Digital Curation and the Cloud

Final Report

Brian Aitken, Patrick McCann, Andrew McHugh, Kerry Miller

Produced by the Digital Curation Centre for JISC's Curation in the Cloud Workshop

Hallam Conference Centre

7th – 8th March 2012

Table of Contents

A draft version of this paper was prepared to inform the JISC Curation in the Cloud workshop on 7th and 8th March 2012.This final version (v1.1) includes revisions and additions that reflect the discussions and outcomes of that event.

The latest version of the document is available from the DCC web site at:

http://www.dcc.ac.uk/resources/publications

1 Executive Summary

Digital curation involves a wide range of activities, many of which may be suitable for deployment within a cloud environment. These range from infrequent, resource-intensive tasks which will benefit from the ability to rapidly provision resources, to day-to-day collaborative activities which can be facilitated by networked cloud services. Associated benefits are offset by risks such as loss of data or service level, legal and governance incompatibilities and transfer bottlenecks. There is considerable variability across both risks and benefits according to the service and deployment models being adopted and the context in which activities are performed. Some risks, such as legal liabilities, are mitigated by the use of alternatives, for example, private cloud models, but this is typically at the expense of benefits such as resource elasticity and economies of scale. The *Infrastructure as a Service* (IaaS) model may provide a basis on which more specialised software services may be provided.

There is considerable work to be done in helping institutions understand the cloud and its associated costs, risks and benefits, and how these compare to their current working methods, in order that the most beneficial uses of cloud technologies may be identified. Specific proposals, echoing recent work coordinated by EPSRC and JISC are the **development of advisory, costing and brokering services** to facilitate appropriate cloud deployments, the exploration of opportunities for **certifying or accrediting cloud preservation providers**, and the **targeted publicity** of outputs from pilot studies to the full range of stakeholders within the curation lifecycle, including data creators and owners, repositories, institutional IT support professionals and senior managers.

2 Definitions of terms

2.1 Cloud

The word 'cloud' has become almost ubiquitous when discussing online technologies and services. One can identify a spectrum of cloud services that ranges from the online word processing provided by Google Docs[1] to the cloud storage and computing provided by the likes of Amazon[2]. However, a definition of 'cloud computing' is not immediately obvious. Cloud involves making use of resources at some remote location across a network. There is some degree of abstraction hiding the actual hardware infrastructure from the user, but when can a web application be considered to be *Software as a Service* (SaaS)? When renting a server from a hosting company, does it qualify as cloud computing if the server is virtual?

[1] http://docs.google.com
[2] http://aws.amazon.com/

The U.S. National Institute of Standards and Technology (NIST) finalised its definition of 'cloud computing' in September 2011[3].

- *On-demand self-service.* Users can access resources automatically as needed.
- *Broad network access.* Resources are accessed over the network using standard tools.
- *Resource pooling.* Resources are shared between users according to demand. Users generally have no or limited awareness of the location of the resources.
- *Rapid elasticity.* Users can easily provision or release resources as needed. The level of resources available can appear to be effectively unlimited to the user.
- *Measured service.* Use of resources is metered, and users are charged on that basis.

2.1.1 Service Models

The above characteristics overlap and are interdependent, but it is easy to see how they fit a service such as Amazon Elastic Compute Cloud (EC2), which allows large numbers of users to create/destroy as many virtual server instances as they wish, only paying for what they use and interacting with Amazon's systems over the internet via their web browser. However, whilst both EC2 and Google Docs are covered by this definition they are clearly very different services. As such, the NIST definition goes on to describe three service models:

2.1.1.1 Infrastructure as a Service

Infrastructure as a Service (IaaS) allows users to access computing resources on which they can deploy software, which can include operating systems. However, users do not have access to the underlying cloud infrastructure. The most prominent example is probably Amazon's growing suite of Web Services (AWS). Amazon EC2 allows users to create and manage virtual server instances, while their Simple Storage Service (S3) provides data storage infrastructure.

2.1.1.2 Platform as a Service

Platform as a Service (PaaS) gives users the ability to deploy applications on the provider's cloud infrastructure using tools supported by the provider. The infrastructure, including operating systems, remains beyond the users' control. Google App Engine[4] is a good PaaS example: users can use it to deploy applications written in Python or Java, making use of a number of application programming interfaces (APIs) built into the platform. Microsoft Windows Azure[5] similarly allows users to deploy applications written in a range of languages.

[3] Mell, P., Grance, T., The NIST Definition of Cloud Computing, 2011, National Institute of StandardsandTechnology, Special Publication 800-145

[4] https://appengine.google.com/start

[5] http://www.windowsazure.com/en-us/

2.1.1.3 Software as a Service

SaaS gives users access to applications running on the provider's cloud infrastructure. Generally accessed using a web browser, these applications often provide functionality that would traditionally be provided by dedicated software on a user's own computer. The aforementioned Google Docs is a one such example; 37 Signals' project management application Basecamp[6] is another. Such applications frequently include collaboration or sharing features that would be more difficult to implement in desktop software.

2.1.1.4 Relationships between the three service models

In principle these service models are agnostic regarding what users do with them. Also, users of a cloud service should be unaware of, or at least have no need to be aware of, the infrastructure on which that service is built. It is possible for PaaS or SaaS services to be built on an IaaS service, or to create SaaS applications on a PaaS platform. The former in particular may be desirable in order to achieve the flexibility of capacity that user-created services may require.

2.1.2 Cloud Deployment models and Provider Characteristics

The NIST report also defines four deployment models:

2.1.2.1 Public Cloud

Public cloud services enable the cloud infrastructure to be used by the general public and they offer users the benefits of rapid scalability and low initial set-up costs. Google Apps is a popular suite of SaaS applications made available through a public cloud.

2.1.2.2 Community Cloud

With community clouds the provided infrastructure is provisioned for exclusive use by a specific community from organisations that have shared concerns. As access is restricted to particular users, community clouds may present fewer security risks than public clouds, however with more limited economies of scale they tend to be less flexible in adapting to user needs and costs can be higher.

2.1.2.3 Private Cloud

A private cloud is owned and operated by an organisation, or a third party on behalf of an organisation. For this reason many of the benefits of cloud computing such as outsourcing IT infrastructure and economies of scale will be less evident, however security risks and data transfer bottlenecks may be mitigated.

2.1.2.4 Hybrid Cloud

Hybrid clouds are a composition of two or more distinct cloud infrastructures (private, community, or public) that although remaining distinct are interconnected to allow data transfer.

[6] http://basecamphq.com/

A private-public hybrid cloud may be used to store sensitive data in-house while outsourcing storage of other data.

2.2 Digital Curation

For the purposes of this document we adopt a broad definition of digital curation, aligned principally to the Digital Curation Centre's Lifecycle model[7], which covers stewardship of data from the point of conceptualisation right up to its eventual disposal. In contextualising the challenges associated with digital curation we look to a range of communities. Traditional memory institutions with mandates to ensure continued accessibility and usability of digital resources are joined by higher and further education institutions, responding to funders' increasingly strict data sharing and management demands.

3 Digital Curation and the Cloud

3.1 Cloud Approaches

3.1.1 The Eduserv Education Cloud

The Eduserv Education Cloud[8] was launched in January 2012 as a community cloud which developed from the UMF Cloud Pilot to serve the higher education community in the UK. The Education Cloud is an Infrastructure as a Service (IaaS) model and will offer participating institutions fast access to cloud storage and computing at times of peak need. The choice of pricing models (pay-as-you-go or virtual data centre) will allow institutions to control their costs and hopefully achieve significant efficiency savings while continuously improving the IT services available to students, researchers, and other staff. However, the Eduserv Education Cloud still has to prove itself and will be closely monitored over the coming months and years to see if it delivers on its early promise. Recent comments from Eduserv's Andy Powell at the JISC Curation in the Cloud workshop[9] suggested an intention to increasingly introduce a PaaS dimension to their portfolio of services. Eduserv can be used at the time of writing, but a lack of self-service interface and billing mechanism (both due for imminent release) means the provision currently falls somewhat short of what most expect from cloud.

A potentially critical aspect of Eduserv for JISC's community is their commitment to store all data within the UK. When data is stored in a public cloud such as Google Apps or Amazon S3 there is no control over its exact location or method of storage. Service level agreements may

[7] http://www.dcc.ac.uk/resources/curation-lifecycle-model
[8] http://www.eduserv.org.uk/hosting/cloud-computing/education-cloud
[9] See http://www.jisc.ac.uk/events/2012/03/curationinthecloud.aspx

limit liabilities and should be considered. For example, the Google Apps terms of service[10] state that if Google Apps damages, destroys, loses or prevents access to data users have no rights to compensation or redress, regardless of disruption caused. Concerns over data security in the cloud, or of jurisdictional implications associated with the cloud's physical location may in some cases make tasks with associated legal implications less well suited to that model. Therefore, tasks associated with sensitive or personal data, data with associated intellectual property considerations or with high compliance or governance requirements may be incompatible with one or more cloud services or service providers. Related to this may be issues of compliance or governance that demand conformity with criteria on information storage, including jurisdiction. By their nature, some cloud services will be implicitly incompatible with such requirements. As noted below, however, the curation community relies upon a number of successful third party hosted and remotely accessible applications for planning, characterisation and validation, which is indicative of a cultural enthusiasm for SaaS and SaaS-like service models.

The Eduserv cloud is operated for the long term benefit of Higher Education and runs directly on the JANET backbone. It has been a challenge to bring prices down to match commercial operations like Amazon. Nevertheless, Eduserv's Education cloud can be considered a good alternative to commercial cloud providers. Services are offered freely on a trial basis and latterly with pay-as-you-go, monthly and annual cost models for processing power and storage. Although its costs appear higher than those of other providers being part of the JANET network may imply cost and time savings for data transfer.

3.1.2 University of Oxford Shared Data Centre

The University of Oxford's Shared Data Centre was established in 2011[11] and provides a private cloud for use by the various colleges of the university. A shared pool of computing resources is offered through an IaaS service model enabling research projects and departments to straightforwardly purchase and manage virtualised computational and storage resources on demand. Although the cloud is only available for internal use at the present there are further plans to expand towards a hybrid model, designating part of the cloud as public, moving non-sensitive data to this potentially third party hosted section and opening up the cloud to other organisations.

Private and community clouds tend to operate from a small number of data centres, often just one, so any problems encountered at a data centre, such as power outages or network failures, will have a significant effect on service availability. Public clouds may be considered more likely to offset such risks by distributing data and services across multiple data centres and ensuring data and compute redundancy. The cloud may be a legitimate choice for off-site redundant storage to complement an otherwise locally established system architecture.

[10] http://www.google.com/apps/intl/en-GB/terms/standard_terms.htm

[11] Curtis, S., Oxford University Builds VMWare Private Cloud, October 2011, TechWeek Europe

3.1.3 Kindura

Kindura[12] was a pilot of a hybrid cloud model combining both internal services and a variety of cloud services, both IaaS and SaaS, brokered by Duracloud[13] (see below). The Duracloud service manages the interoperability of these different services ensuring that end users are presented with standard interfaces and seamless access to resources regardless of where and how they are stored or processed. Commercial cloud and iRODS[14] services are combined, providing a framework to manage storage across multiple providers and access to common preservation services. Storage decisions (local/cloud) are automated according to a set of rules implemented in iRODS. Only binary objects are stored in the cloud; associated metadata and Fedora objects are stored locally. The rules engine uses user metadata, extracted metadata and institutional and provider policies to determine where data should be stored. Rules can be changed without rebuilding code, and can also prompt users to perform migrations, or perform migrations automatically. The hybrid cloud model enabled researchers to make use of the increased storage and computing capabilities of the cloud while keeping sensitive data in-house. Kindura has successfully demonstrated that cost savings can be made using this model and also that services of this type can result in more efficient usage of internal storage, thereby potentially reducing the overheads resulting from retaining internal storage and processing capabilities.

3.1.4 Using the Cloud to complement existing systems

Cloud adoption need not be considered as an all-or-nothing process. Kindura illustrates that different cloud services can be effectively combined, and similarly there are numerous examples of cloud services integrating effectively with offline local provisions. At the recent JISC *Curation and the Cloud* workshop Kris Carpenter explained that the Internet Archive stores its core collection in its own data centres, but locates metadata, indices and various subsets in the cloud, using cloud resources to perform various curation tasks, including automated metadata and link extraction, format migration, aggregation and access. Despite several benefits such as programmatic interfaces for both storage and processing, technical support and operational troubleshooting, and data reliability and availability, there is some danger associated with being reliant on a commercial provider. When considering outsourcing computing resources to a cloud provider the risk that the service may become unavailable must be addressed. Entrenchment or 'vendor lock-in' can be a risk, where lack of standardised APIs and services mean potentially prohibitive costs for changing supplier. The European Union has a Cloud Computing Strategy[15] which highlights the need for standardisation and interoperability to increase uptake of cloud

[12] Stewart, A., Kindura, 2011, JISC InfoNet Case Studies
[13] http://www.duracloud.org/
[14] https://www.irods.org/index.php/What_is_iRODS%3F
[15] http://ec.europa.eu/information_society/activities/cloudcomputing/index_en.htm

services across the EU. The IEEE Standards Associations is also working towards Cloud Interoperability Standards[16].

Carpenter also outlined some of the difficulties associated with bandwidth – it makes great financial sense to host cloud services in Ireland, but for a US organisation this has implications for data transfer. Calculating bandwidth requirements is complex[17], and although the cloud's elasticity is an important benefit, bandwidth to and from the cloud is by its nature much less flexible[18]. It may be necessary to control and prioritise certain types of internet traffic to keep costs down while ensuring that important services are kept available at all times. Ultimately skilled IT networking professionals capable of managing the demands on bandwidth are a necessity to make the most of cloud services[19]. Where curation tasks are dependent on others one assumes inputs and outputs, associated data exchange and in turn possible bottlenecks. Deconstructing preservation workflows into discrete services is tricky, with many visible overlaps, and integrating them across the cloud likewise. Nevertheless, during a recent email exchange on the JISC MRD mailing list, Keith Jeffery of STFC and EuroCRIS pointed out the opportunities for flipping the status quo somewhat, and moving services to where data reside, whereby any overlap could be more comfortably managed. We already see evidence of this in big-data science with instruments deployed to where the data are, and where they are subsequently, at least partially, processed. Even where data cannot be created within a cloud environment (e.g. using SaaS tools) there appear to be few additional complications involved in their transit from a particular remote process or instrument to cloud storage, rather than to traditional storage provisions.

At the University of Hull proposals to use the cloud to complement an existing Fedora repository are gaining traction. The cloud is considered a cheap storage opportunity (more so than SAN) and an appropriate choice for large, seldom-accessed resources. Fears over the cost implications of frequent cloud accesses limit its viability for other content. Datasets are expected to only increase in size and the cloud's possible role is at the forefront of thoughts. For tasks that are infrequent and/or difficult to anticipate and plan for, the cloud is a good fit, offering resource elasticity and a metered charging model. It removes the requirement to maintain systems capable of coping with peak capacity, allowing extra capacity to be provisioned quickly when required. Similarly, batch processing of large datasets including format migration, optical character recognition, image recognition or the creation of search indices are tasks that are perfectly suited to the processing power of the cloud, as is the execution of statistical or analytical tools on datasets. Tasks that can be processed in parallel can take advantage of load balancing, flexible scalability and other optimisation offered by the cloud to accomplish tasks more quickly and cheaply. Interactive, application-style processes are generally less able to

[16] http://standards.ieee.org/news/2011/cloud.html
[17] CGNET, Estimating Bandwidth for the Cloud, March 2011, CGNET Services International
[18] Gittlen, S., Bandwidth Bottlenecks Loom Large in the Cloud, January 2012, Computerworld
[19] Broadhead, S., Cloud Computing: How to Avoid a Network Bottleneck, May 2011, Computer Weekly

capitalise on such opportunities, and are more suited to traditional service models, or the highest level SaaS cloud model. The Washington Post's migration of content from image PDF to OCR text via the cloud is illustrative of its value for tasks such as information transformation and migration. As a general rule, the cloud can only optimise tasks by providing greater quantities of more efficient compute resource (storage, CPU cycles etc) so where bottlenecks occur as a result of necessary human intervention there are fewer opportunities, unless these can be resolved with parallel or concurrent use.

Concerns nevertheless remain at the University of Hull about security and access restrictions, and also durability. When dealing with many tera- or even petabytes of data even near-perfect file durability or uptime implies loss of notable quantities of content or service. In 2011 several cloud providers suffered from high profile outages[20] which resulted in some customers losing all access to their cloud-based data and services for several days. Worth considering is that local (organisation-side) network connectivity issues will likewise restrict access to cloud-hosted resources. In more positive terms, although for example Amazon's service's outage in 2011[21] affected some customers for more than 72 hours, in general cloud providers offer uptime that few in-house IT services can match. Service level agreements (SLAs) of most major cloud providers advertise an annual uptime percentage of around 99.9%, with credit offered if this level is not met.

More generic tasks will of course be better served by widespread cloud infrastructure. More niche tasks that may nevertheless be considered a good fit for cloud deployment may demand the establishment of new cloud services. Tasks such as storage, data migration (otherwise known as format conversion), community watch (customer relationship management) and disposal (deletion) have applicability far beyond the immediate scope of data preservation, and therefore, on that basis at least have more immediate cloud viability. Conversely, more explicit preservation tasks such as content description and representation information definition are not currently served by any explicit cloud service. Existing registry-type resources which have had substantial investment, and offer several cloud-like benefits (such as widespread networked availability) including GDFR[22], PRONOM[23] and the DCC/Caspar tool RRORI[24] may be usefully reshaped as cloud services, but this implies further investment.

[20] Perdue, T., Cloud Computing Service Outages in 2011, 2011m About.com, New Tech
[21] Metz, C., Amazon Cloud Fell from Sky After Botched Network Upgrade, April 2011, The Register
[22] http://www.gdfr.info
[23] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[24] http://registry.dcc.ac.uk:8080/RegistryWeb/Registry/

3.2 Cloud Standards

3.2.1 Technology Platform Standards

OpenStack[25] is an open source IaaS project founded by Rackspace Hosting and NASA, incorporating code from their existing platforms. Over 150 companies have joined the project, including Intel, Dell, AMD, Canonical, Cisco, HP, SUSE, Broadcom, AT&T and Deutsche Telekom. Microsoft have also indicated their support for the project. There are three core components to the project, intended to be of use to institutions or service providers seeking to provide a cloud platform using existing physical hardware:

1. The OpenStack Compute software is designed to manage networks of virtual machines in order to create a cloud platform.
2. OpenStack Object Storage is for creating scalable, long-term object storage using clusters of standardised servers.
3. The OpenStack Image Service provides discovery, registration, and delivery services for virtual disk images.

The APIs used by the OpenStack components are compatible with those used by Amazon EC2 and S3, and the intention is to foster standards and reduce the risk of lock-in via the project's open development model. It is likely that no single provider will remain the most attractive over the lifetime of a cloud-based system, and the widespread implementation of open standards for cloud interfaces would greatly mitigate the sustainability risks associated with cloud solutions by making it easier for users to switch providers and for organisations to provide their own cloud platforms which are compliant with existing applications.

The OpenNebula[26] open source project seeks to develop an industry standard solution for building and managing virtualised data centres and cloud infrastructures. It aims to provide a management layer for the operation of such systems by making use of existing infrastructure. The primary use is in the provision of private clouds, but it provides support for implementation as part of a hybrid cloud solution or in the provision of a public cloud. It offers a choice of interfaces, including one compatible with Amazon's EC2 Interface, and hypervisors (or Virtual Machine Managers). It is used as a component in a range of cloud projects, including StratusLab and SLA@SOI. The use of software such as OpenNebula should help to minimise the cost to an organisation of providing a cloud platform by removing or reducing the need to buy new hardware in order to do so.

[25] http://openstack.org/
[26] http://opennebula.org/

3.2.2 Adapting OAIS to the Cloud

Researchers at the University of Tsukuba, Ibaraki, Japan[27] have been working to explore the viability of aligning the Reference Model for an Open Archival Information System (OAIS), which provides what some consider a definitive functional specification for preservation, to cloud architecture and approaches. Their conclusion is that despite some apparent incompatibilities, notably the often synchronous nature of preservation workflows, there is value in adopting a layered model with aspects of trusted bit level storage/API, information description and function distributed across PaaS and SaaS layers of the cloud architecture. Some of the principle benefits they identify with the cloud model include:

1. The use of a shared platform enables stakeholders in preservation to enjoy continued access to data in order to facilitate its creation, documentation and associated preservation planning;
2. Assuming a trusted storage platform, encapsulation of preserved digital objects plus associated metadata is not required; instead a URI can adequately indicate the location and identity of stored bit streams;
3. Incorporating those aspects of preservation workflow not covered by OAIS (conceptualisation, creation, use) into the cloud model informs other aspects of preservation.

IaaS is disregarded in Askhoj *et al*'s model, as it is considered irrelevant to archivists, beyond the requirement that any chosen infrastructure must provide the basis for trustworthy storage. The model is accompanied by an elaboration of how metadata may be procured or generated. Whether generated at the point of data creation or declaration, pre-registered based on, for example, information about data types contained in representation information registries or arising from preservation events or interactions these processes can be characterised as cloud services, along with storage and virtual packaging.

3.3 Brokerage Services

3.3.1 Duracloud

Duracloud[28] is a software platform and cloud brokering service that enables content to be stored across multiple cloud providers for preservation, with a portfolio of features including backup and replication, cloud-provider renegotiation, integrity checking, synchronisation and content sharing. Their suite of pricing plans ranges from either basic archiving, which amounts to backup and integrity checks and limited sharing, or basic media access which removes integrity checking but adds richer features for content access/transformation (both $375 pm / $4500

[27] Jan Askhoj, Shigeo Sugimoto, Mitsuharu Nagamori, (2011) "Preserving records in the cloud", Records Management Journal, Vol. 21 Iss: 3, pp.175 - 187

[28] http://www.duracloud.org/

annually) to a more complete professional preservation solution which combines all the features of the other packages and adds redundant distributed backup and synchronisation ($599 pm / $7000 annually). The principle behind Duracloud appears to be to simplify preservation to a one-click process. Its storage/backup coverage appears impressive and it facilitates preservation planning with integral tools for image transformation, file identification and sharing. However, any click-and-forget approach will by its nature fall short of the expert stewardship that distinguishes storage from true preservation. Duracloud is a potentially very useful tool for preservation and curation practitioners, but the extent to which it completely frees the institution or individual from more hands-on, expert or context-dependent tasks like preservation planning and selection may be questioned. Its main strengths are the assurances it can offer for data integrity, and the extent to which it relieves the burden of hardware acquisition and maintenance and cloud vendor negotiations. Any risk to the viability of the Duracloud service will ultimately impact on its trustworthiness. Relieving burden from end users also implies abstraction and a diminished opportunity to build relationships with providers, and a continued reliance on Duracloud. Despite the software component of Duracloud being open-source, it is arguably the service component that is most critical for long term preservation. Users willing to utilise cloud provisions for external hosting may be less comfortable in positioning themselves two steps away from their data. DuraCloud asserts that there is no danger of lock-in and specifically that users will continue to have access to their data in the event of the DuraCloud service ceasing to be available, but it is not clear how the apparent tension between this and the brokerage aspect of the service is resolved.

3.3.2 JANET Brokerage

The Janet Brokerage[29] service is aimed principally at UK HE/FE organisations that are intending to undertake large-scale migrations of services to an IaaS Cloud provider. The service offers a procurement framework through which organisations can define their requirements and decide which cloud supplier best suits their needs. The brokerage service has engaged with eight cloud suppliers which are available through the framework, including Eduserv, Dell and Verizon. To be offered through the framework each supplier had to pass through a rigorous selection process to ensure they adhere to quality management and security standards, comply with the data protection act, have documented energy efficiency policies, are available through the Janet Network and have servers located within the EU. The brokerage service is offered free of charge to HE/FE organisations and is sustained in part by a 2% fee levied from the cloud supplier chosen by the organisation.

The brokerage service offered by Janet is markedly different from that offered by Duracloud, providing help and advice on the initial selection of a named cloud provider rather than acting as an ongoing singular point of access through which the services of one or more providers are largely abstracted. The availability of impartial, expert advice during the cloud procurement

[29] http://www.janetbrokerage.ac.uk/

process may be hugely beneficial for organisations and assurances that the providers will have met certain legal obligations and other benchmarks could also prove advantageous.

A big consideration for cloud users is trust. Are services or providers sufficiently trustworthy to reassure users that their faith is not misplaced? Where tasks are considered mission critical, this becomes profound. The preservation community has a number of formal and de facto standards to reference in pursuit of assurances that archives and repositories are providing a safe pair of hands. It would appear appropriate for cloud providers to be exposed to similar scrutiny. Needless to say, services such as the provisions of persistent storage will appear more viable with explicit service level agreements and contractual arrangements that define physical infrastructure, associated disaster planning, escrow, exit and succession arrangements. It remains to be seen whether the Janet Brokerage service will be sustainable and continue to offer support to organisations beyond the initial procurement phase and into the long term but there may be a broader role for such a service in the certification or accreditation of cloud providers, products or terms of service.

3.4 Cloud Applications – the UMF Cloud Case Studies

Supported by the Universities Modernisation Fund, JISC has invested recently in four cloud-based systems that individually illustrate some of the opportunities for embedding several aspects of the research data management lifecycle into the cloud. The following sections summarise each of these, offering an explanation of why they are a good fit for the cloud, and describing which parts of the lifecycle model they aim to satisfy. These examples are illustrative of functions that the cloud appears equipped to provide, but should not be considered an exhaustive list; as described above the diversity evident in different practical manifestations of the curation lifecycle make blanket statements on cloud suitability rather redundant.

3.4.1 *BRISSkit*

The Biomedical Research Infrastructure Software Service (BRISSkit)[30] kit is a prototype open source IT infrastructure for Biomedical Research Informatics. It is being developed by the University of Leicester and the Leicester Cardiovascular Biomedical Research Unit (BRU) based at the Glenfield Hospital in the University Hospitals Leicester NHS Trust. It seeks to share the benefits of the BRU's Biomedical Research Informatics Centre for Cardiovascular Science (BRICCS). BRICCS is a research database linked to clinical data repositories, blood and DNA samples, genomic data and material extracted from documents and image stores. BRISSkit provides cloud-hosted software as a modular service so that researchers can use only those parts of the toolkit that are relevant to them at a particular point in time.

BRISSkit includes two data warehouses, one within the hospital and one maintained by the university. Patient data in the hospital warehouse is anonymised and pushed to the university

[30] http://www2.le.ac.uk/offices/itservices/resources/cs/pso/project-websites/brisskit/brisskit

warehouse so that it can be used in research and combined with data from a range of other sources. BRISSkit also provides a range of interfaces for the creation or receipt of research data from interviews, surveys, biospecimen inventories, genomic data and registries of clinical trial participants. A range of potential sources for clinical data can also be added. The combinations of data available bridge the clinical and university domains and allow sophisticated cohort selection criteria to be applied by researchers.

The data warehouse available to university researchers is hosted on a cloud infrastructure. BRISSkit facilitates the receipt and storage of data as well as access to it for researchers. It also facilitates data transformation in two key ways, firstly in the anonymisation of clinical data and secondly in the combining of data from disparate services for use in research.

3.4.2 *VIDaaS*

The Virtual Infrastructure with Database as a Service (VIDaaS)[31] Project at the University of Oxford aims to deliver a 'Database as a Service' (DaaS) hosted on a hybrid cloud infrastructure. The DaaS is a web-based system which enables users to build relational databases (or import existing ones). Generic data creation, updating and querying interfaces allow users to develop their own web front-ends for DaaS databases. Databases are hosted and maintained centrally and are routinely backed up. Access controls can be applied to define who can edit or view data, allowing users to share data with colleagues or even to make it public. Billed as SaaS, it could be argued that what is actually being offered is a PaaS on which researchers can create their own applications. Certainly, provision of data storage facilities of some form tends to form part of PaaS offerings.

In terms of the DCC Lifecycle Model, the generic interfaces allow it to enable the creation as well as the receipt of data. That it offers storage capability is clear, and the interfaces facilitate accessing data and its reuse. There is potential for data transformation and the tool supports metadata capture.

It is not clear at this point how VIDaaS will be made available to researchers beyond the test group. A range of options is available, from making the VIDaaS software available for others to install to deployment as a national service. The idea of DaaS makes sense, allowing users to quickly construct databases, manage data inserts and then disseminate without having to worry about the underlying infrastructure and resources. However, it could be anticipated that researchers and their institutions, rightly or wrongly, may be reluctant to place data in a service deployed above the institutional level. There could well be concerns about privacy and the security of data. There are limitations as to the kind of research data the service would be suitable for. If work produces large volumes of data, transferring that data across the network would be difficult, and it is not clear how well the infrastructure could cope with that. It may be necessary for the service to be used in conjunction with some sort of file hosting service to

[31] http://vidaas.oucs.ox.ac.uk/

enable the storage of items unsuited to storage in a relational database and/or which are too large to transfer to the database.

3.4.3 *Smart Research Framework (SRF)*

The SRF[32] project brings together a number of existing collaborative data management tools for scientific research and makes them available through the SaaS model. The main focus of the suite of tools is the automated and manual creation and sharing of data from experiments as blog posts. The framework includes broker software that can automatically extract structured data from scientific instruments and publish these measurements on a blog for collaborative reuse. Through the blog interface it is possible to link up processes, data and analysis and formally structure the recording of experiment data and the interconnections between experiments, their data and the equipment used.

Within the curation lifecycle the SRF tools would be principally used for the tasks of receiving and ingesting of data, where experiment data during the process of upload to the cloud would be formatted in a standardised manner, and also the tasks of data storage, access and reuse, with the data being stored within the cloud in a secure manner and being made available to suitable users for direct use and for reuse as the basis for further experimentation.

Collaborative tools such as the SRF are particularly well suited to deployment in the cloud, enabling users from many institutions to work together via different types of internet enabled devices. However, it has yet to be demonstrated exactly how SRF will operate within a cloud environment and details about the costs of storing experiment data and any limitations imposed on data access and reuse are not currently available. As one of the principal aims of this project is to enable the automated upload of experimental data to the cloud there are also further questions about the scale of this data and the frequency of uploads and what implications the speed and cost of data transfer may have on the project.

The cloud provides a particularly good fit for tasks that are by their nature collaborative, although such workflows have been well served by networked collaborative environments to date. We see the benefits for collaboration in the SaaS model in tools such as Google Docs and Virtual Research Environments like MyExperiment. Even repository systems like ePrints and Fedora are compelling for collaborations in data conceptualisation, creation, documentation and dissemination. Another facet of collaboration is multidisciplinary research, and cloud and other distributed forms of data management and availability, such as the GRID, have a useful role and support, for instance, linked open data. Documentation (for example metadata creation) can be particularly enhanced with the adoption of a shared platform supporting information access.

[32] http://www.mylabnotebook.ac.uk/

3.4.4 DataFlow

The DataFlow[33] project at the University of Oxford seeks to develop and promote DataStage and DataBank as free-to-use cloud-hosted systems to facilitate the management, preservation and publication of research materials. DataStage is based on research data management infrastructure developed by the JISC ADMIRAL project. It provides a secure 'local' file management environment for use at the research group level, appearing as a mapped drive on the user's computer. It provides additional Web access and DropBox integration, and can be configured to provide private, shared, collaborative, public and communal directories with simple access controls. It is secured by means of automated daily backup, and makes use of cloud infrastructure to obtain additional storage space as necessary. DataBank is a virtualised, cloud-deployable institutional research data repository based on the databank created by the Bodleian Library at Oxford. Institutions can choose to deploy DataBank in the Eduserv cloud or on their own infrastructure, and it can be used with or without DataStage.

There are a number of cloud storage services, such as the aforementioned DropBox, which map to a local drive on the user's machine for simple integration, but DataStage extends that model with features tailored to the specific needs of researchers. It enables the creation and receipt of research data while DataBank offers a range of functionality expected from an institutional repository including ingest, storage, preservation action (including working with metadata), and access. The elasticity of resources makes a cloud infrastructure a good fit for both DataStage and DataBank.

3.5 Cloud Sustainability and Business Models

3.5.1 Summary of Cloud Costs

Given their scale of operation cloud providers can leverage economies of scale to procure facilities such as bandwidth, storage and administration at a lower cost than would be possible for smaller organisations. The savings can be passed on to end users, who also benefit from pooling of resources in other ways, with acquisition, management and maintenance costs often negated when opting for cloud. Data storage itself can be considerably cheaper than in-house, and in addition to this the tools a user may wish to use in order to process or analyse the data can utilise multiple virtualised servers within the cloud to perform tasks at a substantially faster rate.

Costs vary significantly in the cloud depending on the service model, the scale of the data, the required processing power and the duration of the task in question. Many SaaS businesses offer free trial versions of their applications, but longer term and more established use often requires a monthly or yearly fee per user. For example, Microsoft Office 365, incorporating email, web-based Office suite and Sharepoint intranet software costs $10 per user per month,

[33] http://www.dataflow.ox.ac.uk/

or $24 if online editing of Office files is required. Google Apps for business by comparison costs $5 per user per month or $50 per year, with discounts offered for non-profit organisations and higher education institutions.

With PaaS and IaaS approaches computing resources such as processing power, memory, storage and both incoming and outgoing bandwidth tend to be billed separately, enabling users to plan for a usage level that will be most suitable to their needs. Other extras such as load balancing software, database applications and monitoring facilities may also be purchased. Free trials are also often offered, for example Amazon Web Services offer 12 months of capacity-limited free usage to new users, with any usage over capacity being charged at their standard pay-as-you-go rates.

Establishing a private cloud requires a single organisation to purchase and manage the complete infrastructure associated with cloud services. Private clouds tend to develop out of pre-existing private data centres where many of the requirements of hosting a cloud may already have been met.

3.5.2 Cost Analysis of Cloud Computing for Research

Consultants Curtis and Cartwright's 2012 JISC report into cloud computing for research[34] describes how explicit cloud costings have revealed users' general ignorance about many hidden costs associated with the provision of computing services. When outsourcing to the cloud, overheads like power, buildings, administration, maintenance, backup and software licensing are also transferred and any comparisons between cloud and in-house IT resources should consider this. As a utility, cloud computing resource is never 'owned' by the user and this implies a required ongoing payment. This may present a risk to long-term access and be an issue for relatively short-term projects that may expect an organisation to offer some level of storage and resources beyond the operational existence of the project. David Rosenthal has presented a case to suggest that cloud storage is currently "too expensive for long term use" in comparison with the capital and running costs associated with local storage[35]. Furthermore, deployment on the cloud may present additional costs. The requirement for dedicated technical support staff is unlikely to diminish completely, irrespective of the extent to which the cloud is utilised. The role for those supporting research and associated data curation and preservation activities appears likely to change, but the extent and nature of this change remain somewhat unclear. For tasks with potentially high demands for expert support, such as data creation, interpretation, selection and appraisal, the cloud model will not really offset some of the most considerable costs. Where dedicated support is considered to be proportionately less critical, associated principally with the maintenance of more generic services that can be recreated wholesale on the cloud (for example storage), there may be available resource savings, and

[34] Hammond, M., Hawtin, R., Gillam, L., Oppenheim, C., Cloud Computing for Research Final Report, 2010, Curtis and Cartwright Consulting

[35] See http://blog.dshr.org/2012/02/talk-at-pda2012.html

more compelling resource justification. In addition, bandwidth costs associated with getting data into and out of the cloud should not be overlooked. Although many cloud providers offer unlimited amounts of free incoming data bandwidth most currently charge a per-gigabyte fee for outgoing bandwidth usage and such costs may easily mount up.

3.5.3 Sustainability Implications

The Blue Ribbon Task Force for Sustainable Digital Preservation and Access published its final report in 2010[36], asserting three principle actions required for sustainability; the articulation of a compelling value proposition, the provision of clear incentives to preserve and the definition of roles and responsibilities to ensure an ongoing and efficient flow of resources to support preservation throughout the lifecycle. The cloud approach certainly doesn't guarantee the success of such actions, but likewise does not appear to be a barrier to their accomplishment. Sustainability of individual cloud instances, service providers or applications is far from synonymous with sustainable access assuming their opportunities are exploited in an appropriate fashion.

Despite its core purpose, one might argue that not all intrinsic tasks associated with data preservation/curation have an inherent sustainability requirement. However, whether the issue is explicit, such as in the case of storage (which *must* be expected to persist) or implicit, for example with data transformation (where at the very least a record of what took place should persist) there is likely be some kind of associated requirement. Many institutions accept, almost as an unwritten rule, that at the end of an individual project's lifetime their central infrastructures take responsibility for data management and sustainability. Indeed, this is increasingly demanded of them by funders. If individual projects or researchers are contracting directly with cloud service providers for creation and management of their live project data there will not be the same assumption in favour of free longer term management; it will simply no longer be available, at least until explicit funding is acquired and invested. On one hand researchers may welcome the greater transparency of pricing that the cloud implicitly provides, but they may be wise to be careful what they wish for, since everything with a price tag also has a cost.

A potential barrier that relates very closely to the need for clarity of roles and responsibilities emerges from the fact that the cloud requires very different skills to those required for managing in-house IT services and its adoption may require IT professionals to move towards often dramatically different roles, such as facilitation or brokering. The extent to which the skills of traditional system administration staff are easily transferable to the cloud model will be influential in evaluating issues of sustainability. Similarly, institutions moving to the cloud may be left with redundant hardware and software and potentially continuing service contracts. Notwithstanding benefits of cloud for data processing and storage, transferring large datasets between cloud-based and local storage systems may be costly and problematic.

[36] http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Another direct issue of sustainability relates to the continued availability of the data resources that one purports to preserve. Cloud models appear focused on acquisition, and shortcomings in network bandwidth for example can be very easily overcome with the adoption of so-called "FedEx" models, where the one-off bandwidth capacity of physical data transit trumps that available across most networks. At an indeterminate point in the future, when a project or institution wishes to remove its collections from a cloud provider, or transfer to or integrate with another, there may be no comparable model. Lack of standardisation across cloud providers and APIs conjures similar fears of technological entrenchment.

3.6 Technological Considerations

Many of the tasks involved in data curation involve the processing, manipulation or analysis of data objects, and such tasks would generally be performed by a computer with varying levels of input from the user. Examples include data creation, data validation, format migration and assigning metadata. When considering whether or not to employ cloud technologies to accomplish these tasks, there are a number of factors to be considered.

3.6.1 High Performance Computing in the Cloud

A case study in 2010[37] benchmarked the execution of high performance computing experiments within three different IaaS cloud systems and compared this with experiments using the 'Abe' supercomputer cluster offered by the NCSA[38]. The study concluded that the performance of the public cloud systems was comparable to that offered by the NCSA for experiments that did not rely on inter-process communication over the cloud's network. However, when processes were required to communicate in parallel, cloud performance was significantly worse than the NCSA cluster due to the limitations of the connectivity between virtualised servers. The results demonstrated that for the time being at least a public cloud infrastructure is more suited to experiments using 'cloud-friendly' applications that require less inter-process communication. For such applications a public cloud may offer a better alternative to supercomputer clusters as cloud providers are likely to offer server configurations that have more memory and newer processors than in-house systems.

The extent to which tasks efficiencies can be increased with additional memory or computing resources will inform the choice of whether the cloud is an appropriate choice. Lightweight tasks may not be significantly affected, but more intensive tasks may be significantly slowed by a lack of available computational power. With complex tasks, there is even the potential for outputs to be affected by a lack of resources. What volumes are involved? Both storage and transit

[37] Qiming He, Shujia Zhou, Ben Kobler, Dan Duffy, and Tom McGlynn. 2010. Case study for running HPC applications in public clouds. In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HPDC '10). ACM, New York, NY, USA, 395-401. DOI=10.1145/1851476.1851535 http://doi.acm.org/10.1145/1851476.1851535

[38] http://www.ncsa.illinois.edu/UserInfo/Resources/Hardware/Intel64Cluster/

capacity are critical considerations. The cloud model allows for the rapid provisioning of resources as and when needed, with the user only paying for what is needed, but this ability to quickly gain access to powerful computing resources suitable for intensive tasks is offset by the difficulty of transferring the significant volumes of data on which such tasks would typically be performed. A 2009 study illustrated that physically delivering tapes of data performed better in terms of transfer time than a high speed 20 Mbps WAN link[39]. The issue of transfer time may be less critical if the task can be automated and the large volume being transferred is made up of a number of small objects on which the task is to be performed. In this instance, the task can more effectively be performed in parallel with the transfer.

Common planning tasks such as preservation planning and data management validation (associated with tools such as Plato[40] and CARDIO[41]) are less technologically resource intensive and benefit less from the cloud's properties of scalability. However, in many cases these imply collaborative workflows and benefit from remote availability. The success of these and many other web accessible tools and resources is illustrative of a cultural enthusiasm for using remotely hosted, third party maintained and operated services. While not perhaps conforming to most cloud definitions such applications demonstrate many of the characteristics typically associated with the SaaS model.

3.6.2 Technical Dependencies

Where tasks are dependent on any particular platforms, applications, formats or workflows it may be more difficult to transition to a cloud service. For example, instrumentation may produce data in a proprietary format which requires specialist software in order to process it. The exception to this is that IaaS services may facilitate a large range of operating systems, as well as allowing for the customisation of computational resources. For example, Amazon EC2 allows users to create server instances running Windows Server, OpenSolaris or a range of Linux distributions, giving the user a great deal of flexibility when customising them.

3.6.3 Identification and Reuse

Cloud services are accessed over a network, usually via a web browser. Frequently, it is possible to provide URIs for resources placed in the cloud, facilitating sharing and reuse.

[39] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, g., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M., Above the Clouds: A Berkeley View of Cloud Computing, 2009, Technical Report EECS-2009-28, EECS Department, University of California, Berkeley.

[40] http://www.ifs.tuwien.ac.at/dp/plato/

[41] http://cardio.dcc.ac.uk

4 Summary of Cloud Issues for Curation

Given their potential diversity, curation tasks are difficult to characterise with any great certainty. The lifecycle model published by the Digital Curation Centre is a very useful point of reference and highlights the range of issues implicit in data management and curation, but defining "typical" tasks associated with its core action areas is impossible. As a consequence, any attempt to comment on the suitability or viability for particular classes of curation interaction is somewhat futile.

However, it is possible to identify several characteristics of the cloud which may be considered benefits or risks in particular circumstances. The table that follows offers coverage of a range of cloud characteristics (mostly introduced in the case studies above), some description and a note suggesting where they may be relevant within a curation workflow. The list is not exhaustive or universally applicable and some variability is evident depending on the service and deployment models of particular cloud applications.

Characteristic	**Description**	**Relevance for Curation Lifecycle**
Elasticity	Provision of computing resources on the cloud can be handled in a 'pay-as-you-go' manner: a user only pays for resources that are actually utilised and if demand on these resources increases then capacity will seamlessly and near-instantaneously increase to match the demand. Likewise, if demand for the resource falls then capacity can be automatically reduced, ensuring that users are only charged for resources they actually use.	Potentially significant for tasks with high demands especially if infrequent or atypical. **Create or Receive**; **Access, Use and Reuse**; **Preservation Action**; **Migrate;** and **Transform** tasks are particularly relevant.
On-demand self service	With cloud approaches it is possible to deploy a variety of different systems with little initial financial or infrastructural investment required. No hardware or software needs to be purchased and free trial versions of systems are frequently	Tasks performed by individual users are more easily accomplished as a consequence, and this may most obviously include **Conceptualise**; **Create or Receive**; **Description and**

Characteristic	Description	Relevance for Curation Lifecycle
	available. It is therefore much cheaper and easier to experiment with a variety of approaches without the risk of lock-in to a solution that proves to be unsuitable. The decision of which solution to adopt is also placed in the hands of users, assuming that in general terms the cloud service and the terms upon which it is offered are compatible with a given local institutional setting. Likewise requirements to deal with systems via an intermediary may be diminished. This is task dependent and expertise and various support levels may continue to be required between end users and cloud services in many cases.	**Representation Information**.
Widespread accessibility	Shifting business processes to the cloud can greatly increase their flexibility and accessibility. Cloud services can be concurrently accessed anywhere and at any time from a variety of internet-enabled devices, which may enhance productivity. Computer users are progressively less tied to a specific workstation and instead rely on a combination of desktop PCs, laptops, smartphones, netbooks and tablets. SaaS approaches such as Microsoft Sharepoint Online[42] can deliver content to various devices and maintain version control. Risks associated with transporting data on portable media are to an extent negated. These benefits are of course generally dependent on the availability of a reliable and robust connection to the	This implies benefits of accessibility, so those tackling issues of **Access, Use and Reuse**; and **Create or Receive** and anyone undertaking collaborative activities (for example **Appraise and Select**; **Community Watch and Participation;** and **Preservation Planning**) can benefit from this characteristic of the cloud.

[42] http://sharepoint.microsoft.com/en-us/SharePoint-Online/Pages/default.aspx

Characteristic	Description	Relevance for Curation Lifecycle
	networked services.	
Service / data integrity	Public clouds in particular may be equipped to distribute data and services across multiple remote sites in order to ensure data and resource redundancy. Nevertheless, media reports of outages have been relatively commonplace. Most providers' terms of service limit liabilities associated with data loss, and terms of service may offer little redress to customers that suffer data loss.	All parts of the curation lifecycle demand service and data integrity, but among the most significant are **Store**; **Access, Use and Reuse**; **Ingest**; **Curate and Preserve**; and **Transform.**
Data security	In December 2011 NIST published their "*Guidelines on Security and Privacy in Public Cloud Computing*"[43]. The report suggests that if an institution is primarily concerned with storing and processing confidential, personal, or sensitive data then public clouds may not be suitable and alternatives such as a private cloud may be more appropriate. To the contrary, others have argued that since public cloud providers often have more resources to commit to security they may be able to provide a more secure service than any individual institution. Large cloud providers such as Google or IBM are often at the forefront of security research and therefore have some of the best resources at their disposal. Additionally, the implementation of security measures such as the	Most obviously, the **Store**; and **Access, Use and Reuse** activity will be influenced by data security implications.

[43] Jansen, W., Grance, T., Guidelines on Security and Privacy in Public Cloud Computing, 2011, National Institute of Standards and Technology, Special Publication 800-144

Characteristic	Description	Relevance for Curation Lifecycle
	deployment of patches or upgrades is also cheaper to manage on a larger scale (see *Resource Pooling*, below).	
Resource optimisation	Cloud providers benefit from economies of scale to procure resource comparatively cheaply, and theoretically pass this onto customers. Recent work by David Rosenthal suggests that savings versus offline storage may not be quite as compelling as popularly believed, although there are opportunities to access resources that may be prohibitively costly for single institutions to otherwise acquire. The cloud does imply some additional resource demands – staff retraining and redeployment costs may not be trivial.	Those curation lifecycle tasks with typically high requirements for storage or compute resource include **Create or Receive**; **Access, Use and Reuse**; **Preservation Action**; **Migrate;** and **Transform**.
Delegation of responsibility	The provision of computing resources is expensive and outsourcing some of the associated responsibilities for this to a third party can theoretically cut costs and enable the user to focus on their core functions, leaving selected aspects of IT provision to be managed by a cloud provider that due to economies of scale can offer cheaper and often better services. In order to remain competitive it is in the best interest of Cloud providers to respond rapidly to technological change and offer faster processors and more memory and storage, often for the same or less cost as a few years previously.	Wholly outsourcing tasks such as **Appraise and Select**; **Dispose**; and **Reappraise** appears impossible given the highly specialist human resource requirements these tasks demand which have no obvious cloud surrogate.
Bottlenecks	Estimating computing requirements is a complex challenge	Bottlenecks will occur most obviously at the point

Characteristic	Description	Relevance for Curation Lifecycle
	which the cloud can help to address, but bandwidth constraints are likely to be exacerbated by a reliance on cloud services.	where one lifecycle task interfaces with another - obvious examples are during **Ingest; Create or Receive; Migrate** (particularly where tools are not available on cloud platform) and **Store** (where integrity checking must be done offline due to lack of native provider checking)
Legal liabilities	Contracts, service level agreements, and terms & conditions will all be significantly different to those required for more traditional IT services. Institutions cannot avoid any of their legal duties by using cloud services; they will simply have to find other ways to fulfil them. Jurisdiction can become complex as data may be simultaneously stored in multiple countries outside the UK, all of whom may have different laws governing data protection, intellectual property rights, and the rights of law enforcement agencies to access otherwise confidential data. JISC Legal has created a 'Cloud Computing & the Law' toolkit[44] providing detailed guidance for different stakeholders within FE and HE, such as senior management, users, and IT for processing power and storage. Although the services offered tend to be more expensive than large-scale public clouds, Eduserv is part of the JANET network and may therefore offer universities	Associated legal liabilities may vary by jurisdiction and according to data, but if one assumes that IPR, data protection and other data sensitivities are at the forefront of most minds activities like **Access, Use and Reuse**;**Transformation**; **Create or Receive**; **Dispose**; and **Store** will be most impacted by legal risks.

[44] JISC Legal, Cloud Computing and the Law Toolkit, August 2011, JISC Legal Guidance for ICT Use in Education, Research and External Engagement

Characteristic	Description	Relevance for Curation Lifecycle
	cost and time savings for data transfer.	
Trustworthiness	To invest in the cloud users demand assurances about the sustainability of cloud providers, the continued integrity of data and the ease with which they can move elsewhere in the event of cessation or inadequacy of service. Currently there are few such assurances available; accreditation or certification of providers is limited and most providers appear more focused on the mechanics of ingestion than providing compelling succession opportunities. Brokerage services such as Duracloud only appear to obfuscate the relationship between user and data custodian	Trustworthiness is a necessary commodity for those performing curation services and demanded of those who own or have a vested interest in curated data. In that sense the whole lifecycle can only function completely successfully where trust is assumed. There appears to be an appetite and need for authority organisations to step forward to offer endorsements or accreditations for those providers best equipped to off-set risks to successful curation.

5 Summary and Conclusions

5.1 What types of curation task are more cloud-viable?

There are various features of the cloud model which make it attractive in a range of scenarios. The ability to rapidly obtain and shed computing resources can make it suitable for infrequent, intensive tasks. However, if the task involves the transfer of large volumes of data then the cloud becomes less suitable. The use of cloud services for day-to-day work can facilitate remote working, allow users greater flexibility in the devices they use, promote collaboration and facilitate the sharing or publication of information.

5.2 What service model is the best fit?

In their work mapping OAIS to the cloud, Askhoj et al suggest that IaaS is largely irrelevant to *archivists*. This is true insofar as software solutions must exist to facilitate curation tasks, and that the underlying infrastructure is irrelevant to end users. However, there appears to be a compelling role for IaaS in allowing organisations or groups within them, to provide software (as a service) to researchers and others engaged in data curation. Likewise its potential role in the creation, processing and storage of data cannot be understated. Some of the most frequently cited benefits of the cloud computing model such as its metered payment model, its promises of practically infinite computing resources and the agnosticism of the underlying infrastructure are lost or diminished by the use of a private cloud. However, such an approach can mitigate or remove any increased risk associated with legal liabilities, or concerns over service level or data loss. The sheer range of activities involved in data curation means that there is no single solution. A commercial cloud solution may be of use for infrequent batch operations on data stored elsewhere but it may be inappropriate for the storage of sensitive or personal data.

5.3 What are the risks and benefits?

In general, on demand self-service, networked access and elasticity are clear benefits, while legal liabilities and bandwidth restrictions and bottlenecks present real risks. When considering areas such as service and data integrity, data security, use of resources and the transferring of responsibilities, there are both benefits and risks in using a cloud service. Using a private cloud as opposed to a public one may reduce legal liability issues, but the organisation retains the costs of maintaining the underlying infrastructure and may be unable to offer the sort of elasticity available from large commercial providers. Moving large volumes of data to and from the cloud may be difficult and if that storage is outside the organisation owning the data, and particularly if it is in a different jurisdiction, there may be exposure to significant legal liabilities.

5.4 What has been, and needs to, be done?

Limited work has been carried out to align the OAIS Reference and cloud models, which identified some value in the approach. The UMF cloud pilot projects provide practical implementations of cloud technology in a research environment. It is perhaps worth noting that all support day-to-day work. Duracloud provides a preservation platform that is potentially useful. The recent EPSRC-JISC Cost Analysis of Cloud Computing for Research makes a number of recommendations which could help encourage uptake. Costs need to be more transparent - both those of using cloud solutions and those of working in a more traditional environment. The EPSRC-JISC Analysis suggests that JISC should consider the provision of a cloud cost comparison service and should work to support institutions in calculating their research computing costs in order that they make better comparisons with cloud services. The latter could be extended to cover curation costs, and the DCC may have a part to play in that. In addition, the DCC may also be able to assist with the recommendation that JISC should help institutions to adapt their processes to facilitate access to cloud computing. On the specific topic of trust there may be scope for JISC or an alternative organisation to certify or accredit cloud services for data curation / preservation.

5.5 Does the cloud itself pose preservation problems?

Although somewhat out of scope of this document it is worthwhile to consider that moving any workflows to the cloud will introduce additional risks to data usability and understandability. Naturally, those developing curation workflows will be more aware of such risks and take steps towards their effective resolution, but in general terms, the increased adoption of cloud services across the IT landscape is likely to lead to data loss and inaccessibility in some cases as a consequence of numerous business and technological factors.