Proposal for an Archival Data Storage Comparison Model

Ryan Downie

Morgridge College of Education, University of Denver

RMS 4900: Educational Research and Measurement

Professor Acadia

March 4, 2022

Abstract

This proposal addresses the question of creating a model to illustrate the differences of on-site data storage and cloud storage for archival data in a way presentable to archival professionals. Cloud storage has proven itself to be less understood than traditional, physical data storage, and more difficult to rely upon for long-term preservation needs due to the lack of documented experiences of it. Yet it simultaneously has shown to have ample opportunities for enabling long-term preservation of digital materials and present lower operating costs for archives. There is no black and white dichotomy in the question of physical versus cloud storage, and the unique scenario of every archive means that what may be ideal for one institution will be detrimental to another. Therefore, this proposal seeks to nominate the creation of a model for the SAA and archivists to use to address the differences of physical and cloud storage, with an emphasis on cloud services due to the greater misunderstandings and questions of them. Data for this proposed model is to be collected via surveys and interviews in a mixed methods approach following a stratified random probability sample. Addressed topics include cost, security, long-term viability, and the variability of an institution's size, with the benefits and drawbacks of cloud and physical storage addressed in each from a neutral stance. Furthermore, it is not among the goals of this proposal to endorse either physical or cloud storage, or any one product or service provider. Rather, a neutral view is sought to provide objective results that aim to assist institutions and archivists in their choices.

Topic Under Investigation

Effective storage of archival data has always been among the primary challenges for archives and record-keepers to maximize the lifespan of documents before they degrade beyond usability. Regardless of the protective precautions and procedures provided to physical materials, however, they will inevitably pass a point beyond usability; and, though digital storage has provided an alternative in recent decades that allows for the potential of extending items' usable life spans into hundreds of years or longer, digital objects face many of the same challenges, if not more, to their longevity (Askhoj et al., 2011, p. 175; del Pozo et al., 2010, pp. 291-292; Pandey & Kumar, 2020, p. 28). Consequently, in recent years, cloud storage has increasingly become commonplace among archives as a means to digitally store records at a lower cost while minimizing many of the risks facing records. However, there still is professional uncertainty, novitiate assumptions, and strong discourse on the matters of who, what, when, where, why, and how cloud storage should be used. For this reason an official, endorsed model should be adopted by the Society of American Archivists (SAA) for the benefit of archives and record keepers.

Cloud computing, another term for cloud storage, which is also known as cloud services, is defined by the U.S. National Institute of Standards and Technology (NIST) (Mell & Grance, 2011) as:

> A model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. (p. 2)

An additional, slightly broader definition posited by Askhoj et al. (2011) defines cloud computing as an abstract, scalable service platform in a layered form, accessible remotely via the internet on a pay-per-use basis through an agreement with the service provider (p. 176).

The use of these services for record-keeping by archives has increased rapidly since their inception, currently approaching a near ubiquitousness within the field—and for good reason. One 2018 study on cloud storage usage concluded that the use of cloud storage was expected to grow 14.2% annually over five years, adding onto the more than 50% of the study's recipients who already were using cloud storage (McLeod & Gormly, 2018, pp. 174, 180). Reasons consistently listed for the adoption of cloud computing services by archivists are: security, scalability, back-up copies, maintenance, minimal personnel training, reliability, flexibility, protection against data loss, collaboration, access, disaster protection, and—most notably—reduced costs (Iglesias & Meesangnil, 2010, p. 8; McLeod & Gormly, 2018, pp. 175-181; Mosweu et al., pp. 4-5). These are all aspects of cloud storage that are consistently found advantageous over their corresponding equivalence in on-site digital storage through local hard-drives.

However, the decision over cloud services and on-site storage are not black and white, for although there are many benefits offered by cloud services, there are also numerous reasons against them. Long-term storage in the cloud, for example, is oftentimes found to be drastically more expensive than on-site storage (GFI, 2010, p. 7; McLeod & Gormly, 2018, pp. 175, 178, 186-187; Rosenthal & Vargas, 2012), with one study concluding that, generally, long-term storage of ten years or longer is more expensive using cloud services than on-site systems (Rosenthal & Vargas, 2012, p. 7). Other difficulties encountered with cloud computing include:

internet access, threats of hacking and data breaches, an over-reliance of third-party security, and a lack of interoperability and migration between providers (Mosweu et al., 2019, pp. 7-8).

Due to the wide range of opportunities presented by cloud services, it is likely that most archival institutions could benefit from using cloud computing within individually varying scopes and for varying purposes. For some institutions like smaller community archives, cloud services may be ideal as the primary repository thanks to their lower cost, increased user access, and storage scalability. Yet other institutions might see the most benefit from cloud services only in a short-term, migratory capacity, utilizing the cloud to temporarily hold records until on-site storage is upgraded or transferred.

Research Purpose Statement

The purpose of this research proposal is to put forward to the Society of American Archivists (SAA) the establishment of a standardized model for archives and record-keepers to assist in the decision-making process of adopting cloud services. This is neither intended to promote or detract any one cloud service provider or system against its competitors, nor to endorse any product, albeit data pertaining to specific cloud systems will be utilized. Rather, it aims to highlight that transparency and clarity are essential for institutions and archival professionals to make proper decisions on their methods of storage. Currently, there are few existing models to assist institutions and archivists in identifying the best choices in deciding between on-site and cloud storage, and those models that are available are infrequently used and fail to provide long-term projections of cost (McLeod & Gormly, 2018, p. 188). This absence of a useful aid necessitates a model that will allow institutions to determine whether investing in cloud storage is beneficial for their collection or not, and, if so, of what type, for how long, and where. For an archival institution to have the ability to reliably determine through an SAA

endorsed model if adopting a cloud storage system will be beneficial or detrimental to the institution's goals and its records' continuity is of great importance. Not only will it benefit the archive financially and in the security of the collection, but it can also aid in extending the lifespan of its archived records.

Philosophical Perspective

This research project follows a pragmatic philosophical perspective using mixed-methods. Mixed-methods was selected for this study in order to adequately cover both the financial considerations as well as the personal, real-world experiences and factors obtained by institutions and archival professionals, as neither facet can sufficiently address the topic alone (Bowen, 2018, p. 29). Therefore, no scientific hypothesis to prove or eliminate would be appropriate for this study. Instead, this study is advancing the understanding of applications and solutions for a real-world issue through the use of mixed-methods, a concept which squarely falls under a pragmatic approach (Creswell & Creswell, 2018, p. 10). Creswell and Creswell (2018) define pragmatism as a view that "arises out of actions, situations, and consequences rather than antecedent conditions" concerning "applications—what works—and solutions to problems" (p. 10), and which finds the truth to be whatever works at the time and situation (p. 11). This latter statement is critical to this study, as there is no one-shoe-fits-all scenario for archival storage decisions. Rather, what is appropriate for assisting in the decision making process between on-premise storage and cloud storage systems—and what this study seeks to promote the creation of—is, as put by McLeod and Gormly (2017), "the enhancement of checklists and other guidance documents" (p. 349).

Literature Review

The decision between on-site—otherwise known as on-premise—storage or remote cloud storage is an exceedingly prevalent issue among archives and archivists, and one which has no universal answer. Using cloud computing to store some of an institution's data has become the norm, while using exclusively cloud storage or using only on-premise storage are less common practices (McLeod & Gormly, 2017, p. 355; McLeod and Gormly, 2018, p. 180). Cloud services themselves are rapidly growing in use, with spending rising 17.5% between 2018 and 2019 from 182 billion to 214 billion (Shallcross, 2020, p. 458), and with an estimated five-year compound annual growth rate of 14.2% from 2018 to 2023, by which time it will have surpassed non-cloud infrastructure (McLeod & Gormly, 2018, p. 174). And yet, despite the strong, rapid growth of cloud services and the commonality of them, there is still a great deal of conflicting, unclear, and in some instances little-to-no information about them.

Cost, for example, is regularly cited as the primary factor for adopting cloud storage by institutions and individual archivists alike (Borglund, 2015, p. 115; GFI, 2010, p. 4; Iglesias & Meesangnil, 2010, p. 8; McLeod & Gormly, 2017, pp. 349, 355; McLeod & Gormly, 2018, pp. 180-181). However, some researchers have found that organizations do not accurately recognize the prices of cloud storage systems (McLeod & Gormly, 2017, p. 349; Turner, 2014), others find that "recent literature has brought the financial benefits into question" (McLeod & Gormly, 2017, p. 350), while others, still, plaudit the costs of cloud systems (Borglund, 2015, p. 115; McLeod & Gormly, 2017, p. 355). Such important ambiguity and contradiction should be highlighted for the benefit of all parties looking to potentially utilize cloud storage systems or, if possible, be addressed and clarified by future research.

This literature review will begin by explaining what makes a cloud storage system and the levels they are composed of, then proceed to examine the main factors that can play for or

against cloud services and service providers (CSPs), and conclude by discussing how long- and short-term storage goals can affect the decision-making process.

Structure of Cloud Services

Cloud storage is composed of three service models and four deployment models. Cloud users can have access to software and applications, platforms, or entire digital infrastructures and data centers from their providers, with most larger organizations having enterprise-level storage management strategies with specified drives or folders for employees to access that will then automatically backup for creating copies (Breeding, 2013, p. 25; Stancic et al., 2015, p. 214). By providing these services, cloud storage removes the software licensing costs, new infrastructure requirements, IT work, and often reduces service costs that are otherwise present with on-site storage systems, albeit at the cost of losing control of overall systems and data, potentially higher long-term costs, and entrusting security and possibly ownership to a third party (Askhoj et al., 2011, p. 175; Bowen, 2018, p. 27; GFI, 2010, p. 4; Shallcross, 2020, p. 458).

The three service models of cloud systems are: Software as a Service (SaaS), which allows clients to remotely access software hosted and managed by the service provider through a client interface on multiple devices; Infrastructure as a Service (IaaS) provisions clients with processing, storage, networks, and other computer resources or hardware through the cloud as needed by the client; and Platform as a Service (PaaS) that delivers custom application deployment environments from the provider (Askhoj et al., 2011, pp. 178-181; Badger & Grance, 2012, pp. 13-14; Mosweu et al., 2019, p. 3). The four deployment models are: the private cloud, which is designed for a single organization that can own, manage, and operate the private cloud infrastructure; the community cloud, created by a community of consumers that can be owned, managed, and operated by one or more of the consumers; the public cloud that is

open to the general public; and the hybrid cloud, which is a combination of two or more of the previously stated models (Badger & Grance, 2012, p. 14; Hurley, 2016, pp. 142-143; Mosweu et al., 2019, pp. 2-3; Stancic et al., 2015, p. 214).

Cost

As previously mentioned, cost is the most commonly highlighted factor for why institutions adopt cloud systems, not least of all because of widespread inadequate funding (Pandey & Kumar, 2020, p. 26). Many organizations will save money by switching to cloud services based on their needs and goals (Iglesias & Meesangnil, 2010, p. 8; Mosweu et al., 2019, p. 4), but, at the same time, using the cloud will be far more costly than on-site storage for others. For this reason, institutions and individuals need to not only establish the price estimates of cloud systems appropriate for their needs, but also determine the costs of their current on-premise storage systems. Key cost differences between cloud and on-site storage are: storage lifespans, storage acquisition costs, redundancy copies, storage utilization, personnel costs, infrastructure, maintenance, the price of data migration, lower energy use, lower acquisition costs, and costs associated with equipment failure are carried by the CSP (GFI, 2010, p. 4; McLeod & Gormly, 2017, p. 355; McLeod & Gormly, 2018, p. 178; Mosweu, 2019, p. 4). Not only do the costs of these factors need to be established, but they must also be considered for whatever duration of time the data will be stored for (Rosenthal & Vargas, 2012, p. 5). It can be easy for some to be lured into migrating to cloud services by the low monthly fees associated with cloud systems versus the hardware, personnel, and maintenance expenses needed for on-site storage—yet storage is not inexpensive anywhere (Breeding, 2013, p. 25). Some studies have found that, in practice, there is “little difference between the cost of in-house servers and cloud storage,” and

encourage institutions to consider possible organizational change, operation scale, and data redundancy in their decisions (McLeod & Gormly, 2018, p. 178).

Advertised cloud system prices can also be misleading to their total cost, whether for short- or long-term needs. Although hardware costs that an institution may presently be paying might be eliminated or reduced by moving to the cloud, the expenses lost could very easily just be transferred to the increased cost of copies (McLeod & Gormly, 2018, pp. 186-187). Even still, archive professionals consistently report that they "could be more certain about the costs of their contract with the CSP as opposed to in-house costs, which tended to be unclear" (Borglund, 2015, p. 123). Thus a greater understanding of the costs associated with on-site storage should be a foremost action taken by institutions in their decisions.

Frequently the costs of hard drives are focused on as key expenses to be lost by dropping on-site storage, but this can be a critical error. Because they have been around for decades, the efficiency-to-price ratio is well studied, and has moved at a stable rate—called Kryder's Law—with a sustained average of 30% drop in prices per-year over a period of 30 years, whereas cloud storage prices have dropped at most 3% a year (McLeod & Gormly, 2017, p. 178; Rosenthal & Vargas, 2012, p. 5). Consequently, the rate at which prices of these two competing options drop is vital in considerations where long-term storage is the goal (Rosenthal & Vargas, 2012, p. 5). This is not, however, implicating that on-site storage is universally preferable in terms of cost for archival institutions. It is, instead, exemplifying that the institution's collection size and time span can drastically alter the expected cost. Unfortunately, though, there is yet to be a proven, effective model for ascertaining the costs involved with long-term preservation (Bowen, 2018, p. 28).

Large scale projects may also be far more expensive than anticipated, as despite the low costs per unit and large scale data storage that cloud service centers provide, they can still be far more expensive than on-site storage. Vanderbilt University's Television News Archive, as one example, had approximately 150 terabytes of video assets in its collection storage in 2013, a time when hard drives of multiple terabytes were sold for under $100 at one-time costs, while the price of storing that same collection in the cloud was estimated at $20,000 per month (Breeding, 2013, pp. 25-26). Understandably, this collection was retained in on-site hard drives, but low cost, even free, cloud storage systems can fit the needs of some institutions. Small archives, community archives, and even larger archives that only intend to use cloud storage for select items or temporary migration purposes have many low-cost options available. Cloud service providers offer free storage of varying sizes, while notable providers such as Dropbox offer unlimited storage for $180 per user and Google Drive provides unlimited storage to institutions of more than five users at low costs (Goldsborough, 2015, p. 59). Free services might be ideal for community archives with comparatively small collections and storage needs. However, free services have less reliability, and regularly have no automated back-up or file recovery, even when it's the fault of the provider (Breeding, 2013, p. 25).

Long- and Short-term Storage

Perhaps second in importance only to cost—which, itself, is heavily affected by the matter of time—is whether an institution is investing in cloud storage for long- or short-term needs. Multiple studies have found that in terms of long-term storage (ten years or more), cloud storage is consistently more expensive and difficult (Breeding, 2013, p. 24; GFI, 2010, p. 8; McLeod & Gormly, 2018, pp. 178, 186; Rosenthal & Vargas, 2012, p. 5). Conflicting information is also presented on the subject of long-term storage. In one study, Bowen (2018)

found that, of the study's participants, none "had yet encountered cloud storage challenges associated with long-term [storage]" (p. 30) as no participant had had cloud storage for ten years or longer. Yet, at the same time, participants noted negative and positive opinions about long-term cloud storage, essentially mirroring the wider negative and positive opinions of cloud storage (Bowen, 2018, p. 31). Other studies, however, performed by McLeod and Gormly in 2017 and 2018 found that 35% and 44% of their studies participants, respectively, were using cloud systems for long-term storage (McLeod and Gormly, 2017, p. 344; McLeod and Gormly, 2018, p. 181). Stancic et al. (2015) posit that remote, digital storage could be made safe "if made compliant with the relevant standards" (p. 211), which would be beneficial not only for economic reasons, but also because organizations "cannot afford data loss due to improper long-term preservation techniques" (Bowen, 2018, p. 27), which Breeding (2013) cites as being "the real challenge" (p. 24).

Security

Security is another consistent argument used both for and against cloud services. As more and more archival institutions and professionals move their repositories into the cloud, they entrust the security and preservation of their collections to the service provider (Borglund, 2015, p. 115). For some institutions, organizations, and community archives, security issues alone can be all-consuming, troublesome issues for on-site storage, whereas security provided by the CSP can be an improvement, or, at the very least, be overshadowed by other benefits (Breeding, 2018, p. 10; McLeod & Gormly, 2018, p. 175; Mosweu et al., 2019, p. 4). Cloud computing systems can also decrease threats of data loss from hacks, viruses, cyber-terrorism, and worms, as well as automatically create back-ups of individual items that are stored, digitally, on geographically diverse servers to prevent data loss (Mosweu et al., 2019, pp. 2, 5).

At the same time, however, cloud services themselves have become targets for hackers and other digital attacks as a consequence of their collection of private data, which may draw more malicious attention to archival collections than they would otherwise attract on their own on-site storage systems (Borglund, 2015, p. 115; Mosweu et al., 2019, p. 7). This risk emphasizes why secure, well-known service providers should be preferred, and why it may be theoretically, if not practically, ideal to proactively retain duplicates outside of the main CSP through back-up hard drives or through another CSP, although these additional safety precautions are costly (Bowen, 2018, p. 27; Breeding, 2013, pp. 22, 24-25; Goldsborough, 2015, p. 59).

Institution Size and Scalability

Due to how high prices for larger collections can rise into impracticability, as exemplified by Vanderbilt University's collection, larger institutions may primarily find their optimal cloud services in small, specific collections with a goal of outreach and accessibility, or with temporary and migratory purposes. Smaller institutions and organizations, however, are much more likely to find benefit in cloud storage than larger institutions. Community archives, especially, are "commonly considered the least prepared to undertake digital preservation activities," but "are enabled by cloud-computing solutions to begin acquiring and preserving born-digital records" (Hurley, 2016, p. 129). Notable motivating factors for adopting cloud storage—such as lack of resources for building infrastructure, software licensing costs bing too expensive, lack of resources, and a lack of IT (GFI, 2010, pp. 3-4)—are most clearly exposed in community archives, but exist in institutions of every size as well. Unsurprisingly, cloud systems are very attractive for small organizations.

The cloud also creates scalability, so that community archives have the capability of creating consortia clouds, wherein multiple community archives "pool resources to store their

records in a centralized private cloud" (Hurley, 2016, p. 145). Larger archival institutions can utilize this same process, whether in consortiums or individually, to upscale or downscale as needed (Mosweu et al., 2019, p. 4). As such, an institution or organization can increase or decrease its cloud footprint to fit its needs and goals for the short- or long-term.

Other Factors

In addition to cost and trust in security, there are numerous other factors that make cloud storage beneficial to some institutions and detrimental to others, such as accessibility, flexibility, business continuity and disaster recovery, and trust.

Accessibility

Access can be greatly improved with the adoption of cloud storage, particularly for small-scale and community archives. Public, community, and hybrid cloud service models allow for collections to easily be accessed remotely not only by collection staff, but also by its users—whether they be public or private—and stakeholders (Hurley, 2016, p. 131; McLeod & Gormly, 2018, p. 175; Mosweu et al., 2019, p. 4). Collaboration with other archival institutions and organizations can also be simplified with a shared cloud storage, allowing for all parties to work collaboratively at any time (Mosweu et al., 2019, p. 5). However, the inherently intangible location of cloud storage can also make collections inaccessible for some users as internet access is limited, if not absent, in many places and populations (Mosweu et al., 2019, p. 7).

Flexibility

An undeniable benefit of cloud storage that any institution would benefit from is increased flexibility. For institutions saving money on cloud storage, that extra money can be invested elsewhere. Data can be accessed from multiple devices, whether on-site or remotely,

and with less equipment and training, it allows employees to be more flexible and available for other tasks (Iglesias & Meesangnil, 2010, p. 8; Mosweu et al., 2019, p. 4).

Continuity and Recovery

With data stored in at least one remote server belonging to a CSP, cloud storage is far more resilient to data loss from disasters than on-site storage (Mosweu et al., 2019, p. 5). However, as previously mentioned, this also creates the risk of data loss from a CSP error, closure, or hack. Even still, the risk of losing some data to CSP mistakes is arguably far outweighed by the redundancy protection that cloud storage provides against physical disasters.

Trust

Trust in cloud storage and cloud providers is a common detraction of cloud systems (Borglund, 2015, p. 115). Issues that have been experienced by clients of CSPs include lost records, records retained when they were supposed to have been destroyed, providers going bankrupt or disappearing, failed back-ups, items ending up in wrong servers, difficulties ensuring legal privileges as well as other legal issues, questions over jurisdictions, and unfavorable contractual terms, all of which has degraded trust in cloud storage to the point that it's regularly questioned (Borglund, 2015, 115, 119; Bowen, 2018, p. 28; GFI, 2010, pp. 4-5; McLeod & Gormly, 2017, p. 355, 358-359, 364; Stancic et al., 2015, p. 210). Perhaps most worryingly, some CSPs have failed to take responsibility for lost or corrupted data even when fully at fault (Breeding, 2013, p. 24).

Research Question

The issue in question which this proposed research study seeks to address is: how will a model presenting the benefits and drawbacks of both physical and cloud storage assist archives? To achieve such, this research aims to promote to the SAA the establishment of an official model

to assist archives in the decision-making process of choosing between on-site physical storage and remote cloud storage services—for which there is presently no helpful model. As previously stated, the question is not a matter of which method of data storage is better than the other as a whole, but rather elucidate which method may be better for a given institution's goals, parameters, and needs. Because cloud storage systems are, comparatively, a novelty when contrasted with on-site storage and, consequently, have many more misconceptions, misunderstandings, and uncertainty around their long-term viability, the focus of this research addresses cloud storage directly.

Design and Methods of Data Collection

Following the qualitatively focused mixed-methods approach of this study, quantitative and qualitative data will be obtained collectively through the use of surveys and follow-up interviews. Because the purpose of this research is neither restricted by archival size, type, or purpose, nor geographic location, the participant pool will include any archive and archivist within the United States. In order to find participants and present the digital surveys to them, contact will be made to the Society of American Archivists (SAA) with a request to share the surveys to archives and archivists on the SAA listserv. Through this method it is hoped that respondents representing small, mid-sized, large, and community archives can be reached. Anonymity will be ensured, with no survey questions asking for the name of individual archivists or institutions, and only U.S. states and territories used as geographic information.

Due to the potential of receiving too many responses to properly process with this method, however, a stratified random probability sample will be utilized. Responses will be categorized into four strata—community archive, small archive, medium-sized archive, and large archive—with the respective scopes of these strata determined by the participants themselves and

with each stratum proportional in size to its respective total in the respondent population. Albeit, the study cannot rely on all participants agreeing on the definitive borders between the different sizes of archives—most especially the vague borders around where medium transitions to small or large—yet it is less important for the purposes of this study what size a participant considers their archive to be than the participant's experiences with storage systems. Furthermore, quantifying definitive definitions of archive sizes is a research study in and of itself, and this study does not wish to add bias into the matter by presenting predetermined size definitions. During analysis of survey results, a greater and more accurate depiction of archive sizes will be determinable based on answers provided.

Participants contacted through an SAA listserv will be presented with the survey, as well as an email address to return the survey to. The survey, as discussed in more depth below, will be composed of three parts: an initial set of introductory background questions, a quantitative section of close-ended questions, and a qualitative section of open-ended, experience focused questions. Participants will be presented with the option of volunteering for follow-up interviews.

Introductory Questions

Initial survey questions will inquire as to the size of the participant's archive—including if it is a community archive—what state or territory the archive is located in, and whether or not the institution is using cloud storage. This will segway into the remaining questions of the introductory phase. Some exemplary questions are:

- If using cloud storage systems, is it used for short- or long-term preservation?
- If not using cloud storage already, is your archival institution/organization planning on or considering adopting cloud storage?

- Has your institution/organization used cloud storage in the past?
- Is your archive a community archive?

Quantitative

Following introductory questions, a section of close-ended questions will be presented for obtaining numerically based data and simple yes/no answers. These quantitative questions will primarily focus on the size of the participants' collections—this will, also, allow for a more accurate categorization of archive sizes during analysis—as well as costs experienced with cloud storage services and on-site physical storage, and, when applicable, how long the participant has used cloud storage for. This section is meant to be more in-depth than the prior set of introductory questions, but less involved than the following qualitative section. Some example questions are:

- What is the approximate size of your collection (in terms of megabytes, gigabytes, or terabytes)?
- Does your archive use on-site storage, cloud storage, or both?
    - If both, approximately what percentage is in the cloud?
- If using cloud storage, what are the average monthly or annual costs?
- If using on-site storage, what are the approximate total associated costs (e.g., purchasing of equipment, maintenance, staff training, energy costs, etc.)?
- If using cloud storage, for how long has your archive used cloud services?

Qualitative

The cornerstone of this proposed research study lies in the qualitative information section of the survey, in which survey participants will detail the real-world experiences, benefits, challenges, and questions they have of cloud storage. In-practice experience is paramount for

fully grasping the wider picture of storage systems, as the up-front costs paraded by cloud service providers is only a portion of the reality. Mistakes, unexpected challenges, and disappointments are also important to highlight, as are fortuitous surprises, unexpected benefits, and satisfaction, as archives considering cloud storage may find familiarity in an archive similar to their own and can learn from their experiences. Furthermore, identifying exemplary archive experiences with cloud storage can aid other archives of similar size, need, and goal in their decision-making processes. A few sample questions of this portion of the survey are:

- What were the motivating factors for or against adopting cloud storage services?
- What notable positive experiences and opinions do you have of cloud storage? What negative views do you have?
- If applicable, what did and didn't work with adopting and merging into cloud storage?
- What are your archive's short- and long-term goals and needs, and how does cloud storage help or hamper them?

Follow-up Interviews

Participants will be presented with the choice to provide their contact information to opt-in for follow-up interviews for more in-depth discussions with the study's researchers after the analysis of survey data. Similar to the stratified random probability sample used for the surveys, interview participants will be chosen on a basis as proportional to their represented archive type as can be reached, though at a smaller number. Interview participants will be notified through their contact information, though retain anonymity through pseudonyms and no identifying details. Some of the questions they will be asked will be similar to those of the qualitative section of the survey, in addition to new questions with the aim of receiving more in-depth answers and participants' immediate thoughts on questions.

Population, Sampling, and Participants

As previously mentioned, the participant pool for this research study will be drawn from archivists and archives who are on the Society of American Archivists' (SAA) listserv. Any archivist within the U.S. will be eligible to participate, with it predicted that geographical locations will have minimal impact on the data collected. Of greater importance is to find participants of varying archive sizes, participants involved with community archives, and participants whose archives have different temporal goals and economic needs.

Participants will be granted anonymity, with no identifying data presented in the research. There will be no preference against individuals' levels of experience or time in the profession, or their type of archive. As previously stated, in order to compensate for the potentially high number of responding participants, and to ensure that community, small, mid-sized, and large archives are equally represented, a stratified random probability sample will be used. Four strata representing the aforementioned represented sizes of archives will be established, with an equally representative number of participants in each strata. This sampling method was selected to ensure equal input from the four chosen strata of participants, and to compensate for the possible high number of survey responses.

Ethics Statement

Study respondents who participate in the survey and follow-up interviews will be given confidentiality and anonymity with no identifiable information recorded beyond the participant's U.S. state or territory and their archive's approximate size. Furthermore, participant information will not be used for any other purpose and can be deleted at any time at the request of the participant. To ensure security, a trusted service such as Qualtrics will be used to conduct the online survey, in which informed consent will be iterated before beginning the survey, and

reiterated prior to submission. It is expected that the full survey will take approximately twenty minutes to complete, and the optional interviews will be of a similar time frame.

References

Askhoj, J., Sugimoto, S., & Nagamori, M. (2011). Preserving records in the cloud. *Records Management Journal, 21*(3), 175-187. 10.1108/09565691111186858

Badger, M. L., & Grance, T. (2012). Cloud computing synopsis and recommendations. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-146

Bowen, D. A. (2018). Challenges archivists encounter adopting cloud storage for digital preservation. *Proceedings of the International Conference on Information and Knowledge Engineering*, Athens, 27-33.

Borglund, E. A. M. (2015). What about trust in the cloud? Archivists' views on trust. *Canadian Journal of Information & Library Sciences, 39(2)*, 114-127. 10.1353/ils.2015.0017

Breeding, M. (2013). Digital archiving in the age of cloud computing. *Computers in Libraries, 33(2)*, 22-26.

Breeding, M. (2018). Up in the air: Cloud computing and library systems. *Computers in Libraries, 38(10)*, 9-11.

Creswell, J. W., & Creswell, J. D. (2018). Chapter 1: The selection of a research approach. In: *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed., pp. 3-22). SAGE.

del Pozo, N., Long, A. S., & Pearson, D. (2010). "Land of the lost": A discussion of what can be preserved through digital preservation. *Library Hi Tech, 28*(2), 290-300. 10.1108/07378831011047686

GFI Max (2010). *On-premise vs. cloud-based solutions* (White paper). Retrieved from http://docplayer.net/10287524-Gfi-max-white-paper-on-premise-vs-cloud-based-solutions.html

Goldsborough, R. (2015). The changing world of cloud storage. *Teacher Librarian, 42(5)*, 59.

Hurley, G. (2016). Community archives, community clouds: Enabling digital preservation for small archives. *Archivaria, 81*, 129-150.

Iglesias, E., & Meesangnil, W. (2010). Using Amazon S3 in digital preservation in a mid sized academic library: A case study of CCSU ERIS digital archive system. *Code4Lib Journal,* (12), 1-10.

McLeod, J., & Gormly, B. (2017). Using the cloud for records storage: Issues of trust. *Archival Science, 17(4)*, 349-370. 10.1007/s10502-017-9280-5

McLeod, J., & Gormly, B. (2018). Records storage in the cloud: Are we modelling the cost? *Archives and Manuscripts, 46*(2), 174-192. 10.1080/01576895.2017.1409125

Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-145

Mosweu, T., Luthuli, L., & Mosweu, L. (2019). Implications of cloud-computing services in records management in Africa: Achilles heel of the digital era? *South African Journal of Information Management, 21*(1), 1-12. 10.4102/sajim.v21i1.1069

Pandey, R., & Kumar, V. (2020). Exploring the impediments to digitization and digital preservation of cultural heritage resources: A selective review. *Preservation, Digital Technology & Culture*, *49*(1), 26-37. 10.1515/pdtc-2020-0006

Rosenthal, D., & Vargas, D. (2012). *LOCKSS boxes in the cloud*. Stanford University Libraries. https://web.stanford.edu/group/lockss/resources/2012-09_LOCKSS_Boxes_in_the_Cloud.pdf

Shallcross, M. (2020). Trusting records in the cloud. *The American Archivist, 83(2)*, 458-462. https://doi.org/10.17723/0360-9081-83.2.458a

Stancic, H., Rajh, A., & Brzica, H. (2015). Archival cloud services: Portability, continuity, and sustainability aspects of long-term preservation of electronically signed records. *Canadian Journal of Information & Library Sciences, 39*(2), 210-227. 10.1353/ils.2015.0012

Turner, S. (2014). Capitalizing on big data: Governing information with automated metadata. *Journal of Technology Research, 5*, 1-12.