Digital Preservation at Big Data Scales

Proposing a step-change in preservation system architectures

Dr. David Gerrard – Cambridge University Library – *dg509@cam.ac.uk*
James Mooney – Bodleian Libraries, University of Oxford – *james.mooney@bodleian.ox.ac.uk*
Dave Thompson – *digitally.dave@gmail.com*

This is the author submitted version: submitted 27th November 2017, accepted 21st December 2017
The version of record is:
Gerrard, D., Mooney, J., Thompson, D. (2017), "Digital Preservation at Big Data Scales: proposing a step-change in preservation system architectures", Library Hi Tech. Available at: https://doi.org/10.1108/LHT-06-2017-0122

Structured Abstract

Purpose

To consider how Digital Preservation system architectures will support Business Analysis of large-scale collections of preserved resources, and the use of Big Data analyses by future researchers.

Design / methodology / approach

Architectural reviews of existing systems. Experimental surveys of large digital collections using existing Digital Preservation tools at Big Data scales. Design of a proposed new architecture to work with Big Data volumes of preserved digital resources - also based upon experience of managing a collection of 30 million digital images.

Findings

Modern visualisation tools enable Business Analyses based on file-related metadata, but most currently-available systems need more of this functionality 'out-of-the-box'. Scalability of preservation architecture to Big Data volumes depends upon the ability to run preservation processes in parallel, so indexes that enable effective sub-division of collections are vital. Not all processes scale easily: those that don't require complex management.

Practical Implications

The complexities caused by scaling up to Big Data volumes can be seen as being at odds with preservation, where simplicity matters. However, the sustainability of preservation systems relates directly to their usefulness, and maintaining usefulness will increasingly depend upon being able to process digital resources at Big Data volumes. An effective balance between these conflicting situations must be struck.

Originality

Preservation systems are at a step-change as they move to Big Data scale architectures and respond to more technical research processes. This paper is a timely illustration of the state of play at this pivotal moment.

1 Introduction

Collections of preserved digital resources are becoming vital sources of Big Data for researchers. This paper describes how the architectures of digital preservation systems might evolve over the coming years to cope with new requirements driven by the use of Big Data in research. The need for preservation systems to support management decisions regarding very large sets of data is also covered. While it is also a very important, emerging topic, this paper does NOT discuss the long-term preservation of Big Datasets themselves.

Requirements related to the use of large-scale collections of preserved digital resources have emerged from research into Digital Preservation at two internationally-important university libraries: Cambridge University Library (CUL) and Bodleian Libraries, Oxford University (Bodleian). Furthermore, the architectural direction we propose here is based upon experience of managing a collection of nearly 30 million digital images at Wellcome Library. The purpose of the research was to develop business cases for funding staff and systems for preserving both digital resources and their technical and descriptive metadata. Setting strategies, policies and guidelines for Digital Preservation were also key outcomes.

To this end, surveys were undertaken at both libraries, involving a mix of interviews, online questionnaires and automated analyses of the digital collections. The surveys covered digitised materials, research outputs (publications and, increasingly, research datasets), 'born-digital' papers, manuscripts, and university records, and materials published in digital or AV formats. This paper focuses upon digitised images and research data specifically. It is thus a conceptual paper regarding the future relationship between Big Data and Digital Preservation that has been guided by hands-on research activity, and further directed by engagement with practitioners, researchers, and vendors from the Digital Preservation domain.

1.1 Working definitions of Business Analytics and Big Data

This paper discusses both the Business Analytical requirements related to preserving, managing and sustaining very large digital collections, and the emerging need to use such collections as Big Datasets for future research. Hence in this paper:

1. *Business Analytics* is defined as using (often very large) datasets, mostly generated by IT systems, to provide insight into the process improvement of the organisations where those data are collected (Laney, 2001; Manyika *et al.*, 2011; Normandeau, 2013). A crucial aspect of Business Analytics is the ability to use trend analyses to extrapolate, predict and hence plan for the future (Back *et al*, 2013). While the discussion of this topic in libraries (and beyond) often tends to focus upon analysis of website usage data (e.g. Black, 2009; Fagan, 2014), there have been some attempts to use data from other sources such as Library Management Systems and entry systems (e.g. Cox and Jantti, 2012; Collins and Stone, 2014; Showers and Stone, 2014).
2. *Big Data Analysis* is defined as using scalable computing infrastructure to store very high volumes of data, often in semi-structured or unstructured formats, and then processing those data with a variety of sophisticated algorithms to reveal insights about them (Chen *et al*, 2012). As such, Big Data Analysis is an inductive process which was even suggested to represent 'the end of hypothesis' (Norvig, 2012) during its early usage phase, though this has been criticised since (e.g. Kitchin, 2014). One library-focused use of Big Data analysis is the use of bibliometric data and semantic network mapping algorithms to reveal insights into the use of literature in research (Thelwall, 2008), though the use of digital collections in the Social Sciences (e.g. Karpf, 2012; Edwards *et al.*, 2013) and Digital Humanities (e.g.: Manovich, 2012; Kaplan, 2015; Terras *et al.*, 2017) perhaps represents the greater focus of Big Data Analysis in library-related domains.

2 Theoretical background

2.1 Introduction to key Digital Preservation concepts

This section focuses narrowly upon Digital Preservation concepts used in this paper. For a comprehensive overview of the topic, sources such as Brown (2013), Giaretta (2011) and Corrado and Sandy (2017) are recommended.

The UK's Digital Preservation Coalition define Digital Preservation as: "…managing digital resources over time and the issues in sustaining access to them (Digital Preservation Coalition, 2015)." A similar working definition created for the research described here is:

> *Sourcing digital resources worthy of preservation, getting those resources under control, and then maintaining the usefulness of those resources for the long-term.*

The concept of 'maintaining usefulness' is preferred as it implies *actual use* of preserved resources, rather than the *potential* for use suggested by terms such as 'usability' or 'accessibility'. The importance of being able to account for actual use of resources is vital: in order to justify the investment required to maintain often complex and expensive systems, the value provided by those systems must be constantly re-stated (Smith-Rumsey, 2010). This has obvious implications for the Business Analytical functions of preservation systems. Examples of how the information from such functions should be used are included in the Roadmap from the EU-funded 4C project, which recommends assessing the value of digital assets, the efficiency of preservation systems, and the cost of preserving digital assets across their lifespans. All such activities depend upon Business Analytical systems running constantly as a key part of preservation infrastructure; indeed, the 4C Roadmap states that: "…despite the long-standing tradition of human appraisal of assets (i.e. deciding what to retain), for many organisations data has grown to such an extent that it is no longer feasible for this to be done by a person. Appraisal has to be (at least) semi-automated to be scalable and 'value' is an essential concept that will need to be algorithmically defined (4C Project, 2015:6)."

Undoubtedly *the* key model that has shaped the direction of Digital Preservation since the late 1990s is the *Open Archival Information System (OAIS) - reference model* (Consultative Committee for Space Data Systems, 2012). This model contains three key sub-models, the Submission Information Package (SIP), responsible for managing the ingest of data into a preservation system (the 'getting things under control' part of the process), the Archival Information Package (AIP), responsible for the long-term management of materials, and the Dissemination Information Package (DIP), which enables materials to be found, accessed, and used.

The 'original' digital materials themselves, and the metadata extracted and created in relation to those materials, will become valuable Big Data sources. For example, technical metadata related to digital images can be used to track trends in image digitisation over the years, and thus (potentially) start making predictions about and planning future processes. Technical metadata, through the use of an open schema such as PREMIS (PREMIS Editorial Committee, 2015), could provide a research resource in its own right by supporting the profiling of comparative collections. Similarly, metadata about Research Data submissions will be of interest to future scholars, in the way that the ephemera surrounding scientific figures of the past has become a source of information for today's historians.

For preservation purposes both descriptive and technical metadata are created, ideally, at the time of acquisition. Thus the events that occur during digital resources' lifetimes can be recorded from the start, and these metadata will become of interest not only to the managers and funders of the preservation systems, but also to future researchers. Inconsistencies in the way that descriptive and technical metadata are (or are not) collected and/or created will affect the management regime a collection is subject to, and the overall value of the collection. Thinking about Digital Preservation in the Big Data age suggests new models in which 'information about resources' will often be more important than the objects themselves, as such information is going to be both input into, and output from, sophisticated analytical toolsets.

Another factor in 'getting things under control' is the technical process of recording the characteristics of digital resources. This 'characterisation' process often depends upon tools developed by digital preservation community members, some of which were used in the surveys described here. One key characterisation process is format recognition, in which the information required to tie resources back

to the software needed to use them is discovered. Extracting information about specific files (e.g. the colour profiles of image files, or the text encoding of written documents) is another important characterisation step which helps support future use.

2.2 Architectures of Digital Preservation repository IT systems

The following section contains a short review of three Digital Preservation systems. The systems reviewed were:

1. ExLibris's Rosetta [i].
2. Artefactual's Archivematica [ii].
3. The E-Ark Project [iii].

The systems were selected because their architectural documentation is in the public domain, and because aspects of their architectures contribute to the discussion of scalable systems. However, other scalable systems exist.

Rosetta is a proprietary system built from various components arranged across three tiers: storage, application and database. These components are responsible for various functions informed by the OAIS model, such as ingesting content and delivering it to users. The sub-component arrangement enables scalability: for example, one could replicate delivery components and load-balance traffic between them in a high-volume end-usage scenario (Ex Libris Ltd, 2017).

For Business Analytics, Rosetta has an integrated reporting database which super-users can access using open source Business Intelligence Reporting Tools (BIRT) (Kutner, 2015). Another relevant component is Rosetta's index, based on Apache Solr. Metadata about files stored by the system are indexed, and custom fields can be added. 'Descriptive' metadata fields in particular might be a fruitful Big Data source for researchers, if such metadata were to include, for example, transcripts of digitised manuscripts. Again, load-balanced servers can be dedicated to indexing, enabling scale-out to Big Data volumes. This would currently be a by-product of Rosetta's Solr index, however, not its intended function.

Archivematica is an Open Source preservation management system with a design based in part upon the micro-services architectural pattern. According to Lewis and Fowler (2014), this pattern promotes designing components that are: "… independently replaceable and upgradeable". One advantage Archivematica gains from its use of micro-services is flexibility, as it enables users to customise various preservation workflows using third-party components wrapped in micro-services, but micro-services may also be scaled-out independently, and scheduled to run in parallel, too. Archivematica was one of two preservation systems (alongside Preservica [iv]) selected by the UK's Joint Information Systems Committee (JISC) to provide preservation for their Research Data Shared Service project [v]. As a result of this, at time of writing, its micro-service components were being wrapped in individually deployable Docker containers for use with cloud infrastructure.

Archivematica also generates an index of metadata in Elasticsearch. As with Rosetta / Solr, this indexing function has primarily been created to enable searching for stored content, but could also be used as a source of Big Data.

Unlike the two previous examples, the *European Archival Records and Knowledge Preservation* (EARK) Project is not production software, but is instead a demonstration application. One of the core objectives of EARK was to model the OAIS Information Packages (SIP, AIP and DIP) in ways which supported contemporary use cases, including data mining. EARK was also developed using enterprise-scalable platforms such as Apache Hadoop, HBase and Solr, and a task manager to orchestrate processes across server clusters (Alföldi *et al.*, 2014).

A key EARK Big Data experiment was the Data Mining Showcase (Schmidt *et al*, 2017), which explicitly leveraged the Solr index created with AIP data as a Big Data source *and sink*. An experiment was conducted using a Named Entity Recognition (NER) tool (i.e. the sort of Natural Language Processing tool used in Digital Humanities research) across the descriptive metadata in the index, and the results from this experiment were *added back to the index*. Another key aspect of this work was that, due to the underlying infrastructure, the NER process could be scheduled to run in multiple parallel processes to increase throughput.

3 Questions about large-scale Digital Preservation

Three questions regarding maintaining the usefulness of preserved digital resources at scale are outlined below. The research conducted by the two libraries provided an opportunity to address these.

Firstly, the surveys at both libraries raised questions regarding management of large-scale digital collections. For both digitised images and research data, the surveys provided an opportunity to ask: how might information about digital collections (e.g. the rate of file creation, and the increases in required storage capacity) be captured and presented in ways that enabled predictions about future infrastructure? Then, in the case of digitised materials, the surveys provided an opportunity to extract file characteristics and review properties such as file size, resolution and bit sample rate. In the case of research data, key extra information was sought about the formats researchers had used.

The second question concerned collection scanning tools: how might third-party tools for preservation tasks such as characterisation and format recognition be run across large collections?

The third question concerned how future researchers might want to analyse preserved digital resources at scale. What sort of tools and techniques might researchers want to use in future? How might these tools scale to work with large collections? And how might research results be fed back into the system, so that the work could be properly understood, and the results re-used in future?

4 Experimental methods used

Two principal techniques were used to address the questions above: file scanning and data analysis and visualisation. File scanning used the tools DROID [vi] and JHOVE [vii], which are open source tools developed by the Digital Preservation community. These tools are often used to manage ingesting digital

materials into preservation systems, but they can be used to scan large collections of materials to collect file-level metadata, identify file formats and extract more specific pieces of technical metadata. In CUL, very nearly all of the collected digital materials (both images collections and research data) were stored on large-scale file storage which was relatively easy to access and scan, but at Bodleian, the image collections were stored on tape, meaning 100 terabytes of scratch space was needed to restore the data to before scanning.

Scans with both tools produced sets of metadata for import into a data analysis and visualisation tool to attempt to produce useful information for preservation business cases. The tool chosen for this purpose was Qlik [viii], an application for rapid development of data dashboards that enables users to generate dimensions and measures from tabular data structures 'on the fly'.

The scans had the following intention:

- To use file metadata, in particular modified dates, to establish the rates at which files had been created, and (if possible) extrapolate these to predict future storage requirements.
- To use file format identification to establish the formats, and format versions, of research data files.
- To extract technical metadata from digitised image files such as width, height, colour space etc, and search for errors in the digitisation process such as files being saved in older versions of the TIFF format, or with the wrong colour depth, or at the wrong resolution.

DROID was used for file format identification, while JHOVE was used to extract technical metadata from the images. DROID works with format signatures from the PRONOM file format registry [ix] and searches both metadata from file headers and tell-tale bit patterns from file bodies to identify formats. New signature knowledge bases are released regularly, and users are encouraged to contribute signature information to these. Hence there is a requirement to re-scan collections with every new DROID release to check if files with previously-unknown formats are now identifiable. DROID also records basic file metadata such as the date the file was last modified, which was of use for Business Analysis.

Scanning with JHOVE also provided an opportunity for Bodleian to address the issue of running preservation tools at scale. Bodleian has invested in a 'private-cloud' infrastructure that allows virtual machines to be spun up and run 'on the fly', and for resources such as processor cores and RAM to be scaled-up for these resources as demand dictates. Bodleian's collections survey required 522,000 high-resolution TIFF files to be validated and characterised in depth, i.e. as much information about file size, resolution, TIFF version and colour-depth as possible needed to be recorded. The TIFF characterisation module in the JHOVE tool has three analysis settings, the most intensive (and therefore slowest) of which was required for full file characterisation, so scanning at this intensity provided an opportunity to see if and how such scans might be run across sub-sections of the collection in parallel using GNU Parallel [x] across multiple cores. Would this approach scale linearly as extra cores were added?

How preservation systems might support future digital research methods had been addressed in part by the EARK's Data Mining Showcase (discussed in 2.2 above), but this focused exclusively on textual analysis. Bodleian also had an opportunity, as part of their image scanning work, to assess how an image hashing, or 'fingerprinting' algorithm might be applied directly to the digitised images themselves, in order to discover duplicate, or near duplicate images. The Python ImageHash library (Buchner, 2017) generates a 64 bit hash 'fingerprint' from an image by scaling it into an 8x8 pixel greyscale image. One of

four options can be used to derive an average pixel value in the fingerprint from across the corresponding 1/64th section of the original. Fingerprints produced this way can be compared quickly with those of other images, hence images that are similar to each other can be found. This method was tested on a subset of 1000 images retrieved from Bodleian's collection and the feasibility of using it at larger scales assessed.

5 Results of experimentation

File scanning discovered 1.4 million 'master' TIFF files and 1.3 million JPEGs in CUL's digitised image collections. The scans also indicated that the DROID tool was slow to process large collections, but it was actually more 'heavyweight' than was necessary to capture the simple file metadata needed for file creation trend analyses in preservation business cases. A simpler, quicker Perl script was written for preliminary collection scanning, which revealed a steady accrual of images from which it was easy to extrapolate (Figure 1).

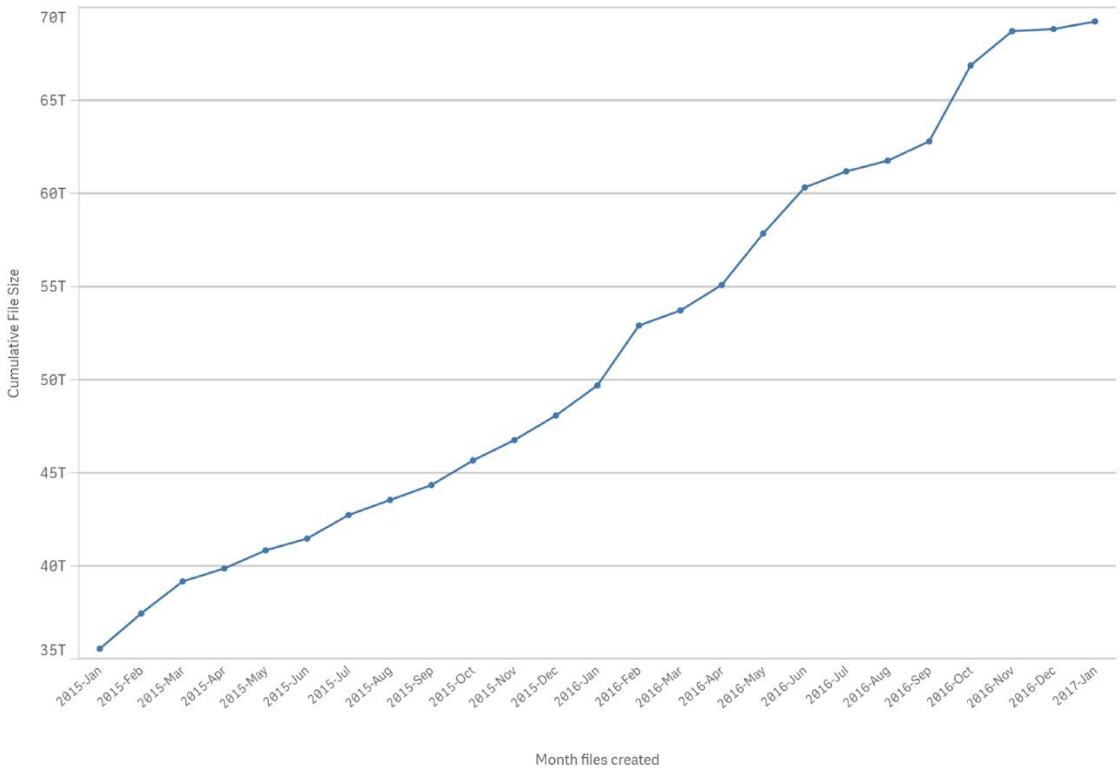

Figure 1 Rate of digital image accrual at CUL

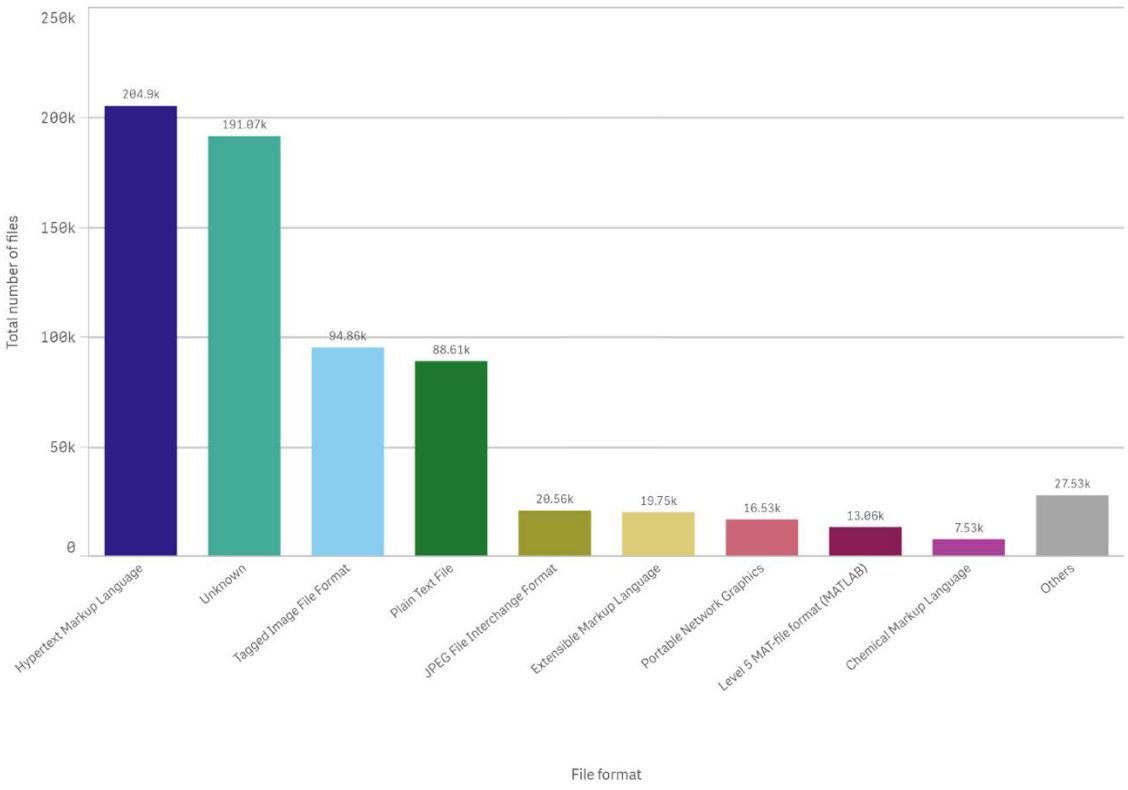

Figure 2 Variety of file formats in CUL's research dataset submissions

Using DROID to scan file systems where research data was stored enabled similar trend analyses to those shown above, and similarly contributed to business cases for Research Data Management (RDM). One particular RDM business planning issue concerned 'step changes' in data volume to which RDM is particularly vulnerable. Analysis enabled by using Qlik to link file scan information to submission data from CUL's DSpace Research Repository showed that one individual dataset submission of 126 GB accounted for approximately two-fifths of the overall total size of submissions, while the largest submitted dataset at Bodleian was approximately 0.5 terabytes. RDM business models need to account for such large submissions, perhaps by mandating extra deposition fees.

DROID's more heavyweight approach was more appropriate for scanning research datasets, as this content consisted of more varied and less well-known file formats. Scanning research datasets also revealed a high proportion of unknown file formats: 44% at CUL, 52% at Bodleian, and represented by the second bar in Figure 2. Unknown formats threaten the long-term usefulness of research data, which can only be mitigated by better understanding of the formats used in research, contributed to format registries such as PRONOM. This also mandates the constant re-scanning of files with DROID as new signatures are released, which has particular implications for preserving research data at scale.

Experiments in scaling-up TIFF characterisation with JHOVE scans indicated that the process would indeed scale as more processing power was added, though performance did start to level out. Figure 3 shows how processing volume of nearly 1200 images per minute could be achieved by running JHOVE in parallel, but also shows how performance tailed off as more cores were added. Memory profiling

undertaken as the process ran indicated that JHOVE was light on memory requirements, hence an initial allocation of 4 GB of memory did not need increasing. The performance levelled out, however, as a bottleneck related to file I/O was approached; this despite results being written to 16 separate output files and amalgamated at the end. It would be interesting to see if using a distributed, clustered file system such as the HDFS platform used by EARK would remove this bottleneck. Overall, running JHOVE in parallel reduced the overall time required to process 522,000 files from 2.5 days to 4.5 hours, and revealed circa 200 errors in the digitisation process such as thumbnail images being copied over master files, and images produced with inconsistent colour profiles. However, it is vital to note that the parallel processing that enabled this scaling was only possible because it was easy to sub-divide the collection into 16 self-contained chunks.

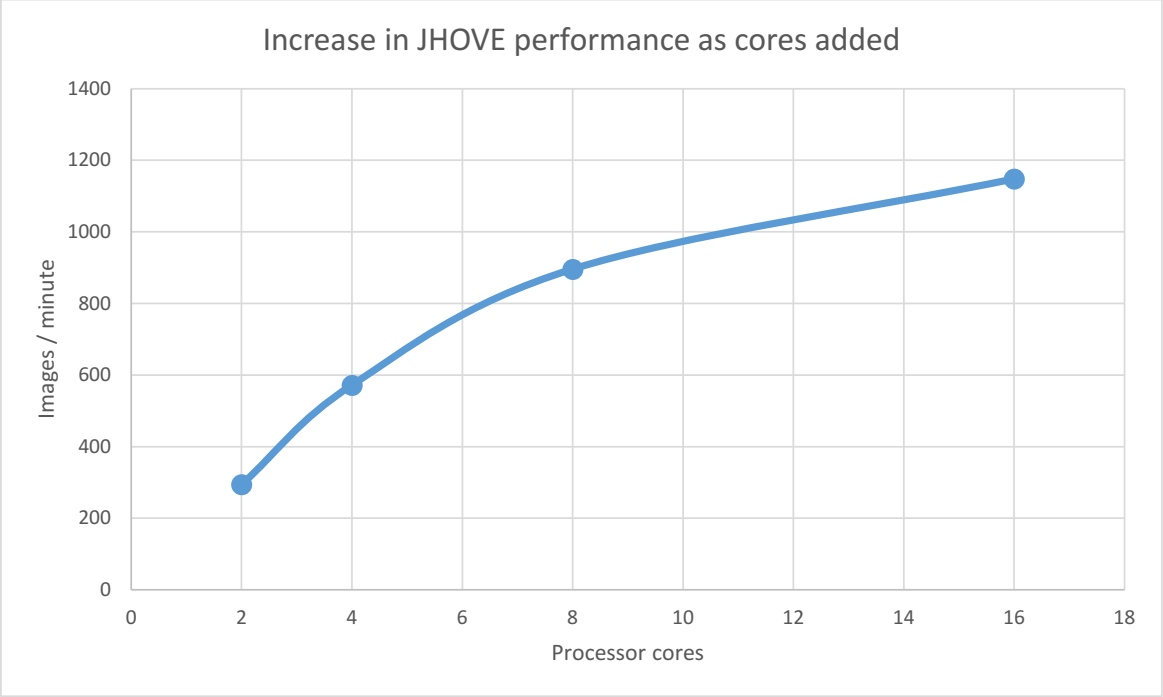

Figure 3 Increase in JHOVE image characterisation performance as cores added

Initial experiments with image hashing and fingerprinting at Bodleian revealed a much higher rate of image duplication than expected in the test set of images used. Figure 4 shows two near duplicate images; these were associated with two different digitization batches in addition to having different file names, creation timestamps, file sizes, and MD5 checksums. While the content of these images was similar, one had its tone and brightness adjusted.

However, one key difference between characterisation and the image comparison processes is that the latter cannot be parallelised in the same way as the former. The final step of the process of finding similar images is to compare each image with all the other images in the collection, hence the processing required *per image* increases in line with the overall collection size. This would obviously be mitigated by adding each image's fingerprint to its metadata, and is hence a good example of the

potential benefits of recording the output of digital research about the collection back with the collection itself, but this does not avoid the fact that pairwise comparison of each hash will still increase as the collection expands. Scalability of similarity measurement could be improved by generating hierarchical clusters of similar images, then comparing new images with exemplars of each cluster, and drilling down into the hierarchy to get closer to a match, but the effectiveness of this would depend upon the image content and how well clusters of similar images formed. Both libraries contain large collections of digitised book and manuscript pages that are likely to have similar fingerprints, and hence form very large clusters. Also, the overall effort of managing the collection increases as such algorithms increase in complexity.

Digitisation batch	AAG	AAH
File name	johnson-aag-0247.tif	johnson-aah-0382.tif
Creation timestamp	2001-08-13 14:15:54	2001-08-13 16:40:44
Resolution	3330 x 2511 pixels	3330 x 2511 pixels
File size	25090234 bytes	25089760 bytes
MD5 checksum	0d73e35e638c730923c61a9d1194b666	82a3db98923f93b4e1471c160f626433
Image hash fingerprint	ff5563e069a0965f93ba90cf834a6c3907a6e8938d822b2e39998f6cf9c38b88	ff5563e069a0965f93ba90cf834a6c3907a6e8938d822b2e39998f6cf9c38b88

Figure 4 Comparison of two images reviewed with image hashing / fingerprinting

6 Discussion

These findings indicate that high-volume Business Analytics of digital collection data can be of immediate use to building business cases for Digital Preservation, by indicating the accrual rate of valuable assets and enabling future predictions, both tasks considered critical by the 4C Project. The surveys also pointed directly at areas where processes could be improved. However, various issues were uncovered by these experiments, such as:

- Different operating systems handle dates inconsistently, making date-related file metadata unreliable. For instance, the varieties of Linux OS commonly have three dates associated with files, the date the file was last read, and two modification dates, one related to the content of the file, and one related to its 'i-node' (i.e. its manifestation on disk, which broadly equates to the file metadata itself). None of these dates represent the 'file creation date', and all may change as copies of files are made, backups restored, permissions changed, etc. In repository scenarios where files are, generally, created and then left unmodified, the i-node date equated closely enough with 'created' for the initial survey work described here, but preservation metadata standards like PREMIS have in part been developed to compensate for such shortcomings. Indexing more accurate date metadata based on PREMIS would provide more confidence regarding dates than extracting the metadata of stored files.
- Step-changes in the amount of data submitted makes trend analysis difficult. Research data are particularly susceptible to this.
- Certain 'heavyweight' processes such as detailed format analysis, or migration to new formats, will need to be re-run across collections constantly in order to maintain the usefulness of those collections. Thus, as collections scale, such processes will need to scale with them. This depends upon the ability to process in parallel, so it is vital that collections can be sub-divided into independent sections. Indexing is crucial to this.
- Processes such as format analysis, which operate independently upon individual files, will scale reasonably effectively, but factors such as file I/O on 'traditional' storage start to cause bottlenecks. More experimentation with highly-scalable file systems is required.
- Some processes, however, such as the experiment in image comparison described here, do not scale effectively. Algorithms can be developed to improve performance, but the more complex these become, the harder they are to manage.
- Opportunities to add the outputs of large-scale research analyses across large collection areas to the collection index should be considered at all times. Preservation managers should be aware of the research being undertaken upon their collections, and should strive to ensure that the collection as a whole benefits from such research, not just individual researchers.

Fundamentally, the OAIS, and the systems that have been guided by it over the past two decades, focus predominantly upon access to low numbers of resources by human beings, or potentially machine access via APIs (again at lower volumes). Supporting Big Data use will require both architectures and access policies that are significantly different to those currently available. This points towards a clear direction for next-generation preservation system architectures.

6.1 Next generation, Big-Data-friendly preservation system architectures

Figure 5 proposes a next-generation Digital Preservation architecture that has scalability to Big Data volumes of data management and processing as its core requirement. The architecture assumes hosting in the Cloud (or upon a private Cloud-like platform), where memory management is distributed, all components can be scaled-out horizontally to meet peaks in demand for various micro-services, and all parts of the system are decoupled, with orchestrated scheduling in-between. At the core of the architecture are three major components: multi-tiered storage upon which the 'original' digital content

is stored, a metadata database (which would comprise various technologies such as triple-stores and full-text indexes, alongside 'traditional' relational databases). A reporting warehouse is also shown, to support Business Analytics. As with all components, these databases could be scaled-out horizontally as data volumes increased. This 'scalable data core' at the centre of the system is similar conceptually to the database tier and Solr index in Rosetta's architecture (reviewed in Section 2.2).

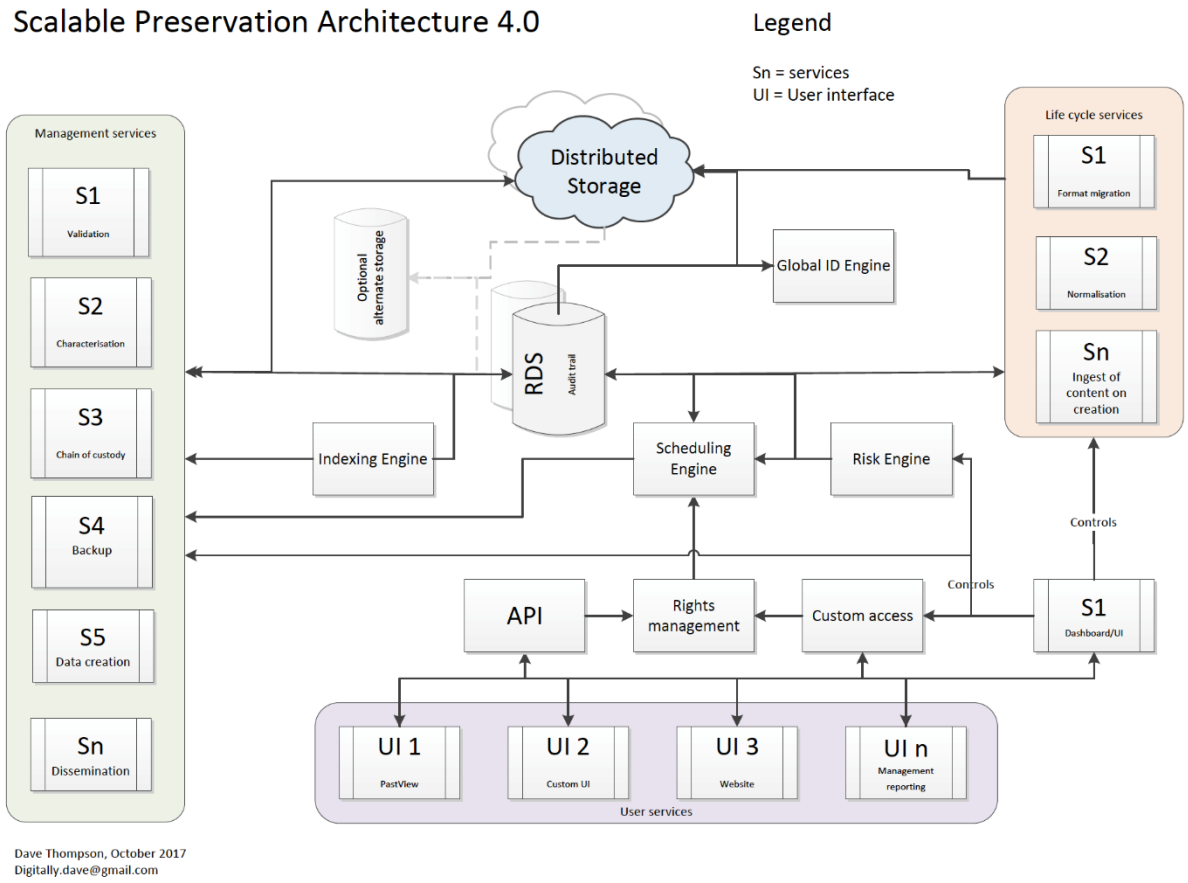

Figure 5 A proposed scalable architecture for a Digital Preservation repository

Surrounding the core are micro-services that support the 'traditional' preservation activities discussed in Section 2. Big Data analyses would also be configured as services 'around' the central core. This could enable various human-readable views upon the preserved data to be developed, and could allow messages about usage of delivery copies of preserved resources to be consumed by the preservation system for Business Analytical Purposes. This 'micro-service-based' approach is similar to Archivematica's approach.

Other aspects of the model that enable management of preserved resources at scale are:

- A Scheduling Engine, enabling multiple instances of the self-contained micro-service processes to be spawned on demand and run in parallel. By moving to containerisation services for the JISC RDSS Project, Archivematica's architecture was a moving towards this at time of writing.
- A Global Identification Engine, which would need to be generate sets of ids for digital resources that could be used to define sub-sections of the collection clearly. Ids would also need to be genuinely 'global'; mapping to externally-recognisable ids such as Digital Object Identifiers (DOIs) would be key to the process of accounting for the usage values of resources.
- A metadata Indexing Engine; not just to enable free-text search of resources, but explicitly as a source of Big Data for research, and a sink to potentially add research findings back into, as per the EARK project. The entire scalability of the system depends upon careful index design that incorporates preservation and technical metadata as well as full-text indexes of descriptive metadata. Large-scale parallel processing depends completely upon being able to divide large collections up according to particular metadata fields.
- A Risk Engine, which would use Business Analytical data from processes such as file format recognition to highlight risks, and manage their mitigation. One such mitigation might be migrating files from one format to another, again by spawning and scheduling self-contained parallel processes. The overall value of the system over time depends upon these mitigations of preservation risks. The key source of information for the Risk Engine would be the kind of data warehouse already provided by Rosetta.

Letting researchers use preserved digital resources, or metadata about those resources, as sources of Big Data has implications regarding access and usage rights management, however. Archives and other content management applications tend to have fine-grained and complex rights associated with their contents, and it would be potentially very easy to let resources slip out into wider usage than they ought to when such resources, or the metadata associated with them, were hosted in a system used as a Big Data source.

Another issue which must be considered is the overall complexity of the infrastructure required. Holding 'preserved' digital resources on infrastructure as complex as a Hadoop File System, for example, could be considered a risk to the sustainability of those resources, if the host organisation cannot easily afford to sustain the skillsets and technology required to maintain those systems. Hence the architecture also includes provision for the use of simpler alternative storage options, upon which resources and metadata could be stored more simply. These simpler storage locations would could also form part of an exit strategy for the system, something that all genuinely sustainable solutions should include.

7 Conclusions

If the resources and metadata managed by Digital Preservation systems are going to be useful sources of Big Data in future research, then the architectures of those systems will need to experience the kind of step-change we have outlined in this paper. Even more fundamentally, if Digital Preservation systems are going to sustain themselves by successfully re-stating their value at all times, then they must support the ability to undertake core Business Analytical tasks upon the large volumes of content they manage. Thus the move to the type of scalable architecture we propose is an inevitability, not an option.

Existing preservation systems such as the three reviewed here are on various stages of the path towards such scalable architectures, but often the developmental emphasis to date has been upon getting high-volumes of content *into* systems at scale, as opposed to working with it at scale once there, with the assumption being that 'once it is in the system, the risks to its existence diminish'. If, however, a preservation system cannot be considered sustainable unless it is usefulness can be restated, this is a false assumption: resources in preservation systems are still at risk unless those systems can also support the monitoring, management and use of those resources at the same high volumes.

The experimental work described here illustrates that coupling modern data analysis tools such as Qlik with standard, out of the box Digital Preservation tools such as DROID and JHOVE can make conducting Business Analysis that contributes to Digital Preservation business cases relatively easy. However, vended Digital Preservation systems should include such tooling out-of-the-box, enabling such Business Analytics to become a standard part of libraries' strategy and planning. Similarly, the OAIS model (undergoing its 2017 review at time of writing) tends to imply the delivery of data to 'Consumers'. This term stands in contrast to the analysis methods described here, where algorithms are pushed into the 'cloud' where the data is hosted to work on it locally. The architecture proposed here thus implies a change in tone in the OAIS model which de-emphasises consumption and instead brings use of resources closer to the core of the preservation system.

The fundamental issue facing Digital Preservation in the age of Big Data is that the bigger and more complex the datasets being preserved, the more metadata one needs. The more one brings in context, the greater the complexity of management and understanding. So it is vital to define, at organisational, research domain, and potentially entire sector levels, where the boundaries around models of Digital Preservation and access lie. Detailed, accurate and understandable descriptions of context related to the creation and use of data are vital here, as such descriptions prevent having to "work on all the data all the time", and enable focus upon the subsets of data of interest in at any given point. Such "queries based on a deep contextual understanding" will be derived from the emergent insights about the data we preserve that future researchers gain from *their* Big Data analyses. It is impossible to predict what these will be, but modelling them, recording them accurately, and re-using them in future research, will be crucial.

8 Acknowledgements

We would like to thank the Polonsky Foundation for funding our research, and the other Polonsky Digital Preservation Fellows, Edith Halvarsson, Somaya Langley, Sarah Mason and Lee Pretlove, for their support.

9 References

4C Project (2015), “Investing in Curation: a shared path to sustainability”, available at: http://www.4cproject.eu/roadmap/ (accessed 28 June 2017).

Alföldi, I., Fülöp, Z., Szatucsek, Z., Borbinha, J., Sípos, A., Billenness, C. and De Lisboa, U. (2014), “European Archival Records and Knowledge Preservation general pilot model and use case definition”, available at: http://www.eark-project.com/resources/project-deliverables/5-d21-e-ark-general-pilot-model-and-use-case-definition/file (accessed 21 October 2017).

Back, W. D., Goodman, N. and Hyde, J. (2013), *Mondrian in Action: Open Source Business Analytics*, Manning Publications Co, Shelter Island, NY.

Black, E. L. (2009), “Web Analytics: a picture of the academic library web site user”, *Journal of Web Librarianship*, Vol. 3 No. 1, pp. 3–14.

Brown, A. (2013), *Practical Digital Preservation: a how-to guide for organisations of any size*, Facet Publishing, London.

Buchner, J. (2017), “ImageHash 3.4”, available at: https://pypi.python.org/pypi/ImageHash (accessed 28 June 2017).

Chen, H., Chiang, R. H. L. and Storey, V. C. (2012), “Business Intelligence and Analytics: from Big Data to big impact”, *Management of Information Systems Quarterly*, Vol. 36 No. 4, pp. 1165–1188.

Collins, E. and Stone, G. (2014), “Understanding Patterns of Library Use Among Undergraduate Students from Different Disciplines”, *Evidence Based Library and Information Practice*, Vol. 9 No. 3, pp. 51–67. Available at: http://eprints.hud.ac.uk/21040/.

Consultative Committee for Space Data Systems (2012), “Open Archival Information System (OAIS) - reference model”, available at: https://www.iso.org/standard/57284.html (accessed 17 May 2017).

Corrado, E. M. and Sandy, H. M. (2017), *Digital Preservation for Libraries, Archives and Museums* 2nd edn, Rowman and Littlefield, Lanham, MD.

Cox, B. L. and Jantti, M. (2012), “Capturing Business Intelligence Required for Targeted Marketing, Demonstrating Value, and Driving Process Improvement”, *Library and Information Science Research*, Vol. 34 No. 4, pp. 308–316.

Digital Preservation Coalition (2015), *Digital Preservation Handbook*. Digital Preservation Coalition, Glasgow, UK. Available at: http://handbook.dpconline.org (accessed 28 June 2017).

Edwards, A., Housley, W., Williams, M., Sloan, L. and Williams, M. (2013), “Digital Social Research, Social Media and the Sociological Imagination: surrogacy, augmentation and re-orientation”, *International Journal of Social Research Methodology*, Vol. 16 No 3, pp. 245–260.

Ex Libris Ltd (2017), “Rosetta System Administration Guide”, available at: https://knowledge.exlibrisgroup.com/%40api/deki/files/57230/Rosetta_System_Administration_Guide.pdf (accessed: 13 October 2017).

Fagan, J. C. (2014), “The Suitability of Web Analytics Key Performance Indicators in the Academic Library Environment”, *The Journal of Academic Librarianship*, Vol. 40 No. 1, pp. 25–34.

Giaretta, D. (2011), *Advanced Digital Preservation* 1st edn, Springer, Berlin.

Kaplan, F. (2015), "A Map for Big Data Research in Digital Humanities", *Frontiers in Digital Humanities*, Vol. 2 No. 1, pp. 1-7.

Karpf, D. (2012), "Social Science Research Methods in Internet Time", *Information, Communication and Society*, Vol. 15 No. 5, pp. 639–661.

Kitchin, R. (2014), "Big Data, New Epistemologies and Paradigm Shifts", *Big Data and Society*, Vol. 1 No. 1, pp. 1–2.

Kutner, O. (2015), "Adding a BIRT Report to Rosetta", available at: https://developers.exlibrisgroup.com/blog/Adding-a-BIRT-report-to-Rosetta (accessed: 17 October 2017).

Laney, D. (2001), "3D Data Management: controlling data Volume, Velocity and Variety", available at: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 17 May 2017).

Lewis, J. and Fowler, M. (2014), "Microservices", available at: https://martinfowler.com/articles/microservices.html (accessed: 17 October 2017).

Manovich, L. (2012), "How to Compare One Million Images?", in Berry, D. M. (Ed.) *Understanding Digital Humanities*, Palgrave Macmillan UK, London, pp. 249–278.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung-Byers, A. (2011), "Big Data: the next frontier for innovation, competition and productivity", available at: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed 17 May 2017)

Normandeau, K. (2013), "Beyond Volume, Variety and Velocity is the issue of Big Data Veracity", available at: http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/ (accessed: 17 May 2017).

Norvig, P. (2012), "On Chomsky and the Two Cultures of Statistical Learning", available at: http://norvig.com/chomsky.html (accessed: 17 May 2017).

PREMIS Editorial Committee (2015), "PREMIS Data Dictionary for Preservation Metadata: version 3" available at: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf (accessed 17 May 2017).

Schmidt, R., Schlarb, S. and Rörden, J. (2017), "E-Ark Data Mining Showcase", available at: http://www.eark-project.com/resources/project-deliverables/90-d63 (accessed 13 October 2017).

Showers, B. and Stone, G. (2014), "Safety in Numbers: developing a shared analytics service for academic libraries", *Performance Measurement and Metrics*, Vol. 15 No. 2, pp. 13–22.

Smith-Rumsey, A. (2010) "Sustainable Economics for a Digital Planet: ensuring long-term access to digital information", available at: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf (accessed 13 October 2017).

Terras, M., Baker, J., Hetherington, J., Beavan, D., Zaltz Austwick, M., Welsh, A., O'Neill, H., Finley, W., Duke-Williams, O. and Farquhar, A. (2017), "Enabling complex analysis of large-scale digital collections: humanities research, high-performance computing, and transforming access to British Library digital collections", *Digital Scholarship in the Humanities*, available at: https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqx020/3789810 (accessed 13 October 2017).

Thelwall, M. (2008), "Bibliometrics to Webometrics", *Journal of Information Science*, Vol. 34 No. 4, pp. 605–621.

[i] http://www.exlibrisgroup.com/category/RosettaOverview (accessed 13 October 2017)
[ii] https://www.archivematica.org/en/ (accessed 13 October 2017)
[iii] http://www.eark-project.com/ (accessed 13 October 2017)
[iv] https://preservica.com (accessed 17 October 2017)
[v] https://www.jisc.ac.uk/rd/projects/research-data-shared-service (accessed 17 October 2017)
[vi] http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/ (accessed 28 June 2017)
[vii] http://jhove.openpreservation.org/ (accessed 28 June 2017)
[viii] http://www.qlik.com/us/ (accessed 28 June 2017)

[ix] https://www.nationalarchives.gov.uk/PRONOM/Default.aspx (accessed 22 October 2017)
[x] https://www.gnu.org/software/parallel/ (accessed 22 October 2017)