Access and Preservation in the cloud: Lessons from operating Preservica Cloud Edition

Kevin O'Farrelly
Preservica
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Kevin.O'Farrelly@preservica.com

Alan Gairey
Preservica
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Alan.Gairey@preservica.com

James Carr
Preservica
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
James.Carr@preservica.com

Maïté Braud
Preservica
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Maite.Braud@preservica.com

Robert Sharpe
Preservica
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Robert.Sharpe@preservica.com

Ann Keen
Preservica
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Ann.Keen@preservica.com

ABSTRACT

The archival community has recently been offered a series of cloud solutions providing various forms of digital preservation. However, Perservica is unique in providing not just bit-level preservation but the full gamut of digital preservation services that, up until recently, were available only to organizations using a system installed on-site following on from a complex, and potentially risky, software development project. This "new paradigm" [1] thus offers a zero capital cost "pay as you go" model to perform not just bit-level preservation but also "active preservation" [2]. This short paper will describe the practical difficulties of providing and operating such a comprehensive service in the cloud.

A cloud system's advantage is to reduce the need for capital costs (since hardware and software are rented not bought up front) and system maintenance (since this is provided by the system's provider). To reduce costs further a system can share multiple organizations' content on a single operational instance. However, this instance must maintain each such tenant organization's isolation (i.e. one organization's content must not be exposed to any others). In addition each tenancy must be able to control its own processes without being able to compromise those of other tenants. This leads to the need for some degree of tenancy administration (without placing on each tenant a large burden of administration that is best handled at the system level).

The need to move bulk content across the internet as part of ingest cannot be avoided but the remaining ingest functionality can be performed either prior to upload (through a downloadable client-side tool) or server-side (through comprehensive workflows). Some ingest steams (e.g., web crawling) in fact can be considerably eased by using the cloud since an organization's local internet bandwidth is no longer relevant.

Other OAIS functional entities (preservation planning, data management, administration and storage) can all be performed without the need to move content across the internet. Access can be provided in a variety of forms including those suitable for archivists and those suitable for the general public. It is also possible to render content server-side to minimize the need for download.

Importantly, it is also possible to export an organization's entire content thereby providing a suitable "end of life" route to move to a different digital preservation system.

General Terms

Infrastructure, communities, strategic environment, preservation strategies and workflows, digital preservation marketplace, case studies and best practice.

Keywords

OAIS, Bit-level Preservation, Logical Preservation, Active Preservation, Cloud

1. INTRODUCTION

There has been a recent trend towards deploying and utilizing software systems in the cloud. In particular, digital archiving and preservation solutions are now available in the cloud. Cloud-based software systems (and digital archiving and preservation solutions in particular) have some distinct advantages and disadvantages over local deployment. This short paper compares and contrasts the experiences of developing solutions both on an organization's site and via a shared tenancy system in the cloud.

Note that in this paper, the term 'the cloud' is used to refer to public cloud instances, where services are made available over a publicly available network. While private clouds (i.e. cloud infrastructure operated solely for one organization) are similar to

public clouds, many of the issues (legal, hardware provision and elasticity in particular) are different.

2. METHODOLOGY

In order to be able to discuss general issues that can occur with cloud systems and how it is possible to address these, it is necessary to have experience. This paper relies on Tessella's experience of developing and running both on-site and cloud-based preservation systems (Preservica). Hence, issues are discussed in general first and then (where appropriate) the Preservica solution to these issues is outlined.

Tessella's on-site preservation system (using the SDB software, recently rebranded as Preservica Enterprise) has been developed over about a decade and is deployed on-site by a number of leading archives and other memory institutions around the world. This allows bespoke functionality to be added to the system's core functionality in order to deliver a system that meets the specific, true needs of the organization.

The cloud-based Preservica service was launched in June 2012 and utilizes the same core software. It is deployed within Amazon Web Services cloud offerings.

3. CHOOSING THE CLOUD

There are a number of features that are important in determining whether or not to use the cloud for a digital preservation system.

3.1 Legal constraints

The use of a cloud solution means that content is stored away from an organization's own site. This may (or may not) be an issue depending on the nature of the content stored, the mandate of the organization and the legislative and regulatory framework in which they operate. The complex topic of intellectual property rights is covered in more details in other places [3].

The single biggest concern seems to be jurisdiction, with, for example, US institutions reluctant to let their content leave the United States and most European institutions reluctant to let their content leave the European Union. To get around this issue Preservica currently (March 2014) is deployed in two separate instances: one on the East Coast of the United States and the other in Dublin in Ireland.

Of course other organizations will have other constraints (e.g., defence contractors are unlikely to be willing to allow their information to be stored in a public cloud) that may prevent them from using the cloud.

3.2 Hardware & Elastic Computing

One of the advantages of cloud systems is that it is not necessary for an organization to purchase or maintain its own hardware. This removes the need for a capital budget and to have to make (often quite technical) purchasing decisions. It also removes the need to have to decide when it is necessary to perform a hardware upgrade (and to pay the capital cost associated with such an upgrade).

Cloud services are usually elastic. This means it is possible to add additional hardware to expand computing capability. In the case of Preservica the core software works by passing the 'heavy loading' tasks to an array of job servers via a queuing system. This means that both on-site and cloud-based systems are known to scale very well. Of course such scaling comes at a cost whether it is via purchased, on-site hardware or rented, virtual servers in the cloud. One of the advantages of the cloud is that it is possible to rent servers for just the time that they are needed meaning that, for example, it is possible, to use the servers needed to process a backlog or a temporary ingest surge and then stop paying for them after that point.

In the case of buying a cloud-based service each user is sharing processing resources with other users. Thus, it is the responsibility of the provider to ensure that sufficient resources are available to cope with steady loads and to deal reasonably with peak demands. Typically this will be monitored via a service level agreement (SLA) determining not just availability but also reliability whilst also specifying any limitations on, say, processing load that the tenants cannot break without sufficient prior agreement (to allow the service provider time to provision for it) and, potentially, payment.

3.3 Tenancies and Tenancy Isolation

Typically a cloud-based, software-as-a-service offering relies on economies of scale as hardware and administration costs are shared across all clients of the service. However, this means that clients of this service also share the same infrastructure, raising the potential for security breaches.

Hence, each organization utilizing the Preservica service becomes a 'tenant' within a selected instance. It is vital that these tenants remain isolated from each other and are not able to see each other's contents or to be able to tell the workflows etc. run by each other. Preservica has undergone extensive design reviews and a rigorous testing program to ensure tenant isolation.

3.4 Exit Strategy

Another very important aspect to consider in choosing a cloud system is how organizations will be able to move between providers. This is important since the cloud is still young and thus can be expected to evolve quickly. In order to be able to gain advantages from these changes, it is important that organizations don't become locked into arrangements that are very difficult to break either for contractual or technical reasons.

Preservica guards against this by allowing a full export of content with related metadata in a published AIP package format. This export process can be configured to allow alternative metadata schemas to be used and/or alternative packaging approaches. This allows great flexibility in how to export and thus in ability to import into a successor system.

3.5 Capital vs. Revenue Costs

Of course, a lot of decisions need to balance costs with the ideal functionality.

Typically, the cost of owning a full OAIS system in the cloud is much lower than the cost of owning and operating a similar system on site. As well as operational costs there are two big overheads in setting up an on-site system: equipment capital costs and software capital costs.

However, in certain circumstances it is possible for the economics to change in favor of an on-site system, even considering these overheads.

The most obvious of these overheads is the capital cost of hardware, especially storage systems. Generally the cost of renting cloud-based hardware is lower than the cost of buying and running an equivalent system on site. However, at high storage volumes the economics of an organization running its own system begin to be comparable to, or even cheaper than, those of using a

cloud-provided one. When taken together with the simplified exit strategy, this could lead to a decision to use an on-site solution.

Another potential overhead for an on-site solution is the capital cost needed to procure, develop and configure the system in the first place. Although a cloud system removes the need to pay these costs, by its very nature such a system must be generic. An on-site system, in contrast, can be built to meet an organization's exact needs (ideally based off an existing, flexible starting system). For example, many of Tessella's customers have built systems to completely automate the process of ingesting very high volumes of materials using ingest workflows configured to work with the peculiarities of each source (e.g., to interpret the output of a digitization stream correctly and then ingest it). This can reduce the effort needed for ingest significantly and can produce a very high payback over the use of a more generic system that requires a large amount of intelligent user input in order to interpret the sources for each ingest of new material.

Hence, the decision on whether to use the cloud or not, is often a balance between one-off capital costs and on-going revenue costs.

4. STORAGE

Many people associate the cloud with storage. Indeed, a basic requirement of a digital preservation system is to offer bit-level preservation. Cloud-based digital preservation systems allow organizations to make use of the economies of scale offered by storing content using infrastructure beyond the means of most individual organizations. It also means that the operating and administration costs are similarly reduced.

In the case of Preservica, the S3 storage services offered by Amazon Web Services are used by default. These services create multiple copies in geographically separated places and perform their own integrity checking. This allows Amazon to claim 99.999999999% durability, which compares favourably to almost any in-house storage arrangement. However, organizations with a mandate to retain content in perpetuity are, naturally, wary of such claims (not least because even if it is accepted that the technical risk is extremely low there is a probability of the system ceasing to exist for other reasons). Indeed some cloud-based storage services have gone bankrupt and thus no longer exist.

To get around this issue, most cloud-based offerings allow organizations to choose to store copies in alternative storage systems. In Preservica's case this can include the ability to hold a local copy using a 'copy home' storage mechanism (using ftp to write content back to hardware controlled by the host organization).

No system can offer a 100% guarantee. Hence, while it is tempting to continue to add more storage options, the ultimate goal will remain unachievable. Some providers do offer an insurance-backed guarantee. However, even here, it must be remembered that, as with other insurance, while a claim might lead to monetary compensation, this will not recover what has been lost, and it will still be necessary for an assessment of the value of what has been lost to be made prior to any claim being paid.

Ultimately, therefore, the appropriate storage policy is a compromise between costs and risks. Preservica allows this balance to be controlled differently based on appropriate criteria. Hence, a storage policy module allows organizations to choose different strategies for different content files (e.g., for digitization streams it might be appropriate to store the high-resolution master images in a cheaper storage system with low access capabilities, such as Amazon's Glacier offering, while storing low-resolution, access copies in a highly available storage system such as Amazon S3).

Preservica has methods to allow content to be moved to allow for changes of policy, because of a change in the perception of risk, or to cope with a triggered risk (e.g., failure of a provider), or to optimize costs after a change in pricing. In the latter case it is important to weigh any costs of moving content (e.g., in bandwidth charges) against any potential savings.

5. ACCESS

Another important feature of most cloud solutions and digital preservation systems is access to content. The capabilities of systems vary here, but Preservica has two distinct offerings.

The first is an archivist's user interface. This provides search and browse capabilities and offers a detailed view of the metadata of each entity (collections, records, files, and embedded objects within files) in the system. This includes the ability to view the audit trail and provenance of each entity. For records with multiple representations (e.g., those that have been migrated from one set of technologies to another) it is possible to compare the significant properties between each representation.

The second user interface is intended to be used by the general public to get live access to the parts of the collection they are allowed to see. This user interface deliberately only displays a subset of the available information about each entity (e.g., it excludes the audit trail) and only the representations intended for public consumption.

In addition, both user interfaces are capable of providing server side rendering to allow users to view content without needing to download it to their device. This is important in a cloud-based environment since downloads come at a cost and, depending on an individual's internet connection, can be slow. It also allows complex technologies to be rendered (e.g., Preservica will render WARC files using the Wayback machine which otherwise would require a complex server setup to be used once the individual has downloaded such a set of files).

This approach of having two distinct user interfaces and therefore two very different user experiences is an example of the separation of concerns that is a feature of the cloud-based approach. It allows very different user communities to be supported from one system. The on-site approach to this issue has typically been to have separate systems (often from different suppliers) but this is harder in the cloud since the integration is much less efficient if systems are not co-located.

6. OTHER OAIS FUNCTIONAL ENTITIES

While most cloud-based systems just offer bit-level preservation and provide some form of ingest and access, these are only some of the functional entities in OAIS and are thus insufficient to meet its demands. Preservica provides a full OAIS solution in addition to Storage and Access described above. It has come about owing to the increasing maturity of the functionality of the core product. This ability to bring functionality that was previously confined to on-site systems with a large bespoke element and significant capital costs into the cloud has been described as a "new paradigm" [1].

6.1 Ingest

A variety of routes are available including the ability to upload client-created SIPs (which can be created from ad-hoc content via a downloadable tool), create SIPs server-side from upload ZIP files and purely server-side ingest routes (e.g., web harvesting). All ingests pass through rigorous quality controls.

6.2 Data Management

This is highly flexible allowing users to describe the information using a schema of their choice and yet still search, view and edit the information [4]. In addition, it is possible to integrate with some external cataloguing systems.

6.3 Preservation Planning

This includes 'Active Preservation' [2] and includes the ability to perform both technical and conceptual characterization, determine which material is at risk either during ingest or at a later date, determine the most appropriate preservation plan and then perform validated format migration at scale. This is controlled via a technical registry [5].

6.4 Administration

If a cloud service is used it is not necessary for an organization to maintain its own technical administrative staff. This is especially valuable to smaller organizations since such tasks are often hard to resource. Even larger organizations find it hard to recruit, manage, train (and ultimately retain) technical staff such as database administrators. Sometimes such administration is out-sourced to a parent organization (e.g., a regional archive might rely on the central IT provision of the region's government). In these cases it can be hard for the needs of the smaller, client organization to be heard and understood by the administrators. Hence, for small and medium sized organizations, at least, there is a distinct advantage in buying a cloud-based service where the administration is performed by skilled and trained administrators who understand the needs of the system.

However, organizations still want (and need) to have some element of control. Hence, Preservica again separates the concerns and distinguishes system-level administration from tenant-level administration.

System-level administration involves managing availability, performing database backups, adding new patches and functionality etc. This is the responsibility of the service provider (Tessella in the case of Preservica Cloud).

The tenant-level administration (i.e. configuring functionality for an organization, determining which local metadata schemas to use etc.) needs to be controlled by the tenant and Preservica provides intuitive browser-based user interfaces to do so. This means that each organization can have control without having the burden of complex system administration.

7. CONCLUSION

This paper has presented some of the advantages and issues of running digital preservation services in the cloud. It shows that it is possible for this approach to offer a much-reduced entry barrier to organizations performing digital preservation without the need to compromise on demanding a full OAIS solution (i.e. both logical and bit-level preservation).

There are a number of technical challenges that have been overcome in the development of a cloud-based digital preservation service. They include:

- Enabling a carefully considered exit strategy.
- Allowing multiple storage options driven by an automatable storage policy.
- Allowing different access functionality for different classes of user especially avoiding the need for download where possible.
- Providing full OAIS functionality on top of storage and access (i.e. not just bit-level preservation).
- Separating system-level administration (carried out by the supplier) from tenant-level administration (carried out by the tenant organization).

8. REFERENCES

[1] Adrian Brown. 2013 Practical Digital Preservation. Facet Publishing, London, UK.

[2] Sharpe R and Brown A. Active Preservation. Lecture Notes In Computer Science, 2009, Proceedings of the 13th European conference on Research and advanced technology for digital libraries, Corfu, Greece, Pages: 465-468.

[3] Andrew Charlesworth. 2012. Intellectual Property Rights for Digital Preservation. DPC Technology Watch Report 12-02

[4] Alan Gairey, Kevin O'Farrelly and Robert Sharpe. 2012. Towards seamless integration of Digital Archives with source systems. In *Proceedings of International Congress on Archiving* (Brisbane, Australia, 20-24 August 2012).

[5] Maïté Braud, James Carr, Kevin Leroux, Joe Rogers and Robert Sharpe. Linked Data Registry: A New Approach to Technical Registries. Submitted to iPres 2014.