Curation, Cultural Heritage and the Cloud

1. About this document

WK was invited to speak on the broad topic of 'Cultural heritage, Cloud and Curation' at an invitational workshop on 'Curation and the Cloud' organised by Neil Grindley and Torsten Reimer of JISC in March 2012. The event had been further facilitated by the provision of a briefing paper in advance by Patrick MacCann and Andrew McHugh (of HATII / DCC). Instead of providing notes as normal from the event, this short report is the manuscript for WK's presentation. It is provided to support discussion about Cloud and preservation among DPC members and as a contribution back to the organizers of the event who were in the process of planning further activities.

A number of DPC members were present in their own right including Neil Grindley and Torsten Reimer (JISC), Kevin Ashley (DCC), Andrew McHugh and Patrick McCann (DCC/ HATII), John Zubrycki (BBC), Adam Rusbridge (UKLA), Keith May (EH), Pip Laurenson (Tate), Jenny Mitcham (ADS / UoY) Laura Molloy (HATII) and Nicky Whitsted (Open Univ). Members reading this text may be interested in following up with those present to understand the wider discussions at the event, or to retrieve the '#curatecloud' Twitter tag which was used throughout the event.

2. Introduction

I want to start by thanking Neil, Torsten and the other organisers for inviting me to speak today and in particular to thank Andrew and Paddy for the briefing paper which we have just been discussing which I found incredibly useful. For my part I claim no particular expertise in cloud computing, nor for that matter can I claim a comprehensive overview of 'cultural heritage'. It occurs to me that 'cloud', 'curation' and 'cultural heritage' are among the three more diffuse concepts that you could hope to find. But I am well placed to ask some questions and perhaps provoke some discussion. It's great to see so much expertise in the room from cultural heritage institutions and from the cloud computing community: perhaps by asking some naïve questions I am able to help you understand each other slightly better and thus use this as a platform for others to engage in more specific and detailed discussion.

3. This presentation

This is actually a pretty simple presentation. I want to start by saying something about cultural heritage institutions and how they think about the world, and what they think about digital preservation more specifically. If I can get you all to think a little like managers and planners in cultural heritage agencies then you will understand better the next part of the presentation which offer six questions about cloud computing and what it might mean for these institutions. There's a slightly off-topic seventh thought which we'll get to in time also. But by the end of this presentation I simply want to ask you – if you were a cultural heritage institution what would you see as the strengths and weaknesses of cloud computing, and in particular what this might mean for your own particular needs for long-term preservation.

4. Institutions?

Let me start by saying that this presentation takes an institutional view about cultural heritage. The gist of the argument is mostly going to apply to libraries, archives, museums and historic environment agencies that are concerned with the preservation of stuff into the long term. That 'stuff' may form a collection in its own right and be perceived to have some intrinsic value – email archives for example. Or it may represent data that in some way supports the key function of preservation which in fact pertains to some physical artefact – 3d scans of carved stones that help us conserve the historic environment for example. Both types of collection are important and create similar problems for the managers of these institutions. It's also important to remember that the institutions almost always have other reasons for existence over and above collection management, and you should not assume that collection management is necessarily the first priority that these agencies may have. So preservation actions or any decisions about preservation infrastructure are likely to be subject to a range of other priorities at any given time.

5. Incredibly diverse

That short definition works to distinguish these institutions from the very many other sectors present in the room, but it doesn't help me unify an incredibly diverse field. It includes some agencies that are very large – Culture and Leisure Services in Glasgow had around 6,000 staff when I joined it – while others have a really very small staff complement. It includes publicly and privately funded organizations as well as organisations that celebrate their openness to the public and others which are definitely not public facing. Some are well equipped and have a lot of expertise: some have neither. Some are interested and dynamic, some are not.

6. Preservation not just about institutions

You could be fooled for thinking that this large and varied set of institutions was responsible for all of the preservation actions that really matter. But that's never been the case and it's even less likely to be true in the digital age. Most of the really important and early preservation decisions are made outside of such agencies – in family archives and community groups. Faith groups, professional bodies, companies, sports clubs and the like are much more significant to the early preservation of cultural heritage, even if preservation is ancillary to their mission. Such collections exist not so much for the purposes of teaching, learning and research which might interest JISC, so much as the identity and belonging which they affirm and the way they embody personal narratives and world views.

I am labouring this point because it would be reckless to underestimate the importance of these groups or the quality and significance of material in their possession. Anyone who has run a local history project will understand what I am saying. You might not be familiar with the story of Marianne Grant, a Jewish grandmother who attended a community history event about the experience of Glasgow's Jewish community and the holocaust. Little by little it became apparent that she had an extraordinary story to tell of internment at Theresenstadt, Birkenau and Auschwitz. As a teenage girl she drew what she saw of life during confinement in the Theresentadt ghetto, a skill that saved her life when she was transferred to a concentration camp, and secondly meant she was given pencils and paper to illustrate research for the camp doctor, Josef Mengele. This included drawings of Allied forces relieving Birkenau. Even more remarkably, she had managed to keep these drawings through her rescue from the camp and Red Cross transfers to Sweden then Scotland. That is to say her family archive included 77 drawings which form a primary source for the Holocaust created by a Jewish victim on site at the time. The collection is of world-class significance. They were stored in a trunk which she donated to her adopted City of Glasgow. So a cultural heritage institution is now involved in preserving this collection, but only relatively late in the story and had no role in the most remarkable elements of preservation.

So the question arises, how might cloud computing change the relationship between cultural heritage institutions and the agents from whom they collect. It seems unlikely that current models of research data lifecycle management model such a relationship accurately enough to provide guidance on this issue.

7. Cultural heritage agencies – a policy perspective

Returning to the core consideration about cloud computing, curation and cultural heritage agencies, it is worth remembering a couple of salient points of policy. As noted the mandate to preserve is normally very strong, but because the implications of decisions are so long, decisions can be relatively slow to emerge and policy is slow to adapt to new conditions: this conservatism is in some senses a bulwark against short-termism. Professional practice is detailed and well described in many cases with agencies like ICON, ARA, IFA, MA and CILIP offering a variety of professional standards, codes of practice and ethics for institutions and individuals alike. Services may be outsourced but it is very unusual for responsibility to be outsourced. That means that trust matters very greatly. Metrics and professional practice for trust are only partially developed for digital preservation making this a tricky and developmental issue for cloud providers. Professional practice for analogue collections is designed to ensure things like provenance, chain of custody and context remain intact and that no conservation action is carried out which is not reversible. The cloud environment may need to model such approaches. Related issues like copyright and uniqueness are relevant to many of the collections and sensitivity review and compliance to data protection rules are a high priority. Unless these latter points are managed effectively and unless their management is spelled out clearly in advance, many agencies would simply exclude any engagement with cloud.

8. Cultural heritage agencies in practice

In many cases – such as local government or universities – the cultural heritage agency is a small section of a larger organisation and consequently is dependent on corporate services. It is not unusual therefore for there to be a dependence on corporate IT and procurement which in turn tends to drive economies in high volume services like desktop computing but is less able to respond to niche or specialist applications. A general drift to shared services can be observed in recent years with functions like personnel, payroll, finance and facilities management being handled by larger but increasingly remote providers who are designed to bring economies of scale but are in many cases also less flexible. Cloud services, if framed as 'shared services' may benefit from such a drift but if so it is not a given that this would support the specialised functions of the cultural heritage agency. The highly specialist requirements of some relevant topics – such as object conservation – mean that it is not unusual for even quite delicate or high profile operations to be outsourced for physical collections care. Outsourcing is politically sensitive though as it can be a challenge to existing staff and their skills, and moreover it requires a degree of subject knowledge to manage the resulting contracts. In practical terms, the cultural heritage community has seen a contraction in recent years. Core budgets - seldom large - have been squeezed and the major external discretionary funding of the lottery which had a profound impact on cultural heritage access and digitisation in the late nineties and early two thousands has been diverted to the Olympics.

9. Cultural heritage agencies and digital preservation

Cultural heritage agencies have shown a mixed response to digital preservation. Although there is clear leadership from some, and although there appears to be a widespread acceptance of the problem and willingness to engage, the capacity to respond is not always sufficient. The responsibility for digital preservation is in many cases diffuse and the mandate has not always been supported by funds or with skills. Practical barriers to preservation include cost, training and access to infrastructure. There is also an uncertainty about the long-term value of digital collections while the standards associated with digital preservation, especially trust metrics, are immature. But the sector is rich with content, partly on account of some very large audio-visual collections, web archiving and in particular on account of digitisation. In many cases, institutions seek revenue from new kinds of engagement with their digital collections. Cultural heritage agencies have been responsible for some of the most impressive digital outreach projects which have had a real impact on the public imagination, especially where they intersect with broadcast media. This shows a particular willingness to collaborate and innovate with digital technology where there is an explicit and immediate public benefit.

10. Six questions about the cloud and the cultural heritage community

Now that I have introduced you to the sector and sketched out some of the concerns that differentiate the cultural heritage community from others with interest in the cloud, I would like to ask six (perhaps seven) questions of uneven depth which occur to me as I consider the implications of cloud computing and digital preservation on this community.

- Big data is here, but it is not quite what we might have expected. It has arrived as audio-visual archives and its web-archives. Has cloud computing helped big science with big data and are there lessons we can learn?
- We have seen that there are barriers in terms of cost, expertise and infrastructure for cultural heritage agencies in addressing digital preservation. Does the elasticity and ubiquity of cloud computing help us over those barriers? Or is this just an illusion?
- We have noted the issue of trust, sensitivity, compliance and provenance which are of high importance in the cultural heritage sector and the allied issues of copyright. Do cloud providers really understand these and how would we measure the kind of trust that we need?
- What are there implications for collecting from the cloud?
- Does preservation 'as a service' in the cloud mean we need to alter the research and development roadmap for digital preservation?
- From where will change come? Is it inevitable? What are the drivers for migration to cloud services?

Underpinning these six questions are two related issues. I wanted to provoke debate by asking 'if you were a cultural heritage agency, what would you do?' Perhaps that could be phrased more rhetorically as 'Is the repository dead?' as the locus for preservation. Let's take these questions one at a time.

11. Big Data, Cultural Heritage Agencies and preservation in the Cloud

A decade ago the eScience community got quite excited about the 'big data' which would result from live data streaming from scientific instruments, and the preservation risks that it faced and how we might address them. It's clear that this promise has been fulfilled at facilities like CERN where vast quantities of data are

created, managed and accessed by large numbers of researchers in many different countries at once. But big-data is not just about big-science. The cultural heritage community has collections as big as any in the sciences and the problems of managing, accessing, preserving and undertaking collection-scale analysis are as great for the humanities as for the sciences: in some senses even more so. Audio-visual productions and web harvests can induce truly colossal data sets, while the capacity to analyse entire digitised corpora in one single, uniform process creates new opportunities and new challenges. The collections are large and getting larger, complex and getting more complex, valuable and becoming more important. A decade ago the promise of e-science was that instead of moving vast quantities of data around that we instead moved the processes to the data and simply transfer the results. Grid computing also offered large scale, elastic and pervasive computing resources and virtual organizations interfacing with the real world through virtual research environments. This all sounds a bit familiar and it sounds as if the ideas of the eScience programme can be matched pretty closely to the aspirations associated with the cloud. So the simple question arises – what happened to the promise of eScience? If it succeeded, does the cloud offer anything better; if it failed to deliver, why would the cloud be any different.

12. Barriers to digital preservation

For some time the costs of preservation have been perceived as a challenge. Although commercial offerings such as Tessella's SDB or Ex Libris's Rosetta show that it is possible to develop a genuine business out of preservation services, it still seems that digital preservation requires a complex process of specification, configuration and testing before solutions can be deployed. This might be called bespoke tailoring, while what is required might be termed 'pret a porter preservation' – tools for the mass market which can be quickly deployed to good and obvious effect and readily understood. A parallel concern is the desire to ensure that preservation tools can deal with the scale of the data which is coming our way – and which is already here in many cases. A decade of research and development has given digital preservation many great tools, demonstrators and theoretical processes: but our concern has been on fundamental problems and perfect answers: sufficiency at scale has not been much celebrated. Cloud computing seems to offer many solutions here though it implies an architecture which is highly dependent on services and their configuration. The cloud – with its stress on elasticity and scalability – seems particularly well suited to these problems. Moreover the capacity to deploy a range of services including virtualisation suggests that we can see economies around the deployment of solutions that might otherwise be too uneconomical to build for only local uses and too complex to deploy at scale. However this almost certainly means that a range of existing tools and services need to be prepared for the cloud while new approaches to versioning, service dependency analysis and 'replay' or 'exhumation' will almost certainly be needed. The attractiveness of cloud-based preservation-services probably include the way that it could enable some agencies to by-pass their in-house IT services. Given the focus on core desktop applications it is unlikely that many corporate networks will encourage or support the rather specialised and at times truly exotic architectures necessary for preservation: in my experience corporate IT would rather train its efforts on mainstream applications, but they are mostly content to support experimentation on other people's networks. So there is a subtle but mutually beneficial aspect of Cloud applications for corporate IT and cultural heritage agencies.

Lowering the barriers to preservation is probably the single most important contribution that we can make to the wider community. From this perspective, the Cloud looks like the silver bullet that digital preservation has been looking for in order to into the mainstream. But the 'trust' issue needs out attention before we hoist the flag, declare victory and wind up the DPC.

13. Trust and the Cloud

Trust in preservation is an enduring problem. For cloud-based preservation it probably breaks down into three related themes: corporate sustainability; tests of policy and practice; and the implications for the effectiveness of preservation action. In some senses a failure on any of these issues is likely to means failure in all of them.

Corporate sustainability: Digital preservation has historically feared 'lock-in' to proprietary technologies and systems and the same is true of lock in to any service provider. In a nutshell, what happens if a cloud provider goes bust or decides to discontinue a service? These questions are not unique to the cloud of course but they do explain why agencies might be slow to trust a cloud service. Moreover the economics and costs of digital preservation remain open to debate and without greater certitude around the costs, and greater clarity about the costs of any given service, it is hard to know whether the service provider has a truly viable business plan. The solution is easy in principle: there simply needs to be a realistic disengagement plan so that data can be moved from one cloud to another, which should not simply be about data but possibly also about services.

But interoperability of clouds is not a given, and though I have looked, I don't believe I've ever seen a plan for how to get things out of a cloud in the event of a service closure.

Policy and practice: As already mentioned, digital preservation operates in a world which is at times highly regulated, or has quite demanding requirements of professional ethics. Sensitivity review of bulk data is very slow and agencies tend to be highly risk averse. Regulatory issues like data protection, copyright, or data export fuse with professional concerns about custody, fixity and provenance, which means that 'public' cloud solutions are unlikely to be appropriate, and even private clouds may need to provide much more thorough audit trails in terms of object tracking. It may not be enough to recover a document and prove its fixity via a checksum: it may be necessary to gather the entire history of access and even the physical location of storage at any given time. It may be necessary to sample checksums periodically to ensure consistency through time and provide early warning of problems. All of this reduces the elasticity which the cloud offers and probably points to some genuine development work for the services providers. And once the high level policy work is completed then there are the local politics to resolve. Remember that shared storage for cultural heritage agencies can be politically sensitive.

Effective Preservation: And there is also the small matter of not yet being entirely clear what success looks like in digital preservation. We have standards like OAIS and increasing recognition of the need for evidence based planning that turns broad policy objectives into specific actions. Professional practice is increasingly being codified for review purposes in a number of audit and certification tools such as LOTAR, ISCO 16363, Data Seal of Approval and DIN 31644, though there is still some uncertainty about which standards should apply in which setting. A memorandum between the authors of the three main standards points to a formal articulation that supports the different approaches, but it is unclear what an audit of a cloud based digital preservation service would look like. Audit is likely point to a need for succession planning which in turn leads back to the issue of interoperability discussed above. Finally the dependency the cloud preservation would place on distributed services suggests that an additional (alternative?) type analysis will be needed: as well as assessing a 'repository', we must surely have a more meaningful analysis of 'services'. Cloud-based preservation implies that the 'trusted digital repository' – as the locus of storage and stewardship – may no longer be a useful category. What remains when you remove the repository from preservation are the skills, policies and services which support it. If repositories become an assemblage of services, then surely the measure of trust one places in a repository is really just an aggregate of the trustworthiness of each service in turn and how effectively they are deployed to a given policy goal?

14. Collecting the cloud

As well as presenting opportunities for undertaking preservation, the Cloud also creates new ways of thinking about collecting. Put simply, it seems obvious that the Cloud will shortly contain collections which are of some cultural significance. If so, cultural heritage agencies will need to consider how to collect from the cloud – or in the cloud – it's not at all clear where this particular metaphor takes us. What might that be like? What would ingest be like? On one hand, if collections are in the Cloud already, and if this architecture brings the benefits and opportunities that are supposed, then it would seem naïve to move content out of the cloud to preserve it. Perhaps the model is more about taking ownership or moving collections across different configurations of service – say from public cloud to a private one (or vice versa)? Or is this more like web archiving? If the cloud is a large, persistent and online environment are we really simply talking about a new generation of specialist web harvesting to take periodic snapshots of relevant content and store it somewhere safe? Also relevant in this discussion, and perhaps generically useful to cloud services, are the sorts of 'roll back' which tools like memento make possible for content management and wiki services. If such tools were widely deployed and if they were reliable, then perhaps the next generation of digital preservation tools will be designed with a very different functionality in mind.

15. Development

The previous sections hint at some interesting new problem which would have an impact on the current development path for digital preservation services. Tools to manage and track authenticity, fixity and provenance have already been widely discussed in digital preservation but the elasticity of the Cloud are likely to make this research increasingly important. Similarly, the complex interdependencies which we already recognise as a problem for the maintenance of files will surely become a more immediate concern. Dependencies on services and the ability to freeze, exhume and replay such services will become more important, either as concern in the management of preservation as a service, or as a preservation action in its

own right. Moreover, the cloud puts some preservation approaches – which might otherwise be considered too exotic to be viable – within reach. This is perhaps most obviously true of virtualisation and emulation but it is also true of concepts like 'migration on demand'.

16. Where does the demand come from?

We have discussed many of the pros and cons of the Cloud as these might apply to the cultural heritage sectors' interests in preservation. But there remains a question in my mind as to where the demand for cloud services is coming from. We can see why practitioners might be interested, in order to access the elasticity of resources and economies of scope; and we can see why strategic planners might be attracted to the cloud as part of a more generalised drift to shared services. But we can't ignore the corporate hype too. There is a risk that a series of promises are being made which will either prove illusory, or which could result in proprietary 'lock in'. My real concern is that, while researching this presentation, there seemed to be too few case studies of preservation in the cloud, too few specific examples of how the cloud has fixed problems that could not be fixed in other ways. This could well be my own lack of insight and I look forward to hearing about these case studies later in the workshop.

Higher education has invested heavily in repositories and for all their faults and failings there is clearl a degree of institutional support for this model, and a growing amount of expertise. It's hard to believe that this enthusiasm will wane completely in favour of cloud solutions, and it's only really likely that some kind of hybrid solution will prevail in the medium term. The role of the research assessment exercise (and now research excellence framework) has been a particularly important driver for the development of research data infrastructure in the UK. But the broader cultural heritage sector has been less sure focussed on the development of repositories and the very great diversity means the drivers are more diffuse and there is probably less shared wisdom too. So it's not impossible that elements of the cultural heritage sector could simply by-pass the repository and go 'straight to solar' – that is put all their preservation activities in the cloud.

17. What would you do?

I started this presentation by asking you all to think like cultural heritage institutions, then encouraging you to ask a series of questions which should help you form a view of the potential for cloud services and preservation from the perspective of cultural heritage agencies. Let me end by asking 'what would you do?' It's time to invite your comments and suggestions, but let me be more specific and set the question in the historical development of digital preservation architectures in cultural heritage institutions:

- If I have a cloud for storage and can access the services that I need, what's the use of a repository?
- Cultural heritage agencies have not generally got institutional repositories. Would you recommend they acquire one?
- Do we need at least need a new metaphor instead of 'repository'? Surely the services matter more than the thing?

What would you do?

18. About this document

Version 1	Text drafted to support powerpoint sketch	01/03/2012	WK
Version 2	Presented at 'Curation in the Cloud' workshop in London	07/03/2012	WK
Version 3	Posted for comment on DPC website and to JISC event organisers	15/03/2012	WK