Challenges Archivists Encounter Adopting Cloud Storage for Digital Preservation

Debra A. Bowen
Department of Computer Science and Information Systems, National University, San Diego, CA, USA

Abstract – This paper provides information to those who design, develop, manage, and provide storage and archival services by answering the question, what are the most challenging aspects adopting cloud storage for long-term digital preservation? There was a gap in whether current challenges from the literature (cost, obsolescence, security) were specific to the archival community or if there were other challenges unique to the profession. Members of the Society of American Archivists (SAA) were the population for this study. Data was collected using three rounds of surveys in which participants identified, ranked, and described challenges. The 10 most challenging aspects of cloud storage included all three challenges from the literature. The only challenge unique to the preservation of digital information was related to fixity checking.

Keywords: archival services, archivist, cloud storage, digital information, digital preservation

1 Introduction

People and organizations are storing massive amounts of digital information in the cloud without fully understanding the long-term benefits and/or risks. The introduction of the Internet has led to such a rapid increase in digital information that estimating the actual amount is challenging and often inaccurate. Storage technologies continually evolve to store larger amounts of information, faster, more cost effective, and for longer periods. Additionally, information technologies have altered the way humans produce, store, and retrieve information about themselves and the world around them. As a result, organizations are struggling to select, store, and maintain valuable digital information so that it will be available when needed. However, just storing digital information to a cloud solution is not enough to ensure it will be accessible, usable, and trustworthy in the future.

The goal of archiving is to retain information that has a long-term preservation need due to historical, cultural, or legal value. Archivists, who are part of the archival community in charge of preserving our most valuable digital information for future access, are in the process of adopting cloud storage for long-term digital preservation [1][2]. The archival community places its trust in the Information Technology (IT) field to develop and manage the technology that ensures information is archived and protected so that it can be accessible when needed. Organizations can now afford to store unlimited amounts of data. However, most cannot afford data loss due to improper long-term preservation techniques. Preserving digital information for long periods involves the likelihood of making sure that accessibility to authentic, usable, and reliable information endures for centuries, without knowing what storage technologies of the future will exist. In addition, archivists need to do this while maintaining both the authenticity and context of the information.

This paper identifies the top 10 challenges archivists encounter when adopting cloud storage for long-term digital preservation. Although the practice of archiving has been around for thousands of years and an archivist has many responsibilities beyond archiving, the focus of this study was only on the digital preservation aspects of archiving.

1.1 Theoretical frameworks

The theoretical frameworks for this study were archival theory and the Open Archival Information System (OAIS), which is a theory of digital preservation. As IT crosses into almost all other fields, the same could be recognized for the discipline of archival science. Almost all disciplines outside of the archival community have a growing interest in archiving and preservation.

1.1.1 Archival theory

Archival theory is a collection of interconnected ideas that encompass archiving, with the objective of guiding archivists in their work. In a digital environment, an archive is not the same as a backup. Backups are often used for recovery purposes in the event that the original data becomes corrupt or lost. An archive’s purpose is the long-term preservation of digital data selected because of its evidential or historical value. Archives, in the context of archival theory, entails everything about a document that is deemed valuable enough to archive, including its relationship to other documents. Maintaining a relationship to other documents is an important responsibility of the archivist and is necessary when proving an archive’s authenticity, regardless of format or media. This fundamental principle of archival theory is referred to as provenance.

Provenance refers to the information about the origin and ownership history of an item or collection of items with the intent of assuring authenticity; however, provenance does not end when authenticity has been captured. Once the authenticity of an item has been determined, an archivist has the responsibility of ensuring the security of the item so that its authenticity is not damaged or lost. Provenance and technology
are inseparable for archiving digital information [3]. In an article by [4], the authors propose an approach for using secure provenance as a foundation for providing a trustworthy cloud. Provenance is an important construct of archival theory, especially now that archives are moving away for the archivist and into the cloud.

The use of provenance as a principle of archival theory underlying this study is not intended to be an indication of expertise about provenance or the nature of archival theory. Archival theory encompasses many other important principles, such as appraising, acquiring, preserving, and providing access, that were not included in this study due to limited scope and focus.

1.1.2 Open Archival Information System (OAIS)

The Open Archival Information System (OAIS) is a theory of digital preservation that has been extensively accepted by the archival community [5]. The idea of an OAIS reference model originated in 1995 from the Consultative Committee for Space Data System (CCSDS). The CCSDS' function at the time was to support the study of terrestrial and space environments. The OAIS is a high-level model for archival repository systems that preserve, manage, and maintain access to digital information, which requires long-term preservation. However, the OAIS does not specify any hardware, software, database, language, or platform requirements to be an OAIS-compliant archive. The model is meant to provide standards for digitally preserving an archive without controlling the method for doing so.

The verbiage in the OAIS reference model is identical to the International Organization for Standardization (ISO) 14721:2012, titled Space data and information transfer systems - Open archival information system (OAIS) - Reference model. An OAIS is, "an archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a designated community" [6, p. 1-1]. In addition, the word “open” in Open Archival Information System is not a reference to allowing open access into an archival system, but rather to the fact that recommendations and standards to the OAIS have been, and will continue to be, discussed and created in open forums.

Efforts into digital preservation has been occurring for almost four decades in a mostly uncoordinated manner. However, because of its focus on digital information the OAIS model has now become a framework for digital preservation in many non-archival disciplines, such as big science, medical, and design and engineering. This acceptance throughout disciplines has made the OAIS model the most widely used framework for digital preservation systems.

2 Challenges from the literature

Preserving digital data long-term is becoming more important because of increased cultural and economic dependence on digital data. Archivists must rely on an IT organization to manage cloud storage to ensure proper preservation of data so that it is available when needed. Exploring the challenges archivists encountered when using cloud storage advances the knowledge base of archival and information technology professionals by providing a clearer understanding of what to expect so that the archival community can successfully prepare for cloud adoption challenges. In addition, it provides the IT field information, unique to the archival community, when designing, developing, managing, or offering cloud storage products or services for long-term digital preservation.

There was limited research that examined the challenges of adopting cloud storage by the archival community. The research literature indicated that the archival profession relied on IT organizations to manage its digital information, not only to support compliance but also to ensure information is stored and protected [1]. While researching cloud storage for long-term digital preservation, three challenges recurred in the literature, (a) cost [1][7], (b) security [12], and (c) obsolescence [1].

2.1 Cost

From the research, it was apparent that the cost involved in long-term preservation was a consideration that had yet to have a proven model. In an article by [1], the authors identified quantifying the value of long-term digital preservation as a future study area. They justified research in this area so that organizations could have some type of decision criteria to measure their digital preservation IT investments. Turner [7] believed this lack of a specific cost model might be, in part, due to organizations not recognizing the value of digital assets.

2.2 Security

Digital information security, especially for sensitive information, was one of the most cited concerns of information professionals looking to adopt cloud storage [8]. Security threats such as data loss, data breaches, and authentication and access vulnerabilities were not favorable, especially for long-term digital preservation. The archival community continues to address cloud security in research papers and projects [8][9].

2.3 Obsolescence

One of the more challenging aspects of long-term digital preservation was obsolescence. This happens when hardware, software, storage media or file formats do not last long enough to provide long-term accessibility of digital information. Shortly after disk storage became popular, the challenges of obsolescence began to appear in archival journals. In an article by [1], the authors concluded with a suggestion for future research that included how to minimize information loss due to obsolescence. Through reviewing the literature, obsolescence appeared to be an on-going digital preservation challenge regardless of the IT storage technology utilized.

Every day, valuable digital information becomes unreliable or inaccessible because of obsolescent storage technologies [10]. Hardware, software, storage media, and file formats become outdated over time, causing preserved digital information to become unusable if not periodically managed. The IT field
continues to be the driver for storage solutions used by archivists and cloud storage is the next solution that archivists will be utilizing, essentially entrusting digital information to the IT field for safekeeping.

3 Research methodology

3.1 Research question

This study targeted archivists to get their perspective on the unique challenges they encountered using cloud storage for long-term digital preservation. The following research question provided a guide for this study:

Q1. From an archivist’s perspective, what are the most challenging aspects when adopting a cloud storage solution for long-term digital preservation?

3.2 Research design

The design for this qualitative study was a modified Delphi technique. The Delphi technique made it possible to elicit challenges experienced by certified archivists so that they could rank and describe their experiences anonymous from each other to further an understanding of their challenges when adopting cloud storage for long-term digital preservation. The main aspects of the Delphi technique are (a) it is a process in which experts respond to the same concept at least twice, (b) panel participants are anonymous from one another, (c) controlled feedback provided by the researcher, and (d) a final group response in which all participant opinions are included.

In addition to panel members providing challenges they experienced, they had the option of selecting three reoccurring cloud storage challenges derived from the literature, which were (a) cost, (b) obsolescence, and (c) security. Because this study utilized preselected challenges during round 1, a modified Delphi model was utilized. This model has been used in other studies providing preselected list items [11][12], which is not a characteristic of a classic Delphi study.

The Delphi technique was an appropriate method for the proposed question and phenomena because collecting data from experienced archivists that have already adopted cloud storage was a way that was convenient for participants and researcher. In addition, archivist are concerned with specific principles when archiving digital information, such as provenance and authenticity and, unlike many cloud storage users, may never use the data they are archiving. A quantitative study that measures factors that influence adoption to cloud storage would not have been as appropriate for archivists, as it might have been for other groups, because some cloud storage factors for archivists were not quantifiable, but rather something understood in the framework of their professional responsibilities.

Subsequent rounds of this modified Delphi study narrowed and ranked the challenges archivists experienced and allowed participants to describe how selected challenges were resolved or managed. Using a convenient online survey tool and collecting data from professionals that directly experienced the phenomena, provided richer feedback than would have resulted from other conventional methods [11].

3.3 Target population and sample

Although digital preservation activities can be found in many non-archival disciplines, the population focus of this research study was on archivists due to preservation being one of the key responsibilities of their profession.

The target population of this study was the members of the Society of American Archivists (SAA). The SAA is a professional not-for-profit organization. The SAA provides a certification method for the archival profession. All members must pass a certification exam covering archival knowledge, skills, and responsibilities. In addition, members hold at least a master’s degree and have a minimum of one year of archival experience. To maintain a certification, members must provide proof of ongoing education, experience, and participation in the profession. The challenges SAA certified archivists encountered when adopting cloud storage for long-term digital preservation was the focus of this study.

The sample was members of the SAA, from the United States, who were 18 years or older. In addition to members agreeing to the informed consent form, the inclusion focused on members that actively worked in an archival related role during cloud adoption. Another inclusion criterion was the availability of panel members to participate in multiple rounds of survey questions. Any member that did not agree to the informed consent form was exited from the study. In addition, any member that did not complete the round 1 survey questions was excluded from further rounds.

Because panel participants needed specific archival knowledge, skills, and experience adopting cloud storage for long-term digital preservation, a purposive criterion sampling method was utilized to intentionally select individuals from a group that was most likely to have experienced the phenomenon personally for this study.

3.4 Data collection

Once a participant clicked on the button agreeing to the informed consent, they were linked to the round 1 survey in SurveyMonkey. The data collection procedures from this point were as follows:

1. Demographic questions began the round 1 survey. Other than the question asking about a participant’s age range, only questions relating to archival experience and adopting to cloud storage were asked. This was intentional to protect participant identity.

2. In addition to panel participants being able to provide open-ended responses, identifying up to three challenges from their own experience, they were also given the opportunity to agree or disagree to having encountered challenges from options derived from the literature.

3. For round 1, a list of three preselected challenges were provided, which came from the literature. The participants were asked if they encountered any of the
three challenges, their response options for each was either yes or no. In addition, there were three text boxes for participants to add their own additional challenges.

4. The round 1 survey request was sent from SurveyMonkey. The email allowed potential participants to continue to the survey or to disagree with the informed consent form and exit the study.

5. SurveyMonkey was used for the online data collection of all questions, for all rounds.

6. The round 1 demographic data and all participant entered challenges were downloaded and analyzed within Microsoft Excel to remove participant identifiable information. Then the data was imported into NVivo Pro 11 qualitative data analysis software tool to create a final list of challenges for round 2.

7. Round 2 was used to narrow the list of 19 challenges to a more manageable size. This was accomplished by having panel participants select the top 10 challenges from the original list of 19 challenges from round 1. The 10 challenges with the highest percentage of participants that ranked them was the final list of challenges for ranking in round 3.

8. Round 3 had panel participants rank the final list of 10 challenges, in order from most challenging. This would be repeated until there was a level of agreement of ranked answers. A final question gave the participants an opportunity to share their thoughts about the long-term aspect of cloud storage.

9. After rounds two and three, Kendall’s Coefficient of Concordance, or Kendall’s W was calculated to determine if a level of agreement had been archived (W > .5) [13]. This would have been repeated until an agreement of W > .5 was achieved. However, W > .5 was attained for round 3, so data collection was completed.

Once Kendall’s W determined a level of agreement of ranked answers, the final outcome was a list of 10 challenges, ranked from most to least challenging, that archivists encountered adopting cloud storage for long-term digital preservation. In addition, how challenges were resolved or managed and opinions about the long-term aspect of cloud storage were documented.

4 Results

This study was guided by one research question, “From an archivist’s perspective, what are the most challenging aspects when adopting a cloud storage solution for long-term digital preservation?” To answer this question comprehensively a ranked list of top 10 challenges was identified and descriptions of those challenges were analyzed for themes. In addition, how those challenges were resolved or managed was also identified. The participant responses during rounds one and two did not include any phrases related to long-term. To capture this aspect of the research question a specific open-ended question was added to the round 3 survey to solicit opinions about long-term, in relation to cloud storage, from the experience of participants. Because no participant’s long-term response described an actual experience, it was deduced that none of the panel participants had yet encountered cloud storage challenges associated with long-term.

Members of the SAA were the sample for this study’s panel due to their credentials and experience with the study’s topic. To ascertain the top 10 cloud storage adoption challenges, three rounds of surveys were employed to obtain a ranked list of challenges provided by a panel of 23 members. Round 1 produced a list of 19 challenges; three were derived from the literature and 16 provided by participants. During round 2, 16 participants narrowed the challenges to 10 and ranked them from most challenging. The top five ranked challenges included the three challenges derived from the literature, (a) cost, (b) obsolescence, and (c) security. Also, four themes emerged from the participants challenge descriptions, (a) staffing, (b) technology, (c) obsolescence, and (d) security. Kendall’s W did not meet the target of W > .5 during round 2, so round 3 was conducted.

During round 3, 12 participants ranked the 10 challenges (Table 1 shows the final list of challenges, in ranked order).

Table 1 - Final Ranked Challenges

<table>
<thead>
<tr>
<th>Final Rank</th>
<th>Challenge</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cost (for equipment, resources, education, other activity, etc.)</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>Obsolescence (hardware, software, file format, media, etc.)</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>Lack of ability to easily do fixity checking</td>
<td>47</td>
</tr>
<tr>
<td>4</td>
<td>Extracting content from the systems in which it was originally used</td>
<td>48</td>
</tr>
<tr>
<td>5</td>
<td>Security (data loss, data breach, authentication, etc.)</td>
<td>55</td>
</tr>
<tr>
<td>6</td>
<td>Connectivity – speed &amp; reliability of access to cloud &amp; file uploads/ transfers</td>
<td>77</td>
</tr>
<tr>
<td>7</td>
<td>Technical knowledge/lack of qualified staff</td>
<td>79</td>
</tr>
<tr>
<td>8</td>
<td>Understaffed IT department</td>
<td>88</td>
</tr>
<tr>
<td>9</td>
<td>Coordination with IT staff</td>
<td>103</td>
</tr>
<tr>
<td>10</td>
<td>Trust of the provider</td>
<td>105</td>
</tr>
</tbody>
</table>

Note. 1 = most challenging. The score is the sum of all rankings for the challenge.

Again, the top five included the three challenges derived from the literature. The lack of ability to easily do fixity checking was also ranked in the top five during round 3. This was the only challenge in the top 10 that was specific to digital preservation, opposed to data that is simply stored. Fixity checking is a digital preservation term that refers to the process...
of checking the integrity of a digital item by verifying that is has not become unexpectedly altered [6]. Fixity ranked two places higher than security in the final ranked list. Noting that there may have been a terminology difference, fixity checking and security could reasonably be the same challenge, however, the security theme that was emerging in round 2 did not develop during round 3 to provide more support for this assumption.

Kendall’s W was calculated at W = .61, which exceeded the target of W > .5. The analysis from challenge descriptions indicated two themes continued from round 2, technology and obsolescence. Also, cost and human expectation, emerged as new themes. Round 3 also solicited participant opinions about the long-term aspect of cloud storage, because it was not addressed in previous rounds. Participant opinions of long-term, in relation to cloud storage, were equal among the category types of positive, negative, and unsure (see Table 2 for participant responses).

Table 2 - Opinions on the Long-Term Aspect of Cloud Storage

<table>
<thead>
<tr>
<th>Participant</th>
<th>Category</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>P05</td>
<td>negative</td>
<td>&quot;There has not been a real long-term digital solution that I know of, just different technologies. Isn't the cloud just a bunch of hard drives over the network - what is long-term about that?&quot;</td>
</tr>
<tr>
<td>P14</td>
<td>negative</td>
<td>&quot;The more data that is moved around the more potential for it to be changed or become corrupt. To be a real long term solution cloud vendors will need to be more aware of integrity issues with archives and how to deal with them&quot;</td>
</tr>
<tr>
<td>P26</td>
<td>negative</td>
<td>&quot;This is problematic to entrust ownership/control of valuable digital materials to a third party. We must find a way of retaining more autonomy for the holding institution.&quot;</td>
</tr>
<tr>
<td>P02</td>
<td>positive</td>
<td>&quot;It is key part of our overall strategy. We have no choice due to cost and management considerations. Especially at this institution's rate of growth.&quot;</td>
</tr>
<tr>
<td>P18</td>
<td>positive</td>
<td>&quot;being non-profit the cloud really helps us to be part of a bigger community, so I hope it proves to be a long-term solution&quot;</td>
</tr>
</tbody>
</table>
| P21         | positive | "I think that as cloud storage becomes increasingly cheaper, it will also become increasingly useful. Ideally, we would store our data using two different storage providers for redundancy and security, but we can't afford to at the time. Open source systems for ingesting content into cloud storage and checking fixity with little to no programming knowledge (such as an open source version of something like Cloudberry Backup, and improved tools from resources like AVPreserve) would also help more institutions adopt cloud storage as a preservation solution."
| P08         | unsure   | "have not really thought about the cloud in terms of long-term, it is fairly a new process and we are still working out all the issues" |
| P11         | unsure   | "I wonder if the issues we had with the older systems will become the same issues we will have with the cloud in 20 or 30 years" |
| P15         | unsure   | "My thought is that it is too soon to tell..." |

As expected from participant challenge descriptions, the technology and obsolescence themes were consistent between rounds. In addition, obsolescence was one of the challenges derived from the literature. Although cost was not a theme in round 2, it was ranked as the fourth most challenging aspect of adopting cloud storage. On the final ranked list from round 3, cost was ranked as the most challenging aspect of adopting to a cloud storage solution. As one participant, P08, stated, “most of the challenges stem from not enough funding.”

An unexpected theme during round 3 was human expectation. Participant P05 mentioned this as the reason why extracting content from the systems in which it was originally used was such a challenge.

As noted previously, the concept of long-term was not included in any of the participant responses during round one or two. When participants were asked to share their thoughts about the long-term aspect of cloud storage, the responses were not specific to any actual experiences. Because long-term for this study represented a period of 10 years or longer [8][14], it is probable that panel participants had not yet encountered any long-term cloud storage challenges.

The results of the study did answer the research question. In addition, the final ranked list of challenges included the three challenges derived from the literature (cost, obsolescence, and security). The long-term aspect of cloud storage had not been experienced by any of the panel participants, but they were able to provide opinions based on knowledge from their overall cloud storage adoption experiences.

The three challenges from the literature were also ranked as top challenges by the panel participants of this study because the challenges of cost, obsolescence, and security are still ongoing issues in the adoption of cloud storage. By taking into account the unique requirements of the archival community, this study’s results provide the IT field a perspective of cloud storage challenges by an industry outside of IT.

The findings yielded by this study could have also resulted partly because the archival community has been managing the challenges of cost and obsolescence for as long as they have been using digital storage. Those two challenges, not being unique to cloud storage, may have been selected because of an association with previous storage technology experiences.
5 Limitations

A limitation of the modified Delphi design was that there were no clear guidelines for determining sample size. In an article by [15], the authors suggested that for a homogeneous group, as the one used for this study, 10 to 15 participants might be sufficient. However, the attrition rate made it difficult to determine an initial sample size so that 10 to 15 participants completed the final round.

Another limitation was that only one professional archivist organization (the SAA) was surveyed for this research study. Even though this group is believed to be a distinguished and knowledgeable group of archivist, there were other organizations whose members would have been able to contribute to this study, but were not engaged due to the limitations of scope, time, and cost.

Several areas of archiving were not mentioned in this study, such as appraising, acquiring, arranging, and describing. Archiving digital information to a cloud storage solution is only one specific concern of the archival community.

6 Recommendations for further research

A recommendation for future research would be to use the list of ranked challenges from this study as a preselected list for a new modified Delphi study, while also expanding the targeted population to a broader area of the archival community. This would potentially provide a richer description of challenges, with more diverse approaches to resolving or managing the challenges. In addition, including the IT community in describing and explaining how they would resolve or manage these challenges, in relation to the archival community, may prove beneficial to both communities.

Using a modified Delphi method to reach a level of agreement for a list of ranked challenges was convenient and easy to implement, however, the focus on ranking and the multiple rounds needed to reach a level of agreement may have detoured potential participants from participating. One design recommendation that could strengthen a similar study would be to use a case study approach to gather challenge descriptions and resolutions and to remove the ranking component. This would be more time consuming; however, fewer participants would be needed and they could potentially provide richer details into the challenges encountered when adopting a cloud storage solution.

7 Conclusion

Using a modified Delphi technique, a panel of SAA members identified, ranked, and described the 10 most challenging aspects they encountered when adopting a cloud storage solution for long-term digital preservation. The top five ranked challenges did include the three challenges derived from the literature; cost, obsolescence, and security. Ranked third was lack of ability to easily do fixture checking, which was the only challenge identified as being specific to digital preservation. The survey rounds were concluded when Kendall’s W was calculated at W = .61, exceeding the target of W > .5.

The challenges from the literature were confirmed by the panel participants as still being relevant. This indicated that these challenges were on going or that they were so prevalent in the archival community that any storage technology would have likely received similar rankings for cost, obsolescence, and security. The need for the archival and IT communities to work together to resolve preservation issues created by new storage technologies is echoed throughout participant responses.

8 References


