Technical aspects from the Polonsky Digital Preservation Programme - The story so far at The Bodleian Libraries and Cambridge University Library

James Mooney
The Bodleian Libraries
Oxford OX2 0EW
james.mooney@bodleian.ox.ac.uk

Dr. David Gerrard
Cambridge University Library
Cambridge CB3 9DR
dg509@cam.ac.uk

ABSTRACT

As part of the two-year Polonsky digital preservation research project[21], the Bodleian Libraries (the University of Oxford) and Cambridge University Library (CUL) are researching and developing requirements for digital-preservation-specific services. Part of the project concerns gathering technical requirements for long-term digital content repository systems; this has included reviewing our own infrastructures, surveying our digitized collections and existing repositories, visiting a number of other institutions and assessing software from various vendors.

Our poster highlights the different challenges both institutions face with their current systems, the collaboration relating to auditing and reporting software used, the work which has already completed, and what is planned for rest of 2017-2018.

KEYWORDS

audit, dspace, droid, fedora, jhove, review, repository

1 THE BODLEIAN LIBRARIES

At Oxford, our initial review has focused upon four main areas within our institution: Imaging Services, Oxford's Research Archives (ORA and ORA-Data [7]), Digital.Bodleian [2] and project-driven websites.

Our imaging services department have created over six million images and use Goobi workflow management software [6] to manage their day-to-day workflows, allowing for preservation actions to be incorporated such as the creation of checksums and validation checks. For 2D digitization the TIFF 6.0 format is currently used; previously using 5.0. These master TIFF files are stored to tape, along with a MD5 checksum. Due to legacy file naming conventions for digitized content additional software is used to track linkages between file names and associated metadata. Utilizing a repository environment to manage master image files in the future will help mitigate this preservation risk.

As part of our survey we reviewed this digitized content, one challenge was restoring 100+ Tb of content from tape to spinning disk. As part of ongoing infrastructure improvements, a 600 Tb RAID disk array was available to store the retrieved data, allowing us to then use parallel JHOVE [3] processes characterize, validate the large number of TIFF files. Transforming the JHOVE XML output with XSLT allowed us to import into Qlik Sense [10], a data visualization application with which reports such as cumulative file size over time could be created to help us predict future storage requirements. In addition SHA256 checksums were calculated outside of JHOVE for ongoing fixity checks along with image fingerprinting with Python image hash libraries to identify near duplicates.

Oxford University Research Archives (ORA) is our institutional repository for scholarly research output. ORA preserves an array of research publications, journal articles, conference papers, working papers, theses, reports and more. There is also an additional repository, ORA-Data, designed to help researchers archive, share and cite research data. Both these repositories based on the Fedora repository software, with ORA having previously migrated from Fedora 2 to 3, one major challenge is migrating to the latest version of Fedora [13].

By using regular automated DROID [12] scans of the repositories, we have also identified the current file formats in use and built a format risk registry. We have established that approximately 52% of content within ORA-Data is unknown, which is understandable as this is research data output. We will engage with our research community to help create new file signatures for these and submit to the PRONOM file registry [20] , so we would like to see the number of unknown formats decrease. SHA-256 checksums generated with DROID have let us report on duplicates, and comparing checksums with previous scans will help to identify any fixity issues.

Digital.Bodleian is a repository bringing together our discrete digitized collections under a single user interface. This service currently hosts approximately 700,000 images online. Originally Digital.Bodleian utilized customized repository software based on Fedora 2, and like our Research Archives, Digital.Bodleian will migrate to latest Fedora repository system. Standardizing on the same repository software will enable more code reuse and many of the same preservation actions will be used across to both systems.

We also have approximately 40 project-driven websites containing digitized content from the last 20+ years, using a variety of different frameworks, databases, metadata formats and user interfaces. There is no common search interface, making content discovery and navigation across collections difficult. Only a few collections offer a machine-readable interface, or any way to link their data with similar data in other Bodleian collections, or with collections at other institutions. These project sites vary in size from 127 million compressed JPEG2000 files, which were created as part of a Google Books project [4], to earlier projects which contain just over thousand JPEG images. Approximately 2.5 million other digitized images have also been made available via these sites. This presents a major challenge, the sites need to continue to be supported until such time that the content can be migrated to Digital.Bodleian, however we face some problems with extracting various forms of metadata from the legacy sites.

One of the next areas of focus will be around the current storage infrastructure. Our services currently have a minimum of three copies of data, one on ultra-resilient spinning disk storage which is mirrored between two data centers, then an additional two copies stored on tape in two locations. However, at present these copies all lie within Oxford's city boundaries. Therefore we will recommend that for our repository environments, two copies will be stored on spinning disk ideally on different technology stacks in more geographically diverse sites, along with further copies of these stored on tape [11]. Due to the extensive infrastructure already in place at Oxford, we have not yet considered 3rd party cloud storage, however this is one of a number of options we are still reviewing.

2 CAMBRIDGE UNIVERSITY LIBRARY

Like our colleagues at Oxford, the Polonsky Fellows at Cambridge University Library have also used DROID to scan the University of Cambridge's DSpace [5] research data repository (Apollo [8]) to gather information about the files stored there. In October 2016, there were 415,000 files linked to 700 'dataset' type submissions, taking up 308 Gb of storage space. These accompany nearly 20,000 research publications (almost exclusively in PDF format). Our analysis indicates that a similar 44% of the files stored within research datasets in Apollo also have unrecognized formats.

To contribute towards the business case for digital preservation, the CUL team also conducted two searches of archival and bibliographic databases. These searches used regular expressions to search for approximately 250 different terms related to digital and audio-visual media carriers. From these queries, we estimate that there are approximately 80,000 - 100,000 unique pieces of digital and AV media in our collections of published materials, and circa 2500 records in our union archival catalogue (Janus [14]) that refer to digital or AV materials. Analysis of the results indicated that the (comparatively) simple approach of using regular expressions achieved a precision measure of 0.8, which was (just) performant enough to base effective management statistics upon.

As with our colleagues at the Bodleian, however, a major area of focus was upon our digitized image collections. Thus far, we have scanned our central storage system with a custom-written Perl script (faster and lighter-weight than DROID, but designed to give more of an overview rather than all the detail that DROID provides). These scans registered over 1.4 million TIFF files and a similar number of JPEGs. They also helped identify several clusters of 'legacy' content and provided information that can be used to bring older content under the same degree of control as the material we produce today.

We have also worked directly with our photography and digital library content and system development teams to analyse their digitization and content development workflows. We are now preparing some more in-depth tasks related to digitized materials, which are:

1. To assist with an infrastructure upgrade to provide photographers with local redundant / robust, but fast 'scratch' storage, and to work with them to develop storage management plans and policies. We may also trial Versioning Object Storage (e.g. Swift [16].
2. To follow the Bodleian's approach, by conducting more in-depth scans, using characterisation tools such as JHOVE, to find out more about our legacy objects and prioritise cleaning them up and bringing them under control.
3. To help the photographic, content creation and development teams improve their workflows. This is partially about tools - so we will experiment with Goobi and Apache ActiveMQ / Camel [1] - but it is also, fundamentally, about working with staff to help them improve their processes.

Our survey work has also highlighted a similar need to consider the underlying storage architecture as that encountered by the Bodelian: all our data is currently stored and backed up within Cambridge, with no mirrored versions or backups held further afield.

Finally, all the above has been conducted alongside a review of preservation systems and vendor companies. Part of this has been to join the JISC Research Data Management Shared Services Pilot [15], where we are one of three UK universities (alongside St. Andrews and Lancaster) who are testing both Preservica [17] and Archivematica [9]. We also plan to test three other systems (Fedora 4, Rosetta [19] and Roda [18]) during a programme of case studies lasting until spring 2018.

3 ACKNOWLEDGMENTS

We would like to thank our colleagues Edith Halvarsson, Sarah Mason, Somaya Langley and Lee Pretlove for their help and advice, and the Polonsky Foundation for funding research of digital preservation at Oxford and Cambridge (DPOC) 2016-2018.

REFERENCES

[1] 2015. Apache Camel. (2015). http://camel.apache.org/
[2] 2015. Digital Bodleian. (2015). http://digital.bodleian.ox.ac.uk/
[3] 2015. JHOVE: Open Source file format identification, validation and characterisation. (2015). http://jhove.openpreservation.org/
[4] 2015. Oxford Google Books Project. (2015). http://www.bodleian.ox.ac.uk/dbooks
[5] 2016. DSpace is a turnkey institutional repository application. (2016). http://www.dspace.org/
[6] 2016. Goobi overview. (2016). https://www.intranda.com/en/digiverso/goobi/goobi-overview/
[7] 2016. ORA Help and Information. (2016). http://www.bodleian.ox.ac.uk/ora
[8] 2017. Apollo: University of Cambridge Repository. (2017). https://www.repository.cam.ac.uk/
[9] 2017. Archivematica: open source digital preservation system. (2017).
[10] 2017. Data Visualisation Software: Qlik Sense. (2017). http://www.qlik.com/us/products/qlik-sense
[11] 2017. *Digital Preservation Handbook* (2 ed.). http://handbook.dpconline.org/
[12] 2017. Download DROID: file format identification tool. (2017). http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/
[13] 2017. Fedora Repository. (2017). http://fedorarepository.org/
[14] 2017. Janus Home Page. (2017). https://janus.lib.cam.ac.uk/
[15] 2017. JISC: Research Data and related topics. (2017). https://researchdata.jiscinvolve.org/wp/
[16] 2017. Object Versioning in Swift. (2017). https://docs.openstack.org/developer/swift/overview_object_versioning.html
[17] 2017. Preservica Digital Preservation. (2017).
[18] 2017. RODA: an open source digital repository designed for preservation. (2017).
[19] 2017. Rosetta: Digital Management and Preservation. (2017). http://www.exlibrisgroup.com/category/RosettaOverview
[20] 2017. The technical registry: PRONOM. (2017).
[21] Sarah Mason, Lee Pretlove, Edith Halvarsson, Somaya Langley, James Mooney, and David Gerrard. 2016. Digital Preservation at Oxford and Cambridge. (2016). http://www.dpoc.ac.uk