Joint Archives Service
Feasibility and costs of digitising and de-accessioning or removing existing collections

Terminology:

- **Digitisation**: the creation of a digital facsimile of a hard copy item or object
- **Digital preservation:** the process of ensuring the authenticity and future accessibility of digital content

Introduction

Archive services have for many years undertaken digitisation projects related to their collections. Usually this is to facilitate improved access (via online platforms) or to provide alternative means of access for damaged items. In all such instances as far as is known, the digital copy has been kept alongside and complementary to the hard copy archive, not instead of. The question is sometimes posed as to whether it would be possible and/or desirable to digitise hard copy archival collections and to then dispose of or de-accession the originals, thereby saving space and thus (in theory) reducing costs. This idea has been analysed a number of times in different places. Nowhere is it possible to identify an archive service which has taken this approach. Even if a digitisation strategy were approved and adopted, there is no 'quick fix' – the process of converting hard copy archives into digital surrogates would be both expensive and involved. Beyond the financial costs, there are both practical and ethical reasons why relying purely on digitised content may not be deemed the right approach.

Summary of likely resources required to digitise JAS collections.

The workings below are based on the principle of 'best endeavours' and currently available information. A much fuller survey for collections would be necessary to obtain more granular detail.

- There are over **15 million pages** and **4500 hours** of audio-visual material currently held in the repositories at DHC.
- Total cost of digitising 1 box including processing and 60 years' digital preservation (equivalent of the full life cost of a building): **£12,000**. To put this in context, 1000 boxes (an average year's intake at DHC) would cost: **£4.5 million** and take **3 years.**
- These take all formats from single pages or photographs to large volumes, complex maps, cassette tapes and film material
- The estimated cost of digitising all of these is **£154.6 million**
- The resulting digitised files would make up **1.34 Petabytes** of data. The estimated cost of digital preservation of this data is approximately **£485,000 per annum.**

- The data stored in the digital repository would generate **906 tonnes of carbon emissions per year**.
- The time required to conserve, digitise, process and upload to the digital repository would be in excess of **149 working years.**

Considerations

Digitisation is a resource intensive activity. Before any image capture occurs, the physical condition of archive documents needs to be assessed and sensitive information and copyright considerations identified.

The method of capture depends entirely on the physical form of the document – its size, format and previous storage conditions all have an influence.

As well as the initial cost of capturing a digital image, an ongoing commitment is required to ensure the preservation and future accessibility of the image. This involves adding metadata to the image and preserving the file in an active preservation environment. The steps involved in a large-scale digitisation project are detailed in Appendix 2.

An illustrative example of the scale of digitisation is that of the British Library. Despite being a well-resourced organisation undertaking an extensive digitisation programme for many years, the British Library has only digitised 1% of its holdings.

If digitisation as the sole form of archival record keeping were to be agreed, it is likely that it would have to become the primary focus of the service to the exclusion of most other activities.

Ethical

Historic archives are items with an intrinsic heritage value related to the medium and manner in which they were produced as well as the information they contain. They are also the ultimate guarantor of authenticity. To consider this another way, it is unlikely that museum objects would be scanned, digitally captured and then destroyed, so the same question would need to be asked of paper and parchment collections. In general, but not absolute terms, if material is transferred in hard copy form, it is retained in that form – born digital material likewise stays in its digital format. There is a case to be made for the scanning and digital preservation of some modern records, but only on a case-by-case basis e.g. where microform or paper copies have deteriorated to the point where informational loss is likely to take place.

Ownership

Whilst the JAS has custody of thousands of separate collections, it does not have title to many of them. Deposited collections would have to be returned to their owners.

Conservation

Experience would suggest that there would be significant work required in order to prepare documents for digitisation as many of them are in formats or a condition that do not easily lend themselves to being digitised effectively. This would be light-touch remedial conservation in order to facilitate the production of a useable and authentic copy of the original document. For example, a parchment roll which has been kept rolled for centuries will need conservation treatment to relax it and clean it before it can be digitised with an overhead camera.

Digital preservation

If destruction of the original archives were to occur there would need to be a robust process in place to ensure that the newly generated digital copy will be accessible in hundreds of years' time. Digital files are by their nature much less stable than paper and parchment held in appropriate conditions. The oldest document in the collection, from AD965, has now survived for over 1050 years. Digital information has been in existence for less than 100 years. Risks to digital information such as obsolescence of file types, corruption and loss would have to be mitigated.

The digital repository in use by the JAS, Preservica, is an active preservation system designed to safeguard digital information long into the future. A substantial increase in volume of digital material to be preserved would require re-assessing whether JAS's needs would be better suited by an on-premise edition of Preservica rather than the cloud hosted edition currently in use.

Environmental impact

All digital activity relies on a digital infrastructure which has a heavy impact on the environment. From the mining of the materials to the use of resources until the disposal of our computers. Undertaking any large-scale digitisation project would therefore have a significant environmental impact in terms of equipment required to capture the images and the ongoing storage of the resulting files and metadata in energy-intensive data centres.

Both Bournemouth, Christchurch and Poole Council and Dorset Council are committed to being carbon neutral by 2030 and 2040 respectively. Additional investment and consideration would be required to ensure such a large digital project would not undermine these goals.

Legal issues

Aside from ownership of the collections, other factors influence what can be digitised and made available for public viewing.

- Copyright – identification of intellectual property rights and the protections necessary for some content.
- Data Protection legislation – significant attention needs to be paid to this, to ensure that confidential or embarrassing information is not divulged as a result of its new digital format.

Confidential disposal of the records

In a 'digitise and destroy' project all archival records would need to be destroyed confidentially due to potentially sensitive information contained within the records. Records not owned by JAS would need to be returned. This is an enormous administrative task in itself and has not been included in the calculations presented here.

Reputational damage

If a 'digitise and destroy' policy were pursued, it would create an exceptionally unusual precedent – and something likely to lead to adverse publicity for the service and its funding councils. It is likely that The National Archives would get involved with a probable loss of Accredited status for the service. It is possible that depositors would be dissuaded from offering collections to the JAS if they felt that they could be destroyed following digitisation.

Benefits

Archives create digital surrogates of analogue material for two key reasons:

1. To facilitate long-term preservation of original documents: Very fragile or deteriorating items are digitised to allow future access without further damaging the original. For example, DHC digitised 8000 packets of negatives suffering from vinegar syndrome, an irreversible process of decay, thanks to grant funding and a successful crowdfunding campaign.

2. To improve access to popular collections: some documents are consulted very regularly and so copied to reduce handling of the originals and enhance access. For example, DHC already has 0.5million pages of the most popular Dorset records relating to family history available via Ancestry.

More information on JAS' approach to digitisation can be found in its Digitisation Policy (2020).

Methodology – some caveats

The figures in this document are for illustrative purposes only. It would be an enormous undertaking to make an accurate assessment of the costs of digitising and processing the entire archives. This is due to their variation, which incidentally is the reason that many people find archive collections so fascinating. The archive collections contain huge quantities and considerable variation in type, format and size. Each variation presents different challenges in the digitisation process relating to handling and readability, resulting in varying rates of digitisation. The calculations are based on best guess extrapolations from the collections currently held by the Joint Archive Service.

The cost of cataloguing collections has been limited to basic metadata applied during digitisation. To make best use of the resulting digital files additional cataloguing work would need to be undertaken.

The cost and timescale of confidential destruction of archive material has been excluded from the calculations.

The costs of staff administrative work to package and track documents being copied off-site have been excluded from the calculations.

Conclusions

Digitisation is a complex and time-consuming process rather than a 'quick fix'. It provides many benefits and as a form of additional or enhanced access provides an essential part of the archive service's ability to reach current and new audiences.

As has been demonstrated, the costs of digitisation and the activities flowing from it are potentially very large. The potential cost of creating sufficient storage space for incoming collections for a further ten years through a process of 'digitised and remove' would be approximately **£8.5 million**.

To view digitisation as a single or sole solution for Dorset's archives would be to place the JAS in the position of being the only service of its type to do so and would have consequences which are not yet fully understood. To make best use of available resources it is recommended that the JAS continues to prioritise collections for digitisation according to preservation and access needs.

Joint Archives Service

February 2021

Appendix 1: Calculations

A. Amount to be digitised

1. Number of boxes in the repositories at DHC
    - a. 56,000 total box spaces – 3,600 currently empty = 52,400 occupied box spaces
2. Items in an 'average' collection (nb, there is so such thing as an average collection, they are all by their nature unique)
    - a. D-BKL Bankes of Kingston Lacy Collection – approx. 900 boxes. This collection was chosen as it has been catalogued in sufficient detail to calculate quantities. It also contains a variety of sizes, formats and media.
    - b. Using catalogue data – there are 85746 individual documents and 1739 volumes in the collection. Assumption of average of 160 pages per book (including both sides of each page). Individual documents would require both sides digitising. Total pages to digitise would be 449,792.
    - **c.** 449,792 pages across 900 boxes gives an average of 499.7 pages per box.
    - **d.** Estimated number of pages across the whole repository (52,400 x 499.7) = **26187889.78**
3. Audio-visual collections: 6000 items identified as audio-visual in current listings.
    - a. Assuming an average length of 45 minutes per recording
    - b. Total hours to digitise would be 4500.

B. Cost of digitisation

1. Paper and parchment collections
    - a. 2010 report: digitisation of archives can cost between 4-15Euros, adjusted for inflation = 4.54 to 17.02. £4 to £15.03.
    - b. 2015 blog by digitisation company: digitisation costs between £0.20 to £3.90, adjusted for inflation = £0.20 to £4.35
    - c. The average of these is £5.90
2. Audio collections
    - a. These would have to be out-sourced as DHC does not have the specialist skills and equipment to digitise complex formats.
    - b. Previous project to digitise Dorset Sound Archive (DSA) oral history collection: £21.33 per hour of recording
    - c. Assume balance of audio recordings in total of audio-visual material is 20% = 900 hours
    - d. Total cost = £19200
    - e. DSA took 0.154 working days per hour of recording. This multiplied by 900 = 138.6 working days
3. Audio-visual collections (excluding audio only)
    - a. The cost of digitising these materials varies significantly depending on original format and condition
    - b. Previous digitisation of AV material cost between £30 and £55 per hour of recording. We have used the higher figure for calculation.
    - c. Total = £198,000
4. Volunteers
    - a. JAS has achieved several previous digitisation projects with the assistance of volunteers.

   b. The use of volunteers has not been included in these calculations and could be considered as a way to bring costs down. However, the management of volunteers also has a cost in staff time to train, supervise and quality check their work.
   c. The use of volunteers would considerably increase the time required to complete a digitisation project due to volunteer working patterns.

C. Total cost for digitisation of all archives currently held

1. Paper and parchment collections: **£ 154,377,610.24**
2. Audio-visual material: £217,200
3. Combined total = **£ 154,594,810.24**

D. Files sizes for preservation

1. For paper and parchment collections it is estimated that 20% may be suitable for saving as PDF/A – if they are flat, black and white printed documents.
2. For the remaining 80% it is assumed we would need to save a high-resolution image file, TIFF, as the preservation copy.
3. Using data from digitised collections already held in the digital repository, Preservica:
   a. Average size of a PDF/A is 66KB
   b. Average size of a TIFF is 45MB
   c. Multiplying these by the number of pages previously calculated in the proportions mentioned results in a total of 942Terabytes (TB) of data.
4. For audio-visual collections audio files would be saved as WAV and audio-visual files as MXF, in accordance with current best practice recommendations
   a. Average size of an hour long WAV audio file is 988.8MB
   b. Average size of an hour long MXF video file is 112.5GB
   c. Multiplying these by the number of hours previously calculated results in a total of 405.55TB of data
5. The combined total size of all files digitised is **1.34Petabytes** (1348TB)

E. Cost of digital preservation of digitised files

1. 2020-2021 pricing for Preservica is £5403 for the first TB in S3 (immediate access storage) and the first TB in Glacier (delayed access storage). Subsequent TB are £517 per TB in S3 and £195 per TB in Glacier
2. The cost of preserving the digital data in S3 would be £702,396.91 per annum
3. The cost of preserving the digital data in Glacier would be £268,292.32 per annum
4. A combination of both S3 and Glacier storage options would be used so the actual cost would be between these two figures. The current division is approximately 50% in each. Halving each annual cost and adding on the annual fee brings the total to **£485,344.65 per annum.**

F. Time to digitise

1. Based on data recorded from previous digitisation projects
   a. Average time to digitise one page of a loose document: 12.65 minutes
   b. Average time to digitise one page from a volume: 6 minutes
   c. Multiplied by amounts identified in section A above this equals 18078 hours for documents and 27720 hours for volumes, or **26.8 years** of one person working full time.

2. Audio-visual material must be digitised in real time. Including preparation of materials the time required to digitise sound and video material would be 5400 hours, or **3.16 years** of one person working full time.
3. Total time to digitise all material = **29.95 years**

G. Time to ingest digital repository

1. Based on current experience ingesting files to Preservica Cloud Edition
   a. With an average connection speed of 827mbps would take 16925 working days, or **73.3 working years** to upload
   b. Such a large ingest would need to be broken into sections. Staff administration time per TB of data is conservatively assumed at 30 minutes. Multiplied to the total amount of data it would require **91 working days** of staff time.
2. Total time to process and ingest digital files = **73.7 working years**.

H. Environmental impact

1. The energy required to save 1 GB of data to the cloud and store for 1 year is 7kWh. Multiplied by the data requiring storage in the digital repository, this would use 9437055kWh per year.
2. Converted to carbon emissions based on the UK Grid's emission factor this equates to **905.7 tonnes of carbon per year.**
3. Work to reduce the energy use of the physical repository over the past 3 years has resulted in a 259% reduction in energy use from 210378.9kWh in 2016-2017 to 58622.8kWh in 2019-2020.
4. Converted to carbon emissions based on the UK Grid's emission factor the physical repositories currently emit **5.6 tonnes of carbon per year**.

I. Cost per box

1. Where the cost of one box, 1000 boxes (1 year's space) and 10,000 boxes (10 years' space) has been given, these include the cost of digital preservation of the files created during digitisation for 60 years. In reality, this would be an ongoing financial commitment.

Appendix 2: What is involved in a large-scale digitisation project?

There are many stages to a digitisation project:

1. Assess the condition of the collection: ascertain which collections are suitable for each method of digitisation. For example, a single sheet of standard-sized paper could be easily scanned. A parchment roll which has been kept rolled for centuries will need conservation treatment to relax it and clean it before it can be digitised with an overhead camera.
2. Invite quotes for out-sourcing digitisation services: this may involve considerable work to identify the exact scope of the project. The size of the project is likely to be prohibitive to external digitisation organisations who would have to dedicate years to completing it.
3. Raise funds to complete the project: the budget for this work would need to be identified and secured before the project commenced.
4. Assess copyright ownership in individual items: if copyright ownership is unclear, or not assigned to Dorset History Centre, resources would need to be dedicated to identifying copyright holders and seeking permission before digitisation could legally commence.
5. Assess access status of individual items: records containing sensitive information subject to Data Protection legislation needs to be identified and a plan implemented to ensure its secure handling.
6. Assess ownership status of collections: approximately 50% of DHC collections are deposited, meaning legal ownership resides elsewhere.
7. Prepare collections for digitisation: every document needs to be labelled, listed and packaged appropriately to be transferred to the digitising organisation.
8. Capture to standard output formats: to maximise the long-term accessibility of the digitised images they need to be captured in a high-resolution, open source, non-proprietary format (e.g. TIFF, PDF/A)
9. Add metadata: information needs to be captured at the time of digitisation to ensure future accessibility of the image. This includes a catalogue reference number, brief description, dates, copyright status, access conditions.
10. Post-processing: application of tools to ensure the image meets required standards, e.g. cropping, colour editing. For large items captured in several parts, such as maps, images need to be stitched together to make one complete facsimile.
11. Quality assurance: staff time would need to be dedicated to checking sample batches of digitised images to ensure the expected high standards are met.
12. Transfer to digital preservation system: the digital master files and their accompanying metadata needs to be ingested into the digital repository
13. Any details to catalogue: to enable searching, details of all items digitised will need to be added to the archive catalogue system.
14. Apply access conditions: any items requiring restricted access due to Data Protection, copyright restriction or other legislation to be exempt from public access
15. Provide public access: as well as making digitised collections available online, archive staff will be needed to provide support to enquirers potentially overwhelmed by the volume of material available.

Appendix 3: Further reading

Joint Archives Service Digitisation Policy (dorsetcouncil.gov.uk)

The Cost of Digitising Europe’s Cultural Heritage, Nick Poole, November 2010: digiti_report.pdf (nickpoole.org.uk)

How much does digitisation cost? TownsWeb Archiving, 2015 How much does Digitisation cost? (townswebarchiving.com)

Why don’t archivists digitise everything? Region of Peel Archives, 2017 https://peelarchivesblog.com/2017/05/31/why-dont-archivists-digitize-everything/

Why don’t archivists digitise everything? University of Westminster, 2020 Why don't archivists digitise everything? - Finding and Using Digital Archives during COVID-19 - LibGuides at University of Westminster