The National Archives

CASE STUDY: BODLEIAN LIBRARY, UNIVERSITY OF OXFORD

January 2014

ABSTRACT

This case study covers the Bodleian Library and the University of Oxford, and their provision of a "private cloud" local infrastructure for its digital collections including digitised books, images and multimedia, research data, and catalogues. It explains the organisational context, the nature of its digital preservation requirements and approaches, its storage services for research data, the technical infrastructure, and the business case and funding. It concludes with the key lessons they have learnt and future plans.

ORGANISATIONAL CONTEXT

The Bodleian Library at the University of Oxford has been a legal deposit library for 400 years. In addition to its books and journals it holds a wide range of archives and special collections. For example, the BEAM service within the library looks at personal archives that are hybrid collections of analogue and digital material.

The Library was involved with the Google Books digitisation project and a OS map digitisation project, and is currently part of a project with the Vatican to digitise maps.

The Library's role in supporting researchers includes services for managing their research data. The University's research data management strategy involves a top level, multi-agency approach - the Library, Research Services, Oxford e-Research Centre, and central IT services have all been involved. The Research Support Portal offers a central reference point for Research data management to Oxford academics. In general, the central IT service provides operational support and the Library deals with archiving of data.

The Library, Oxford e-Research Centre, and central IT service all offer their own storage services that have been described as"private cloud": IT services offers conventional storage for research data on a monthly rental basis; the Library offers archival storage for deposits (usually from humanities scholars) which may or may

not have a related web presence: and the Oxford e-Research Centre offers storage with access to the academic Data Grid. Data for long term deposit goes into the archival storage, but currently there is a one-off single payment which covers on-going storage costs. They are also looking currently at alternative cost models for small deposits and unfunded research.

The options for storage and archiving are set out via the Research Support Portal. The Library will provide a quote for archiving digital material generated from a research grant, costed on the standard basis for a "Small Research Facility". For larger/more complex datasets they can collaborate with other institutions, e.g. the Oxford eResearch Centre.

DIGITAL PRESERVATION:

Significance

Digital Preservation is of major importance to the Library. As a legal deposit library "that's what we do".

The University also needs to comply with Research Council mandates for retention of data. The requirement is typically for ten years preservation of most digital research data, but in practice storage costs diminish so much that after ten years it may be more costly to review the data than to continue preserving it.

For digitised materials they are working mostly with humanities, but born digital research datasets are coming from different disciplines such as sciences and social sciences, which don't have the same metadata standards, traditions or expertise, and therefore this needs new skills in the Library.

They are at a transition stage where over the next 4-5 years the largest collections will be born digital rather than digitised.

Current approaches

For hybrid archives the material is very varied. They are given old devices from which they have to extract the data and do format translation/migration. If digitising themselves, they have more control, and so far have only had to migrate from TIFF to JPEG 2000, a choice which was driven more by engineering and costs than

preservation imperatives. The Library use their own platform based on simplified FEDORA. The FEDORA system doesn't actually do preservation processes, so they do checksums and use JOVE, Open Planets etc. for characterisation and other preservation functions. For hybrid archives/personal digital materials, they use Forensic Toolkit software on incoming data – this is costly software but identifies valuable files and the formats for possible migration.

There is a diversity of research data formats which are often machine-specific and proprietary so they can't be migrated easily. They have looked at bit stream preservation and keeping context information for these formats.

How they would want this to change over the next 3 years

Currently they have different approaches to different material. They would like to see more packaged tools that could be applied across the board and more automation, e.g. for the workflow involved in characterisation. Currently, they have long workflows with lots of dependencies. Cloud Virtual Machines could help speed up dealing with processing and checking images at ingest, etc.

Range of content types and volumes of digital material

Currently they have 300 Tb of which a small amount is born digital material.

Over the next three years they are planning to start to archive data at a Petabyte scale. Currently, the content is overwhelmingly images and text but that range is expanding with the addition of AV and other file formats over time.

Private Cloud Storage for Digital Preservation

Oxford has been managing its own local virtualized storage for preservation and describes this as private cloud. They are not procuring cloud storage from an external service. They do not view most external cloud providers as a viable proposition for them currently, as their ongoing revenue payment models do not match the research funding model of fixed duration payments for projects, and are therefore potentially too costly over time. They are also worried about getting data out and any costs for this being affordable. Some external specialist archiving services are interested in proper long term storage and might be a better match. For example they might consider Arkivum if the cost profile was favourable.

The Bodleian are actively looking for partners for a consortium approach but the infrastructure is not there yet. They have some discussions with Cambridge which would be an obvious partner, and also have strong ties with Stanford. They can see preservation advantages in having a copy of the data elsewhere, stored in a geographically distant location with a preservation partner institution.

Technical Infrastructure

They use FEDORA, home-grown software, PRONOM, and DROID. They also use Forensic Toolkit and are looking at HYDRA tools. They don't plan on having a single ingest or delivery route as the material is too diverse.

The private cloud is based on VMware ESX running across a number of clustered servers.

Business case and Funding

The main issues in a business case for cloud storage for digital preservation in the University are the cost and risk profiles.

To fit in with fixed term research project grant funding they need the ability to pay up front for perpetual and/or 10+ year storage of research data rather than by annual payments. Most third-party cloud services have an ongoing revenue cost that is harder to fit in to this funding model (Arkivum is currently the main exception to this as it offers a single payment option for 10-20 years). Local storage however can charge capital costs to a grant proposal and provide a degree of upfront payment for the costs. However, this is not the sole funding model as services are also needed for unfunded projects and some current and future costs may be funded or partly funded via overheads on research grants.

They feel there can be asymmetric risk in using an external provider: the risk can be higher for the client than the provider. If the provider goes bust there may beno mechanism to get data out in good time. The exceptions to this are cloud providers that have an escrow copy as part of their service, or where an institution can operate two separate cloud service providers in parallel to mitigate the risk (but this has costs and management overheads).

KEY LESSONS THEY HAVE LEARNT

- Archiving and digital preservation is often much less well developed than people think.
- The Open Archival Information System (OAIS) Reference Model is a mixed blessing – fine as model but in practice not necessarily scalable, and too rigid if adhered to slavishly;
- It is not simple, if like a university you are working with very large data volumes. At scale you need to have the network infrastructure and bandwidth to use the cloud;
- The balance of risk when using external providers can be asymmetric; there is no guarantee of longevity for commercial providers and in the event of data loss, far more damage would be done to the institution's reputation than to that of the cloud provider. You need to mitigate these risks in your chosen implementation and exit strategies;
- Using a shared private cloud (perhaps similar to the Digital Preservation Network in USA) or a consortiums of like-minded institutions might be more viable for archives. The US and Australia are at a more advanced stage in this.

FUTURE PLANS

The growth in research data and the need to archive it will require a significant re-engineering of provision at all levels across the University in coming years.

FURTHER INFORMATION

Research Data Management at Oxford - http://www.admin.ox.ac.uk/rdm/

Bodleian Library -http://www.bodleian.ox.ac.uk/bdlss

VMware ESX - http://www.vmware.com/files/pdf/VMware-ESX-and-VMware-ESXi-DS-EN.pdf

Arkivum: http://www.arkivum.com/

The Digital Preservation Network - http://www.dpn.org/

The Open Archival Information System (OAIS) Reference Model - http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=57284