Hull digital archive infrastructure

Step 1 – The starting point for processing digital collections is a secure and safe store. Following internal assessment at the University of different storage options for this purpose, we are making use of Box (Links to an external site.) for this. Additional storage options are being explored, including OneDrive, to support different starting points, and the overall system can be adapted to meet requirements.

Step 2 – Once stored the files can be appraised using BitCurator (Links to an external site.), a widely used open source tool for appraising digital collections and identifying what a collection holds. For instance, this can help to clarify which files do not need to be preserved (e.g., duplicates) and which files hold personal information that needs to be managed sensitively or removed.

Step 3 – Once appraised, files require a basic descriptive metadata record to inform their management. Experience has shown that the more metadata that can be captured at this 'pre-ingest' stage, the better. This can vary from bibliographic and permission metadata, created manually, as well as technical metadata captured through automated means. This task can be done manually using Excel, or through scripts that extract relevant file metadata from a collection.

Step 4 – The files are now ready for processing for preservation. We are using the open source Archivematica (Links to an external site.) service for this, a tool that effectively combines a host of preservation micro-services so that they can be applied together. Files within Box are shared with Archivematica, and the resulting outputs provide archival and dissemination packages combining files and associated metadata.

Step 5 – It is at this point that the files are either archived, or processed further to manage them for access. Archival packages are pushed to a long-term store in the cloud (see also later), while dissemination packages are unpacked for ongoing processing.

Step 6a – The files and all associated metadata are ingested into a Hyrax (Links to an external site.) digital repository. This open source repository solution can be used for a range of digital content management tasks, including which files are within which collections, and the management of permissions for access. Any amendments to metadata to support ongoing management of the files can also be carried out.

Step 6b – A subset of metadata describing the files is also pushed to our archival cataloguing system, CALM (Links to an external site.). The relevant archival collections are set up within this system, and the initial metadata is then further developed to generate full archival records and findings aids. Connections are maintained with the files held in Hyrax through appropriate linked identifiers.

[**Workflow** – whilst not a step in itself, the system automates the movement of the files and their metadata between the different systems via the Hull Synchronizer (Links to an

external site.) workflow application, a dedicated open source development created by CoSector (Links to an external site.) and Cottage Labs (Links to an external site.) that enables the combination of the best of breed components used.]

Step 7 – Records held within CALM are exported for access through the Hull History Centre catalogue (Links to an external site.), built using the Blacklight (Links to an external site.) discovery system. Archival records within the catalogue that have an associated digital file will have a link added back to the Hyrax repository, and the files displayed to the end-user, where possible, via use of the Universal Viewer (Links to an external site.) plugin to the catalogue.

NB. The default presentation of archival files will be through the HHC catalogue interface. Files can also be surfaced through other interfaces according to need.

Step 8 – Enhanced access to collections, or a subset of them, can be provided through the development of an online exhibition, using the Spotlight (Links to an external site.) plugin to the catalogue. This takes individual files and displays them alongside specifically composed text to showcase particular themes and ideas.

Deployment and the cloud

The open source components of the system overall are designed to be deployed either locally or in the cloud, connecting to the commercial services via APIs. The different functional components are packaged into Docker (Links to an external site.) containers, which have been tested for use on Amazon Web Services and Microsoft Azure. We are currently operating the system on Microsoft Azure (Links to an external site.)'s cloud platform, and this also provides the store for the archival packages created by Archivematica.

Chris Awre
c.awre@hull.ac.uk