ACHIEVING SUSTAINABLE DIGITAL PRESERVATION IN THE CLOUD

ADRIAN BROWN, CHRISTOPHER FRYER
PARLIAMENTARY ARCHIVES, UNITED KINGDOM

INTRODUCTION

The Parliamentary Archives of the United Kingdom has been an early adopter of cloud services for the purposes of digital preservation. This paper describes how Parliament is using the cloud as part of its digital repository infrastructure, the rationale for doing so, and how we have addressed the various challenges and opportunities arising from this approach. We conclude that the cloud can form part of a sustainable digital preservation service, and indeed can yield a number of benefits, provided the risks are correctly identified and appropriately mitigated.

BACKGROUND AND CONTEXT

The Parliamentary Archives provides a records management and archives service for both Houses of the UK Parliament. It preserves and provides public access to an archival collection of international significance, including many of the core constitutional records of the UK. It also supports present-day business information management activities within Parliament. A major challenge for the Archives has been to develop the capability to preserve digital records alongside its traditional paper and parchment collections. Since 2010 a staged project has led to the successful implementation of an operational digital repository, which enables Parliament to preserve authentic born digital records and digital surrogates, ensuring access for current and future generations, and mitigating threats such as cultural and technological change and the inherent fragility and mutability of digital information. Parliament's approach to digital preservation is described in detail in Parliamentary Archives (2009).

The Parliamentary Archives acquires growing volumes of parliamentary digital content from a broad range of sources, from born-digital business records managed in Parliament's Electronic Document and Records Management System (SPIRE) and the digital outputs of both Houses' substantial publishing activities, to digitised surrogates of iconic parliamentary documents. The Archives is also responsible for archiving the parliamentary web estate, which consists of the main parliament.uk domain, along with many other sites and third-party channels such as social media. As of July 2014, the Archives had ingested over 10 Terabytes (TB) of records into the digital repository, with at least 80 TB of priority content identified for ingest within the next 5 years. This volume will, of course, grow year by year. For example, with audio-visual content the collection could quickly grow to Petabyte levels. Although the majority of its collections are open to the public, a proportion of material is closed; the digital repository must therefore also enforce access controls.

As part of its new digital repository infrastructure, the use of cloud storage services has enabled Parliament to provide a rigorous preservation storage capability which is flexible, scalable, and cost-effective. The decision to use the cloud in this case must be understood in the context of a wider drive to using cloud services within the UK public sector, and the development of the G-Cloud Framework. The G-Cloud is a purchasing framework established by the UK Government in 2012, to facilitate adoption of cloud services across the public sector. It enables public sector organizations to purchase these services through call-off arrangements, without the need for full tendering processes. Services which are available through the framework can be discovered through an online CloudStore,[1] and encompass Software-as-a-Service, Platform-as-a-Service, Infrastructure-as-a-Service, and specialised services including digital archiving. The G-Cloud framework enabled the Government to introduce a 'Cloud First' ICT procurement policy across the UK public sector in 2013. This policy, which is mandated to central government and strongly recommended to the wider public sector, requires organizations to consider and fully evaluate potential cloud solutions first when procuring new or existing services, although they remain free to choose an alternative to the cloud if they can demonstrate that it offers better value for money. It has been predicted that, as an international trend, governments will spend proportionally more on cloud computing than the private sector, with expenditure projected to increase 35% year-on-year until 2018.[2]

Although not subject to this mandate, Parliament has also chosen to operate a 'Cloud First' policy. While it was decided that the digital repository management system should be managed in-house, the storage platform for the repository content was identified as a candidate for using the cloud. After a thorough options review, it was decided that open content would be stored with cloud storage providers, whereas sensitive closed content would be stored on an internal storage platform. Storage services were procured through the G-Cloud Framework, the first time this approach had been used by Parliament.

ARCHITECTURE

The high level architecture of Parliament's digital repository is illustrated in Figure 1:

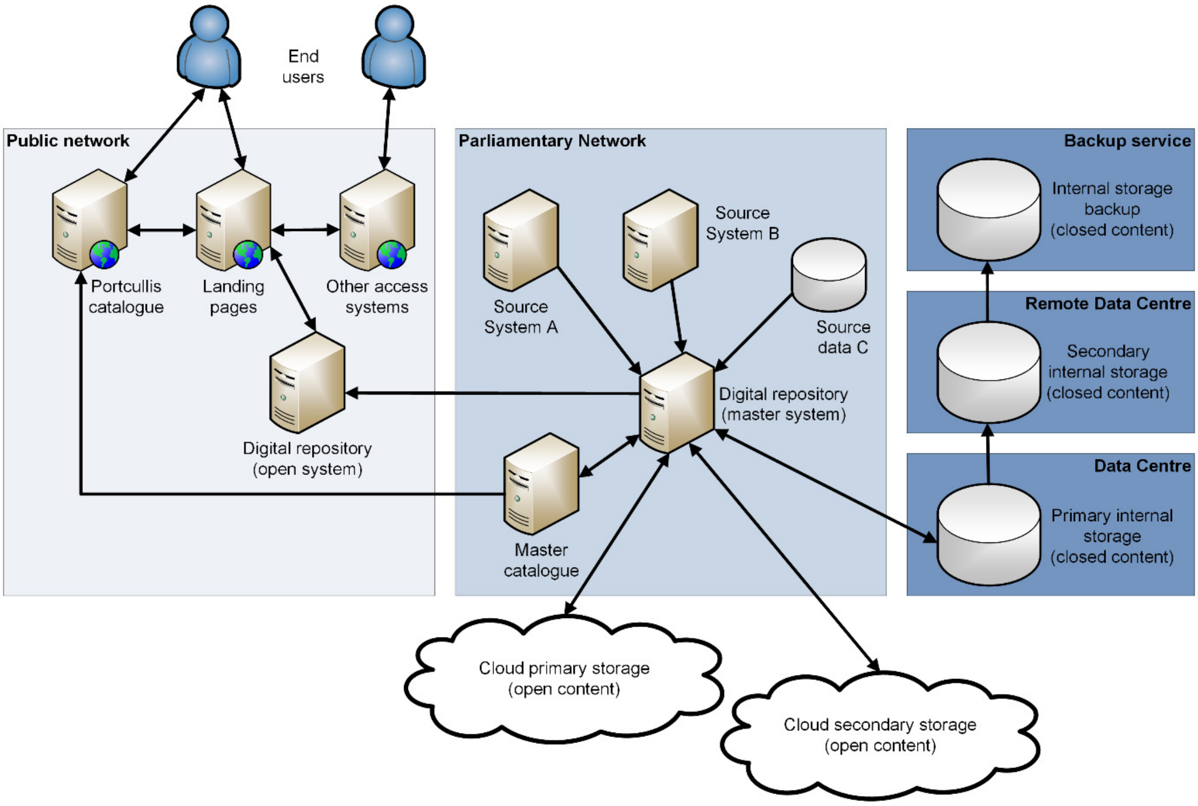
End users
Public network
Portcullis catalogue
Landing pages
Other access systems
Digital repository (open system)
Parliamentary Network
Source System A
Source System B
Source data C
Digital repository (master system)
Master catalogue
Backup service
Internal storage backup (closed content)
Remote Data Centre
Secondary internal storage (closed content)
Data Centre
Primary internal storage (closed content)
Cloud primary storage (open content)
Cloud secondary storage (open content)

Figure 1: High level architecture of Parliament's digital repository

The heart of Parliament's digital repository is a commercial software platform (Preservica Enterprise Edition (formerly Safety Deposit Box) from Preservica, part of the Tessella Group[3]). The repository ingests content from a wide variety of internal systems and data sources. Descriptive metadata about digital objects is managed within the Archives' catalogue management system, alongside descriptions of the physical collections.

Descriptive metadata is mirrored to the public, web-based front-end for the catalogue, called Portcullis,[4] and open content is replicated across to a second, public-facing instance of the repository. Public users can discover this content via the catalogue, from which they are directed to a landing page system which describes how the content can be accessed. In most cases, this provides direct access to view or download the object, but it may also link to an existing access system. In the case of copies which are not available online or where access is chargeable, users are directed to order copies.

Open content is stored in the cloud, with copies of all content mirrored between two different cloud service providers. The two cloud storage services operate on different technology platforms, one being based on EMC Atmos,[5] the other on Amazon S3 Web Services.[6]
Closed content is stored on internal disk-based storage, mirrored between two data centres and with a traditional backup service.

CHALLENGES AND OPPORTUNITIES

Using the cloud as part of a digital preservation strategy creates both challenges and opportunities. The potential benefits of cloud services for the cultural heritage sector certainly seem appealing. In particular, they offer a low technical and financial barrier to entry, which may provide opportunities for smaller organisations to implement digital preservation capabilities which would otherwise be unachievable. Balanced against these opportunities are a number of commons areas of concern held across the sector, with significant issues needing to be understood and addressed. Beagrie, Charlesworth and Miller (2014) provide a recent summary of the issues surrounding the use of the cloud for archives in the UK. The following sections examine how Parliament has mitigated the challenges and embraced the opportunities.

COMPLIANCE AND DATA SOVEREIGNTY

One key concern was to ensure Parliament's ability to comply with relevant legislation, and in particular Freedom of Information and Data Protection laws, as well as addressing data sovereignty. In both cases, these concerns were addressed quite straightforwardly by ensuring that the relevant requirements were clearly defined and subsequently incorporated into contracts. Each contract clearly states the respective roles and obligations of Parliament and the service provider with respect to legal compliance and data sovereignty. It was a specific requirement that all data must be hosted entirely within the European Economic Area, to ensure that it is covered by European privacy legislation and not subject to other jurisdictions, and hence to legislation such as the US Patriot Act.

INFORMATION SECURITY

Information security, including the proper compartmentalization of data, access controls, and monitoring and notification of security breaches, was another important issue. Concerns around information security in the cloud tend to centre more around questions of trust than technical issues – there is no inherent reason for cloud providers to be less secure than in-house infrastructure. Two particular areas of risk specific to the cloud are, firstly, that customers may be indirectly impacted by attacks, such as a denial-of-service, targeted at other, higher-profile customers of the same service provider and, secondly, that compartmentalization of data between customers may be compromised, allowing one customer to access another's data.

Parliament addressed the information security risks in three ways. Firstly, the information security challenge was inherently lessened by the decision to only use the cloud for storing open data which is already in the public domain, with closed material being stored on separate, internal infrastructure. Secondly, as with compliance and data sovereignty, many were primarily addressed through the definition of clear contractual requirements, for example, around monitoring and reporting potential and actual security breaches, security clearances of supplier staff, and compliance with forensic investigations in the event of a breach. Finally, in framing our requirements, we were also able to invoke relevant UK Government information security standards. Services provided through the G-Cloud can be accredited through the UK Government's Pan Government Accreditation Service as meeting the requirements for managing information up to a specified Business Impact Level (a form of protective marking). Thus, some services are only accredited to store information at the lowest impact level (i.e. open data), while others can store much more sensitive material. This provides G-Cloud customers with wide-ranging assurance on information security issues, and greatly simplified the formulation of our information security requirements. It is also possible that Parliament may, in future, choose to store closed content on appropriately-accredited cloud storage platforms.

AUTHENTICITY

The ability to preserve authentic records is, of course, one of the most fundamental requirements of any digital repository. In the context of storage, durability is the most significant aspect of authenticity to consider. Durability may be defined as a measure of the integrity of stored content over the duration of the service period, and encompasses not only the protection of data from unauthorised and deliberate alteration and from bit rot, but also ensuring that it is safeguarded in the event of disaster, supplier failure, or a decision to change provider. The problems of managing the integrity of large volumes of data over extended time periods have been subject to much discussion and analysis. David Rosenthal, co-founder of the Lots Of Copies Keeps Stuff Safe (LOCKSS) digital preservation system[7] has discussed the challenges associated with cloud storage adoption extensively. He states, in relation to the digital preservation problems of anticipating and countering cloud-rot, "At scale this is an insoluble problem, in the sense that you are never going to know that you have solved it."[8] By using the example of the simplest possible model of long-term storage, a black box within which a Petabyte of data is stored and then retrieved in 100 years, Rosenthal extrapolates that this would require a level of durability far exceeding that stated for Amazon's S3 service, the most widely-used cloud storage platform (Rosenthal, 2014). Rosenthal's calculations bring home the inevitable conclusion that a certain degree of loss must be anticipated, given a large enough volume of data stored over a long enough time period. As he says:

> "We are going to lose stuff. How much stuff we lose depends on how much we spend storing it; the more we spend the safer the bits. Unfortunately, this is subject to the Law of Diminishing Returns. Each successive 9 of reliability is exponentially more expensive." (Rosenthal, 2014)

Furthermore, determining the durability of a storage system is problematic. Although storage providers will often cite durability ratings (for example, Amazon S3 quotes 99.999999999% durability, meaning that for every 10,000 objects stored, one will typically be lost every 10,000,000 years[9]) it can be difficult to discover any scientific basis for these claims and, in any event, since they are not typically related to a contractual commitment, their real value is debatable. However, this issue does need to be set in context. Firstly, it is not one which is unique to the cloud: integrity management becomes increasingly problematic at scale however data is stored. Secondly, it is just as much a challenge for physical collections, albeit typically over longer timescales – any archival collection of sufficient age will have suffered losses and damage, and archivists and conservators accept that their role is to minimise such loss, but that it cannot be eliminated altogether. The same mind set is required for digital preservation.

Other, non-technical durability challenges are more specific to the cloud. In particular, there are real risks that data might be lost if a supplier were to go out of business, in the event of a contractual dispute, or at the end of a contract. Despite these challenges, the Parliamentary Archives has taken certain pragmatic mitigations to counter threats to durability. Instead of relying upon a single cloud storage provider, Parliament has procured two, which operate in parallel with all content duplicated between them. Each provider maintains multiple copies of all content, duplicated in at least two geographically-separate data centres, and uses techniques such as erasure coding to provide additional levels of durability. The two providers also operate on entirely distinct technologies, which offers a further degree of resilience and insulation from threats associated with a specific technology. Given that many cloud suppliers are actually reselling services, and often operating out of the same data centres as their competitors, particular care was taken to ensure that the two suppliers were technologically, geographically and organisationally distinct. The use of two suppliers increases the durability of Parliament's storage (although quantifying this remains a future challenge), and provides insulation from supplier failure. The two contracts have also been deliberately offset in time, to minimise the risks associated with any future changes of supplier.

The use of in-house repository software with cloud storage also raises questions about how best to perform integrity checking of stored content. Although Preservica provides a full integrity-checking service, Parliament has chosen not to implement this for our cloud storage, for reasons of technology, cost and performance. Each cloud provider stores multiple copies of each object, but these are not exposed to the end user; hence, there is no means for Preservica to check the integrity of individual copies, only the checksum value which the provider holds for the entire set. Thus, the repository can only ever check between the two providers. Secondly, in order to perform the integrity checking, Preservica would have to retrieve each file from storage. Cloud providers charge for the predominant direction of traffic – ordinarily, for Parliament this would be upload at ingest but, with integrity checking there would be a charge for every download, which would increase the overall cost very substantially. In addition, it quickly becomes impractical to regularly download the entire contents of an archive which will soon be 10s or 100s of TBs in size, across an internet connection, in any reasonable timescale. Instead, Parliament has chosen to integrity check with Preservica on ingest, and whenever an item is retrieved from storage, but not in between. The cloud providers perform regular integrity checking of their internal copies, and we have satisfied ourselves that these methods were robust. In addition, having copies stored with two separate providers, on different platforms, reduces the likelihood of an unrecoverable integrity failure affecting the same object, at the same time, with both providers, to a vanishingly small value.

We do apply Preservica's integrity checking to the internal, closed storage. This is a scheduled task, which runs daily. In order to avoid overloading the network we limit each check to 1,000 files, with each file checked every 30 days, but this can easily be configured.

ELASTICITY AND PORTABILITY

In an environment where the volume of digital content is ever increasing, often at unpredictable rates, the flexibility which cloud services offer is very appealing – it is fast and simple to begin using services, and to change providers, and services can easily be scaled to meet actual demand, without incurring unnecessary costs for unused capacity. This elasticity proved critical in the decision of Parliament to procure cloud storage solutions for the storage of open content. However, it should be noted that some aspects of the cloud's elasticity are more relevant to the context of digital preservation than others. For example, the low barrier to entry and cost model are undoubtedly very attractive for many memory institutions. On the other hand, since most repositories will only ever grow in volume, the ability to scale down as well as up is less significant. There can also be a perceived tension between cultural heritage organisations' mission to safeguard digital material for the long term, and the perceived short term benefits gained through implementing cloud services. As Neil Beagrie comments:

> "The cloud can rarely be beaten as a way to get something up and running quickly, affordably, and with a minimum of fuss. But some of the most compelling attributes of the cloud are best suited to ephemeral or (relatively!) short-term use cases, whereas archives are there for the long-term. Are they compatible?" (Beagrie, 2014)

However, this tension exists throughout the digital domain. The ephemeral nature not only of digital content, but also of the infrastructure required to manage it, is a constant challenge for long term sustainability - the cloud simply highlights this issue. Digital preservation requires constant and proactive management, irrespective of whether it takes place within the cloud or some other environment.

Portability refers to the ease of moving data from one storage environment to another. In the context of the cloud, this is primarily an issue when moving from one service provider to another. Although every provider will have facilities for 'on boarding' and 'off boarding' data, and any cloud contract should include clear provisions relating to their responsibilities in this regard, the logistical challenges of transferring large volumes of data between suppliers are daunting. The data volumes which the Parliamentary Archives expects to store within the next 5 years, while comparatively small compared to many memory institutions, would nonetheless be challenging to move. With currently available bandwidth, it would not be feasible for Parliament to transfer it across an internet connection; the use of physical storage media would be the only option. However, with current technologies, and allowing for integrity checking at either end, such a process would be likely to take weeks or months. One of the principal advantages of the cloud is its flexibility - the ability to easily move between service providers in order to ensure the best value and fit with requirements. However, unless content can be moved between providers within a timeframe which matches that for the change of service, this flexibility becomes a moot point for digital repositories, and the risk of vendor lock-in increases. In the short term, Parliament is managing this through the use of two suppliers, with deliberately offset contract terms – this increases the time window for moving data from one supplier. Our best hope in the long term must be that data transfer technologies, whether online or offline, will keep pace with growth in volume, or that a sufficiently competitive market will emerge to drive service providers to offer better solutions.

ECONOMICS

Cloud services usually operate a ‘pay-as-you-go’ pricing model: the customer typically pays a monthly charge based on the volume of data stored, as well as the number of data uploads and downloads. This is very different to the traditional model for IT infrastructure, which involves periodic upfront capital investment, depreciated over the life of the system, together with annual operating costs. The obvious advantages of the cloud model for customers are that it does not require a substantial initial investment, and costs are directly related to usage, avoiding the risk of purchasing unused capacity. This can be very attractive, especially to smaller organizations which may be unwilling or unable to commit to major upfront investment. However, it can create a much greater degree of uncertainty around medium and long term costs: in a cloud environment, it becomes critical for institutions to be able to accurately forecast future storage volumes and ingest rates and this can represent a major challenge, especially for collecting archives. It is a manageable challenge for Parliament which, as an institutional archive, is generally able to predict future accessions with reasonable confidence. In particular, the vast majority of ingests by volume arise from digitisation programmes, which produce highly predictable outputs. Nonetheless, the move to cloud storage has required a new approach to financial planning, and close liaison between the Parliamentary Archives and Parliamentary ICT to ensure accurate forecasting and financial planning.

Understanding the long-term economic implications of cloud versus traditional models remains a significant challenge. Parliament undertook a cost modelling exercise which indicated that, over an 8 year period, the cloud would be significantly cheaper than in-house storage for the digital repository. However, this calculation inevitably included a number of assumptions regarding both internal and external factors, and its validity therefore rests upon their accuracy, which has yet to be fully tested. Furthermore, the longer-term economics remain much more uncertain. Work by researchers such as David Rosenthal (see Rosenthal, 2014) suggests that the cloud may prove a more costly option for archival data storage over the long term, but further analysis is required in this area, alongside research into sustainable models for funding digital preservation over time – decisions about the economics of the long term data storage need to be based on a thorough understanding of how preservation activities will be funded over the same timescales.

INTEGRATION WITH PARLIAMENT’S ICT INFRASTRUCTURE

In many ways, cloud storage has proven very straightforward to incorporate within the repository infrastructure. Preservica uses the concept of ‘storage adaptors’ - interfaces to different storage environments. An Amazon storage adaptor already existed, and Parliament developed an EMC Atmos adaptor. Preservica can easily be configured to determine which adaptor(s) particular content is stored with, and also provides the means to move content between adaptors.

The main infrastructure challenges have arisen primarily from more generic issues around the nature of Parliament's network. For example, the network makes use of proxy servers, and this required modification of the storage adaptors to make them 'proxy-aware'. In addition, it was discovered that the way in which some cloud providers manage the IP address ranges of their services was incompatible with Parliament's network policies, requiring fundamental changes to elements of Parliament's network infrastructure.

The other challenge was simply one of scale: it was essential to devise an architecture which would allow timely ingest of large data volumes, but without having an adverse impact on other systems within the Parliamentary network. This has required the use of dedicated servers and internet connections. However, although the technical challenges of integrating cloud services with an existing organizational IT infrastructure should not be underestimated, they have all proven manageable.

TRANSPARENCY AND TRUST

Our experience thus far suggests that using the cloud, as Parliament has done, to provide one specific element of the digital repository infrastructure, is in many ways little different to providing that element in-house. However, it does tend to bring the risks and issues associated with digital repository storage in general to the fore, and has ensured that we have fully considered the risks and identified appropriate mitigations, both with respect to the cloud and our in-house infrastructure. This can only be a positive outcome.

Perhaps the most fundamental change necessitated by use of the cloud is the delegation of certain responsibilities to a third party, which requires issues of trust and transparency to be addressed. In particular, it requires customers to ensure that roles and responsibilities are defined with absolute clarity, so that there can be no doubt where the boundaries lie between the obligations and expectations of the customer and the supplier. Allied to this, clear, practical and appropriate service level agreements are essential.

CONCLUSION

The Parliamentary Archives' experience of implementing cloud storage demonstrates that it is feasible to use cloud services within a digital repository environment. The emergence of the cloud can help to lower the barriers to entry to digital preservation, especially for smaller organizations, and can provide a robust, scalable infrastructure for repository storage. As with any approach, it is essential to fully understand the associated risks and ensure that appropriate mitigations are in place. However, the risks arising from the cloud are often similar to those pertaining to other storage technologies, and strategies for managing them are available. In the longer term, questions about economics and portability remain, but these are not, in themselves, reasons to avoid using the cloud.

It should be noted that Parliament's use of the cloud has been limited to one specific facet of the digital repository. The increasing adoption of purpose-built, fully-fledged digital preservation cloud services, such as Preservica Cloud and DuraCloud, undoubtedly raises additional challenges and opportunities beyond those discussed in this paper.

NOTES

[1] See http://govstore.service.gov.uk/cloudstore/.
[2] See Curtis, 2014.
[3] See http://preservica.com/.
[4] See www.portcullis.parliament.uk.
[5] See http://uk.emc.com/storage/atmos/atmos.htm.
[6] See http://aws.amazon.com/s3/.
[7] See http://www.lockss.org/.
[8] See Rosenthal, 2014.
[9] See http://aws.amazon.com/s3/faqs/.

BIBLIOGRAPHY

Beagrie, N. (2014). *Cloud storage and archives: a match made in heaven?* The National Archives Blog. < http://blog.nationalarchives.gov.uk/blog/cloud-storage-archives-match-made-heaven/>. [consulted: 18 July 2014].

Beagrie, N., Charlesworth, A. and Miller, P. (2014). *How Cloud Storage can address the needs of public archives in the UK*. The National Archives: London. <http://www.nationalarchives.gov.uk/documents/archives/cloud-storage-guidance.pdf>. [consulted: 18 July 2014].

Curtis, J. (2014). "Government 'to Surpass Private Sector Cloud Spend'". *Computer Business Review*. <http://www.cbronline.com/news/cloud/cloud-saas/government-to-surpass-private-sector-cloud-spend-4254170>. [consulted: 16 July 2014].

Parliamentary Archives (2009). *A Digital Preservation Policy for Parliament*. London. <http://www.parliament.uk/documents/upload/digitalpreservationpolicy1.0.pdf>. [consulted: 16 July 2014].

Rosenthal, D. (2014). "EverCloud workshop". <http://blog.dshr.org/2014/04/evercloud-workshop.html>. [consulted: 16 July 2014].