Contents

Introduction 4
The University of Hull and digital archiving 4
Archive design and sustainability 4
Use case: City of Culture born digital archives 5
Use case: digitisation 5
Our technical platform 7
Introduction 7
System architecture 7
Deployment and the cloud 8
Benefits and impact 9
Our service options 10
Preservation planning consultancy 10
Preparing for preservation – getting digital files in shape 10
Processing digital files for preservation 10
Archival store provision 10
Managing digital files 10
Cataloguing digital archival records 10
Delivering access to digital files 10
Developing online exhibitions 10
Conclusion 11
Next steps and contact details 12
Introduction

One of the great advances enabled by information technology has been the ability to create documents and images, plus many other types of digital file, and to share information and knowledge using these. With continuing software developments, this ability has spread and the relative ease with which digital materials can be created has increased. The flipside of this advance has been a growing requirement for ways to store what has been created, especially over time. Solutions for effective ways to store digital materials are themselves developing, but the ability to proactively manage files over time so that they are not just stored, but also managed to enable access in the future, has been playing catch-up given the scale of digital growth. How can digital materials be effectively archived and have preservation actions carried out on them to ensure the materials continue to accessible and useful?

The University of Hull and digital archiving

The University Archives has had a long-standing interest and involvement in digital archiving. Early engagement in a Jisc project looking at the relationship between records management and digital preservation (REMAP, 2007-9) led to involvement in the Mellon-funded AIMS project (2009-11) that developed an inter-institutional model for stewardship of digital materials. Learning from these projects led to a proposal within the University to develop a digital archive of the City of Culture, once it was announced in 2013 that Hull would the host for this celebration of culture in 2017. A subsequent Jisc project with the University of York (Filling the Digital Preservation Gap, 2015-16), which looked specifically at archiving of research data, helped inform the proposed technical solution for the 2017 archive, and the University has been working in partnership with CoSector at the University of London since 2018 on its development. Access to the City of Culture Digital Archive will be provided from early 2021, with new areas being released over time as the collection (>130,000 files) is processed.

Archive design and sustainability

In setting out to create a digital archive the University was faced with a standard question when implementing a technology solution: do we buy an existing off the shelf system for a one-off solution or do we develop a bespoke solution to meet our needs? We opted for the latter, for the following reasons:

- We wanted to have flexibility in how we developed and built the archive, as we did not know what the collection might present us with and what we would need to do to manage it.
- With digital archiving requiring a range of technical tasks in the overall workflow, we recognised the benefits of combining individual system components that are best of breed for the specific functionality they offer over a single solution approach.
- We wanted to ensure that that the technical solution implemented for the City of Culture was re-purposable for other University digital archiving needs, to manage costs overall and support ongoing sustainability.

This paper describes the use cases we have encountered in further detail, and then looks at the solution that has been put in place and the benefits/impact this can provide. Having incorporated flexibility we are now interested to engage with partners of the University (and other local organisations, as appropriate) to understand whether the technical platform put in place may be of value in supporting digital archiving needs beyond the University. A range of potential service options is presented to help explore needs; feedback, and ideas, on these and/or on what would be of most value is welcome.

1The terms ‘digital archiving’ and ‘digital preservation’ are occasionally used interchangeable, or with one being a subset of the other. Within this document, digital archiving is taken to be the holistic activity of looking after digital materials, which encompasses specific digital preservation actions that inform this.
The idea for the Hull City of Culture Digital Archive was conceived shortly after the announcement in November 2013 that Hull was to be the holder of the 2017 title of UK City of Culture. Knowing that 2017 had the potential to be a complete game-changer for Hull, it was seen as crucial to capture a historical record of the year. There was a strong desire to document this time to guarantee that decisions made, works created, residents engaged, visitors attracted and money spent were chronicled and accessible to researchers in the future.

We approached several depositors; including the Culture Company that delivered the 2017 programme, major partners such as the university itself and artistic and community contributors and over the course of the year collected more than 1.5 TB of digital archive records, totalling more than 130,000 individual files. The most significant deposits in terms of size were:

- >80,000 digital organisational records extracted from the Culture Company’s SharePoint system
- >18,000 high quality image and video files with descriptive metadata extracted from the Culture Company’s Asset Bank DAMS

The majority of the collection is made up of fairly common, well-supported file types that broadly can be grouped into:

- Organisational records e.g. Word documents, Excel spreadsheets, PowerPoint presentations, pdfs etc.
- Audio-visual records e.g. jpegs, tiffs, mp4s etc.

These make up the bulk of the 152 different file formats (i.e. PRONOM IDs) identified within the collection though there are also small amounts of more esoteric files.

The university has steadily been collecting born digital material alongside traditional paper-based records for more than ten years but the City of Culture collection represents the first major collection to be almost completely born digital in format. In order that the story of the 2017 year of culture could be preserved and made accessible for years to come, we were in need of a solution that saw the records through a journey of deposit, archival processing and repository presentation and management.

Whilst the City of Culture collection was the catalyst for starting work on the digital archive infrastructure, it was important that the solution would also eventually be usable for other digital university collections. It was also important to ensure some continuity with existing analogue archival procedures and therefore to be able to integrate with the other systems that we use for example, CALM cataloguing software and Blacklight for online discovery.
Use case: digitisation

A key part of most 21st Century Archive services’ preservation and access strategies will involve the digitisation of analogue material, whether that be paper, parchment, photographs, film or video. The University of Hull is no exception. The output of any digitisation project will of course be digital files. Such files, whether resulting from digitisation or having been “born digital” require the same level of preservation to ensure that they endure.

The majority of in-house digitisation done by the University of Hull is ad-hoc in response to requests from researchers. The digital output is currently saved on a network drive in a folder structure that mirrors the catalogue structure. In addition to this, there have been several items and collections digitised externally, especially film and video. As digitisation of images and video is likely to result in very large files, this digitisation output has been saved on external hard drives and a non-networked workstation in order to free up space on network drives. This presents a risk of data loss as all hardware will eventually fail and so regular backups need to be made.

Historically, the University of Hull archives have been able to provide access to digitised material on request on a very small scale: through email, web transfer services and copying files onto physical media such as CDs and USB sticks. This is time and labour intensive and often relies on researchers having direct contact with archivists to know that digital copies are even available. It does not allow for serendipitous discovery of digitised archives.

All the above considered, we sought a solution that would:

- Allow for digitised material to be archivally processed in a uniform manner
- Offer the same level of security and data integrity as required for born digital records
- Allow us to link digitised material to their catalogue record
- Provide researchers access to digitised material through our existing online catalogue alongside their archival descriptions and other important metadata

This digital archive solution removes the onus on the archivist to manually retrieve and disseminate digital surrogates on request. It allows the researcher to find out instantly if there is a digital surrogate of their desired record - and if there is, to access it immediately. This self-service experience can be accessed anywhere. Whilst the archivist has to learn how to prepare digitised content for ingest into the digital archive system the process of cataloguing and exporting to the online catalogue remains broadly the same. This continuity is welcome in the context of a service having to come to terms with new ways of working.
Our technical platform

Introduction

Taking a collection of digital files and managing them to preserve and provide access to them requires a range of technical functions to cover the different steps needed. Files must be appraised for their suitability for preservation, and have descriptive metadata generated to describe them; the files need to be processed for preservation, enabling relevant preservation metadata and derivatives to be created, and have the correct permissions set against them to guide which files can be accessed, and how; and the files, within appropriate collections, need to be presented in useful ways to highlight their value.

Recognising the breadth of functionality required to fully manage and preserve digital collections, we have chosen to combine the strengths of different systems that best address different parts of the overall process. This provides not only focused capability at each step, but also the flexibility to adjust the overall process to meet the specific requirements of different collections.

One of the key highlights in the development of digital preservation technologies has been the way open source software and services have contributed to what can be achieved. The open availability of micro-services that can process files for preservation, e.g., JHOVE and the file format registries, has been hugely beneficial in the development of broader preservation systems. Building on this, our platform exploits these and other best of breed open source components with relevant commercial services to provide an effective combination of capabilities and an overall preservation and access workflow.
**System architecture**

**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 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, 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.

**Step 4** - The files are now ready for processing for preservation. We are using the open source Archivematica service for this, a tool that effectively combines a host of preservation microservices so that they 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 later), while dissemination packages are unpacked for ongoing processing.

**Step 6a** - The files and all associated metadata are ingested into a Hyrax 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. 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 Hullsync workflow application, a dedicated open source development 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, built using the Blacklight 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 plugin to the catalogue.

NB. The default presentation of archival files is 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 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 containers, which have been tested for use on Amazon Web Services and Microsoft Azure. We are currently operating the system on Microsoft Azure’s cloud platform, and this also provides the store for the archival packages created by Archivematica.
Benefits and impact

Much archiving and preservation over the centuries has been by accident, through objects and documents being hidden away so nothing untoward can happen to them. It is likely that materials created today may well be preserved in some cases through exactly the same process. But whilst this approach can be effective for physical materials, it is highly unlikely it will work for digital materials, not least because files ‘found’ in 50 or more years time may be unintelligible without knowing what format they are, and the context and technical environment in which they can be used. The most effective way of giving future users of the material a chance is to capture information about the materials (the files themselves with useful derivatives and metadata) and archive this. Even physical documents from thousands of years ago needed the ‘metadata’ of the Rosetta stone to help make sense of them, and the same principle applies to digital materials created today.

It is this desire to ensure that the materials from the City of Culture 2017 Digital Archive can be accessed as part of the legacy and history of Hull that has informed the development of the digital archive. The benefits we are looking to accrue from actively preserving the collection, and the impact from these, are listed below:

- **To maintain a record of the digital files in a meaningful and intelligible form**
  
  In the same way that any formal record of an event or a period of time benefits from organisation, description and context, maintaining the files in a usable state will inform future use.

- **To create a record of how to work with and use the digital files being archived.**
  
  An addendum to the first point above, but of specific relevance to the archiving of digital materials where recording the technical information about a file is imperative in ensuring it can be used in the future.

- **Ensuring there are copies of files in suitable file formats to facilitate access**
  
  Many file formats in use today will continue to be in use for some time. But some will not. Ensuring there are copies of files in formats that have a longer anticipated lifetime (e.g., through using an open rather than a proprietary format) will ensure files don’t become inaccessible.

- **Streamlining the archiving of files through appraisal**
  
  Using tools that provide insights into a collection of digital materials can provide information that clarifies which files actually need to be preserved into the archive. Many collections will have duplicate, empty or unnecessary files, that can be excluded and focus effort on where it is needed.

- **Showcasing collections**
  
  Archiving individual files can be of benefit for the reasons already stated. But it is rare that files exist in isolation, with most being within collections. Archiving digital materials lays the ground for showcasing these collections, or presenting new collections that facilitate interpretation.

  Ultimately, digital archiving will also provide transparency of action carried out in the management of digital files. This in itself will convey integrity to the materials and provide trust in what they contain when viewed and used in future years.

---

2 For example, the word processor WordStar was popular in the 1980s. It was overtaken in popularity by other word processing software and now WordStar files (.WSD, .WST, .WMC, .WS2) are a thing of the past.
Our service options

We are keen to understand what your needs for digital preservation are to help inform how we may be able to help. To inform a discussion, we have identified the following services that we believe can be of value in supporting digital preservation needs. These combine the automated and manual tasks that are required to make digital archiving effective, and take their lead from the experience of working in these areas at the University. We do not wish to be limited by these, but present them as the basis for discussion.

**Preservation planning consultancy**

**FORMAT – WORKSHOP / DOCUMENT**

The start of digital preservation is working out requirements and assessing precisely what is needed. This is not a technical conversation, but one that looks to identify what digital preservation can offer, what the goals will be, and how to achieve them. The output would be a preservation plan document to inform consideration of potential solutions.

**Preparing for preservation – getting digital files in shape**

**FORMAT – WORKSHOP / DOCUMENT**

Once it has been decided to preserve a collection, there is a need to understand how to get the digital files into shape so they can be preserved effectively. This requires appraisal of the files and the development of metadata to provide initial description of what is being preserved. The output would be a guide document for how to carry out this work.

**Processing digital files for preservation**

**FORMAT – SERVICE DELIVERY**

Where it has been decided to ingest the files for preservation using the University's technical platform, work is required to assess whether this needs a separate instance of the platform, or whether the files can be managed through the existing implementation the University uses. The output from this is a serviced ingest of the digital files, carrying out of preservation actions against them, and the creation of a repository for long-term management of the files and associated metadata.

**Archival store provision**

**FORMAT – SERVICE DELIVERY**

Aside of the repository creation, the archival packages generated by the ingest will also be made available. These can be supplied for local storage, but can also be stored within the archival store associated with the technical platform, as preferred.

**Managing digital files**

**FORMAT – WORKSHOP / DOCUMENT**

Once ingested, the digital files will need management to ensure they can be used for long-term access. This will cover the management of files within collections, the management of the metadata associated with them, and the setting of permissions to inform access. The output will be a guide document to inform ongoing practice.

**Cataloguing digital archival records**

**FORMAT – WORKSHOP**

The ability to create archive catalogue records will depend on the cataloguing tool available. The focus at this stage is on how to create effective records for digital files, establishing principles that can be applied in the preferred cataloguing environment.

**Delivering access to digital files**

**FORMAT – SERVICE DELIVERY**

With the digital files ingested and managed for preservation, there will be collections that need to have access provided now rather than later. Options for using the HHC catalogue or a separate instance of Blacklight exist that can be applied to the scenario at hand. Alternative access through a different user interface may be possible, dependent on the integration required.

**Developing online exhibitions**

**FORMAT – WORKSHOP**

Going beyond providing access to digital files through a catalogue interface, this workshop will look at what is involved in creating an online exhibition based on a subset of the files available, making use of the Spotlight plugin to the Blacklight user interface.
Conclusion

To conclude, the Library and Archives at the University of Hull has been able to create a technical platform that is enabling it to create a digital archive for the City of Culture. This technical platform combines specific open source and commercial best of breed components together to provide an infrastructure that can be utilised for a wide range of digital archiving needs.

Archiving digital materials proactively using the platform will allow us to preserve the materials for years to come, whilst also showcasing those that can be made accessible over time. Building on this development, the University is now looking to explore how others in the region and beyond can benefit from the platform and associated digital archiving services.

Contact details

For further information or to discuss any of the topics within this White Paper, please contact:

Chris Awre  
Head of Information Services  
University Library  
Brynmor Jones Library  
University of Hull  
Hull HU6 7RX  
Tel: 07707 785223 / 01482 465441  
Email: c.awre@hull.ac.uk

Laura Giles  
Digital Archivist  
University Library  
Hull History Centre  
Worship Street  
Hull HU2 8BG  
Tel: 01482 317506  
Email: l.giles@hull.ac.uk

We are open and interested in conversations within the following areas:

• Managing and organising digital collections
• Preserving digital collections for long-term access and use
• Carrying out preservation actions on digital files
• Cataloguing archived digital files
• Delivering access to digital collections

…but let us know if you would like to discuss other topics related to these.