Technical Aspects from the Polonsky Digital Preservation Programme

THE STORY SO FAR AT THE BODLEIAN LIBRARIES AND CAMBRIDGE UNIVERSITY LIBRARY

As part of the two-year Polonsky digital preservation research project, 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.

SOME OF THE SYSTEMS REVIEWED AT OXFORD...

Repositories
Oxford University Research Archives (ORA) is our institutional repository for scholarly research output. ORA-Data is a seperate repository designed to help researchers archive, share and cite research data. Both repositories are based on Fedora 3. One major challenge is migrating these to the latest version of Fedora.

Digitization
Our imaging services department have created over six million images and use Goobi workflow management software to manage their day-to-day workflows, with preservation actions have been incorporated. For 2D digitization the TIFF 6.0 format is currently used, these master TIFF files are stored to tape, along with checksums. Due to legacy file naming conventions 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.

Legacy Websites
40+ project-driven websites containing digitized content from the last 20+ years, all using a variety of different frameworks, databases, metadata formats and user interfaces. Total approximately 2.5 million digitized images via these sites, excluding our Google Book project (124 million images). This presents a major challenge, supporting these sites until such time they can be migrated to Digital.Bodleian, our online image collection repository.

...AND AT CAMBRIDGE

Repository
Apollo is our DSpace Open Research Data repository managed by the Office of Scholarly Communications. In late 2016, it held nearly 9,000 research publications and more than 700 research datasets.

Digitization
Digital Content and Digital Library Units digitize items for Cambridge Digital Library and others. Reviewing our central storage we have found 1.4 million TIFF files and a similar number of JPEGs. We have also now identified several clusters of legacy content which will need to be brought under the same degree of control as the material we produce today.

Born Digital
Our catalogues of published materials contain over 15.5 million records, of which we estimate circa 100,000 are, or contain, digital or audio visual carriers. Our union archival catalogue (Janus) contains over 560,000 records, of which we estimate circa 2500 - 3000 contain digital or audio-visual media carriers.

FILE SYSTEM SCANS

Major collections scanned using DROID and custom (more lightweight) Perl script.
- Identifying the current file formats and versions in use
- Perl script allowing quick overview of number of files and size of collections
- Identifying duplicates via SHA-256 checksums, regular re-scanning to identify fixity issues

CHARACTERIZATION / VALIDATION / FIXITY

- Parallel JHOVE processes to extract metadata properties, due to quanity of files to review
- XML output being transformed with XSLT for customized output
- Additional SHA-256 checksums created to identify duplicates and re-run for fixity checks
- Image fingerprinting with Python image hash libraries to identify near duplicates
- Confirm that TIFF master image files are valid and well-formed

SURVEYING : METHODS USED AT BOTH INSTITUTIONS

DATA ANALYSIS

Searching union archival catalogues and bibliographic databases (MARC) using digital and AV carrier queries with approximately 250 regular expressions.

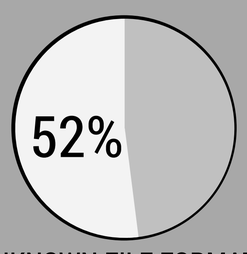
VISUALIZATION

Loading results of all these methods into Qlik Sense, a data visualization application which allows for further analysis of the data collected and gives us the ability to dynamically generate custom reports (file format breakdowns, cumulative file size over time, submissions with high percentage of unknown formats, etc).

SURVEYING : SOME RESULTS SO FAR...

6+
MILLION DIGITIZED TIFF IMAGES TO REVIEW AT OXFORD

1.4+
MILLION DIGITIZED TIFF IMAGES TO REVIEW AT CAMBRIDGE

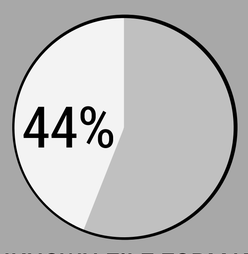


UNKNOWN FILE FORMATS IN ORA DATA (RESEARCH DATA) FROM 290 DATASETS

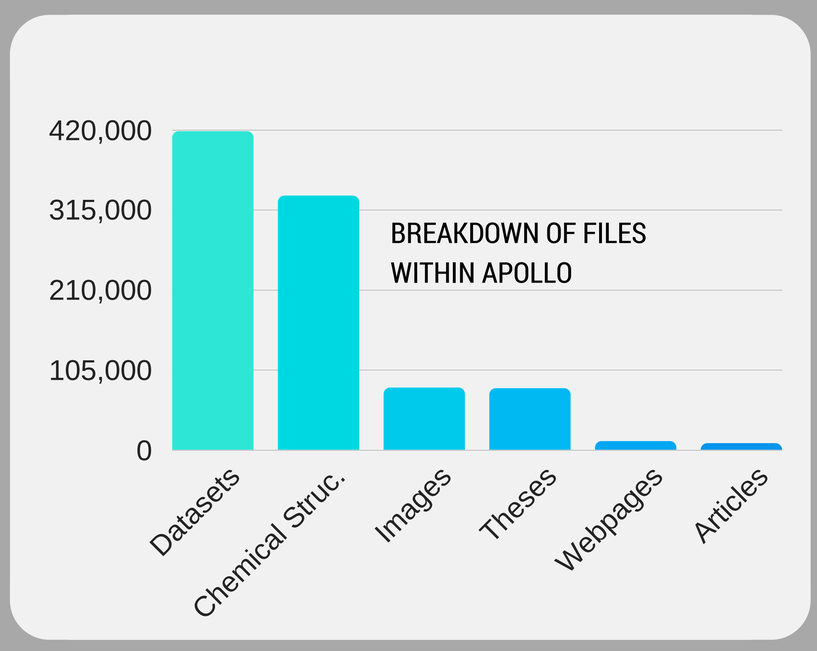
2.2%
DUPLICATES OR NEAR DUPLICATES DETECTED FROM IMAGE FINGERPRINTING AT OXFORD

100,000+
ESTIMATED UNIQUE PIECES OF DIGITAL AND AV MEDIA FROM PUBLISHED MATERIALS AT CAMBRIDGE

2.5
MILLION
ONLINE IMAGES WITHIN LEGACY WEBSITES AT OXFORD



UNKNOWN FILE FORMATS IN APOLLO RESEARCH DATA FROM 700 DATASETS



THE FUTURE 2017-2018

PRESERVATION STORAGE REVIEW AT BOTH INSTITUTIONS ADVISING MOVE FROM CITY BOUND TO MORE GEO-DIVERSE LOCATIONS

BUILDING A FILE FORMAT RISK REGISTRY FROM DROID SCAN ANALYSIS IN QLIK SENSE

3 → 4.7 NEW PRESERVATION WORKFLOWS ADDED TO OXFORD REPOSITORIES WHEN MIGRATING FROM FEDORA VERSION 3

AUTOMATED ANALYSIS OF SCHEDULED DROID SCANNING OF REPOSITORIES AND CHECKSUM FIXITY CHECKS ON DIGITIZED CONTENT

CAMBRIDGE TO INVESTIGATE DIGITIZATION WORKFLOWS WITH GOOBI / APACHE ACTIVEMQ & CAMEL

COLLABORATE WITH RESEARCHERS TO CREATE FILE SIGNATURES TO IDENTIFY UNKNOWN RESEARCH DATA FORMATS AND SUBMIT TO PRONOM

PRESERVATION SYSTEM REVIEW AT CAMBRIDGE INCLUDING: PRESERVICA, ARCHIVEMATICA, FEDORA, ROSETTA AND RODA

MIGRATION OF LEGACY WEBSITE IMAGES AND METADATA TO DIGITAL.BODLEIAN OXFORD'S ONLINE IMAGE REPOSITORY

JAMES MOONEY, BODLEIAN LIBRARIES, OXFORD | **DAVID GERRARD**, CAMBRIDGE UNIVERSITY LIBRARY