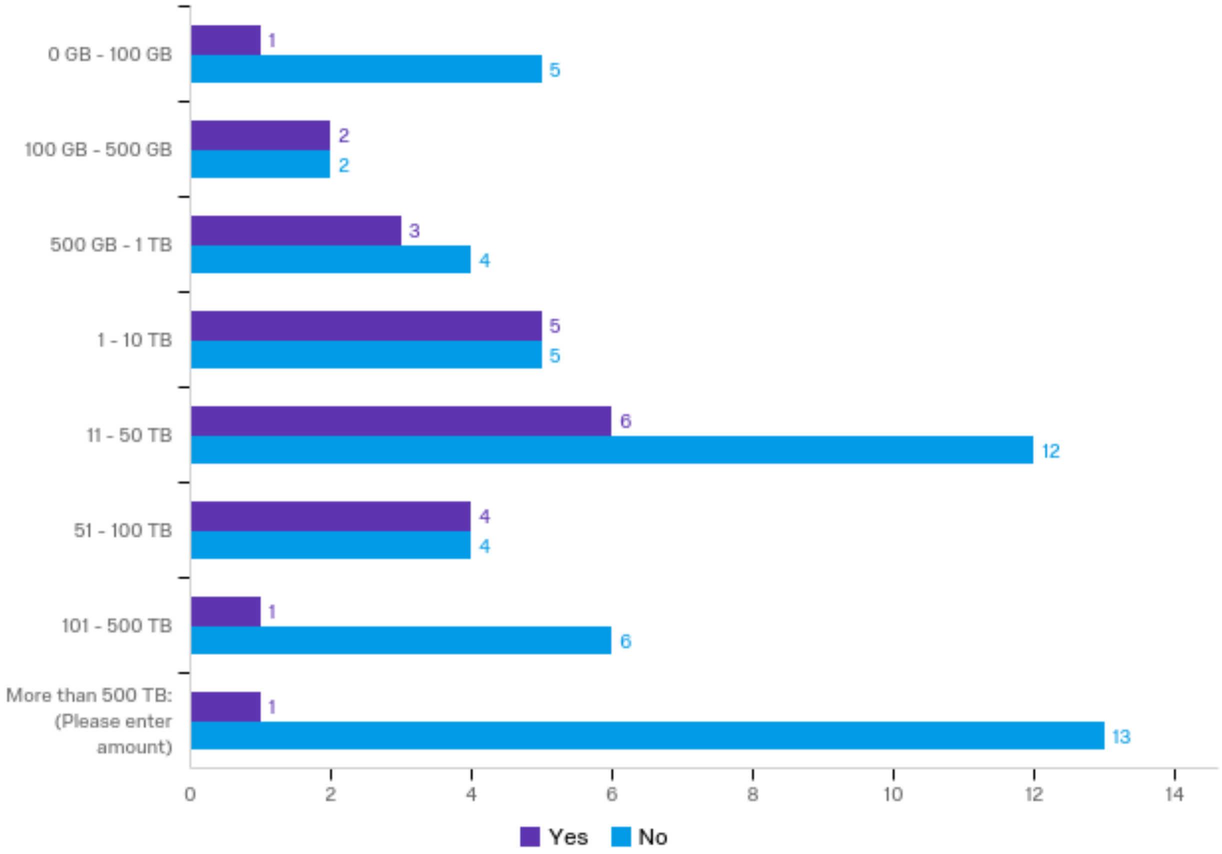


FIGURE 19: Responses to the question “Are you using cloud services that offer fixity services?” displayed by amount of content being managed

The questions that follow in this section were only asked of the 23 respondents who were using cloud services that offer fixity services (i.e., those that answered “yes” to Question 23).

Question 24: Do you have the ability to run your own fixity services on the vendor services?

Of the 23 respondents, a little over half (13 respondents, or 56.5%) indicated they were able to run their own fixity operations on the vended cloud services, while 43.5% (10 respondents) indicated that they were not able to run their own fixity operations on vended cloud services.

Question 25: Do you receive fixity information from vendors that you may use as you see fit?

Nineteen respondents (82.6%) receive fixity information from vendors and are able to use the information as they see fit. Four respondents (4.4%) either do not receive fixity information vendors, or they do and are not able to use the information as they see fit.

Question 26: Do you use the fixity information the vendors are providing?

Of the 19 respondents that receive fixity information provided by cloud vendors, 13 of them (68.4%) indicated that they were able to make use of the fixity information provided to them. While there were not enough free-text responses to the "no, if not - why not" question to see trends emerge, some of the reasons given by respondents who were not using the fixity information provided by vendors included limited staff availability, varying levels of readiness to use fixity information, and preference for alternate checksum options not provided by the cloud vendors.

> *"At this stage we are about to transition to a different cloud (and object) storage option. Although fixity information will be accessible, at this stage it will require redevelopment of our fixity check mechanism, which is on our roadmap."*
>
> *"Unusable in the way they store files. Also, they use MD5, and we prefer to check fixity with SHA-1, as it is stronger."*
>
> *"Don't have available staff time at present to make best use of the information."*

Question 27: Do you record fixity information provided by vendors?

As to whether institutions record the fixity information, 31.6% (six of the 23 respondents) indicated that fixity information is recorded in a software-based management system, such as a collections management or digital asset management system, 26.3% (five respondents) record fixity information outside of a formal management system, and 42.1% (eight respondents) did not record fixity information provided by cloud vendors.

Question 28: Please provide any other information or requirements around using fixity information in conjunction with vendor services, in as much detail as possible.

The final question in the section about cloud services asked people to provide any additional information about the services they used. Of the 23 respondents that utilize cloud services, only 12 respondents answered the open-ended question. Vendor services mentioned include Amazon (three responses), Preservica (three responses), Educopia / MetaArchive (one response), AP Trust (one response), and Openstack Swift (one response).

Section 4: Demographic Information

This section captured basic demographic information of respondents.

Question 29: Types of Organizations

Of the 88 responses to this question, 47.7% of respondents classified themselves in academia (including university entities as well as library/archives). Government institutions (both government entities and national/federal/legal deposit libraries) made up 20.5% of the responding organizations. Another grouping of respondents made up 18.2% and represented other library/archive/museum organizations (including museums, historical societies, public libraries, and independent library/archives). Smaller groupings break down as institutional and research data repositories (2.3%), other kinds of nonprofits (10.2%), and for-profit corporations (1.1%). The chart below shows a more detailed breakdown based on organization types.

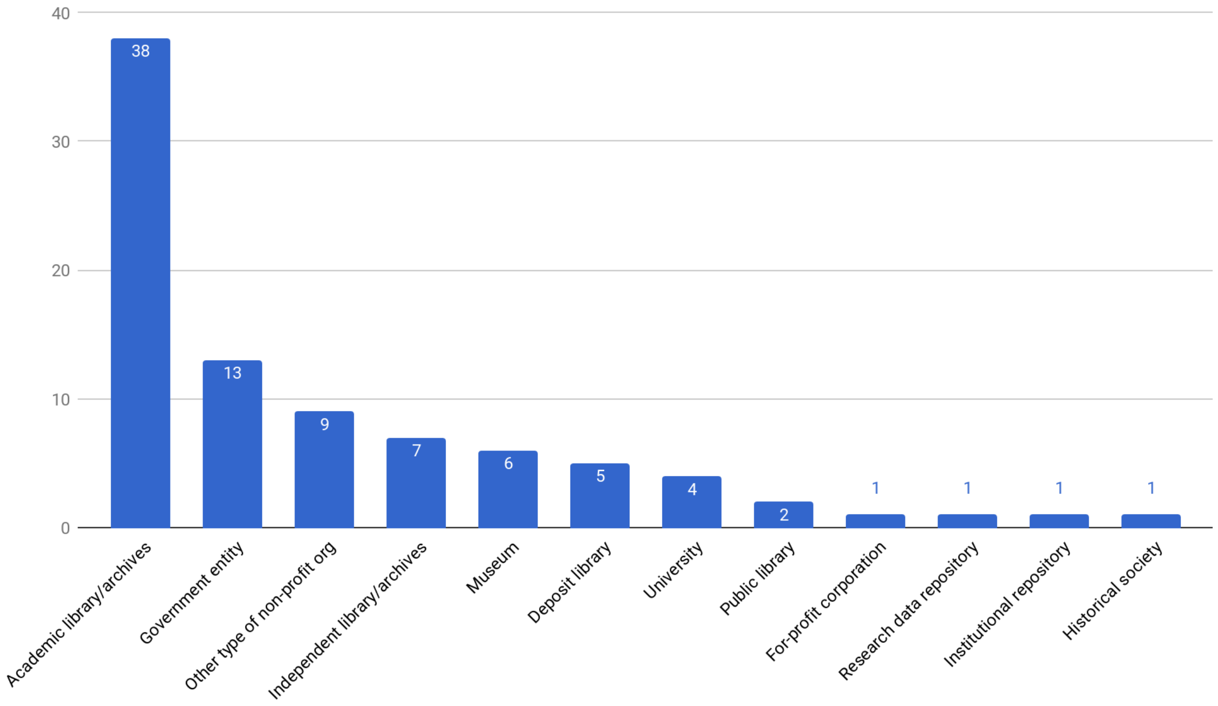


FIGURE 20: Type of organizations represented by respondents

Academia was heavily represented in our respondent group, particularly libraries/archives within academic institutions. Readers working in underrepresented institutions (for-profit corporations and data repositories, for instance) may want to look at specific responses to get a better idea of how institutions similar to theirs operate.

Question 31: Is there anything else you would like to tell us about your practices around fixity?

The 24 comments provided for this question covered several topics and were difficult to categorize, but an overall theme of complexity emerged. Many respondents mentioned that differing policies exist for different storage environments within the same institution, and that workflows and policies were emerging/evolving over time:

> *“Not all of our digital content is managed with the same workflows or in the same system. Some of the content is highly managed and other content not so much.”*

> *“Our practices around fixity are evolving and being established.”*

Several respondents expressed a corresponding desire for policies, tools, and other resources that can handle varied and complex environments, particularly automation:

> *“We cannot store our digital content on any sort of networked storage, therefore eliminating any sort of automated option.”*

> *“It's a bit difficult to tie together the scattered software utilities with other systems and infrastructure into an efficient workflow with very limited staff.”*

> *“We are interested in fixity checking solutions, efficiencies, and checksumming methods that can improve our systems.”*

ANALYSIS

The fixity survey was designed to uncover de facto or common practices for fixity checking and determine if there are any common challenges institutions face when implementing a fixity check routine. This section will discuss collecting and gathering fixity information, fixity checking intervals, the mechanics of fixity checking, fixity checking in cloud storage, and other common challenges.

Collecting and Gathering Fixity Information

Even with the common understanding that “the earlier you have a fixity value the better,” collecting fixity information at the point of acquisition was not a consistent practice. Many respondents reported that they do not receive fixity values from record creators, and even if they required them, they would not expect high compliance.

> *“While we always collect fixity information if it is provided as part of the acquisition package, in practice this information is rarely provided by contributors.”*

> *“Most donors lack the technical knowledge, time, and/or inclination to go to the lengths of creating checksums prior to transfer. The only donors who do create checksums are administrative units that transfer vital record series to our university archives.”*

> *“Would like to do it always, but many people delivering files to us are still not aware of fixity at all. Unfortunately, some are still not even willing to be bothered about it.“*

> *“We do it whenever that's possible. My opinion/what I think is the opinion of the community is that capturing fixity information as early in the process as possible is the first, most basic, and arguably most important step in being able to manage its preservation over time. Collecting fixity information created prior to archival acquisition is of course ideal, but challenging - requires the sort of education and outreach to creators/managers of data that we aspire to but have not yet achieved.”*

This puts the burden on the organizations to create the initial fixity value, which the majority of survey respondents do. They create fixity information upon acquisition in order to ensure file integrity, authenticity, and trustworthiness.

Data transfer was recognized as another key point at which to check fixity. The number of affirmative answers to checking fixity after data transfer suggested that the data transfer process was recognized by most as a risky process warranting the extra caution of fixity verification.

Fixity Checking Intervals

Respondents who reported having a greater amount of total content managed for long-term preservation (> 50 TB) were more likely to report that fixity checking was done on a rolling or continuous basis with interval loops dependent on the size of the preservation corpus. Respondents also overwhelmingly ran fixity checks on the entire amount of their content rather than a sampling of it, indicating that the practice of sampling is not sufficient enough to achieve the assurances that fixity checking provides.

Determining an appropriate fixity checking interval depends on many factors. Many organizations stated in survey comments that how often content was checked depended on a variety of factors:

> *"Checking frequencies vary based on storage location, changes in storage management and storage devices"*

> *"We maintain several systems, each with different fixity strategies and frequency. Some systems perform an ongoing storage check, while others have annual sampling."*

> *"Factors influencing the fixity check interval include: total size of the corpus; total number of collections in the corpus; total size of the collection; time since the last successful check of the collection; time since the last known change to the collection; availability of peer-to-peer partners to conduct the distributed check; availability of computational resources to conduct the check; relative priority of the check compared to other checks and other tasks requiring computational resources; relative priority assigned to the collection compared to other collections in the corpus; randomness as a mitigation of threats posed by malicious peers."*

> *"In answering the survey I found it hard sometimes to choose an answer since not all of our digital content is managed with the same workflows or in the same system. Some of the content is highly managed and other content not so much. I tried to strike a balance in my responses."*

The Mechanics of Fixity Checking

No respondents reported relying solely on hardware-based fixity checking, indicating that software-based fixity checking is preferred. There did not appear to be any clear preference as to a particular type of software or fixity checking method (automated/scheduled software, scripts/custom codes, or manually run software). Almost

43% of respondents reported using the MD5 algorithm to calculate fixity, however SHA1 and SHA256 were also popular algorithms.

Likewise, a similar number of respondents (40%) reported that they recorded their fixity information "in databases and logs. "In object metadata" (30%) and "alongside content" (24%) were also popular answers for how fixity information was stored. Checksums are run at the file level rather than checking one fixity value for multiple files (i.e., a bag, block, folder, etc.) or checking multiple fixity values for one file (as may be done in managing A/V files) by 66% of the respondents. The preference for checking fixity at the file level was consistent regardless of the size of the collection managed. In fact, 65% of respondents who manage over 500 TB of data reported running individual file-level fixity checks.

The method in which respondents are creating fixity information and performing fixity checks is often affected by technical challenges an organization may face. Many of the challenges that were documented in the survey centered around software—choosing an appropriate tool within budget, making the software work in the institution's environment, and determining how tools can fit systematically into workflows—as shown in the quotes below.

> *"We are trying to find a way to do this, and also plan to start soon, but we are basically without the right equipment and tools to do so. We need to try to get the right equipment allocated to us (which comes down to funds available), and also research the right tools, although we have been playing with the AV Preserve tool Fixity."*

> *"It's a bit difficult to tie together the scattered software utilities with other systems and infrastructure into an efficient workflow with very limited staff. If it's easy to do, it gets done. So far there hasn't been time to put a process in place that MAKES it easy to do regardless, so then the ease of use depends upon ready-to-hand tools."*

> *"How we are doing fixity checks is affected by the fact that (1) we cannot store our digital content on any sort of networked storage, therefore eliminating any sort of automated option; (2) the fixity software cannot be set to send me emails if a change is found because IT would not provide the information on the email server, thus requiring that I look at each report to find if a change is indicated."*

> *"We would always collect it as a matter of protocol if workflows and software tools were more in place in a systematic way. This is something that needs to be addressed."*

Fixity Checking in Cloud Storage

The majority of survey respondents (51 of 74 responses, or almost 70%) are not using cloud vendors that offer fixity services. Answers on the questions pertaining to fixity in cloud storage reflect only 30% of the total survey respondents.

Of the relatively small number of survey participants using cloud services that offer fixity services, a little over half (56.5%) are able to run their own fixity operations on cloud services. The majority (82.6%) receive fixity information from the service providers. It is interesting that four respondents appeared to neither run their own fixity operations nor receive fixity information from their cloud storage service providers. Approximately 70% of respondents used the fixity information they received in some fashion. Some respondents who do not use fixity information they receive indicated that staff time is a challenge in implementing a process to do so.

Other Common Challenges

Other themes that came across as common challenges for respondents include lack of time, lack of policy or program development, lack of support, staffing, or understanding of the value of fixity practices, and general technical challenges.

For some respondents who answered "no" to performing fixity checks, they listed lack of available time as a reason they were not actively utilizing fixity practices.

> *"Not yet, but we hope to implement fixity checking in the next year."*

> *"I've downloaded a program called Fixity, but haven't had time to learn how to use it."*

The lack of written policy and programs in development were other challenges respondents faced.

> *“Our institutional practices/program have not yet been developed to the point of needing to have fixity checked. We are working on developing a digital preservation program, but we do not yet have one in place.”*

> *“We are just in the experimental stage and have no official policy yet on which content needs fixity information.”*

Others are just beginning to understand the purpose and value of fixity, and need support in educating themselves and others on its importance.

> *“I would like to attend a crash course in fixity. Perhaps NDSA could create one.”*

> *“The biggest issue we have with fixity is honestly justifying and explaining it to our users, which are our staff members. Moving to better fixity practices would require a cultural shift in the organization, in my opinion.”*

> *“Lack of support from IT department”*

Technical challenges were another issue for respondents.

> *“No bandwidth—although we do take advantage of our cloud storage service's data integrity checks, we don't do any on our own”*

> *“We would always collect it as a matter of protocol if workflows and software tools were more in place in a systematic way. This is something that needs to be addressed.”*

> *“We have witnessed that the larger the file size of individual files, the more often data errors are encountered. Mainly when "moving" data (network, carrier-to-carrier, etc). Since the majority of our files are audiovisual files, the file sizes we are dealing with are rather large.”*

Even with the many challenges facing respondents, it is clear that the organizations are trying their best with what resources (time, financial, staff skills, etc.) they have. Fixity practices at some institutions are still being developed and others are well-polished. Those who are still developing policies and procedures are hungry for more information and guidance.

> *"I will be really interested in your survey results. I think we need more guidance regarding at what point fixity should be done in the workflow, at what level in the technology stack, and which checksums are more reliable indicators than others."*

As a first step, it is hoped that this survey will help organizations understand what others are doing. Future case studies with survey respondents should provide a more detailed view of individual practices.

CONCLUSION

Fixity is a key concept for the long-term preservation and management of digital material for many reasons. Previous scholarship on fixity has shown its vital importance in discovering changes to data and all that this error-checking can imply: authenticity and renderability of files, trustworthiness of institutions, and system monitoring/maintenance. Despite the centrality of fixity to the field of digital preservation, there is little prescriptive guidance on when and how to create fixity information, where to store it, and how often to check it. This absence is not without reason, however: the incredible variety of organizational structures, priorities, staffing levels, funding, resources, and size of collections held by institutions that do digital preservation make it difficult to establish a single set of one-size-fits-all best practices.

The goal of this survey was to learn about the implementation and adoption of fixity practices across the GLAM community; to attempt identification of de facto best practices; and to identify challenges institutions face in creating, managing, and utilizing fixity information. A great deal of variety in the practices of institutions was found, but general themes emerged: collecting or creating file-level fixity values as early as possible, storing them alongside files or within logs rather than within the file itself, and verifying fixity information at an interval that balances needs, collection, size, resources, and other factors. Respondents also talked frequently about improvements they wanted to make, such as looking for new tools, developing workflows, and making other changes to their fixity activities. It may be useful, then, for practitioners to think of their fixity practices within a maturity or continuous improvement model; digital preservation practitioners develop a set of ideal practices for their institution and continuously evaluate methods and the allocation of available resources to get as close as possible to their ideal state.

NEXT STEPS

Future work, such as follow-up surveys and case studies, would ideally gather information about reasoning and logic behind fixity practices; such information could be helpful for practitioners who seek to marry their ideal fixity practices with what they are able to do in reality. It may also be helpful to further investigate challenges in instituting a fixity routine or to determine the decision-making process of practitioners who create fixity information for some collections but not others.

Other interested parties have suggested that more information be gathered on how often fixity checks fail and what is done in those cases. In addition, digital preservation practitioners often desire or require information about vendors' fixity practices; further work to survey vendors to learn more about how they create, store, and verify fixity information for customers would be a valuable resource for practitioners who are seeking to choose service providers.

The results of this survey revealed areas where future work within the field might benefit practitioners. Tools to educate donors and vendors about checksums and to help donors create them easily would benefit many institutions who struggle to receive fixity information as close to the point of creation as possible. Similarly, improved interoperability and portability of fixity information between tools would prevent the duplication of work for many respondents. Work to create a maturity model with good/better/best practices, something similar to or something to compliment the File Fixity and Data Integrity section of the NDSA "Levels of Digital Preservation," would also benefit digital preservation practitioners as they seek to start or improve their fixity practices.

APPENDIX 1: INTERPRETING SURVEY RESPONSES FOR THIS REPORT

This appendix provides information about how the survey data was prepared and analyzed for use in this report. Information is provided by section and specific question as necessary.

Section 1

The questions in Section 1 included many free text fields. When possible, information from the free text fields was analyzed and reviewed to look for common themes. These themes were grouped into categories for further analysis as described below.

- Question 1: Looked at the comments in the “no” answers to look for common reasons respondents were not currently utilizing fixity practices. Common themes of program immaturity, time/labor, tools, knowledge/training, and desire/intent to being using fixity soon were chosen based on the content of the comments. Every point of each comment was reflected in one or more of these categories; this was done to see if there were common reasons for a total lack of fixity checking.
- Question 3: Grouped the free-text responses into the following categories: checksums generated at ingest, checksums provided by vendors, checksums provided by donors, checksums not provided, authenticity, integrity, transfer, and de-duplication. Each free-text response was reflected in one or more of these categories. This was done to try to reveal common logistical/practical and technological concerns that might impact if/how respondents collected fixity info from donors, vendors, etc.
- Question 5: Grouped the free-text responses into the following categories: checked at acquisition, checked at ingest, checked ASAP, checked after ingest, storage, integrity, trustworthiness, transfer, and de-duplication. Each free-text response was reflected in one or more of these categories.
- Question 6: Used Qualtrics reporting functionality to create a list of all responses. There were eight “other” responses to this question. Three of these talked about using checksums to handle multiple versions or copies of files, and two talked about using them for inventory control; no other clear categories were apparent.

Section 2

Working with questions 15-21, the number of total responses for each question was calculated first. For analysis purposes, if the question allowed for multiple responses, the data in each response was separated into individual columns and then moved back into a single column for ease of grouping and creating accurate graphs. The following describes how the text fields were coded to make responses more manageable if necessary.

- In general, “other” free text fields were reviewed then categorized depending on what common answers emerged.
- Question 15: No real consistency with the 12 responses/5 other responses. No need to group responses.
- Question 16: Six other responses. Five SHA versions, One Tokens. No need to group responses.
- Question 17: Twelve other responses. Most are combinations of the options in the chart. No need to group responses.
- Question 18: Seven other responses. Three were the same (tape), the others were all unique. No need to group responses.
- Question 21. All text answers. Grouped answers into categories of Audio/Video, Disk Images, Packages, and NA for evaluation.

Section 3

- Question 28. All text answers. Grouped by named vendor services.

Section 4

- Question 29 (institution types): The seven “other” responses were coded to the most appropriate institution type (provided as options) based on the free-form text respondents provided. Two became “academic library or archives,” two became “government entity,” one became “independent library or archives,” and two became “non-profit organization (not one of the above types).” These were then added into the final groupings for this question in this report.

APPENDIX 2: SURVEY QUESTIONS

This appendix lists the survey text and answer options as originally provided.

Welcome to the NDSA Survey on Fixity Practices!

The goal of this survey is to identify gaps between fixity best practices (as identified in the NDSA Fixity Guidance Report [http://ndsa.org/documents/NDSA-Fixity-Guidance-Report-final100214.pdf]) and real-world implementation, as well as to identify possible reasons that institutions do not meet specific best practices. Participation in this survey is voluntary, and open to employees/volunteers/interns in institutions that manage digital content for long-term preservation.

The responses from this survey will be anonymous—any information explicitly tying survey responses to an individual or institution will be stripped. The aggregate survey data will be shared publicly, and additional case studies that highlight practice at specific institutions may be published in partnership with individuals/institutions that indicate they are willing to participate.

This survey is organized into 4 thematic sections: the basics of how your institution gathers/uses fixity information, details specific to your institutional policies and workflows, questions about cloud services, and a final section asking for details about your institution. You will find further instructions in each section to guide you. Answer each question to the best of your ability, choosing the most appropriate answer available. It is impossible for any survey to fully capture the nuance of local practice, so you will see many free-text fields that allow you to include information not available in the other options. As a general rule, the questions in this survey pertain to preservation copies of files only, rather than access copies or other manifestations. Similarly, the survey questions pertain to digital content that is managed for long-term preservation rather than working files or otherwise unmanaged content.

If you have questions or concerns about this survey, please contact the NDSA Fixity Working Group at NDSA-FIXITY@lists.clir.org.

[Section 1: The Basics; this section addresses the basics of if and why your institution uses fixity information.]

Question 1 Do your organizational practices include utilizing fixity information at any point in time?

- Yes
- No (if no, why not?)________________ (2) *[Condition: No (if no, why not?) Is Selected: Skip to Section 4: About your Institution]*

Question 2 Does your organization collect fixity information (created by another institution or separate entity within your organization) along with digital content at the time of acquisition if it is available?

- No, never
- Yes, very rarely
- Yes, sometimes (25-5-% of the time)
- Yes, frequently (>50% of the time)
- Yes, always

Question 3 Please provide any relevant details about why you collect fixity information as frequently as you do. ____

Question 4 Does your organization create fixity checks for digital content if they are not provided at the time of acquisition? Please indicate how often you collect fixity information:

- No, never
- Yes, very rarely
- Yes, sometimes (25-5-% of the time)
- Yes, frequently (>50% of the time)
- Yes, always

Question 5 Please provide any relevant details about why you create fixity information as frequently as you do. ____

Question 6 [Matrix] What are the reasons your organization collects, checks, maintains, and verifies fixity information? Please rate the importance of each of these items (not important, somewhat important, moderately important, extremely important):

- Determine if the data has been corrupted or altered over time
- Determine if the data has been corrupted or altered during transmission
- To support the authenticity or trustworthiness of the digital objects
- To monitor hardware degradation

- For authenticity: To prove you are providing the digital object that has been requested
- To permit an update to a portion of a content file while proving the other portions remain unchanged (ex: split video files)
- Meet requirements or best practice guidelines
- Help identify systemic or human error in the management of digital content
- Other

[Section 2: Where, When, and How; this section helps to communicate when, where and how fixity is being used in your institution.]

Question 7 Do you check fixity information after transferring data?

- Yes
- No
- Sometimes

Question 8 Do you check fixity at regular intervals - please specify the intervals that your organization uses. Select all that apply:

- Hourly
- Daily
- Monthly
- Quarterly
- Every Six Months
- Yearly
- Every Two Years
- Other: ____________________

Question 9 Please provide any relevant details about how often you check fixity (e.g., differing frequencies based on storage location). _______________________________________

Question 10 How much total content (preservation copies that are managed for long-term preservation only) are you running fixity on? Please provide your answer in total number of TB:

- 0 GB - 100 GB
- 100 GB - 500 GB
- 500 GB - 1 Tb
- 1 - 10 TB

- 11 - 50 TB
- 51 - 100 TB
- 101 - 500 TB (6)
- More than 500 TB: (Please enter amount) ____________________

Question 11 [Matrix] What factors does your organization consider when determining fixity check frequency? Please rate the importance of each of these items (not important, somewhat important, moderately important, extremely important):

- Concern of media failure due to increased use of storage media (e.g. tape)
- Storage media reaching end of expected lifespan
- Throughput limitations (e.g. network bandwidth)
- Number and size of file or objects that require fixity checks
- Reliance on checksums generated by storage providers (i.e. cloud providers and others)
- Regular checks done at the block level via a system
- Available staff time/resources

Question 12 Do you check fixity information at regular intervals of a sampling of your digital content?

- Yes (If so, what is the sample size in %) ____________
- No

Question 13 Do you check fixity information at regular intervals of all your preserved digital content?

- Yes
- No

Question 14 Is your fixity checking done utilizing built in hardware or is it software-based?

- Hardware (part of file system)
- Software
- Both hardware and software

[Question 15 shown if in Question 14 if 'Software' or 'Both hardware and software' Is Selected.]
Question 15 What software, tools, or services are you using to capture/verify fixity information? Select all that apply:

- Scripts / custom code

- Automated/scheduled software
- Manually run software
- Third-party services (if yes, please provide details): ________________
- Other: ________________

[Question 16 shown if in Question 14 if 'Software' or 'Both hardware and software' Is Selected.]
Question 16 What type of fixity checking algorithm does your preservation software use? Select all that apply:

- CRC
- MD5
- SHA1
- SHA256
- Other: _____________

Question 17 Who is responsible for fixity checking (e.g., running manual scans, scheduling automated scanning, etc.). Select all that apply:

- Digital preservation manager
- System administrator
- Software developer / programmer
- Other IT
- User support specialist
- Collection needs analyst
- Policy analyst
- Content analyst / maintainer
- Data manager / processor
- Cataloger / metadata analyst
- Format specialist
- Workflow specialist
- Electronic records archivist
- Archives & special collection curator
- Digital librarian
- Administrator / manager (other than digital preservation manager)
- Outreach specialist / trainer
- Rights specialist
- Usability specialist
- Digital archivist

- Other (please indicate): ______________

Question 18 Where are the preservation copies stored, upon which the fixity checking occurs? Select all that apply:

- In-house online storage
- In-house nearline storage (non-networked)
- Offsite storage (including cloud storage vendors)
- Other: __________

Question 19 Where does your organization record fixity information? Select all that apply:
In object metadata records

- In databases and logs
- Alongside content
- In the files themselves (e.g., stored in the file header of an A/V file)

Question 20 What level of granularity do you utilize when running checksums? Select all that apply:

- Per-block/folder/bag/etc. Checksums (multiple files per checksum)
- Per-file checksums (one file per checksum)
- Partial-file checksums (multiple checksums per file)

Question 21 If you run partial-file checksums (multiple checksums per file), what is your use case? ___________________________________

[no Question 22]

[Section 3: Cloud Services; this section addresses fixity issues specific to using cloud services.]

Question 23 Are you using cloud services that offer fixity services?

- Yes
- No *[Condition: No (if no, why not?) Is Selected: Skip to Section 4: About your Institution]*

Question 24 Do you have the ability to run your own fixity services on the vendor services?

- Yes
- No

Question 25 Do you receive fixity information from vendors that you may use as you see fit?

- Yes
- No

[Question 26 shown if in Question 25 if 'Yes' Is Selected.]
Question 26 Do you use the fixity information the vendors are providing?

- Yes
- No (if not, why not?)___________________

[Question 27 shown if in Question 25 if 'Yes' Is Selected.]
Question 27 Do you record fixity information provided by the vendors?

- Yes, in a software-based management system (such as a collections management or digital asset management system)
- Yes, outside of a formal management system
- No

Question 28 Please provide any other information or requirements around using fixity information in conjunction with vendor services, in as much details as possible.

____

[Section 4: About your Institution; this section provides us with basic demographic information about your institution.]

Question 29 Which of the following most closely describes the type or function of your organization?

- Academic institution department (not a library or archives)
- Academic library or archives
- For-profit corporation
- Historical society
- Institutional repository
- Independent library or archives
- Government entity
- K-12
- Museum
- National, federal or legal deposit library

- Non-profit organization (not one of the above types)
- Public library
- Research data repository
- Research group
- University
- Other (please specify): ____________________

Question 30 Would you be willing to have us work with you to document your fixity practices in the form of a use case? We would like to expand on the survey by providing selected real-life examples. These would be used in a final report about the survey and/or as individual blog posts.

- Yes (please provide your email address) ______________
- No

Question 31 Is there anything else you would like to tell us about your practices around fixity? ________________________________