Archives First: digital preservation

Further investigations into digital preservation for local authorities

Viv Cothey *

2020

* Gloucestershire County Council

Not caring about Archives because you have nothing to archive is no different from saying you don’t care about freedom of speech because you have nothing to say. Or that you don’t care about freedom of the press because you don’t like to read.

(after Snowden, 2019, p 208)

Disclaimer

The views and opinions expressed in this report do not necessarily represent those of the institutions to which the author is affiliated.

Executive summary

This report is about an investigation into digital preservation by (English) local authorities which was commissioned by the Archives First consortium of eleven local authority record offices or similar memory organisations (Archives). The investigation is partly funded by The National Archives.

Archival institutions are uniquely able to serve the public by providing current and future generations with access to authentic unique original records.

In the case of local authority Archives these records will include documents related to significant decision making processes and events that bear on individuals and their communities.

Archival practice, especially relating to *provenance* and *purposeful preservation*, is instrumental in supporting continuing public trust and essential to all of us being able to hold authority to account. The report explains how Archival practice differs from library practice where provenance and purposeful preservation are absent.

The current investigation follows an earlier Archives First project in 2016-2017 that investigated local authority digital preservation preparedness. The 2016-2017 investigation revealed that local authority line of business systems in respect of children services, did not support the statutory requirement to retain digital records over the long-term (at least 100 years).

This follow-up investigation aims to improve the preparedness of Archives in anticipation of local authority line of business systems becoming able to export digital records that need long-term retention. This investigation has therefore considered the practical impacts on Archives as the need for long-term retention of digital records increases. It focusses on records relating to child adoption and democratic services.

The report introduces the notion of *authentic preservation* which whilst second nature to Archivists may be less well understood beyond the Archive.

The investigation is based on a detailed survey of current digital preservation practice coupled with some more practical detailed examinations.

The four principal “take-aways” are,

1. the essential export of child adoption records from line of business systems is still not demonstrable,

2. the current digital preservation supplier base is insufficiently informed about the needs of local authority Archives especially in respect of closed records, that is records that are not fully available for access by the general public,

3. the authentic preservation process that provides the foundation for continued public trust entails the long-term survival of provenential information which is currently bound up with the Archive’s line of business system, and

4. there is no digital preservation “silver bullet” for local authority Archives.

The report proposes an appropriate technological framework or architecture for authentic preservation which addresses the challenge of satisfying a long-term system (100 years) requirement by using an iterative succession of short-term (five year) solutions.

A simple cost comparison for a single five year iteration from each of three suppliers is presented (see section 7.3).

There are six recommendations (see section 7.7),

1. Education and training,

   Archives First members should host a series of workshops in support of learning about authentic preservation,

2. Metadata,

   Archives First members should co-operate and collaborate to produce a draft package metadata standard that can be shared with, at least, Arkivum, Artefactual, Metadatis and Preservica,

3. Component testing,

   Archives First should collaborate to gain experience of working with a variety of preservation systems by testing archival information package (AIP) deposit and retrieval workflow components, for example Exactly, with,

   - the Archives and Record Council Wales digital project,
   - Arkivum,
   - Metadatis,
   - Norfolk Record Office, and
   - Preservica,

4. Mutual support,

   Archives First members (and other local authorities) should co-operate and collaborate to investigate and pilot a mutual support process for storing and discovering AIPs,

5. Pensions records,

   Archives First should collaborate to investigate the authentic preservation of pensions records, and

6. AIP encryption

Archives First should commission a project to develop and demonstrate creating encrypted AIPs within a local authority corporate network. This should be consistent with Arkivum and Preservica workflows and Archivematica's encrypted AIPs.

Contents

List of tables

List of figures

1. Introduction

This report is about digital preservation by (English) local authorities. As such the intended audience spans Archivists, technologists and decision makers both local and national. The report makes no novel claims, indeed most of the evidence and argument has been available for at least a decade. However what may be novel is the attempt to provide a hybrid analysis that bridges a divide between Archivists and technologists. Apologies are due to both groups for any explanations that appear to them as being overly obvious.

The label "archives" is now used in such a confusing variety of ways that some clarification is needed. The myth that an Archive is just a library of old stuff must be disabused.

A recent self-referential description of the Archive's purpose as being "archiving in the public interest" (Department for Digital, Culture, Media and Sport, 2017) is the latest in a modern evolutionary sequence that begins in 1838 with the creation of the Public Record Office, County Record Offices mostly by the 1950s and The National Archives in 2003.

The information retained and organised in Archives protects people and has legal force. It is not an exaggeration to say that users trust the integrity of information managed by archivists and rely upon it "to hold government and organisations to account" (The National Archives, 2016). In a similar vein, Procter (2018) says,

> "[Archivists] are often unaware of … the way in which the characteristics of archives – an ability to provide information and evidence and *sustain rights* – have provided, and continue to provide, the rationale for their maintenance over time." [emphasis added]
>
> (p xv)

That the Archive protects individuals and democratic society is demonstrated most vividly in situations of failure. Archivists and the archival record are routinely at the centre of post-disaster inquiries (for example Hillsborough Stadium and Grenfell Tower). An archival failure is at the root of the Windrush scandal although some victims have been able to exert their rights by using evidence secured by County Record Offices, now often known as County Archives.

As early as 1956 Schellenberg found it necessary to distinguish between Archives and libraries. He did this by carefully contrasting purposes, methods and techniques (Schellenberg, 1956, pp 17-25).

Although superficially similar, for example both Archives and libraries provide information to consumers, libraries, especially public libraries, serve their users by aligning their book stock with users' changing borrowing preferences. Stock not issued (that is not loaned) will be weeded in order to make shelf space for new stock. The principal focus of librarians is their readers (Ranganathan, 1931 and many others subsequently). Book stock in libraries comprises published material which is non-unique. (Lasting copies are retained within the deposit library system.)

Archives do not lend and expect that only a small (estimated less than 1% each year) of their collections will be requested by users. Some material within the collections will be "closed", that is not available to a general consumer, possibly for several decades.[1]

Collections are aligned to the institutional bureaucracies from which material is received and archival appraisal processes regulate the acquisition of new material. The appraisal process considers the potential information needs of future generations which includes the completeness of collection series. This is because the absence of a particular record can be as informative as its existence. Since items in the Archives' collections are unique and their value depends on potential future use not current readership, the collection is not weeded but is maintained in secure environmentally controlled strong rooms. Archivists must focus on maintaining the provenance of their collections.

Archival provenance is discussed more fully in section 5 of this report where it is contrasted with the bibliographic paradigm presumed by library catalogues.

Archives First is a consortium of eleven local authority archives and similar "memory" organisations based in the south of England (see appendix one).

In 2016 Archives First commissioned a previous project (the "digital preservation project") to investigate how digital working has affected the way that information is created and how archivists can contribute to the long-term management of retained digital material (Cothey and Pickavance, 2017). The executive summary from the report of the previous project is given in appendix two.

There are unique purposes, methods and techniques found in Archives which are so well absorbed into professional practice that they are unspoken. During the course of this investigation the project team found it necessary to explicitly describe the Archival setting in order to explain the digital preservation requirement. An understanding of this Archival setting is helpful in understanding this report.

The Archivist is the *custodian* of *authentic* "records" that are *preserved* for the benefit of current and future generations.

The emphasised terms as well as "records" each warrant some explanation especially since their meaning is interdependent.

A "record" is here the smallest unit of coherent information identified by the Archive. For example a folder containing multiple leaves of paper created separately but for a common reason. Individual leaves within the folder only retain their proper meaning when in context. A record is not a database-record, nor is it likely to be a single digital word-processing or image file.

Records may contain personal or commercially confidential information. A common misunderstanding of the Data Protection Act 2018 is that local authorities (and other organisations) must delete all personal information when it is no longer being used for the purpose for which it was collected. The Act contains a provision, "archiving

[1] see Local Government Act 1972

in the public interest” for Archives to carry on collecting personal information as previously.

Records containing personal or commercial information may be “closed” that is access is restricted. Closure i.e. restriction periods vary, typically from 10 years (for example electoral rolls[1]) up to 100 years.

An authentic record here means that the record is faithfully that received by the Archive. Archival authenticity makes no claims regarding the validity of the record merely that the record is as it was when it crossed the Archives conceptual “threshold”. Of course this implies that the record remains complete in the sense that, for example, pages must not go missing.

Authenticity goes hand in hand with Archival provenance. Provenance supports information discovery but it is more than a simple catalogue or index. The provenance of the record situates it within its generating bureaucratic organisation.

As custodian the Archivist is responsible for ensuring the continued authenticity of records (which entails also ensuring the survival of provenance). As well as supporting the current generation of users in holding local government and organisations to account the Archivist is supporting future generations of users.

A common question when discussing record preservation is “How long do you want/ need to keep them?”. Preserving records here means survival in perpetuity, that is for an arbitrarily long time period. Before the introduction of modern Archival practices the survival of historical records had been accidental. Archivists now adopt a more purposeful approach to preservation where survival is no longer based on good fortune but is carefully planned, executed, monitored and tested. Records are prepared by identifying and responding to potentially harmful factors in order to establish appropriately stable material suitable for long-term storage.

Physical records are maintained in secure fire-resistant environmentally controlled vaults. Transaction logs monitor access to records while environmental monitoring evidences operational practice.

As custodian the Archivist is responsible for ensuring the preservation of records that is the Archivist must take all practical steps to ensure their survival *together with the survival of their provenance*.

This Archival setting provides a context from which to address the investigation.

The earlier digital preservation project revealed that digital working by local authorities entails information being encoded within short-term lines of business systems. Such information is especially vulnerable to loss and corruption whenever the business system is periodically renewed.

“Long-term” here means any period longer than the anticipated operational life expectancy of a business system. This life expectancy will be dictated by factors such as hardware/software obsolescence and system re-procurement policy.

[1] See Representation of the People (England and Wales) (Amendment) Regulations 2006

Without an information export facility adoption records that must be retained for at least 100 years (Department for Education, 2013) will have to survive possibly a dozen such system migrations.

The previous project revealed the lack of an information export facility from these lines of business systems that precludes employing orthodox digital preservation techniques. This is because there is no digital record to preserve! (Information can only be retained by preserving the *system*.)

A response to these findings from the earlier Archives First digital preservation project was to recommend that there be a "follow-up" project. We now report the investigations comprising the Archives First follow-up project (see appendix three).

The project team was based at Gloucestershire Archives which is the archive service of Gloucestershire County Council (see appendix four).

Background documents are available from `<url:https://www.gloucestershire.gov.uk/archives/archivesfirst/>`.

The project was funded in part by The National Archives and by the Archives First consortium.

1.1 Goals and research questions

The project brief generated two goals which were agreed at a kick-off meeting during May 2018 (see appendix five). Each goal was operationalised as a research question.

Goal one:

> Identify currently available options for local authority and similar "memory" based institutions to specify appropriate solutions to meet their so-called digital preservation needs.

Research question:

> What digital preservation options are available? In what ways are they similar and in what ways are they different?

Goal two:

> Identify currently available options for exporting Archival Information Packages (AIPs) from systems used by local authorities.

Research question:

> How can AIPs be exported for long-term storage?

An additional goal requested at the kick-off meeting was to provide a supplier cost comparison. It was agreed that an attempt would be made to create a cost model which would allow for a comparison of indicative costs.

1.2 Project deliverables

The deliverables specified at the kick-off meeting were

- an interim report to record the visits and surveys undertaken by the investigators
- a draft final report to document the investigation’s analysis and conclusions
- workshops for consortium members to present the outcome of the investigation, and
- a final report.

The remaining sections of the report address,

2. relevant previous work in the field of digital preservation. This is presented chronologically.

3. the investigative methodology employed by this investigation,

4. investigatory work with lines of business systems to export information (goal two),

5. a detailed qualitative exposition of what digital preservation *is* in the context of local authority records. This section is informed by working with three system suppliers. It introduces the notion of “authentic” preservation (which entails purposeful preservation) and concludes by describing an appropriate process architecture,

6. relevant technological, including supplier, contributions that support the authentic preservation architecture. This is presented alphabetically.

7. the investigation’s conclusions and six recommendations, and lastly

8. acknowledgements of funding support and other contributions to the investigation.

There are ten appendices.

2. Summary of previous work

The previous section, Introduction, identified several unique and defining characteristics of a local authority Archive. The Archivist is the custodian of authentic records and their provenance which is held in trust for future generations. These are the records that allow the people to hold authority to account.

Long-term digital preservation is not a technological problem and it does not have a technological solution. There is no digital preservation “silver bullet”. Rather long-term digital preservation is a management problem that can only be solved by appropriate business processes.

Management has at its disposal a variety of techniques and processes developed across a range of domains especially the forensic sciences, cryptography and computer science that can be usefully appropriated.

The previous work identified below are significant contributions, either direct or indirect, to the authentic preservation architecture which is described later in section five. They are here presented chronologically.

2.1 Cryptographic hash functions and fixity

See appendix six for a discussion of cryptographic hash functions and their associated message digests.

It has long been recognised that a message digest provides a reliable test that a message had not been accidentally or maliciously corrupted. Hence publishing “file signatures”, that is the digest, became standard practice in the early years of the Web in order to validate files (the MD5 message digest specification was published in 1992).

The digital preservation sector picked up upon this application of cryptographic hash functions to develop a rigorous notion of “fixity”. The fixity characteristic of an information package, for example the Secure Hash Algorithm 1 (SHA1) digest of the information package file, can be used to test that the bit-stream representing the information package has not changed.

Information package fixity is the diagnostic characteristic for a package being authentic, that is the package under consideration is the same as the reference package having the same fixity.

2.2 METS and PREMIS

The Metadata Encoding and Transmission Standard (METS) was developed in 2001. It is maintained by the Library of Congress (Library of Congress, 2019a).

METS is an Extensible Markup Language (XML) schema for recording and sharing metadata that describes digital objects that are to be retained over the long-term. XML is used since it is intended that the metadata be processed automatically.

A data dictionary Preservation Metadata: Implementation Strategies (PREMIS) was published in 2005 (Library of Congress, 2019b). PREMIS addresses the need for digital repositories to record their management of the digital objects for which they are responsible.

PREMIS is also expressed using XML. Figure 2.2 shows, for example, how fixity information is documented.

```
      <fixity>
        <messageDigestAlgorithm>sha256</messageDigestAlgorithm>
        <messageDigest>
a3c74f8fcd1f855e4be1566e9ed8488b71ee1ccfc81c814975bac1c7cdf50874
        </messageDigest>
      </fixity>
```

Figure 2.2: PREMIS xml fragment

PREMIS is also maintained by the Library of Congress. It is now standard practice to combine METS and PREMIS within the digital preservation business process (Library of Congress, 2017).

2.3 Excuse me… some digital preservation fallacies?

This reflective article by Rusbridge (2006) attempts to challenge the then prevalent consensus within the digital preservation discourse which supported six claims which he thought fallacious. Four of these claims are noted here.

1. file formats become obsolete very rapidly
2. interventions must occur frequently
3. digital preservation repositories should have very long timescale aspirations, and
4. the preserved object must be easily and instantly accessible in the format du jour

Unfortunately the digital preservation discourse remains focussed on file format obsolescence (Cerf, 2015) which is a misleading distraction.

Software or digital obsolescence is much trumpeted. In one respect the warnings are correct. But in a crucial respect the warnings are fear mongering.

What is correct is that (software) *applications* become obsolete. So, for example even if one could obtain a copy of Microsoft Word for Windows it would not be possible to run it since Microsoft Windows 3.1 and the hardware needed to run it are no longer available.

But this does *not* mean that Microsoft Word for Windows document files cannot be accessed.

It is true that a small number of specialized proprietary file formats have become obsolete and are no longer supported by their creators. However the vast majority of file formats have remained renderable (or readable). Most file formats are not at risk of suddenly becoming no longer renderable.[1]

Some file formats are enduring because,

(a) there are multiple independent implementations of rendering software,
(b) there is non-restrictive publication of all necessary intellectual property to create software to read the format, and
(c) there is a critical mass of users with an enduring interest in maintaining rendering software.

File formats at risk have the inverse characteristics. That is, the format is obscure, or there is only one supplier of proprietary rendering software, or the user base is niche. Such formats can be recognized in good time for a complementary preservation format version to be created.

Given that claim 1 - “file formats become obsolete very rapidly” is a fallacy then claim 2 falls. Rusbridge pointed out that most digital preservation repositories were funded by short time finance and that a succession of different repository agreements would be be the norm. This report concurs with Rusbridge’s conclusion that digital repositories should not have long-term aspirations but makes a more functional argument.

In respect of claim 4, Rusbridge argues that this could be an unreasonable and unachievable objective. However his drawing attention to the consumer receiving information in the format du jour is a key insight.

2.4 Saving the wine not the bottle

In the early 2000’s UNESCO developed a campaign to promote digital preservation especially within the heritage sector. The need for a paradigm shift was identified in order to deflect attention away from a then prevalent approach of conserving removable optical discs (CDs) and magnetic discs.

Abid (2007) presents a new preservation paradigm, saying,

> “Since time immemorial, the methods and practices of documentary heritage conservation have given the highest priority to preservation of

[1] “...the [file format] problem may not be as severe as the digital preservation community perceived it to be some 10 years ago.” (Digital Preservation Coalition, 2019a).

> carriers: paper and ink, the various generations of computer disks, magnetic tapes or emulsions for film, photography or microfilm. In the digital domain, it is the wine that is to be saved not the bottle.”
>
> (p 11)

The digital preservation paradigm is now to focus on the bit-stream and to regard the carrier as being ephemeral.

2.5 Digital preservation, archival science and methodological foundations for digital libraries

Ross (2007) takes the opportunity to use archival science and in particular the essential role played by “authenticity” and “provenance” as key ideas that need to be adopted if digital preservation is to be worthwhile. He says,

> “Digital preservation is about more than keeping the bits – those streams of 1s and 0s that we use to represent information. It is about maintaining the semantic meaning of the digital object and its content, about maintaining its provenance and authenticity, about retaining its ‘interrelatedness’, and about securing information about the context of its creation and use.”
>
> (p 2)

2.6 Gaip (Gloucestershire Archives information packager)

Gaip was first demonstrated in May 2008 (Cothey, 2010). It was a command line tool that created an AIP. Gaip was part of the digital preservation programme at Gloucestershire Archives that promoted practice based learning and advocacy in order to develop the underlying skills base within the local authority.

Gaip was succeeded by GAip which had a graphical user interface. There were small adjustments to the AIP structure which provided BagIt compatibility and the inclusion of a “gaip.xml” file that contained package metadata as submitted by the archivist.

2.7 BagIt

The BagIt file package format proposal was published in September 2008 by the Library of Congress and the California Digital Library in order to formalise then common practices for transferring collections of files (Boyko et al, 2008). The BagIt “bag” format facilitates verifying that the transfer of a file hierarchy is successful.

A minimal BagIt structure is shown in figure 2.5.

```
<base_directory_name>
 |
 +-- bagit.txt
 |
 +-- manifest<algorithm>.txt
 |
 +-- data/
      |
      +--    [payload files]
```

Figure 2.5: minimal BagIt structure

The manifest documents a cryptographic hash for each payload file.

The initial BagIt proposal anticipated that the BagIt structure would be *serialised* using either “tar” (tape archive) (Wikipedia, 2019a) or “zip” (Wikipedia, 2019b). The serialisation process represents the hierarchical tree structure of figure 2.5 as a single bit-stream or file.

By 2016 BagIt was being described as being widely used in digital preservation processes. This was because it was recognised that the BagIt “bag” format works as well for temporal transfer as it does for as spatial transfer. This is especially the case when the BagIt bag includes METS and PREMIS information.

The BagIt format is integral to many digital preservation projects including Archivematica (see section 6.2.1) and E-ARK (E-ARK, 2018 and Digital Information LifeCycle Interoperability Standards Board, 2019a).

In 2018 the BagIt specification dropped serialisation. Serialisation together with any *compression* is now left to the BagIt user. Compression refers to computational techniques of encoding that seek to minimise the number of bits in a bit-stream while retaining all the information. Serialised BagIt bags should not be compressed since many popular file formats are already compressed (for example docx, jpeg, pdf). Compressing an already compressed file increases it’s size and compression exacerbates vulnerability to “bit-rot”. ISO 21320-1 which standardises the zip algorithm prohibits compression.

2.8 Preservation is not a place

Abrams et al (2009) rethink the underpinnings of the University of California’s California Digital Library (CDL). They identify that the long-term survival of digital records can only be achieved by using a sequence of technological systems each being essentially short-term or temporary. They say, when explaining that CDL should now deprecate the notion that digital curation is based on the idea of a repository that,

> "Technical systems are inherently ephemeral, their useful lifespan being constantly encroached upon by disruptive technological change. Rather than pursuing the somewhat illusory goal of long-lived systems, curation goals are better served by concentrating on long-lived content, sustained by an evolving repertoire of nimble, commodified services."
>
> (p 9)

The authors also emphasise that there is more to digital preservation than just storing the "bits". Like Ross (2007) whom they cite and is mentioned above, they stress the requirement for archival provenance and authenticity,

> "...the importance of provenance, the understanding of a [record's] source and relationship to the information content it encapsulates. One of the distinguishing characteristics of digital content over analog forms is its ease of undetectable mutability. By analogy we finally assert that [records] must not only be accessible and usable, but also authentic, that is, they are what they purport to be."
>
> (p 10)

2.9 Parsimonious preservation

Gollins (2009) renews the attack on the myth that hardware/software obsolescence is the major threat to the survival of digital information. He targets instead a lack of attention to securing the digital record.

The parsimonious position advocated by Gollins is relevant to local authorities during the present period of financial austerity. His concluding remarks are worth quoting at length,

> "I have argued that the imminent threats in digital curation for institutions new to the field are other than they might first appear; in particular while the threat of technological (software/data format) obsolescence is real in some particular cases, *a much more imminent threat is poor capture and storage of the original material in a safe and secure way* [emphasis added].
>
> I have observed that the capabilities that many existing institutional IT systems (and their support teams) provide as a part of normal business often address many of the challenges of capture, custody, and integrity facing the new digital curator.
>
> [...]
>
> In short, a series of small, simple and affordable steps can be taken by institutions to ensure the long-term survival of vital digital data, thus

> lowering the barrier to entry for institutions to the interesting and vital aspect of information management.”

This report assumes that parsimony within digital preservation business processes is most definitely a virtue. It is hoped that the report’s conclusions and recommendations are parsimonious.

2.10 Scat is Curation And Trust (SCAT)

SCAT is a graphical user interface based application that offers a workbench type approach to assist an archivist curate or digitally preserve collections of digital files (Cothey, 2010). It is a product of the digital preservation programme at Gloucestershire Archives.

A principal function of SCAT is to be a packager, that is, it creates AIPs. These AIPs are a small evolution on from the GAip AIPs in that the package metadata file, “gaip.xml” makes more use of METS, PREMIS and PRONOM (see appendix seven). PRONOM is a technical registry maintained by The National Archives that is used to uniquely identify file formats (The National Archives, 2006). Also the AIPs are labelled with universally unique identifier (UUID)s (Wikipedia, 2019c).

The use of UUIDs to label information packages has now become commonplace.

The package metadata file that is packaged within the AIP is a key feature. Package metadata includes provenential metadata which is essential for reconstructing the meaning of the preserved content.

SCAT supports the maintenance of an independent authentication database that records the fixity of stored AIPs

2.11 Open Archival Information System (OAIS)

The Consultative Committee for Space Data Systems “works to support […] the establishment of data and system standards” (Wikipedia, 2019d).

In 2012 the committee published a reference model for an open archival information system (OAIS) which has subsequently been adopted as an ISO standard, ISO 14721:2012.

The reference model, now known simply as OAIS, provides a vocabulary and conceptual framework to describe and compare long-term preservation systems (not necessarily digital).

OAIS has become embedded within the digital preservation discourse to such an extent that its usage is taken as read.

OAIS established the concept of an information package and in particular the Archival Information Package (AIP). An AIP is a conceptual container for a particular preservation instance. It must include both content information and preservation description information (such as fixity). The AIP should be self-sufficient in that it includes all material, codes, schemas etc. necessary to make sense of the content.

The AIP is the information package that is preserved long-term. The other information packages, the Submission Information Package (SIP) and Dissemination Information Package (DIP) are only retained while they are being used. The SIP is used to convey content information from the information producer to long-term storage. The DIP is used to convey content from the long-term store to the information consumer.

OAIS does not specify how to realise any long-term information system component.

3. Project methodology

Section two, Summary of previous work, described significant contributions either direct or indirect, to the authentic preservation architecture which is described later in section five.

In this section we describe how the project team undertook the investigation.

The methodology agreed at the kick-off meeting (see appendix five) was that the team would work concurrently in respect of the two principal goals,

- identify digital preservation options, and
- identify options for exporting AIPs from lines of business systems.

In both cases investigations would include detailed discussions with system providers in order to understand the essential mechanisms that support their systems' operations.

It was agreed that the digital preservation function would be decomposed as the four components,

- packaging
- storage
- discovery, and
- presentation.

It was anticipated that so-called digital preservation needs would vary between institutions and that not all of the four components above would be equally emphasised. For example, "digital library" type products are already adept at delivering discovery and presentation/publishing solutions. In instances where preservation is not paramount a digital library may prove to be sufficient to satisfy an institution's needs.

The project's investigative methodology assumed an OAIS approach to the preservation of digital information. That is, retention is "long-term" or beyond the expected life span of any technological component.

It would also be assumed that long-term digital preservation is essentially a management not a technological activity. Hence it is how the preservation tasks are managed that is the most important feature of any long-term digital preservation proposal.

In addition to working with system providers, the project team would also work with system customers and with other similar projects, especially those addressing long-term storage issues, and with other consortium members.

3.1 “100 year” use case

It was immediately discovered that the distinctive role of Archives and especially the distinctive purpose of a local authority archive (see section one) was not well understood by suppliers. It is key to the success of a local authority Archive that the authenticity (that is the fixity) of the digital record can be demonstrated. Users trust the integrity and evidential value of information managed for them by the Archive.

A possible root to this misunderstanding is the myth that an archive is just a library of “old stuff” so that a local authority’s digital preservation needs would be met by a digital library approach (Wikipedia, 2019e), for example Eprints (2019) or Fedora (2019).

The digital preservation sector has developed from the academic/science and business communities as a result of a growing need to access or re-use works existing in a digital format. During this period digital library systems developed rapidly which has possibly re-enforced the library-of-old-stuff myth. The lack of contact between local authority archivists and digital preservation developers has meant that this myth has gone unchallenged.

In order to support a common understanding of need the project team documented a long-term retention use case (see Cothey, 2018a). The use case focusses on adoption records. Although the 100 year period is at the upper end of mandated retention periods, other features are typical. These include,

- information must be accessible throughout the period,
- however general access to the information is not permitted,
- but redacted access (which varies over time and by consumer) may be permitted
- personal information may be covered by the Data Protection Act 2018.

In addition the digital objects of interest or records are “case files” made up of a collection of “documents”and not individual digital files. Also there is a particular need to preserve closed records.

The Statutory Guidance on Adoption (Department for Education, 2013) provides the use case exemplar.

The 100 year use case document was distributed widely across the local authority archives communities during June 2018 as a request for comment and became central to the team’s investigations.

During the project team’s investigations a more detailed solution to the use case was developed. This makes use of “double bagging” and encryption which together make available a wider range of options for long-term authentic digital preservation.

3.2 Supplier/user surveys

The project team undertook a series of meetings and discussions to survey suppliers and users. In several cases there were follow up meetings or correspondence in order to clarify points of detail. The general approach was that of a structured interview which was intended to help the project team establish consistency and to more easily facilitate comparison of findings. The survey instrument is shown in appendix eight.

Because of the dispersed nature of the “suppliers/users” telephone and video were used as well as face to face contact.

A calendar of survey “visits” together with other project activity is given in appendix nine.

The project team emphasised their intended transparent and open investigative approach and that the investigation was not a prelude to recommending a procurement. Comments in the report have been previewed by suppliers/users.

The project team is grateful to all those who participated.

The rest of this section is presented alphabetically.

3.2.1 Arkivum

The project team met with Matthew Addis, Simon Bostock, and Paula Keogh.

An initial meeting discussed the Archives First project and Arkivum’s digital preservation offerings. This discussion prompted documenting the “100 use case” in order to clarify the local authority preservation requirement.

The project team also attended South West Heritage Trust’s “Transforming archive systems” joint Arkivum/Metadatis presentation at Somerset Heritage Centre where an ingest to discovery workflow for digital items was demonstrated.

A follow up meeting discussed an example AIP provided by the project team, in particular how metadata associated with the AIP can be exploited by the Arkivum discovery system. Arkivum successfully used the example provided to demonstrate ingest and discovery.

3.2.2 Artefactual

The project team held a video conference with Erin O’Meara and Justin Simpson.

The discussion was wide ranging but emphasised Artefactual’s organisational culture as being Archivist led. This was evident by their knowledge of “closed” records and references to dark archives. The open-source approach was seen as offering a good solution to the long-term needs exemplified by the 100 year use case.

A project team member had follow up telephone conversation during October 2019 with Justin Simpson to discuss recent developments in particular encrypted AIPs.[1]

3.2.3 Dorset History Centre

A project team member met with Cassandra Pickavance to discuss Dorset History Centre’s experience of using a Preservica product

3.2.4 Llyfrgell Genedlaethol Cymru (National Library of Wales)

The project team held a telephone conference with Sally McInnes, Jenny A’Brook, Liam Tomkins and Oliver Tickner (Conwy Borough Council).

The discussion compared and contrasted the Archives and Records Council Wales digital preservation project (McInnes, 2018) with the Archives First project. Points included the preference for open source, storage repositories, encryption, funding model, workflow and the use of Exactly.

3.2.5 Metadatis

The project team met twice with Charles Care and Rachel Care.

The initial meeting was to discuss the “100 year use case” and fixity monitoring of AIPs.

The project team also attended South West Heritage Trust’s “Transforming archive systems” joint Arkivum/Metadatis presentation at Somerset Heritage Centre where an ingest to discovery workflow for digital items was demonstrated.

[1] <url:https://www.archivematica.org/en/docs/archivematica-1.10/user-manual/archival-storage/archival-storage/#aip-encryption>

This made use of provenential metadata that had been exported from a CALM (Collections management for archives libraries and museums) discovery database and imported by Epexio.

The resulting system functions as a digital library with an archival type catalogue for discovery.

A follow up meeting discussed an example AIP provided by the project team, in particular how metadata associated with the AIP can be exploited by the Epexio discovery system and if Metadatis can support AIP storage.

3.2.6 Norfolk Record Office

Project team members held a telephone conference with Gary Tuson and Ian Palfrey to discuss Norfolk Record Office's experience in digital preservation generally and implementing Artefactual's open source products within a local authority context in particular. Norfolk are migrating their CALM installation to AtoM. The topic of enduring software formats was discussed.

3.2.7 Preservica

The project team met with Gareth Aitken, Peter Anderton, Tracy Broadhurst and Jon Tilbury.

An initial meeting was to discuss the "100 year use case" and fixity monitoring of AIPs. A follow up meeting discussed an example AIP provided by the project team, in particular how metadata associated with the AIP can be exploited by the Preservica discovery system.

A Preservica product is already in use by several Archives First consortium members.

3.2.8 Wellcome Trust

A member of the project team met with Alexandria Eveleigh, Toni Hardy, Victoria Sloyan and Jonathan Tweed, to discuss the Wellcome Trust's experience of both Preservica and Archivematica and digital preservation matters more generally.

Subsequently the Wellcome Trust has become a contributor to the Archivematica software.

3.3 Other

In addition to the supplier/user surveys mentioned above in section 3.2, the investigation is informed by other meetings and discussions noted below.

3.3.1 Archangel: trusted archives of digital public documents

Members of the project team tested aspects of Archangel’s blockchain application (Collomosse et al, 2018).

3.3.2 Archives and Records Association conference 2019, Leeds

Viv Cothey presented *Never mind the technology, what about the exit plan?* (Cothey, 2019b and 2019c).

3.3.3 Children Services (adoption records)

A member of the project team introduced the project to the Caldicott Guardian, Head of Service (Adoption) and the information asset owner.

The project team met with John Deane and Andy Dowden in order to be briefed on Gloucestershire County Council’s use of Liquidlogic’s LCS.

3.3.4 Democratic Services (Council minutes)

The project team met with Stephen Bace in order to be briefed on Gloucestershire County Council’s use of Civica’s modern.gov, and to agree a test information export example.

3.3.5 Digital Preservation Coalition

Viv Cothey attended a briefing on repository migration in York (Digital Preservation Coalition, 2018).

An issue arising is the bulk of information requiring long-term retention that continues to reside in short-term line of business systems rather than being “digitally preserved”.

Heather Forbes attended a briefing in Birmingham on modelling the financial aspects of digital preservation especially the value of preserved data (Digital Preservation Coalition, 2019b).

3.3.6 E-ARK archival information package review

Viv Cothey submitted comments on behalf of the Archives First project team in respect of the review consultation (Digital Information LifeCycle Interoperability Standards Board, 2019b and Brendenberg et al., 2019).

3.3.7 Memory – Identity - Rights in Records - Access (MIRRA) (Hoyle, 2018)

Members of the project team met with Victoria Hoyle in order to share information and experience regarding the long-term retention of child adoption records.

3.3.8 National Digital Stewardship Alliance

Viv Cothey submitted comments regarding the "Levels reboot project" (National Digital Stewardship Alliance, 2019) on behalf of the Archives First project team.

3.3.9 Sopra Steria

Sopra Steria provide IT services to Gloucestershire County Council. Members of the Archives First project met with Chris Murray and Richard Clarke in order to agree how fixity monitoring of stored AIPs might be undertaken.

A proof of concept application was produced and tested by Roz Farr on behalf of the project team. This demonstrated generating fixity information (in a corporate Microsoft networked environment) for AIPs deposited in the corporate storage system.

The briefing note *Long term storage for digital preservation: the role of "fixity"* which was prepared in respect of this exercise is included as appendix ten.

3.3.10 TNA "digital learning set"

A member of the project team attended each of the four workshops for archivists actively involved in digital preservation. The 100 year use case was presented at one of the workshops.

3.3.11 Ubuntu

Ubuntu is a GNU/Linux distribution that supports a range of computing requirements. The operating system and associated application software are open-source.

Members of the project team had previous experience of testing Archivematica. A non-corporate networked Ubuntu machine was used to test the current versions of both Archivematica and AtoM (see section 6.2).

3.3.12 Windows 10

In anticipation of the end of life of Microsoft Windows 7 members of the project team used a non-corporate networked Windows 10 machine to test the availability of several software products that can be used to support authentic preservation.

These were,

- Gpg4win
- Exactly
- Strawberry Perl
- Python, and
- BagIt.

4. Information export

The previous section presented the investigation's methodology. This included meetings with system suppliers, customers and with other similar projects. These discussions informed the development of the authentic preservation architecture described in section five.

Retaining information long-term entails a sequence of steps that can be thought of as,

i. get the bits,
ii. store the bits, and then
iii. retrieve the bits.

This memorable but simplified description begs a more rigorous exposition.

Steps ii and iii are considered more fully in the next section five.

Section four now focusses on the investigation's second goal; "how can AIPs be exported for long-term storage?" This goal is motivated by the critical finding of the earlier Archives First digital preservation project that information of interest is encoded within lines of business systems, typically on-line transaction processing systems, whose function is to support the day to day activities of local government staff.

The problem is that even when there is no longer any need for further activity in respect of a particular individual (or other topic), information remains encoded within the line of business system and can only be accessed via the system. Periodic line of business system replacement in order to overcome system obsolescence threatens the long-term survival of this information.

If such information can be decoded and exported as a coherent static digital record (which might include text, images, audio and video files) then the record can participate in a preservation process. Such a record is an information package.

An authentic preservation process can reliably retain information packages over the long-term.

The project investigated lines of business systems supporting child adoption and supporting democratic governance (that is members' meetings).

Child adoption is a line of business that is currently in transition (Department for Education, 2016). This involves the creation of Regional Adoption Agencies (RAA).

Gloucestershire County Council currently use Liquidlogic's Children's Social Care System (LCS). West Sussex Council use Servelec's Mosaic's case management system. It is understood that Gloucestershire County Council,s RAA (Adoption West) uses Social Care Network Solutions' CHARMS product.

The investigation also received material from OLM Systems regarding the Eclipse child care system.

It is the adoption record instance that gives rise to the 100-year use case discussed previously.

The second line of business of interest to the investigation team is democratic governance. Local authorities need to retain formal records of Council business, typically committee papers and minutes. This is of interest to nearly all of the Archives First consortium. There is a strong consensus for using Civica's modern.gov governance and meeting management system; Gloucestershire County Council is a modern.gov user.

Each of these line of business use cases is now considered with respect to their systems' information export capabilities.

It should be recalled that authentic preservation entails the survival of provenance (see page 3). This is discussed further in section five, Authentic preservation. However the requirement to ensure the survival of provenance is discussed here because provenential information is maintained as provenential metadata driving the Archive's discovery system. This is the Archive's line of business system. It, and provenance, is vulnerable to the same risks as any other line of business system. Given the essential role that provenance plays in authentic preservation the CALM instance of a discovery system as used by Gloucestershire Archives is considered also.

4.1 Liquidlogic LCS

Currently the survival of child adoption records as required by the *Statutory guidance on adoption* (Department for Education, 2013) relies on the survival of the LCS line of business system. The earlier Archives First project reported that such line of business systems must be able to export information for long-term retention (Cothey and Pickavance, 2017).

After discussions with ICT and referring to the Head of Service (Adoption), Information Asset Owner and the Caldicott Guardian, it was agreed that the adoption record export facility from Liquidlogic LCS would be used as a test case for demonstrating information export.

As of the date of this report it has not proved possible to demonstrate any such information export.

4.2 Civica modern.gov

Current practice is to periodically use the meeting management system to assemble and print out paper copies of a sequence of meeting minutes (for example *Pension Committee minutes 2018*). There are two versions. The public version which is also published on the Council's website and the full

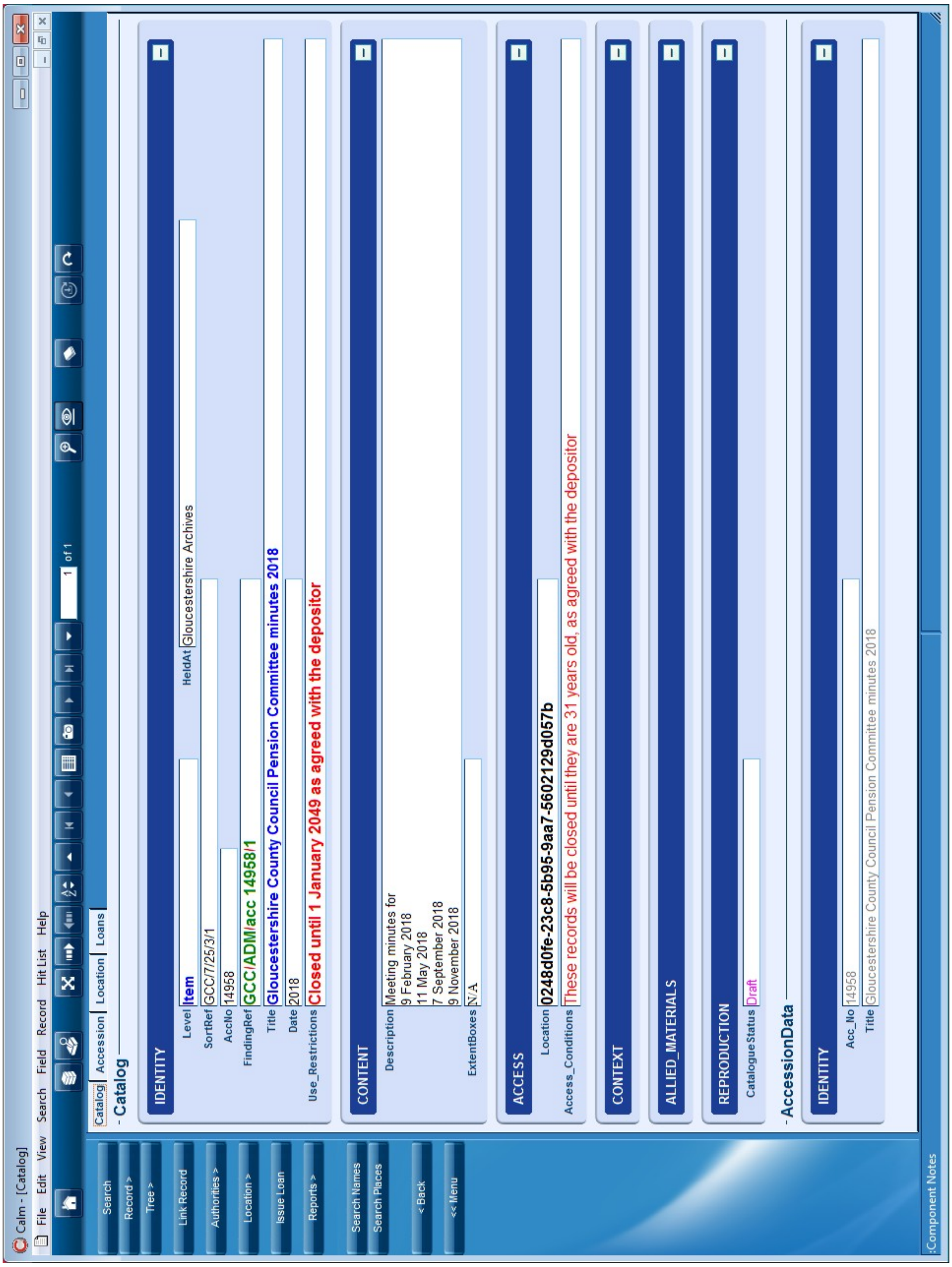
version which includes the “pink pages”. The pink pages are closed because, for example, they contain personal or commercially confidential information. The Archives receive a paper copy of the full version.
After discussions with Democratic Services (the relevant business unit) it was agreed that a digital version of the full version of *Pensions Committee minutes 2018* would be assembled. This comprised a collection of four portable document format (pdf) files.

It was a straightforward task to create an AIP within the corporate network using SCAT which included recording provenential metadata in CALM and package metadata in the SCAT “gaip.xml” file.

A CALM screen shot (GCC/ADM/acc 14958/1) is shown overleaf in figure 4.2.

The “gaip.xml” package metadata (Collins, 2019) is shown in appendix seven.

The AIP together with authenticating fixity information is stored in corporately managed storage.



- Catalog

IDENTITY

Level: Item
HeldAt: Gloucestershire Archives
SortRef: GCC/7/25/3/1
AccNo: 14958
FindingRef: GCC/ADM/acc 14958/1
Title: Gloucestershire County Council Pension Committee minutes 2018
Date: 2018
Use_Restrictions: Closed until 1 January 2049 as agreed with the depositor

CONTENT

Description: Meeting minutes for
9 February 2018
11 May 2018
7 September 2018
9 November 2018
ExtentBoxes: N/A

ACCESS

Location: 0248d0fe-23c8-5b95-9aa7-5602129d057b
Access_Conditions: These records will be closed until they are 31 years old, as agreed with the depositor

CONTEXT

ALLIED_MATERIALS

REPRODUCTION

CatalogueStatus: Draft

- AccessionData

IDENTITY

Acc_No: 14958
Title: Gloucestershire County Council Pension Committee minutes 2018

Figure 4.2: CALM screen-shot for GCC/ADM/acc 14958/1

4.3 CALM (Collections management for archives libraries and museums)

Gloucestershire Archives is a CALM user.

The CALM “catalogue” is a bespoke discovery system that inter alia is the unique record of provenance for the Archive’s collection. Until recently CALM had a near monopoly in the the Archive collections management system market. It was developed originally by DS Information Systems but was acquired by Axiell in 2008.

As with the line of business systems discussed above, CALM collections management system will become obsolete and require replacement. However unlike the previous systems where only information relating to particular “cases” must be decoded and exported in order to participate in a preservation process, it is now the whole catalogue that must be exported.[1]

It was mentioned above that it is provenential metadata that drives discovery systems. This is discussed more fully in section 5.2.1. The project team has been able to demonstrate metadata export for particular catalogue records. It is understood that a general export of metadata and its subsequent re-use is feasible as evidenced by Archive catalogue migration projects.

Section four has dealt with the long-term implications of information that is bound up in lines of business systems where the system has a short-term life expectancy.

Mostly this focussed on two particular lines of business, child adoption records and democratic governance. However the Archives’ line of business system must be considered also since this “catalogue” has a short-term life expectancy yet the provenential descriptions must survive long-term.

Section five next introduces the notion of authentic preservation and with it purposeful preservation.

[1] This is not strictly the case where more than a single “collection” is being described in the catalogue. It is all of the catalogue relating to a collection where records are being authentically preserved that must be exported.

5. Authentic preservation

We now put on one side the failure to make any progress with either Gloucestershire County Council or West Sussex Council being able to demonstrate exporting records from their respective child care line of business system (Liquidlogic LCS or Mosaic).

To continue we assume that, as with Civica modern.gov, a record is presented for preservation.

As mentioned previously, the long-term retention of information entails a sequence of steps that can be thought of as,

i. get the bits,
ii. store the bits, and then
iii. retrieve the bits.

We now develop a more rigorous exposition of tasks ii and iii; this is *authentic preservation* (Cothey, 2019).

Authentic preservation is a business process that has the goal of achieving the *known authentic survival* of retained information. That is,

a) information, including provenance, must survive,
b) surviving information must be authentic, and,
c) authenticity can be demonstrated.

This section concludes by describing a necessary and sufficient architecture for authentic preservation. This is a sequence of replicated short-term storage, discovery and authentication systems which includes encryption of information stored outwith the local authority boundary.

The architecture addresses the store-the-bits and retrieve-the-bits steps above in an Archival setting since authentic preservation requires not only that digital bits (that is the information package) survive in storage, but that when it is retrieved the information package can be authenticated as being just that which was stored, possibly a decade or more previously. Provenential information also survives so that information packages' content can be properly understood.

A preliminary version of the architecture was presented to the Archives First consortium and the Archives West Midlands group at the London Metropolitan Archives in March 2019 (Cothey, 2019).

The authentic preservation process is decomposed to three principal components,

- storage,
- discovery, and
- authentication.

These will now be considered in some detail. In addition four other essential constituents of the architecture will be discussed,

- packaging,
- purposeful preservation,
- encryption, and
- exit plans.

5.1 Storage

There is no such thing as the reliable *long-term* storage of digital bits. Digital bits will inevitably suffer from data corruption ("bit-rot").

In consequence much effort has gone into creating fault tolerant storage systems that embody internal methods of monitoring and then recovering from data corruption (such as RAID (Wikipedia, 2019f) and ZFS (Wikipedia, 2019g). Storage systems can also make use of storage replication such as tape backup.

However *short-term* storage of information in the form of static data where short-term means within the operational life-time of the storage system can be reliable although the system will be still exposed to external risks (for example fire, flood etc.).

It is important to recognise that any particular information package representation within a storage system is a construct of the storage system and its associated technologies. The information package bit-stream can only be regarded as being stable at the interface to the storage system when being either presented to, or reconstructed and retrieved from, the storage system.

5.2 Discovery

In order to function all information retrieval (IR) systems must have a discovery mechanism sometimes called a finding aid. Typically a finding aid is a system which when offered a "search" term responds with a list of all resources that correspond in some way to the search term offered.

Information technology has had a significant impact on IR with the "online public access catalog" (OPACs) for libraries introduced in the 1970s being a seminal IR application. Both professionals and end-users now have access to improved finding aids that, it is assumed, provide more sophisticated discovery systems for locating information resources.

However effective searching as an IR technique is much helped by an understanding of how the finding aid and it's entries (that is searchable terms) have been constructed. As will be seen, the comparative simplicity and standardisation of library catalogues when compared with Archival finding aids better supports resource discovery by library end-users than do finding aids used in Archives. At the root of this contrast is the fundamental difference between a library and an Archive. Schellenberg devotes an entire chapter, *Library relationships*, in Modern Archives to explaining the distinction (Schellenberg, 1956, pp 17-35).

In the context of authentic preservation, the obvious purpose of the finding aid is to discover the identifier of a needed information package. It is assumed that the identifier will be a UUID.

Unlike the majority of libraries, Archives do not provide open access, that is end-users do not themselves retrieve records. Access is mediated by professionals. As part of this mediation special considerations may apply to some records for example where a closure period is still in force.

Provenance in the Archival setting is mentioned in the Introduction and archival provenance versus the bibliographic paradigm is discussed below.

Recall that the provenance of a record situates it within its generating bureaucratic organisation. Also note the first part of the description of known authentic survival above, that is “information, *including provenance*, must survive”.

This means that it is not sufficient for just the provenential metadata relating to a particular AIP to survive. Sufficient provenential information must survive also to allow the AIP content to be validly situated. In everyday language, this means that the “archive catalogue” must survive. As will be seen this can be achieved by replicating the discovery system.

5.2.1 archival provenance versus the bibliographic paradigm

Some fundamental differences between Archives and libraries have not yet been generally recognised by the IT and IR communities. This lack of recognition has spilled over to the digital preservation community. The simpler bibliographic paradigm has been used to model the discovery of both library and Archive information resources.

For over a century libraries have relied on knowledge based classification schemes and bibliographic cataloguing to create finding aids for their resources. The principal purpose of (library) classification is to bring similar resources together for the benefit of users who can then browse similar resources (where similarity is based on what the resource is about).

There are several internationally recognised knowledge classification schemes. Classification is normally hierarchical moving from the general to the particular becoming ever more specific according to the needs of the user community.

An example from the Library of Congress classification class C is shown in figure 5.2.1a.

```
“C”           Auxiliary sciences of history
  “CD”        Diplomatics, Archives, Seals
    “CD921”   Archives
```

Figure 5.2.1a: Library of Congress classification example

For open access libraries such as public libraries the classification assigned to a resource will generally determine where the resource is physically located. Library users are thus able to personally access their needed resource.

Library catalogues rely on bibliographic metadata, for example “author”, and “title”. The catalogue describes and enumerates a library’s resources by assigning a unique identifier to each instance of the resource (multiple instances, that is “copies” are commonplace). In addition the catalogue will usually provide location information, for example shelf or stack marks. Many resources may have the same classification but even duplicate copies of the same resource would have a different unique identifier.

Bibliographic metadata records are available in standardised formats which are shared (that is bought in) by many libraries. The cataloguing in publication projects operated by national libraries maintain this standardisation.

A key feature of resource discovery in libraries is that finding, that is, discovering the location of a needed resource, is based on what knowledge the resource is about. This apparently obvious statement has important ramifications as grasped by Melvil Dewey in 1876 such as labelling the resource so that it can be shelved in its “proper” location. For example, Robert Pirsig’s *Zen and the art of motorcycle maintenance* (1974) is usually classified “917.3” (Dewey Decimal history and geography) not in the 100s philosophy, 200s religion or 600s technology.

Archives have neither bibliographic catalogues nor knowledge based classification schemes. An Archive’s information resources are organised according to the concept of “fonds” (and sub-fonds) that reflect the bureaucracy that created the resource. The Archive’s finding aid is a bespoke hierarchical scheme of provenance that identifies the bureaucratic context (and therefore provenance) of a given resource. There are no multiple instances as with libraries noted above.

Identifying the provenance of an Archival resource is a vital part of establishing its evidential worth. (For example, who wrote it, when, where and why.) It is the systematic maintenance of an Archive's provenential catalogue that maintains the integrity of the information that users (and society more broadly) rely on and trust.

Although two Archives may be similar, for example they may both receive material from a local authority, their finding aids are likely to differ since Archive finding aid entries are not standardised and cannot be shared in the same way that bibliographic records are shared.

Archival discovery is based on how the resource was created or where it comes from. An illustration of the contrast between a provenential and bibliographic approach occurs when an Archive has a book in their collection and hence includes it in their finding aid. For example, the library catalogue entry shown in figure 5.2.1b for Matthews' book about St Ives uses standard bibliographic metadata and reflects the imprint.

```
Title:          A history of the parishes of St
                Ives, Lelant, Towednack and
                Zennor in the County of
                Cornwall
Author:         John Hobson Matthews
Published:      London: Elliot Stock: 1892
Format:         Printed
Classification: 942.375
```

Figure 5.2.1b: Bibliographic paradigm example

Every library that holds a copy this book can use the same metadata values for a full catalogue entry since how to build these values is also standardised. The classification system used here is Dewey Decimal.

The National Archives at Kew also holds a copy of Matthews' book. But their finding aid provides provenential not bibliographic metadata which is shown in figure 5.2.1c.

```
Reference:      ZLIB 19/99
Description:    St Ives, Lelant, Towednack and
                Zennor by J.H.Matthews
Date:           1892
Former reference
in its original
department:     Filed/Room 2C/Bookcase 3
```

Figure 5.2.1c: Provenential metadata example

Bibliographic information is here barely recognisable since the “title” and “author” metadata elements are not used. This bespoke provenential metadata is of no use as a library finding aid entry or to another Archive.

Moreover a normal bibliographic author-title search does not discover the resource.

Digital content may present additional discovery opportunities. Digital or electronic libraries mimic the functionality of traditional libraries but have digital information resources. IR techniques for discovery can be extended to include searching content as well as catalogues. This is known as “full-text” searching. Full-text searching exploits the “bag of words” assumption that what a text is about is represented by the collection of words in the text.

In addition to end users being able to download copies of digital content, digital libraries can allow end users to deposit content, or upload. This digital content would be accompanied by an appropriate bibliographic entry to update the digital library’s catalogue. This is an example of self-publishing or self-deposit which can be a requirement in some sectors. Confusingly the digital library/repository topic area is often linked with the Open Archives Initiative (OAI) which, for example, offers metadata harvesting to create union bibliographic catalogues.

5.3 Authentication

The authentication of a record is being able to demonstrate (that is prove beyond a reasonable doubt) that the record in question is the record that was received by the Archive, possibly decades previously.

When necessary physical records in the Archive are authenticated by both their custodial history and a forensic examination of their material and appearance.

“Fixity” that is a cryptographic hash (or message digest) is a forensic tool used to characterise a digital file. Several cryptographic hashing algorithms are available each computing a characteristic message digest for a digital file (see appendix six). The important property of a message digest is that each different digital file generates a different message digest.[1] If a digital file is characterised on two different occasions and the two message digests are the same then the records have been demonstrated also to be the same.

The challenge therefore is to maintain an independent catalogue of message digests for each stored information package so that whenever the information package is retrieved it can be verified as being authentic.

[1] A hash “collision” occurs when a pair of files generate the same hash value. File manipulation procedures have been demonstrated that can create a hash collision for a particular file. This makes some applications of cryptographic hashing insecure but fixity based on more than one cryptographic algorithm remains unaffected.

In the event of an authentication failure there needs to be an effective recovery or correction mechanism.

The authentic preservation architecture assumes that long-term preservation process is based on a sequence of successful fully authenticated short-term preservation processes.

5.4 Packaging

The notion of “packaging” refers to the OAIS concept of an information package that encompasses all the material necessary to reconstruct the meaning of the package payload or information content.

The BagIt “bag” format (see section 2.6) is recognised as having enduring properties. That is, specifications are public and the software tools and techniques necessary to both read and write a serialised BagIt bag have multiple independent implementations. Users include the Library of Congress.

A defining characteristic of a BagIt bag is the manifest that records fixity information for each file included as payload (or content). Thus although the fixity of each file in an AIP payload is known, the fixity of the AIP itself (which includes adjunct files such as package metadata) is not. Automated verification procedures that confirm payload fixity require access to the payload which may not be desirable.

A “double bagging” procedure overcomes the drawback of not monitoring the AIP fixity. This is shown diagrammatically in Figure 5.4a. The outer bag manifest includes the fixity of the AIP. Also the package metadata file is duplicated in the outer bag. (Note that this metadata must be regarded as being fully publicly accessible.)

```
<base_directory_name>
 |
 +-- bagit.txt
 |
 +-- manifest<algorithm>.txt
 |
 +-- <package metadata file>
 |
 +-- data/
       |
       +-- <Archival Information Package>
```

Figure 5.4a: AIP double bagging

In figure 5.4b the double bag payload is encrypted. Otherwise the double bag in figure 5.4b is as that in figure 5.4a.

```
<base_directory_name>
 |
 +-- bagit.txt
 |
 +-- manifest<algorithm>.txt
 |
 +-- <package metadata file>
 |
 +-- data/
       |
       +-- <encrypted AIP>
```

Figure 5.4b: Encrypted AIP double bagging

The BagIt property is now that the bag manifest records the fixity of the encrypted AIP.

5.5 Purposeful preservation

Purposeful preservation here means taking all practical steps to ensure the long-term survival of retained information.

Making use of replicated independent information systems has been employed since at least 1597 when the procedure to copy parish registers and retain the information off-site (that is Bishops' Transcripts) was mandated (FamilySearch, 2019). Even then it was recognised that having off-site back up did not replace proper security measures protecting the primary record. Double, then triple lock access controls for the parish register and the use of most durable media was also mandated.

As mentioned earlier, Archival records are unique. Their long-term survival is no longer accidental where purposeful preservation has applied relevant conservation measures and provided secure environmentally controlled storage conditions. Archival strong rooms are constructed to have not only good access controls but also to have "four hour" fire resistance.

Digital records are especially fragile and vulnerable to corruption. However they benefit from being very easy to replicate. It is thus practical (and as will be seen, essential) to establish replicated records for their authentic preservation.

Purposeful preservation must extend to preserving provenance. That is the Archives' provenential "catalogues" must also benefit from purposeful preservation. This requirement was mentioned above, section 5.2.

It will be seen that purposeful preservation includes "exit planning".

5.6 Encryption

Encryption is discussed here since whenever personal information as well as other classes of confidential information are exported beyond the (local authority) corporate boundary all practical steps must be taken to assure its confidentiality. It follows that, if confidential information *can* be encrypted then it *must* be encrypted.

Modern cryptographic systems are ubiquitous. The notion of a cryptographic “key” be that a password or pass-phrase that is necessary to either encrypt or decrypt a digital message is probably understood. The un-encrypted message is referred to as the plain-text.

Less well understood is the distinction between symmetric and asymmetric encryption. Symmetric encryption uses the same secret key to both encrypt and decrypt the message. The challenge therefore is to distribute the secret key (called “key-exchange”) to the intended recipient of the encrypted message without it being intercepted by an adversary. It is assumed that the encrypted message itself will be intercepted; that is why it is encrypted.

Asymmetric encryption, sometimes called public key infrastructure (PKI) uses different keys for encryption and decryption. The key-exchange problem is solved in PKI because the cryptographic system expects the decryption key to be published that is, it is not secret. (But still only the intended recipient can decrypt the message!)

Information encryption in transit is assumed to be a default practice. The practice is embedded in so-called secure transmission protocols available across the Internet. Users of these protocols are often unaware of the implicit multiple cryptographic systems being used since they operate “silently”.

If information is transferred without using a secure internet protocol, for example by sharing physical media, then it might be necessary to implement a secret key-exchange in order to decrypt information after it has been encrypted.

While encryption in transit is presumed, many regard information encryption at rest as being incompatible with information preservation. Encryption at rest refers to information being *stored* in an encrypted state. The objection is because of the existential risk arising from the threat of loss of encryption keys. That is, the survival of information must not be predicated upon the long-term survival of an encryption key.

The objection is overcome when two conditions are satisfied,

- encryption is not long-term so that decryption does not depend on the long-term survival of an encryption key, and

- in any event, a plain-text version is retained.

It can be seen immediately from the preserved information replication requirement of purposeful preservation that, for example, maintaining an in-house (short-term) preservation system for plain-text together with an encrypted version in an independent short-term preservation system elsewhere satisfies these conditions.

The short-term survival of an encryption key is discussed later in section 5.8.1.

The authentic preservation architecture satisfies the requirement that personal and other confidential information is encrypted when exported beyond the corporate boundary.

5.7 Exit plans

Long-term authentic preservation can only be achieved as an effective sequence of short-term authentic preservation processes. This is because (by definition) long-term is beyond the life expectancy of the IT systems or components that support the preservation process.

It has been seen that enduring file formats can be expected to have long-term life expectancies (for example, National Archives of Australia, 2019).

However most systems applications, particularly line of business systems, are intended to have short-term life expectancies (measured in only a few years). There is an “end of life” date after which there is no further development or support. Procurement policies can also require that application systems be replaced periodically and thus define an end of life date.

Migrating from an expiring system to its successor can range from a version upgrade to a complete system replacement. All migrations put at risk any digital information that the system hosts.

It should be noted that whenever a system migration is undertaken the opportunity should be taken to export any no longer operational information that is to be retained long-term. This is because of the system specific encoding of information by expiring systems which is not replicated by replacement systems. This leads to an inevitable long-term loss of information.

Migrating information when replacing a system and testing that a migration has been successful is a major undertaking which requires the active assistance of system suppliers. Managing this undertaking is referred to as an “exit plan”.

Purposeful preservation requires that systems involved in the authentic preservation process have exit plans that are tested prior to the systems becoming responsible for hosting any information.

Local authorities present a broad spectrum of arrangements regarding IT support, development and infrastructure ranging being 100% “in-house” to 100% “outsourced”. Even when IT is “in-house” there is usually a complex combination of external providers that contribute to system delivery.

Two exit scenarios can be identified. The more straightforward gives rise to an “orderly exit” plan for system end of life that is based on receiving support from the expiring system supplier(s) in order to achieve a successful migration.

The second scenario gives rise to a “disorderly exit” plan which has to assume that crucial support to carry out an orderly exit is now absent. Disorderly exits occur outwith expected end of life events. They are unexpected and sudden.

The Carillion plc failure (liquidation) in 2018 (Wikipedia, 2019i) provides an example of disorderly exit. An exacerbating feature was the requirement to lock out all access to sites where Carillion had been working so that sub-contractors could not recover their tools. This was a comprehensive but lawfully necessary “denial of service attack”.

Disaster recovery (DR) plans do not usually apply in Carillion type situations since DR effectively winds the clock back to a recovery point immediately prior to a disaster event. It is assumed that all planned system support can be obtained in order that systems can be reconstructed.
Disorderly exit plans must work when DR has failed, such as MySpace losing 13 years worth of users data (Colbron, 2019) and the Kings College, London data loss (Martin and Corfield, 2017). In neither case was information preserved.

A recent paper, Adair et al. (2019) makes the case, with supporting empirical evidence, for *early* exit planning and identifies that “Risk management is not well represented in current digital preservation literature” (p 3). *Early* here refers to exit planning being undertaken as part of the procurement process.

The disorderly exit plan for authentic preservation is to ensure that there are at least two completely independent preservation systems. That is, there is no single point of failure. The PLATTER report (DigitalPreservationEurope, 2008) provides good prompts to think about this including organisational and institutional “failures” (which includes mergers and acquisitions) as well as key personnel and encryption keys.

5.8 Architecture

This final part of section five describes a long-term authentic preservation architecture that is based on a sequence of interlinked short-term authentic preservation business processes. These are independently replicated in order to support purposeful preservation. Since at least one copy of each AIP is stored beyond the local authority boundary, encryption must be used.

System migration between the successive short-term authentic preservation business processes is discussed in section 5.8.1.

Figure 5.8a illustrates how long-term authentic preservation, in this case for 100 years, is situated within its overall OAIS producer to consumer long-term retention framework.

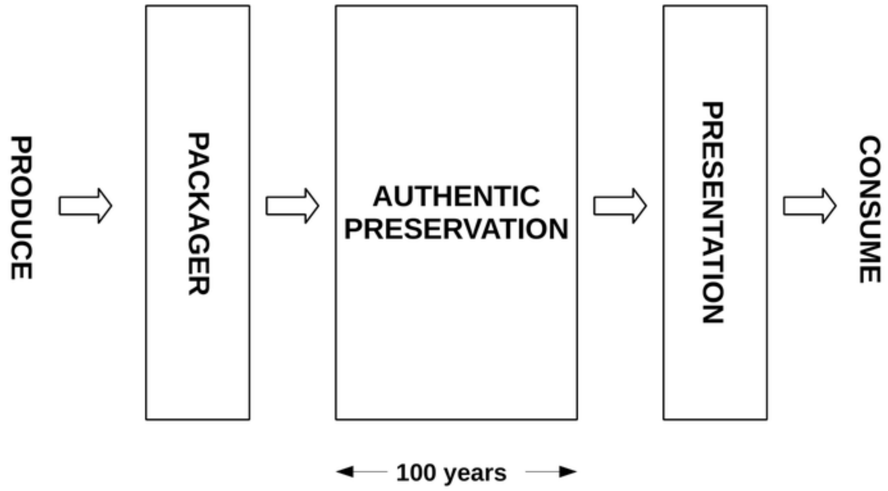


Figure 5.8a: business process architecture for long-term retention

The architectural elements illustrated are,

"produce" is an information export process for a line of business system,

"packager" is the process that creates an Archival Information Package (AIP) which contains the exported information. The packager system need be only ephemeral but must create the AIP conforming to an expected enduring format.

"authentic preservation"
provides the known authentic survival of the AIP,

"presentation" is the inverse process to "packaging". The presentation system similarly need be only ephemeral but must accept AIPs conforming to an expected enduring format. The presentation process provides whatever file format conversion might be needed to transform a retrieved AIP to a dissemination information package (DIP) using "formats du jour".

"consume" is whatever processes are required by the end-user to access the formats du jour. These might include virus checking.

Note that while "produce" and "packager" are contemporary processes that apply to *all* information being retained, the "presentation" and "consume" processes apply to only a small proportion (estimated < 1%) of AIPs at some unpredictable time (decades) in the future.

The three sub-elements of authentic preservation are illustrated in figure 5.8b.

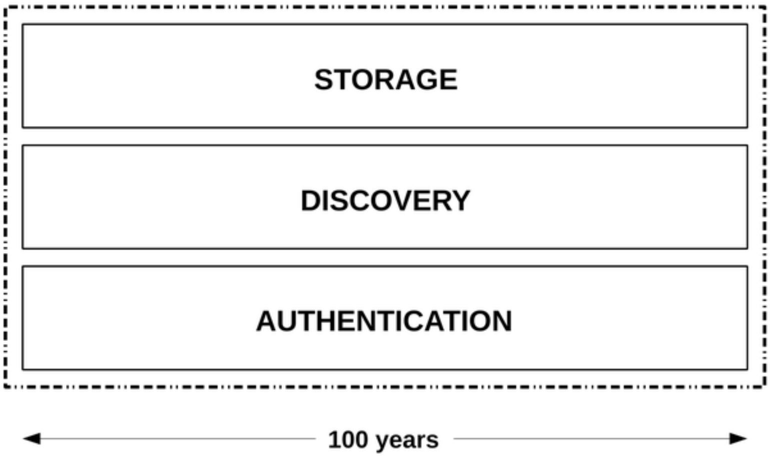


Figure 5.8b: authentic preservation

Since each of the "storage", "discovery" and "authentication" system sub-elements of authentic preservation must survive long-term, they need to be implemented as a sequence of short-term systems.

Exit plans for the sub-elements must ensure their successful periodic system migration, see section 5.8.1 below.

A short-term sequence, in this case every ten years for 30 years, of storage, discovery and authentication is illustrated in figure 5.8c overleaf. A ten year frequency for system replacement is considered to be at the upper limit of anticipated system life expectancy.

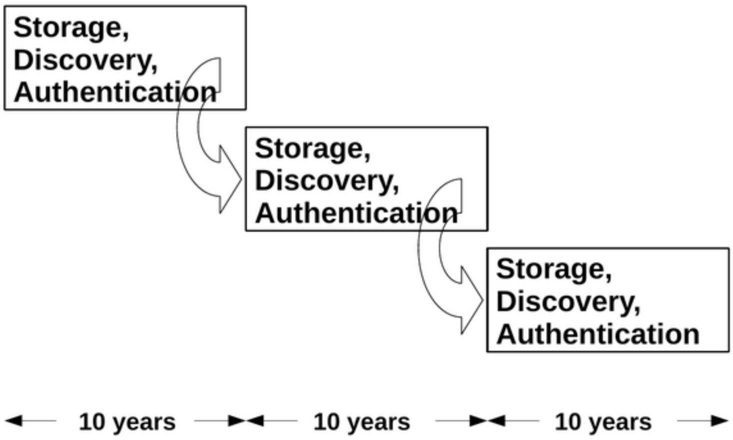


Figure 5.8c: sequence of short-term authentic preservation systems

Figure 5.8d illustrates the last architectural step for authentic preservation by illustrating multiple (at least two) independent replicated systems for purposeful preservation.

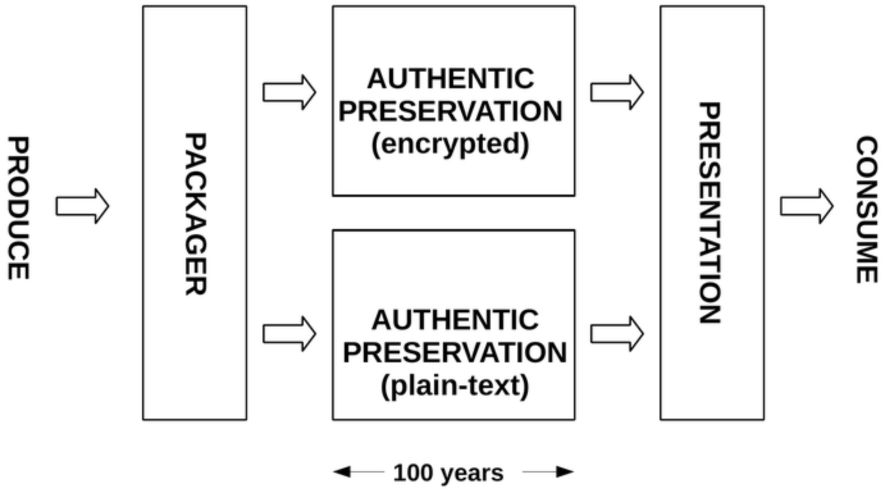

PRODUCE
PACKAGER
AUTHENTIC PRESERVATION (encrypted)
AUTHENTIC PRESERVATION (plain-text)
100 years
PRESENTATION
CONSUME

Figure 5.8d: purposeful authentic preservation

Systems storing AIPs beyond the local authority boundary host encrypted AIPs while AIPs stored within the local authority remain as plain-text.

Special attention should here be paid to the essential replication of the discovery system. This must not only facilitate discovery of stored AIPs by UUID, but also establish the provenance of the material by reference to a catalogue of the relevant collection.

5.8.1 authentic preservation system migrations

Authentic preservation is built from replicated short-term storage, discovery and authentication systems that are regularly migrated as they each reach their anticipated end of life.

The authentication system is a simple “database” of fixity information in respect of each stored AIP. Fixity information is replicated in at least two independent databases.

As each AIP is migrated to the new storage system its fixity can be verified by reference to both authentication databases. Any fixity failures can be resolved by reference to the replicated AIP. A new fixity database can be generated as a side effect. Thus the authentication database is ephemeral in comparison to the arbitrary long duration survival of AIPs. The authentication database must survive for at least a complete migration cycle.

Each migration provides the opportunity to refresh any encryption. Hence particular encryption keys need survive only over the short-term. Just like the authentication database there is only a short-term survival requirement. It is understood that sufficient manual clerical procedures exist within local authorities to ensure key security.

Migrating discovery information is more challenging since this entails the export and then re-import of provenential metadata between discovery systems that may not share the same metadata schema. (Recall that provenential metadata for the entire collection (or fonds) must survive not just individual metadata for the AIP.)

There needs to a lossless conversion so that no information is lost in translation. This is assisted by using standardised schemas and round trip testing (Wikipedia, 2019i). A backstop safeguard is to ensure that provenential metadata is retained as “package” metadata within the AIP.

Section five has addressed the issues of “storing the bits” and “retrieving the bits” within an Archival setting which is characterised by authentic preservation.

Authentic preservation achieves the known authentic survival of retained information. That is,

a) information, including provenance, must survive,
b) surviving information must be authentic, and,
c) authenticity can be demonstrated.

Section six next, considers the “players” and “products” that contribute to realising the authentic preservation architecture just described.

6. Digital preservation system contributors

Section two, summary of previous work, identified significant contributions either direct or indirect, to the authentic preservation architecture which was described in section five. It was pointed out that long-term digital preservation is a management problem not a technological problem.

Section five, authentic preservation, is informed by the investigations described in section three, project methodology. Authentic preservation is built from replicated short-term storage, discovery and authentication systems that are regularly migrated as they each reach their anticipated end of life. AIPs that are stored beyond the local authority boundary are encrypted.

In this section we identify the “players” and “products” that support the authentic preservation system architecture for local authorities and similar “memory” based institutions. These are organised alphabetically.

6.1 Arkivum <`url:https://arkivum.com/`>

Arkivum is a UK based company that has its origins in information science research undertaken at the University of Southampton. Arkivum was founded in 2011. They say,

“Arkivum provides long-term digital preservation by integrating open source preservation and access software with archival storage and replication capabilities. These are powered by Artefactual for preservation workflows, AtoM for access/discovery, MongoDB and Kafka for database and workflow options and a number of other open standards. This includes the option of either tape and UK data centre archival storage, a fully cloud hosted environment (typically in Amazon Web Services) or a combination of these options in the hybrid offering. All of these endeavour to ensure the survival of information packages created using the Perpetua Preservation Module, powered by Archivematica.”

(Arkivum, 2019)

Arkivum is distinct in providing long-term preservation solutions based on open source software. Since customers can independently acquire copies of the software it can be argued that long-term survival risks are mitigated.

Their products support an end-to-end workflow from “ingest” to discovery and access.

Arkivum services the pharmaceutical, financial, higher education and heritage sectors. In the UK customers include South West Heritage Trust.

It is understood that Arkivum can support the proposed authentic preservation architecture. In particular,

a) an encrypted AIP that has been “double bagged” can be accepted,
b) the (outer) bag can be validated,
c) metadata (submitted in an agreed comma separated value format) can be used for discovery,
d) the encrypted AIP can be retrieved by its UUID, and
e) at the conclusion of a short-term preservation period
   - AIPs can be returned in bulk, and
   - discovery metadata can be exported.

Key steps in achieving this were demonstrated to the project team.

Arkivum offer a data escrow facility. In the event of a disorderly exit a customer can fully recover all deposited data. (Arkivum reported that this facility had been tested by one of their customers.)

Arkivum operate on a subscription basis. This can be either individual or as a consortium.

Subscription cost per consortium member depends on the size of the consortium, from three members to 20 or more members and the amount of storage consumed. The cost comparison model developed by the project team is discussed in section 7.1.

6.2 Artefactual Systems <url:http://www.artefactual.com>

Artefactual was formed in 2001. Key software products are Archivematica and Access to Memory (AtoM). AtoM is a discovery system that optionally links with Archivematica.

Artefactual is based in Vancouver.

6.2.1 Archivematica

> “Archivematica is an integrated suite of open-source software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model. Users monitor and control ingest and preservation micro-services via a web-based dashboard. Archivematica uses METS, PREMIS, Dublin Core, the Library of Congress BagIt specification and other recognized standards to generate trustworthy, authentic, reliable and system-independent Archival Information Packages (AIPs) for storage in your preferred repository.”

(Archivematica, 2019)

Artefactual and in particular Archivematica developments were much assisted by UNESCO who supported “the aggregation and development of an open source archival system, building on, and drawing together existing open source programs” (Bradley et al, 2007).

The Archivematica software produces an information package (AIP) using the BagIt format.

In the UK users include Llyfrgell Genedlaethol Cymru (National Library of Wales), Norfolk Record Office, University of York and Wellcome Trust. Arkivum products are based on Archivematica.

Recent Archivematica developments include encrypting AIPs.

6.2.1 AtoM

“AtoM is a fully web-based archival description application that is based on International Council on Archives standards.”

(UNESCO, 2019)

AtoM is open-source. It provides a discovery system or “catalogue” that can be coupled with Archivematica in order to retrieve AIPs.

As a straightforward catalogue AtoM provides a digital library type function that, for example, enables end-users to find, inspect and download documents and images.

There is a large user base particularly in North America. In the UK users include Llyfrgell Genedlaethol Cymru (National Library of Wales), Norfolk Record Office and the University of York. Arkivum products optionally make use of AtoM.

6.2.3 Format Identification for Digital Objects (FIDO)

Artefactual also maintain FIDO. This is a command line software tool for accessing a PRONOM registry database in order to determine what file format is present. The central PRONOM file format registry was created and has since been supported by the UK National Archives.

FIDO/PRONOM inspects a digital bit-stream and provides a best estimate for the digital preservation community of the intrinsic format represented as opposed to any external claims regarding format. It is vital to have this file format identification in order to manage the long-term survival of information when formats may no longer endure.

6.3 Axiell/CALM

Axiell is a Swedish based supplier of proprietary collections management systems. It has grown to be the largest provider of such specialist information systems in Europe, mainly through acquisitions. In 2008 Axiell acquired DS who provided CALM (Collections management for Archives Libraries and Museums). CALM is used extensively within the local authority Archive sector.

Gloucestershire Archives uses CALM.

6.4 BagIt

Section 2.3 refers to the BagIt file package format proposal or specification.

The BagIt specification has been implemented as an open source Python library and a command line software tool known as bagit-python. This is supported by the Library of Congress.

bagit-python is a packager in that it creates a bag or package containing the user specified content.

6.4.1 Exactly `<url:https://www.weareavp.com/products/exactly/>`

Exactly is a graphical user application for creating BagIt conforming packages. Exactly is open source and is available for both Microsoft Windows and Apple. A feature of Exactly is that it simplifies the task of including user supplied (that is package) metadata. Exactly is produced and maintained by AVP a US based information management company.

The Exactly application includes an optional facility to send completed packages to another computer using standard Internet protocols.

6.5 GNU Privacy Guard (GnuPG) `<url:https://www.gnupg.org/>`

“GnuPG is a complete and free implementation of the OpenPGP standard as defined by RFC4880 (also known as PGP).”

(GnuPG project, 2019)

That is, GnuPG is a set of tools that carry out a range of cryptographic tasks. In particular this includes a command line utility “gpg”.

gpg supports symmetric encryption where the same secret key is used to both encrypt and decrypt a file with a selection of cryptographic algorithms including Advanced Encryption System (AES).

AES is an international standard and is available in many different encryption products. When appropriately implemented AES is recognized as being the most secure algorithm available.

6.5.1 Gpg4win

Gpg4win is the GnuPG product that runs on Microsoft Windows. Its creation was initially funded by the German Federal Office for Information Security.

Gpg4win includes gpg.

6.6 Metadata

Metadata is what drives discovery systems. It provides a formal specification of the object being described. As shown previously bibliographic metadata emphasises the imprint, that is publishing information such as “title”, “author”, “publisher”, “place of publication” while provenential metadata emphasises bureaucratic context.

In order to support inter-working, since the 1960s there has been considerable national and inter-national effort to develop and maintain metadata standards. These consist of dictionaries of metadata elements (that is “tags”, terms or field names), such as “title” and “creator” together with their associated definitions.

The special value of standardised metadata element dictionaries is that they provide a stable long-term definition for what a term means and how it should be used.

Digital information file formats include metadata showing, for example, creator name, creation time, software version used, and much more. Image files are particular rich in metadata. Metadata from digital files can be extracted (and possibly modified). Such metadata is sometimes referred to as *technical* metadata.

As part of their ingest processing both Archivematica and Preservica scan the files that have been presented to them and extract as much technical metadata as they can find.

The term *package* metadata is here used to refer to the user supplied provenential metadata which relates to the AIP itself rather than to individual files that are part of the packages content.

6.6.1 ExifTool `<url:https://www.sno.phy.queensu.ca/~phil/exiftool/>`

ExifTool is open-source cross platform software written by Phil Harvey to read, write and edit file metadata. The target application is image files created by digital cameras. The software has been in uses for over 15 years and is now widely used to read metadata from many other file formats.

A key feature is the way in which ExifTool uses Extensible Metadata Platform (XMP) to create and populate a document containing a copy of the technical metadata. This document can be easily edited and supplementary metadata elements, together with their associated values, added.

6.6.2 Dublin Core (DC)

DC is a limited set of 15 bibliographic metadata elements that were developed for use in Web documents and other published electronic resources. As seen previously, despite their popularity, these terms have limited value in provenential metadata.

6.6.3 General International Standard Archival Description (ISAD(G))

ISAD(G) provides a set of 26 elements of which only six are mandatory. The intent is to provide a descriptive framework rather than a rigid format and it relies on an understanding of the “level” of description. This is an implicit reference to the hierarchical nature of provenential metadata. Example levels are shown in figure 6.6.3a.

fonds
sub-fonds
series
file
item
piece

Figure 6.6.3a: Provenential metadata levels

Since it is an ISAD(G) principle that there should be no duplication of information and that the metadata should identify the level of description problems of simple discovery can arise.

This is illustrated by an example based on an item from Gloucestershire Archives and shown in figure 6.6.3b.

fonds	Abenhall, St Michael
sub-fonds	Incumbent
series	Parish registers
item	Burials (1813 - 1922)

Figure 6.6.3b: Provenential metadata example

The ISAD(G) mandatory metadata representation is shown in figure 6.6.3c. It can be seen that the title contains information derived from several provenential metadata elements.

```
Reference   =  PI/IN/1/11
Title       =  Register of burials for the
               parish of Abenhall, St Michael
Creator     =
Date(s)     =  1813 – 1932”
Extent      =  one volume
Level       =  item
```

Figure 6.6.3c: ISAD(G) metadata

This is not sufficient to support discovery. In order to facilitate a user finding the resource in response to the reasonable search phrase, “Abenhall burial registers” the hierarchical entries in the provenential metadata are adjusted as shown in figure 6.6.3d.

```
fonds       Abenhall parish records
sub-fonds   Incumbent
series      Parish registers for Abenhall,
            St Michael
item        Register of burials for the parish
            of Abenhall, 1813 – 1922
```

Figure 6.6.3d: Modified provenential metadata

This now no longer complies with the ISAD(G) no-duplication rule but does support discovery.

6.6.4 Metadata Encoding and Transmission Standard (METS)

METS was discussed previously in section 2.2.

Archivematica include a “mets” file in their AIP.

6.6.5 Preservation Metadata: Implementation Strategies (PREMIS)

PREMIS was discussed previously in section 2.2.

Archivematica include PREMIS metadata within their “mets” file

6.7 Metadatis <url:https://metadatis.com/>

Metadatis is a UK based company that specialises in discovery systems for archives making full use of provenential metadata. They say,

> "Combining expertise in software development, information management, library science, and data science, we build cutting-edge, data-centric applications."
>
> (Metadatis, 2019)

Metadatis offer Epexio, a proprietary cloud based archive oriented discovery system that provides publishing like access to digital resources. Data is stored in the Cloud (Amazon S3).

Epexio is distinct in that its use of provenential metadata overcomes the difficulties described in section 6.5.3 when ISAD(G) metadata is used bibliographically.

UK customers include South West Heritage Trust and University of Warwick Modern Records Centre.

It is understood that Metadatis can support the proposed authentic preservation architecture (although this has not yet been demonstrated). This support draws on their expertise in managing the storage aspects of Epexio.

In particular,

a) an encrypted AIP that has been "double bagged" can be accepted,
b) the (outer) bag can be validated,
c) metadata (submitted in an agreed XML format) can be used for discovery,
d) the encrypted AIP can be retrieved by its UUID, and
e) at the conclusion of a short-term preservation period
    - AIPs can be returned in bulk, and
    - discovery metadata can be exported.

It is assumed that in the event of a disorderly exit some form of Amazon bucket key sharing would facilitate continued access to data.

Metadatis operate on a subscription basis.

The cost comparison model used by the project team is discussed in section 7.3.

6.8 Preservica <url:https://preservica.com/>

Preservica is a UK based company that came to prominence following its work with The National Archives in the early 2000s. They say that,

> “Ensuring the accessibility and authenticity of digital information over successive technology cycles and custodians requires a different approach to traditional backup, archiving, storage and content management.
>
> Preservica’s standards-based (OAIS ISO 14721) active preservation software combines the critical capabilities of successful long-term digital preservation into a single integrated platform. It keeps content safely stored, makes sure it can be found and trusted, provides secure immediate access, and automatically updates files to future-friendly formats.”
>
> (Preservica, 2019)

Preservica offers *active* preservation curation using their proprietary system. Data is stored in the Cloud (Amazon S3 and Glacier or Microsoft Azure). Preservica is distinct in providing long-term preservation solutions based on file format migration; it is this that provides its active curation.

UK customers include, Dorset History Centre, West Sussex Records Office and Oxfordshire History Centre.

It is understood that Preservica can support the proposed authentic preservation architecture (although this has not yet been demonstrated).

In particular,

a) an encrypted AIP that has been “double bagged” can be accepted,
b) the (outer) bag can be validated,
c) metadata (submitted in an agreed XML format) can be used for discover
d) the encrypted AIP can be retrieved by its UUID, and
e) at the conclusion of a short-term preservation period
    - AIPs can be returned in bulk, and
    - discovery metadata can be exported.

Preservica offers a data recovery facility in the event of a disorderly exit. This is based on sharing the Amazon bucket key.

Preservica operates on a subscription basis. This can be either individual or as a consortium.

Subscription cost per consortium member depends on the size of the consortium, from three members to 20 or more members and the amount of storage consumed. The cost comparison model used by the project team is discussed in section 7.3.

6.9 SCAT (Scat is Curation and Trust)

SCAT was discussed previously in section 2.7.

SCAT is used by Gloucestershire Archives.

7. Conclusions and recommendations

The investigation's research questions and deliverables are set out in section 1. These are reported below, sections 7.1 to 7.4. Section 7.5 discusses the missing link – AIP encryption; section 7.6 considers potential options for authentic preservation by local authorities by developing section 5.8 in particular figure 5.8d: purposeful authentic preservation. Six recommendations are set out in section 7.7.

A positive general finding is that the digital preservation sector is ready, willing and able to offer products to local authorities and similar "memory" based institutions. Already, for example, University of Warwick Modern Records Centre, South West Heritage Trust and Oxfordshire History Centre (Picture Oxon) have installed solutions from Arkivum, Metadatis or Preservica.

However it must be noted that although these examples provide sophisticated electronic library functionality, it is presumed that they do not provide digital preservation as here described. That is, there is no purposeful preservation, authentication or disorderly exit plan. The presumption is based on suppliers not citing these installations when authentic digital preservation was discussed. In addition it is presumed that these installations do not include closed records (which institutions would manage in some other way).

Section 5.8 describes an authentic preservation architecture that can deliver digital preservation as predicated by the 100 year adoption record use case. (Other less demanding use cases can be subsumed within the 100 year adoption record use case instance.)

Archival preservation of born-digital records by local authority Archives, that is authentic preservation, ignores nearly all of the marketed benefits of the existing services and products suppliers (Arkivum, Metadatis and Preservica) because records are closed and must have been encrypted. This is a niche customer requirement for the digital preservation services and products sector.[1] Indeed the suppliers questioned whether they could provide any benefits at all when AIPs were encrypted!

The key benefit offered by all of the suppliers surveyed (Arkivum, Artefactual, Metadatis and Preservica) is a well developed understanding of the essential organisational culture and technical approach necessary for *long-term* digital retention. This understanding is reflected in their practice.

Elements of a "pre-packaging" fix (see also figure 5.8a) which would allow local authorities to use existing services and products to authentically preserve closed and encrypted AIPs have been tested by the project team. This is part of a local authority digital preservation "missing link" described in section 7.5 that overcomes the closed and encrypted difficulty that has been identified. Preliminary discussions (and experiments) with suppliers suggest that a fix along the lines suggested is straightforward and easily accomplished.

[1] While this report was being prepared Justin Simpson (Artefactual) identified the Simon Fraser University use case <url:https://groups.gourl:ogle.com/forum/#!msg/archivematica/Kvl6E6xLocw/BLCZKGJmCAAJ>.

A material challenge is the local authority and more precisely local authorities' management of its IT support. This is exacerbated by a widespread lack of attention by local authorities generally to their responsibilities in the area of digital preservation. This is prolonged and widespread. Anecdotal evidence suggests that the project team's direct experience is by no means unique.

Focussing on the 100 year adoption record use case, the local authority has to do just two things. It has to ensure that adoption records can be/are exported from a line of business system (in the investigation's instance, Liquidlogic's LCS or Mosaic). And then it has to ensure that essential digital preservation tools are available.

Following the earlier Archives First report (Cothey and Pickavance, 2017) the project team agreed the objective of demonstrating the export of adoption records by mid-June 2018. It was discovered that this objective could not be achieved until the Liquidlogic test environment had been upgraded since it appeared that the current version of LCS was not up to the task. Also some detail of configuration was needed from the Information Asset Owner.

At the time of writing adoption record export has still not been demonstrated either by a Liquidlogic installation or a Mosaic installation (West Sussex).

By way of contrast the project team were very quickly provided with test records from OLM System's Eclipse children's social care system.

The project team needed to ensure that tools considered necessary for digital preservation would work and did so within a corporate network environment. These tools are Exactly (section 6.4.1) and Gnu4win (section 6.5.1). The original thinking behind the Exactly request was that it could provide a default packaging function. Gnu4win (or a functionally equivalent product) is required in order to encrypt the AIPs for storage beyond the boundary of the corporate network: Gnu4win provides "gpg". The lack of gpg (or its functional equivalent) would be a proverbial "showstopper".

Exactly was requested January 2019 and Gnu4win was requested May 2019. At the time of writing neither tool is available within the corporate network environment.[1]

The investigation's answers with respect to the three questions set out in section 1.1 are given below.

7.1 What digital preservation options are available? In what ways are they similar and in what ways are they different?

There are currently no available options for local authorities to carry out authentic digital preservation. This is because there is no provision for long-term authentication of AIPs.

[1] Both have now (November 2019) been made available.

Section 7.6 describes options that would become available when the missing link fixing the issue of closed records and encryption has been resolved.

Given this and assuming that compliance testing in respect of authentic preservation is successful, then the effect of all the options described in 7.6 is broadly similar. Closer investigation of issues such as user-interfaces, training, documentation and support, system performance etc. might reveal a small difference subject to users preference but these issues have not been considered.

The only identified distinction between the options comes from how provenential metadata, discovery and discovery system replication is addressed. Our understanding is that the Epexio system is the only discovery component that eschews the bibliographic paradigm and thereby more completely supports provenance.

7.2 How can AIPs be exported for long-term storage?

7.2.1 Adoption records

Gloucestershire County Council/Liquidlogic have not been able to demonstrate to the project team the export of adoption records from the line of business system (LCS).

West Sussex Council/Mosaic have not been able to demonstrate to the project team the export of adoption records from the line of business system.

OLM have provided the project team with a test example showing the export capability of Eclipse. The test adoption record comprised a collection of .pdf files which are amenable to preservation.

7.2.2 Council minutes etc.

Gloucestershire County Council/Civica were able to provide the project team with a test example of committee meetings minutes from the modern.gov line of business system.

The test example was amenable to preservation and has been “hand” processed using packaging software installed on the corporate network (SCAT). The resulting AIP together with authenticating fixity information is stored in-house.

7.3 A supplier cost comparison model

An expansion of the initial project brief was to provide a supplier cost comparison. The project team have therefore devised a simplistic cost model which is based on the authentic preservation architecture, that is long-term authentic preservation being achieved by a sequence of short-term authentic preservation increments, typically on a five year cycle.

Cost depends on quantity. The model assumes that 1TB (terabyte) of AIP data is deposited in the short-term preservation system per year for five years. After five years the cumulated amount of 5TB is withdrawn.

Cost also depends on the subscription model. Both Arkivum and Preservica offer a consortium based arrangement where individual member subscriptions depend on the size of the consortium.

CAVEAT Suppliers have provided budgetary costs (excluding VAT) only. The model makes no attempt to adjust for the inevitable apples and pears effect of comparing products that differ in detail.

The model does not include any ancillary costs such as staff time incurred by an Archive.

The table shows the cumulated non-discounted five year cost for a single five year short-term authentic preservation increment.

	Consortium of three members	Consortium of 20+ members
Arkivum*	£39,500	£24,500
Metadatis	£16,500	£16,500
Preservica**	£35,775	£25,555

Table 7.3: Supplier cost comparison

* Arkivum has an additional set-up charge of £2,000 per member which has not been included above.

** Except by Preservica some costs may be charged for data egress at the conclusion of the short-term preservation cycle.

7.4 Project deliverables

The project's deliverables are specified in section 1.2.

7.4.1 interim report

An interim report to the Archives First consortium was delivered December 2018. This was supported by a working paper (Cothey, 2018).

7.4.2 draft conclusions

Draft conclusions were shared with both the Archives First consortium and the Archives West Midlands group at a meeting in March 2019.

The conclusions were supported by a paper describing authentic preservation, disorderly exit plans and the co-operative model for preservation business processes (Cothey, 2019).

7.4.3 workshops

The meeting mentioned above included workshop exercises to explore digital preservation issues.

The learning set meeting in Gloucester provided an opportunity to share and explore digital preservation experience.

Exit planning was discussed at the Archives and Records Association conference.

It is proposed (see section 7.7.1) to hold further workshop events as part of this investigation in order to share the practice of Archives First members as regards exempt/closed records and the encryption of records.

It is hoped that a representative of the Information Commissioner's Office will attend.

7.4.4 final report

The project's final report is this document.

7.5 The missing link (AIP encryption)

For ease of writing the missing link is here described with reference to Arkivum's preservation system but with minor modifications the description is understood to apply equally to Preservica's system.

Arkivum's user front-end provides a packaging function workflow that allows a user to assemble a collection of digital files that then comprise the content of an AIP.

The AIP is implemented as a zipped BagIt bag labelled using a UUID which is presented for both storage and discovery. Provenential metadata can be included in the bag.

On receipt of the AIP the preservation system's default action is to explore and index (in order to support full-text search) all the AIP's content. Additionally, although the fixity of the content is verified, the fixity of the AIP itself is ignored. This introduces both privacy and authentication issues.

The double-bagging with encryption procedure discussed in section 5.4 addresses both of these issues.

But in order to make use of the current Arkivum workflow the encrypted AIP (including provenential metadata) needs to already exist. When this is the case the collection of digital files assembled by the Arkivum front-end user is just the encrypted AIP. (Note that some UUID labelling and discovery metadata details have been ignored here in order to simplify the description.)

So the missing link is pre-packaging software that creates an encrypted AIP.

Demonstrations of pre-packaging components have already been developed as part of the project team's investigations. These have involved using SCAT and Exactly to carry out bagging and gpg to carry out encryption. The demonstrations have been in non corporate network Linux and Windows 10 computing environments.

The pressing need is to develop a proof of concept demonstration of creating an encrypted AIP that can be implemented within a local authority corporate network.

It is important that the work is shared with Arkivum, Artefactual, Metadatis and Preservica.

It is understood that the result of an annual manual clerical procedure to verify the secure continuing existence of the short-term encryption key could be included in the Annual Governance Statement.[1]

7.6 Potential authentic preservation options

All of the options presume that the AIP encryption fix is in place which itself presumes that a local authority has made available pre-requisite software as discussed previously.

[1] Accounts and Audit Regulations 2015

And, of course, none of the options will exist unless lines of business systems such as Liquidlogic LCS actually export preservable records!
As described in section 5.8, architecturally, each long-term option comprises a sequence of short-term systems there being periodic managed migrations that address the issues of authentication and encryption. The periods should overlap so that not both systems that are paired are being migrated at the same time (for example at the end of the same calendar year). Recall that the fundamental requirement of authentic preservation is the known authentic long-term survival of records. The survival of provenance must be similarly assured.

All of the options are predicated on a local authority associating itself with at least one other partner in order to achieve the essential independent replication of each short-term system to mitigate the disorderly exit risk.

Partners can be considered to be one of three types. Firstly there is a commercial arrangement, for example with Arkivum, Metadatis or Preservica. Then there is a mutual co-operative arrangement with another local authority. And lastly there is a shared interest arrangement with a self-supporting institution such as within Higher Education or the national library system.

The type of partner chosen is likely to determine the amount of training/support and how “smooth” a work-flow, including user interface, can be created. Commercial arrangements will be at one end of a spectrum with mutual co-operative arrangements at the other.

Arkivum and Preservica provide a more complete solution than Metadatis but to date, Metadatis have not been requested to provide a complete solution.

Mutual co-operative and shared interest arrangements are as yet unexplored by local government in England. However examples of such collaborations include Archives and Records Council Wales’ digital preservation project (Arcifau Cymru, 2019) and Florida’s Dark Archive In The Sunshine State (Caplan, 2010).

The underlying procedure that provides a potential authentic preservation option is,

step one,

package the record as an AIP including package metadata,

label the AIP with a UUID and include in the usual finding aid,

store the AIP in the usual corporate secure storage, and

maintain a fixity database with message digests from at least two cryptographic hash algorithms.

This procedure was completed in the case of the *Pensions Committee minutes 2018*, see section 4.2 above.

Step two, the AIP must now be encrypted, for example the command,

```
gpg --passphrase “pass phrase” --symmetric <AIP_file_name>
```

creates the file `<AIP_file_name.gpg>` which is a symmetric encryption of the AIP using AES.

Step three, double bag,

package the encrypted AIP including a copy of the AIP package metadata,

label the AIP with the original UUID and “.gpg” extension,

And step four,

deposit encrypted AIP with at least one partner of choice,

update associated discovery systems, authentication databases and replicated provenential metadata systems.

If no secure corporate storage (step one) is available then step four must be replicated with at least two partners.

7.7 Recommendations

There are six recommendations,

Any significant progress depends on developments under the control of the local authority. These recommendations assume that work is also being carried out to ensure that,

- lines of business systems such as Liquidlogic LCS can be used to export records,
- gpg (or a functionally equivalent) application is available within the GCC corporate network, and
- Exactly (or a functionally equivalent) application is available within the GCC corporate network.

7.7.1 Education and training

The purpose and practice of local authority Archives has to include an approach to digital preservation, here called *authentic preservation*, that runs counter to established norms within a digital preservation world that focusses on widespread (bibliographic) access to heritage.

There is therefore a need to undertake an education and training program for Archivists that helps them to more fully understand,

- the role of provenential metadata, its migration, replication and its inclusion as package metadata,
- the need for explicit systems of record authentication,
- how a sequence of short-term systems can create a long-term system,
- the role and use of encryption, and
- how to create and test a disorderly exit plan.

Recommendation 1

Archives First should host a series of workshop events in support of learning about authentic preservation.

7.7.2 Metadata

Following on from recommendation 1, an important development for local authority Archives would be a standardised approach to package metadata that could be used both by Archives and systems providers to assist package (AIP) deposit and discovery. Any standardisation would also have to support system migration. Appendix seven illustrates package metadata used by Gloucestershire Archives. This makes use of existing standards, METS, DC and XMP.

Recommendation 2

Archives First members should co-operate and collaborate to produce a draft package metadata standard that can be shared with, at least, Arkivum, Artefactual, Metadatis and Preservica.

7.7.3 Component testing

Recommendation 3

Archives First should collaborate to gain experience of working with a variety of preservation systems by testing AIP deposit and retrieval workflow components, for example Exactly, with,

- the Archives and Record Council Wales digital project,
- Arkivum,
- Metadatis,
- Norfolk Record Office, and
- Preservica.

7.7.4 Mutual support

Multiple replication including off-site copies is a fundamental strategy whenever digital survival is required. *Independent* replication is essential for disorderly exit planning.

Given that a local authority's IT provider already supplies an off-site replication function then a remaining vulnerability is a disorderly failure of that provider (or within the provider's supply chain) or the Archive.

Independent replication can be achieved by a local authority sharing their existing infrastructure (storage and "catalogue") with another local authority.

Recommendation 4

Archives First members (and other local authorities) should co-operate and collaborate to investigate and pilot a mutual support process for storing and discovering archival information packages.

7.7.5 Pensions records

The long-term retention of pensions records at Gloucestershire County Council has emerged as another authentic preservation use case.

Given that the number of cases is likely to far exceed that of adoptions and the issues of information export from Liquidlogic LCS are no nearer resolution, then the pensions records use case is a more promising example for study.

Recommendation 5

Archives First should collaborate to investigate the authentic preservation of pensions records.

7.7.6 Missing link (AIP encryption)

Recommendation 6

Archives First should commission a project to develop and demonstrate creating encrypted (symmetric, AES) AIPs within a local authority corporate network. This should be consistent with Arkivum and Preservica workflows and Archivematica's encrypted AIPs.

8. Acknowledgements

The project was partly funded by The National Archives, UK and by the Archives First consortium.

The author wishes to thank the other members of the project team for their critical contributions and especially Claire Collins for her insights and enthusiasm.

Thanks also go to the suppliers without whose contributions this investigation would not have been possible and to many reviewers who provided clarifications and corrections. All errors that remain are the author’s responsibility.

References

Abid A. (2007). Safeguarding our digital heritage: a new preservation paradigm. In de Lusenet Y and Wintermans V (editors) *Preserving the digital heritage: principles and policies* The Hague: Netherlands National Commission for UNESCO.

Abrams S., Cruse P. ans Kunze J. (2009). Preservation is not a place. *International Journal of Digital Curation* 1(4), pp 8-21.

Adair A., Esteva M. and Chang B. (2019). Early exit strategies in digital preservation. In *iPres 2019: 16th International Conference on Digital Preservation, Amsterdam, 16-20 September 2019*. *, Amsterdam*. Amsterdam.

Archivematica. (2019). *Archivematica*. [online] available from <url:https://www.archivematica.org/en/>.

Arcifau Cymru. (2019). *Digital preservation*. [online] available from <url:https://archives.wales/archives-and-records-council-wales/arcw-projects/digital-preservation/>.

Arkivum. (2019). *Personal communication*. Arkivum.

Boyko A., Kunze J., Madden L. and Littman J. (2008). *The BagIt file package format (v0.93)*. [online] available from <url:https://tools.ietf.org/id/draft-kunze-bagit-00.txt>.

Bradley K., Lei J. and Blackall C. (2007). *Towards an open source repository and preservation system: recommendations on the implementation of an open source digital archival preservation system an on related software development*. Paris: UNESCO.

Bredenberg K., Faria L., Ferreira M., Nielsen A. B., Rörden J., Schlarb S. and Wilson C. (2019). *E-ARK archival information package (AIP)*. [online] available from <url:https://earkaip.dilcis.eu/pdf/aip-specification.pdf>.

British Broadcasting Corporation. (2015). *Google's Vint Cerf warns of 'digital Dark Age'*. [online] available from <url:https://www.bbc.co.uk/news/science-environment-31450389>.

Cabinet Office. (2017). *Better information for better government*. [online] available from <url:https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/589946/2017-01-18_-_Better_Information_for_Better_Government.pdf>.

Caplan P. (2010). *DAITSS, an OAIS-based preservation repository*. [online] available from <url:http://daitss.fcla.edu/sites/daitss.fcla.edu/files/DAITSS%20in%20ACM%20rev_0.pdf>.

Cerf V. (2015). The future of the Internet: meaning and names or numbers. In *American Association for the Advancement of Science annual meeting, San Jose CA, 13 February 2015*. See also British Broadcasting Corporation (2015).

Colbron K. (2019). *What we can learn from the MySpace data loss?* [online] available from <url:https://www.jisc.ac.uk/blog/what-can-we-learn-from-the-myspace-data-loss-26-mar-2019>.

Collins C. (2019). *gaip.xml: package metadata.* [online] available from <https://www.gloucestershire.gov.uk/media/2093889/gaip-xml-package-metadata.zip>.

Collomosse, J., Bui, T., Brown, A., Sheridan, J., Green, A., Bell, M., Fawcett, J., Higgins, J. and Thereaux, O. (2018). Archangel: trusted archives of digital public documents. In *Proceedings of the ACM Symposium on Document Engineering, Halifax NS, August 28-31, 2018,* pp. 31:1-31:4, ACM, New York: Association for Computing Machinery.

Cothey V. (2010). Digital curation at Gloucestershire Archives: from ingest to production by way of trusted storage. *Journal of the Society of Archivists*, 31(2), pp 207-228.

Cothey V. (2018a). *Digital preservation for local authorities: the 100 year use case*. [online] available from <https://www.gloucestershire.gov.uk/media/2086051/100-year-use-case-20180625.pdf>.

Cothey V. (2018b). *A use case approach to archival digital preservation: an analysis*. [online] available from <https://www.gloucestershire.gov.uk/media/2087712/use-case-approach-to-digital-preservation-4.pdf>.

Cothey V. (2019a). *Retaining digital information over the long-term*. [online] available from <url:https://www.gloucestershire.gov.uk/media/2087704/retaining-digital-information-over-the-long-term.pdf>.

Cothey V. (2019b). *Never mind the technology: abstract*. [online] available from <url:https://www.gloucestershire.gov.uk/media/2093885/never-mind-the-technology-abstract.pdf>.

Cothey V. (2019c). *Never mind the technology: presentation slides*. [online] available from <url:https://www.gloucestershire.gov.uk/media/2093886/never-mind-the-technology-presentation-slides.pdf>.

Cothey V. and Pickavance C. (2017). *Archives First: digital preservation project*. [online] available from <url:https://www.gloucestershire.gov.uk/media/18083/201709-archivesfirst-digital-preservation-final-report.pdf>.

Department for Digital, Culture, Media and Sport. (2017). *Archives: public interest: written question 111381*. [online] available from <url:https://www.parliament.uk/business/publications/written-questions-answers-statements/written-question/Commons/2017-11-03/111381/>.

Department for Education. (2013). *Statutory guidance on adoption: for local authorities, voluntary adoption agencies and adoption support agencies*. [online] available from <url:https://www.gov.uk/government/publications/adoption-statutory-guidance-2013>.

Department for Education. (2016). *Education and Adoption Act 2016: Section 15*. [online] available from <url:https://www.legislation.gov.uk/ukpga/2016/6/section/15>.

Digital Information LifeCycle Interoperability Standards Board. (2019a). *Archival Information Package (AIP)*. [online] available from <url:https://dilcis.eu/specifications/aip>.

Digital Information LifeCycle Interoperability Standards Board. (2019b). *Specifications*. [online] available from <url:https://www.dilcis.eu/specifications/9-specifications/26-2018-specification-review>.

DigitalPreservationEurope. (2008). *DPE repository planning checklist and guidance*. [online] available from <url:https://digital.library.unt.edu/ark:/67531/metadc799759/m1/3/>.

Digital Preservation Coalition. (2018). *So long, and thanks for all the bits: migrating your data between repositories*. [online] available from <url:https://dpconline.org/events/repo-migration-briefing-2018>.

Digital Preservation Coalition. (2019a). *Digital preservation handbook: file formats and standards*. [online] available from <url:https://dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards>.

Digital Preservation Coalition. (2019b). *Counting on reproducibility: tangible efforts and intangible assets*. [online] available from <url:https://dpconline.org/events/past-events/counting-on-reproducibility>.

E-ARK. (2018). *About E-ARK*. [online] available from <uri:https://www.eark-project.com/about>.

EPrints. (2019). *EPrints repository software: welcome to demoprints*. [online] available from <url:http://demoprints.eprints.org>.

FamilySearch. (2019). *England Bishop’s Transcripts – FamilySearch historical records*. [online] available from <url:https://www.familysearch.org/wiki/en/England_Bishop’s_Transcripts_-_FamilySearch_historical_Records>.

Fedora. (2019). *Fedora: about Fedora*. [online] available from <url:https://duraspace.org/fedora/about/>.

Gollins T. (2009). Parsimonious preservation: preventing pointless processes! In *Online information 2009: proceedings London 1-3 December 2009*. [online] available from

<url:https://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf>.

GnuPG project. (2019). *GnuPG*. [online] available from <url:https://gnupg.org>.

Hoyle V. (2018). *Introducing the MIRRA project*. [online] available from <url:https://blogs.ucl.ac.uk/mirra/2018/06/12/introducing-the-mirra-project/>.

Library of Congress (2017). *Guidelines for using PREMIS with METS for exchange*. [online] available from <url:https://www.loc.gov/standards/premis/guidelines2017-premismets.pdf>.

Library of Congress (2019a). *Metadata encoding and transmission standard*. [online] available from <url:https://www.loc.gov/standards/mets/>.

Library of Congress (2019b). *Preservation metadata: implementation strategies*. [online] available from <url:https://www.loc.gov/standards/premis/>.

McInnes S. (2018). *Preserving the Welsh record: a bit at a time*. [online] available from <url:https://www.dpconline.org/blog/a-bit-at-a-time-arcw>.

Martin A. J. and Corfield G. (2017). *KCL external review blames whole IT team for mega-outage, leaves managers unshamed*. [online] available from <url:https:/www.theregister.co.uk/2017/02/23/kcl_external_review/>.

Metadatis. (2019). *Metadatis*. [online] available from <url:https://metadatis.com>.

National Archives of Australia. (2019). *Long term file formats*. [online] available from <http://www.naa.gov.au/information-management/managing-information-and-records/preserving/long-term-file-formats.aspx>.

National Digital Stewardship Alliance. (2019). *Levels of digital preservation*. [online] available from <https://ndsa.org/activities/levels-of-digital-preservation>.

Pirsig R. M. (1974). *Zen and the art of motorcycle maintenance: an inquiry into values*. London: Bodley Head.

Preservica. (2019). *We are Preservica: leaders in active digital preservation*. [online] available from <url:https://www.preservica.com/about>.

Procter M. (2018). Introduction to the English edition. In Delsalle P. *A history of archival practice* [Translated and revised by Margaret Procter] London: Routledge.

Ranganathan S. R. (1931). *Five laws of library science*. London: Goldston.

Rivest R. (1992). *The MD5 message-digest algorithm*. [online] available from <url:https//:www.ietf.org/rfc/rfc1321.txt>.

Ross S. (2007). ‚Digital preservation, archival science and methodological foundations for digital libraries. *Keynote Address at the 11th European Conference on Digital Libraries (ECDL), Budapest, 17 September 2007*.

Rusbridge C. (2006). Excuse me… some digital preservation fallacies. *Ariadne 46*. [online] available from <url:http://www.ariadne.ac.uk/issue/46/rusbridge>.

Schellenberg T. R. (1956). *Modern archives: principles and techniques*. [Midway reprint] Chicago: University of Chicago Press.

Snowden E. (2019). *Permanent record*. London: Macmillan.

The National Archives. (2006). *PRONOM*. [online] available from <url:https://www.nationalarchives.gov.uk/PRONOM/>.

The National Archives. (2016). *Consultation on a new strategic vision for the archives sector*.

UNESCO. (2019). *UNESCO Archives AtoM Catalogue*. [online] available from <url:https://atom.archives.unesco.org>.

Wikipedia, (2019a). *tar (computing)*. [online] available from <url:https://en.wikipedia.org/wiki/Tar_(computing)>.

Wikipedia, (2019b). *Zip (file format)*. [online] available from <url:https://en.wikipedia.org/wiki/Zip_(file_format)>.

Wikipedia, (2019c). *Universally unique identifier*. [online] available from <url:https://en.wikipedia.org/wiki/Universally_unique_identifier>.

Wikipedia, (2019d). *Consultative Committee for Space Data Systems*. [online] available from <url:https://en.wikipedia.org/wiki/Consultative_Committee_for_Space_Data_Systems>.

Wikipedia, (2019e). *Digital library*. [online] available from <url:https://en.wikipedia.org/wiki/Digital_library>.

Wikipedia, (2019f). *RAID*. [online] available from <url:https://en.wikipedia.org/wiki/RAID>.

Wikipedia, (2019g). *ZFS*. [online] available from <url:https://en.wikipedia.org/wiki/ZFS>.

Wikipedia, (2019h). *Carillion*. [online] available from <url:https://en.wikipedia.org/wiki/Carillion>.

Wikipedia, (2019i). *Round trip format conversion*. [online] available from <url:https://en.wikipedia.org/wiki/Round_trip_format_conversion>.

Appendix one

Members of the Archives First consortium

Berkshire Record Office, West Berkshire Council.

Dorset History Centre, Dorset County Council.

East Sussex Record Office as lead partner in The Keep (East Sussex County Council, Brighton & Hove Council and the University of Sussex).

Gloucestershire Archives, Gloucestershire County Council.

Hampshire Record Office, Hampshire County Council.

Isle of Wight Record Office, Isle of Wight Council.

Portsmouth History Centre, Portsmouth City Council.

Southampton Archives, Southampton City Council.

Surrey History Centre, Surrey County Council.

West Sussex Record Office, West Sussex County Council.

Wiltshire and Swindon History Centre, Wiltshire Council.

Appendix two

Executive summary from

Archives First: Digital preservation project, 2017

Archives First is a consortium of eleven local authority records keeping services across the south of England.

During late 2016 and early 2017 Archives First undertook a project to determine what added value archivists could provide in the context of so-called “digital working” since digital working “...completely change[s] the concept of information and records, as well as what constitutes effective information management” (Cabinet Office, 2017, p 6).

In an era of fake news, users trust the integrity of information managed by archivists and rely upon it “to hold government and organisations to account” (The National Archives, 2016).

The project aims to understand how this trust can be maintained into the future and in particular to identify how archivists can contribute to the long-term management of preserved digital material.

Following a survey of the eleven local authorities into how digital working has affected the way that information is now created, the project concludes that:

- an urgent paradigm shift is needed that focusses local authority archivists' attention on the *long-term* preservation of information in digital format rather than on their traditional role relating to the *permanent* retention of information,

- Archives First should influence the debate within the archives and associated information technology communities regarding the long-term management of digital material.

- the current generation of computer systems developed to provide for digital working has ignored the need for the long-term archival preservation of information. It is vital to recognise that most information is now assembled temporarily from disparate items of structured data and does not exist as a (digital) document entity.

- Archives First should emphasise the intellectual added value of the catalogue which is much more than a mere list of contents. It is the organisational and descriptive power of the catalogue that underpins the achievement of archival provenance and integrity.

There are four recommendations:

1. archivists should adopt a leadership role in respect of issues connected with the long-term preservation of digital information. In particular they should engage with both information asset owners and technologists to become involved in system procurement. They should also identify all information which is to be retained for ten years or longer.

2. individual archivists should exploit opportunities to become familiar with digital preservation issues, terminology and practice by, for example, supporting small scale digital preservation projects and providing training opportunities. Innovative practice should be shared within Archives First and more broadly.

3. Archives First should carry out a follow up project to investigate archival information package export specification and functionality in respect of:
    a) Liquidlogic Children's Social Care System (i.e. adoption cases), and
    b) modern.gov (i.e. committee minutes).

    (Archival information packages provide the basis for long-term information preservation.)

    It is anticipated that the investigation here will include liaising with archivists in Scotland and will support further work aimed at specifying mandatory functional requirements to be included in future local authority system procurement exercises.

    It is also anticipated that the outcome of such an investigation will be shared with other relevant consortia.

4. Archives First should carry out an investigation to determine the minimum requirements of a long-term storage system for archival information packages and identify available options for local authorities. It is anticipated that the outcome of such an investigation will be shared with other relevant consortia.

Appendix three

Project funding proposal (to The National Archives Sustainability Fund)

1. Project background

We would like to run a second digital preservation project as a follow-up project from the first Archives First digital preservation project, 2016-2017.

The full report of the first project is published online and the executive summary of findings is included as appendix two of this document.

The report included four recommendations for following up the initial work.

1. archivists should adopt a leadership role in respect of issues connected with the long-term preservation of digital information. In particular they should engage with both information asset owners and technologists to become involved in system procurement. They should also identify all information which is to be retained for ten years or longer.

2. individual archivists should exploit opportunities to become familiar with digital preservation issues, terminology and practice by, for example, supporting small scale digital preservation projects and providing training opportunities. Innovative practice should be shared within Archives First and more broadly.

3. Archives First should carry out a follow up project to investigate archival information package export specification and functionality in respect of,

   a) Liquidlogic Children's Social Care System (i.e. adoption cases), and

   b) modern.gov (i.e. committee minutes).

   (Archival information packages provide the basis for long-term information preservation.)

   It is anticipated that the investigation here will include liaising with archivists in Scotland and will support further work aimed at specifying mandatory functional requirements to be included in future local authority system procurement exercises. It is also anticipated that the outcome of such an investigation will be shared with other relevant consortia

4. Archives First should carry out an investigation to determine the minimum requirements of a long-term storage system for archival information packages and identify available options for local authorities. It is anticipated that the outcome of such an investigation will be shared with other relevant consortia.

It also became clear during the collaborative work in our first project that in order to meet recommendations 1 and 2, it would help to do more work on recommendation

4 to include those record offices where it has not yet been possible to procure a digital preservation solution.

We are keen to work alongside TNA to help deliver the strategic vision for the archives sector, *Archives Unlocked* with associated business plan *Archives Inspire.* Also to make a modest contribution to section 9 (Digital research) of TNA's Digital Strategy, 2017-2019.

We plan to work in collaboration with Archives West Midlands who are doing a more in-depth investigation into Preservica and Archivematica from the end users' point of view.

We believe that our two projects are entirely complementary. We plan to keep in touch throughout our respective projects (if funded) and run a free joint workshop in Birmingham at the end to share results with staff from all participating offices.

2. Project overview

We are very keen to build on the success of our first collaborative project, to keep the momentum going within Archives First and to move forward with the digital preservation agenda – our most challenging priority.

All services involved are very willing to share data, analytical information and to support the next steps of the research proposed.

2.1 Focus on recommendation 4

To investigate minimum requirements of a long-term storage system for archival information packages and identify available options for local authorities. The aim is to produce a report with sufficient technical detail to facilitate local government and similar archives users to produce quality specifications if they do get the opportunity to procure a solution.

This would involve looking at what the market currently supplies in relation to key elements of digital preservation and how this matches up to the various needs of local government archives services.

- Packaging (e.g. Archivematica),
- Storage (e.g. Archivematica),
- Discovery (e.g. AtoM, CALM, Metadatis), and
- Presentation (e.g. Preservica).

2.2 Move forward on recommendations 1 and 3

To work with Liquidlogic (children's records) and modern.gov (for local government minutes) to examine archival information package export options. This will involve working with information asset owners. This work will in turn help identify key requirements when asset owners commission new systems holding local government records (i.e. the business applications for live data, not the e-preservation storage system for semi-current or archived data). We have extended this to include Mosaic (another major local government framework contractor supplying systems for children's records) used in West Sussex and Berkshire. This will help us examine how similar Liquidlogic and Mosaic are or if lessons learnt from studying one are transferable to others.

3. Rationale

- Within Archives First getting to grips with digital preservation remains our highest priority.

- We see collaboration as the most effective means of tackling this area of work.

- Being able to understand the pros and cons of the technical solutions on offer is key to drawing up a quality specification and engaging in active e-preservation.

- The General Data Protection Regulations specify privacy by design. We need to engage with council information asset owners to ensure that long term archival/historical research requirements are also built in.

- A consortium approach supported by The National Archives is likely to be carry more weight when approaching system suppliers.

- Appointing an ICT business analyst and project staff who will remain available beyond the end of the project so the knowledge gained is not lost.

4. Key objectives

- Gathering quality evidence to make an informed case for moving forward digital preservation, either individually within our own authorities, or collectively.

- Drawing up a shopping list of essential and desirable requirements of system providers (both for e-preservation systems and for local government systems containing our core records).

- To contribute to TNA's commissioned investigation of requirements to set up an Active Learning set.

- Sharing what we've learnt so far. The process itself will be a valuable learning outcome in itself, both within the Archives First grouping and more widely across the local authority sector.

- Raising the profile of long term digital preservation amongst key local government system suppliers at a time when they are focused on the related issues of making their software GDPR compliant.

- Reinvigorate and strengthen the Archives First partnership.

5. Out of scope/limitations

At present the proposal is limited to those archives services signed up to Archives First (to meet the requirements of this fund). If feasible we would like to extend to other local government services in the SE and SW. We have already been working with many of these services in sharing the results of our initial project and inviting them to an Archives First Preservica workshop.

6. How will the project work in practice?

Project to be led by Gloucestershire Archives but provisional results to be tested out in other local authorities and results shared in collaborative workshops.

To re-employ Viv Cothey (mentor for initial Archives First project, now based in Gloucestershire and Cornwall) as a mentor to provide strategic vision and specialist technical knowledge.

To recruit a business analyst to undertake supplier visits, obtain relevant information and write up results.

7. Costings

Expenditure

- Business analyst £#### (based on business analyst with relevant background experience in local government ICT systems @ £#### per day and travel expenses.

- Project Mentor: £#### temporary register payments for Viv Cothey's time spent supporting project @ c. £#### per hour (including on-costs).

- Project Manager: £#### backfill payment for Heather Forbes (County Archivist) and Claire Collins (E-preservation archivist) – mostly in-kind support but a small contribution to backfill will be necessary due to existing commitments.

- In-kind contribution of at least 10 days' officer time from Gloucestershire Archives, four days' officer time from West Sussex Record Office (banking and to support the Mosaic investigations) and an average of two days' officer time from other participants, £####.

- £#### towards joint workshop with West Midlands consortium to share results. Includes travel for two pre-meetings to share interim results and organise workshop.

Total £24,277.

Income

- £#### cash from Archives First (subscriptions from the participating record offices).

- £#### in-kind support from contributing partners (project management, financial management, contribution to analysis from offices with existing systems, and work with information asset owners in different authorities).

- We are looking for The National Archives Sustainability Fund to provide £#### towards the cost of the salaries.

Total £24,277.

8. Project partners

- Berkshire Record Office (West Berkshire Council)
- Dorset History Centre (Dorset County Council)
- Gloucestershire Archives (Gloucestershire County Council)
- Hampshire Record Office (Hampshire County Council)
- Isle of Wight Record Office (Isle of Wight Council)
- Portsmouth History Centre (Portsmouth City Council)
- Southampton Archives (Southampton City Council)
- Surrey History Centre (Surrey County Council)
- West Sussex Record Office (West Sussex County Council)
- Wiltshire Record Office (Wiltshire Council)
- East Sussex Record Office as lead partner in The Keep (East Sussex County Council, Brighton and Hove Council and the University of Sussex)

If feasible, we would also like to invite participation from other local authority record offices in the south east and south west not currently part of a collaborative network. However, we have not included them at this stage due to the lack of an appropriate governance instrument.

9. Benefits of working collaboratively

We remain convinced that the only way the sector and the profession are going to be able to develop robust approaches to digital preservation is through collaboration. Our first project proved this approach worked so we are keen to build on this;

- efficiencies and savings through pooling resources and expertise,
- maximising use of expertise within the region,
- contributing to the overall national picture by focusing efforts on areas not being addressed by others, and
- developing skills through active participation and sharing learning in a mutually supportive way.

10. What will the project deliver?

- Technical evaluation of e-preservation solutions available with shopping list of issues to include in a technical specification.

- Recommendations for information asset owners when commissioning new local government systems with records that require long term preservation.

- Outputs for Archives First: a) published report and b) final workshop.

- Contribution to TNA's consultancy on e-preservation action learning sets, and

- Learning and expertise will be shared across the two regions.

11. Timescales

April 2018 – March 2019.

Review, workshop, report and sharing findings, April – June 2019

12. Project organisation

- Gloucestershire County Council will provide the project manager/s, employ the business analyst, project mentor and backfill staff, and set up the detailed project plan and proposed work packages. This will be shared with collaborators prior to project start.

- West Sussex County Council will act as banker for project funds and reimburse Gloucestershire in accordance with expenditure.

Project personnel

Steve Askew (IT consultant)
is a business analyst and advised the investigation from the perspective of (local authority) corporate IT systems integration and procurement.

Claire Collins (Collections development manager)
is a professional Archivist and has over ten years experience of digital preservation practice; she leads on digital preservation for Gloucestershire County Council. In addition to research and documentation she co-ordinated the project's investigation.

Viv Cothey (Principal investigator)
is a qualified librarian and has a PhD in information science from the University of Bristol. He is the author/developer of the Gaip, GAip, SCAT progression of digital preservation tools.

Heather Forbes (County Archivist, Gloucestershire)
is a professional Archivist and has fulfilled a leadership role within the sector for over twenty years. She has been actively engaged in digital preservation matters for over ten years.

Appendix five

Brief for kick-off meeting

Archives First: Digital preservation, follow up project

Goals

- Identify currently available options for local authority and similar “memory” based institutions to specify appropriate solutions to meet their so-called digital preservation needs.

  Research question 1: What digital preservation options are available? In what ways are they similar and in what ways are they different?

- Identify currently available options for exporting Archival Information Packages (AIPs) from systems used by local authorities.

  Research question 2: How can AIPs be exported for long-term storage?

Methodology

Work in respect of the two goals will be carried out concurrently.

In both cases the investigations will include detailed work with system providers in order to understand the essential mechanisms that support the systems’ operations.

The digital preservation function will be decomposed as four components:

- packaging,
- storage
- discovery, and
- presentation.

It is anticipated that so-called digital preservation needs will vary between institutions and that not all of the four components mentioned will be equally emphasised. For example, so-called digital libraries are already adept at delivering digital discovery and presentation solutions. In some cases this may prove to be adequate for an institution.

The investigation’s methodology will assume an OAIS approach to the preservation of digital information that is “long term”. It will also be assumed that long-term digital preservation is essentially a management not a technological problem. Hence it is how the technology is managed that is most important feature of any long-term digital preservation option.

In addition to working with system providers, the investigators will also work with system customers, other similar projects, especially in respect of trusted long-term storage issues, and consortium members.

Deliverables

- an interim report to record the visits and surveys undertaken by the investigators
- a draft final report to document the investigation's analysis and conclusions
- workshops for consortium members to present the outcome of the investigation, and
- a final report.

(Dr) Viv Cothey

Principal investigator

14 May 2018

Appendix six

Cryptographic hash functions and fixity

A cryptographic hash function takes an arbitrary message and generates a fixed length message "digest". The important cryptographic property is that given a message it is easy to generate the digest. But given the digest it is not possible, that is, it is computationally infeasible, to discover the message.

Hash function algorithms are "open", that is they are available to cryptanalysts in particular to inspect and to improve.

Early algorithms include MD5 which was published in 1992 (Rivest, 1992).

Given that, amongst other things, message digests are used to protect the security of user credentials (passwords), hash collision is an intense area of research. A hash collision occurs when a message is created such that its digest is the same as that of a given digest. (This means that if a database of user credentials is obtained then password security can be by-passed.)

This research has led to a series of improved cryptographic hash functions, MD5, SHA, SHA-256, SHA-512.

From the outset it was recognised that a message digest provided a reliable test that a message had not been accidentally or maliciously corrupted. Hence publishing "file signatures", that is the digest, became a standard practice in the early days of the Web.

The digital preservation sector picked up upon this application of cryptographic hash functions to develop a rigorous notion of "fixity". The fixity characteristic of an information package, for example MD5 digest of the information package file, can be used to test that the bit-stream representing the information package has not changed.

Information package fixity is the diagnostic characteristic of a package being authentic, that is the package under consideration is the same as the reference package having the same fixity.

Appendix seven

‘gaip.xml’ package metadata

```
<!-- METS package metadata file generated by 'SCAT' -->
<!-- File created 2019-02-15T12:59:44 -->
<mets xsi:schemaLocation="http://www.loc.gov/METS/
  http://www.loc.gov/standards/mets/version111/mets.xsd">
  <metsHdr CREATEDATE="2019-02-15T12:59:44">
    <agent ROLE="CREATOR" TYPE="ORGANIZATION">
      <name>Gloucestershire Archives</name>
    </agent>
    <agent ROLE="CREATOR" TYPE="INDIVIDUAL">
      <name>Claire Collins</name>
    </agent>
  </metsHdr>
  <dmdSec ID="description_0">
    <mdWrap LABEL="RDF-XMP-DC" MDTYPE="OTHER">
      <xmlData>
        <rdf:RDF>
          <rdf:Description rdf:about="">
            <dc:creator>
              <rdf:Seq>
                <rdf:li>Claire Collins</rdf:li>
              </rdf:Seq>
            </dc:creator>
            <dc:date>
              <rdf:Seq>
                <rdf:li>2019</rdf:li>
              </rdf:Seq>
            </dc:date>
            <dc:description>
              <rdf:Alt>
                <rdf:li xml:lang="x-default">
Meeting minutes for 9 February 2018, 11 May 2018, 7 September 2018, 9 November
2018
                </rdf:li>
              </rdf:Alt>
            </dc:description>
            <dc:format>application/zip</dc:format>
            <dc:rights>
              <rdf:Alt>
                <rdf:li xml:lang="x-default">
                  All rights reserved (en)
                </rdf:li>
              </rdf:Alt>
            </dc:rights>
            <dc:title>
              <rdf:Alt>
                <rdf:li xml:lang="x-default">
            Minutes of the Pensions Committee, 2018
                </rdf:li>
              </rdf:Alt>
            </dc:title>
            <dc:type>
              <rdf:Bag>
                <rdf:li>information package</rdf:li>
              </rdf:Bag>
            </dc:type>
          </rdf:Description>
          <rdf:Description rdf:about="">
          <dcterms:spatial>Gloucestershire</dcterms:spatial>
          <dcterms:temporal>2018</dcterms:temporal>
        </rdf:Description>
        <rdf:Description rdf:about="">
          <xmp:CreateDate>2019-02-15T12:59:44</xmp:CreateDate>
          <xmp:CreatorTool>'SCAT' is Curation And Trust</xmp:CreatorTool>
          <xmp:Identifier>
            <rdf:Bag>
              <rdf:li>GCC/ADM/acc 14958/1</rdf:li>
            </rdf:Bag>
          </xmp:Identifier>
          <xmp:MetadataDate>2019-02-15T12:59:44</xmp:MetadataDate
        </rdf:Description>
        <rdf:Description rdf:about="">
          <xmpRights:Owner>
            <rdf:Bag>
              <rdf:li>Gloucestershire Archives</rdf:li>
            </rdf:Bag>
          </xmpRights:Owner>
          <xmpRights:UsageTerms>
            <rdf:Alt>
              <rdf:li xml:lang="x-default">All usage reserved (en)</rdf:li>
            </rdf:Alt>
          </xmpRights:UsageTerms>
        </rdf:Description>
      </rdf:RDF>
    </xmlData>
  </mdWrap>
</dmdSec>
<dmdSec ID="description_1">
  <mdWrap LABEL="RDF-XMP-DC" MDTYPE="OTHER">
  <xmlData><rdf:RDF>
      <rdf:Description rdf:about="">
        <dc:format>application/pdf</dc:format>
      </rdf:Description>
    </rdf:RDF>
    </xmlData>
  </mdWrap>
</dmdSec>
<dmdSec ID="description_2">
  <mdWrap LABEL="RDF-XMP-DC" MDTYPE="OTHER">
    <xmlData>
      <rdf:RDF>
        <rdf:Description rdf:about="">
          <dc:format>application/pdf</dc:format>
        </rdf:Description>
      </rdf:RDF>
    </xmlData>
  </mdWrap>
</dmdSec>
<dmdSec ID="description_3">
  <mdWrap LABEL="RDF-XMP-DC" MDTYPE="OTHER">
    <xmlData>
      <rdf:RDF>
        <rdf:Description rdf:about="">
          <dc:format>application/pdf</dc:format>
        </rdf:Description>
      </rdf:RDF>
    </xmlData>
  </mdWrap>
</dmdSec>
```

```
  <dmdSec ID="description_4">
    <mdWrap LABEL="RDF-XMP-DC" MDTYPE="OTHER">
      <xmlData>
        <rdf:RDF>
          <rdf:Description rdf:about="">
            <dc:format>application/pdf</dc:format>
          </rdf:Description>
        </rdf:RDF>
      </xmlData>
    </mdWrap>
  </dmdSec>
  <fileSec>
    <fileGrp>
      <file DMDID="description_1" ID="file_1">
        <Flocat
          LOCTYPE="OTHER" xlink:href="data/Pension Committee 2018/07092018 -
Pension Committee confidential.pdf"/>
      </file>
      <file DMDID="description_2" ID="file_2">
        <Flocat
          LOCTYPE="OTHER" xlink:href="data/Pension Committee 2018/09022018 -
Pension Committee confidential.pdf"/>
      </file>
      <file DMDID="description_3" ID="file_3">
        <Flocat
          LOCTYPE="OTHER" xlink:href="data/Pension Committee 2018/09112018-
Pension Committee confidential.pdf"/>
      </file>
      <file DMDID="description_4" ID="file_4">
        <Flocat
          LOCTYPE="OTHER"
          xlink:href="data/Pension Committee 2018/11052018- Pension Committee
confidential.pdf"/>
      </file>
    </fileGrp>
  </fileSec>
  <structMap>
    <div TYPE="information_package">
      <fptr FILEID="file_1"/>
      <fptr FILEID="file_2"/>
      <fptr FILEID="file_3"/>
      <fptr FILEID="file_4"/>
    </div>
  </structMap>
</mets>
```

Appendix eight

Structured interview topics

1. Packaging

- is material packaged (AIP)? If so the to what standards etc?
- is (catalogue type) metadata included with the package?

2. Storage

- are deposits fixity checked?
- is there encryption in transit?
- how is the authenticity/fixity of deposits monitored?
- what are the backup/recovery arrangements? When are they tested?
- is there encryption at rest? What are the key management arrangements?
- what are the access controls? Can packages be deleted? Can a single individual delete all copies?

3. Discovery

- what are the cataloguing arrangements?
- how is a package (or similar) discovered?
- how is the catalogue preserved?
- can the catalogue be reconstructed?

4. Production

- is material packaged (DIP)?
- is there authenticity/fixity verification?
- is there encryption in transit? What are the key management arrangements?
- how are presentation formats defined and can they be varied over time?
- what are the audit arrangements, for example, to know who requested an item and when?

5. General

- how long is the contract?
- what is the exit plan?
- what are the escrow arrangements?

6. Social (children) services

- is there a relationship, for example, is there a digital preservation champion within Social Services?
- Are Social Services aware of digital preservation issues? Do they have a plan?

7. Anything else not mentioned that is relevant?

Appendix nine

Calendar of project activities

Abbreviations:

ACW	Archives Council Wales
AF	Archives First
ARA	Archives and Records Association
AWM	Archives West Midlands
CC	Claire Collins
DPC	Digital Preservation Coalition
GCC	Gloucestershire County Council
HF	Heather Forbes
MIRRA	Memory Identity Rights-in-Records Access
RF	Roz Farr
SA	Steve Askew
SWHT	South West Heritage Trust
TNA	The National Archives
VC	Viv Cothey

Date	Activity	Personnel	Resources
16/5/2018	Project kick off meeting	SA CC VC HF Victoria Hoyle (MIRRA) Sam Johnston (AF) Lisa Snook (AWM) Jo Terry (AWM) Wendy Walker (AF)	Kick off briefing note
29/5/2018	Visit to Dorset History Centre	CC Cassandra Pickavance	Structured interview questionnaire
7/6/2018	Visit to Arkivum, Reading	SA Simon Bostock CC VC Paula Keogh	Structured interview questionnaire

Date	Activity	Personnel	Resources
8/6/2018	TNA action learning set Gloucester Presented overview of project and key risks, record export from line of business systems and lack of exit planning.	Lizzie Baker (Tyne & Wear Archives) Heidi Bellamy (TNA) CC VC HF Jo Pugh (TNA) Jo Terry (Staffs RO)	
25/6/2018	Circulated 100 year use case for comment	ACW AF AWM GCC records mgmt. Tim Gollins Gary Tuson Wellcome Trust	100 year use case
19/6/2018	Meeting with GCC ICT re Liquidlogic LCS	SA CC John Deane Andy Dowden HF	Kick off briefing note 100 year use case
3/7/2018	Attendance at DPC event, York 'So long and thanks for all the bits' migrating your data between repositories	VC	
9/7/2018	Visit to Preservica, Abingdon	Gareth Aitken Peter Anderton SA Tracy Broadhurst CC VC Jon Tilbury	100 year use case Structured interview questionnaire
10/7/2018	Video conference with Artefactual	SA CC VC Erin O'Meara Justin Simpson	100 year use case Structured interview questionnaire
18/7/2018	Consent from Caldicott Guardian, to use adoption records and Liquidlogic LCS as a case study.	Tim Browne HF Julie Miles Tammy Wheatley	

Date	Activity	Personnel	Resources
07/2018	Completed fixity proof of concept work	Chris Murray Steve Hawkins RF	Briefing note *Long term storage for digital preservation: the role of "fixity"* see appendix ten
12/09/2018	Visit to Wellcome Trust	CC Alexandra Eveleigh Toni Hardy Victoria Sloyan Jonathan Tweed	100 year use case Structured interview questionnaire
26/10/2018	Meeting with GCC Democratic Services re Civica modern.gov	SA Stephen Bace CC VC	Structured interview questionnaire
5/11/2018	Meeting with Metadatis	SA Rachel Care Charles Care CC HF	100 year use case
9/11/2018	Received output from modern.gov	Stephen Bace	
7/12/2018	Interim report delivered at Archives First meeting, Hampshire Record Office	AF members SA CC VC	100 year use case *A use case approach to archival preservation: an analysis* (Cothey, 2018b)
10/12/2018	Telephone conference with ACW	SA CC VC Sally McInnes Oliver Tickner Liam Tomkins	100 year use case
16/1/2019	TNA Digital Learning Set conference	HF	Summary to date
31/1/2019	Telephone conference with Norfolk Record Office	Gary Tuson Ian Palfrey	100 year use case

Date	Activity	Personnel	Resources
28/2/2019	Review profile for E-ARK AIP	VC	E-ARC aip review
7/3/2019	Joint AWM/AF workshop event at London Metropolitan Archives	AF members AWM members CC VC HF	*Retaining digital information over the long term* (Cothey 2019a)
18/3/2019	“Transforming Archive systems” event hosted by SWHT at Somerset Heritage Centre, Taunton	SA CC VC HF	Presentations from Arkivum and Metadatis
5/4/2019	Create test AIP including associated metadata	CC	*gaip.xml: package metadata* (Collins, 2019)
5/4/2019	Received dummy output from OLM	HF	
17/4/2019	Distribute “round two” note Distribute test AIP	Arkivum CC VC Metadatis Preservica	
29/4/2019	Meeting with Arkivum, Reading	Matthew Addis CC VC Paula Keogh	Round two note Test AIP
30/4/2019	Meeting with Metadatis	Charles Care Rachel Care CC VC	Round two note Test AIP
29/5/2019	Attendance at DPC event, Birmingham ‘Counting on reproducibility: tangible efforts and intangible assets’ ’	HF	
12/6/2019	Meeting with Preservica, Abingdon	Peter Anderton Tracy Broadhurst CC VC Jon Tilbury	Round two note Test AIP

Date	Activity	Personnel	Resources
29/8/2019	ARA Conference, Leeds	VC	*Never mind the technology: * (* Cothey, 2019b, 2019c)
11/2019	Draft comments feedback	ACW Arkivum Artefactual Metadatis Norfolk Record Office Preservica	Draft comments
21-22/ 11/2019	TNA workshop on Bayesian networks	HF	Draft report

Long term storage for digital preservation: the role of “fixity”

Introduction

An important characteristic of systems that store digital files is that files are unchanged. This property, that is the property of being unchanged or unchanging, is called fixity and is a key feature of systems that offer long term storage for digital preservation.

“Long term” in this context may be many decades and is certainly longer than the lifetime of any particular storage technology. The challenge therefore is to provide a storage technology-independent mechanism that can be used to demonstrate fixity over decades.

The purpose of this paper is to describe such a mechanism in sufficient detail that it can be used in a real-world trial. The field of cryptography has already addressed parallel problems. Several families of cryptographic hash algorithms have been developed which when used in an appropriate protocol provide, for example, digital signing. A cryptographic hash algorithm or message digest is an essentially one-way function that generates a fixed length output from a potentially large input (or file). The one-way property implies that the output cannot be undone to discover the input. Also, the hash output is unique in the sense that it is resistant to collisions. A collision is where two different inputs result in the same output.

The message digest of a file therefore provides the basis of a fixity mechanism. If over a period of time the output value remains the same then it can be accepted that the input is unchanged. And, if the output value is different then it is certain that the input is changed.

Popular, that is widely used, algorithms all belong to the so-called MD4 family of message digests. The algorithms are non-proprietary and have multiple software implementations. Examples include MD5, SHA-1 and SHA-256.

Like all algorithms used in cryptography, message digest algorithms are routinely attacked in order to identify any weakness. Although collision attacks have been demonstrated for some older algorithms, the attacks rely on some very particular conditions and are algorithm specific.

The paper identifies three stages in the digital preservation information cycle and defines in outline a fixity management protocol for each. The three stages are;

- deposition, a file is deposited in the storage system,
- curation, the file is maintained securely by the storage system,
- production, a copy of a stored file is provided on request.

The fixity management protocols rely on each party, that is the depositor on the one hand and the store on the other, separately computing the values of message digests for the file in question. These values are referred to as fixity values or just fixity . If the fixity found agrees with the expected fixity then the file is accepted as being unchanged. Otherwise a fixity failure has been discovered and remedial action is indicated. It should also be clear that while the store is securely maintaining preserved files, the depositor needs to securely maintain a set of expected fixity values. This is not a significant challenge since this is the day to day storage of an operational file.

For the avoidance doubt it should be noted that issues such as security, backup, disaster recovery etc. are not discussed here since they are a component of any storage system and should not be confused with the particular requirements for long term digital preservation.

Deposition

This fixity management protocol for the deposit of a file into a storage system assumes a secure communication channel from depositor to the store and vice-versa.

The description refers to a single cryptographic hash algorithm only but in practice several would be used. Note also that the deposit may be a batch of files rather than a single file.

Action on failure is not described. It is likely that this will involve manual intervention.

Depositor:	Store:
Identify the source file	
Create a temporary copy of the source file	
Compute the source fixity	
Compute the temporary copy fixity	
If the fixity values agree then proceed	
Create depositor's fixity report *	
Upload the source file to the store *target* file	
	Receive the target file
	Compute the target fixity
	Create store's fixity report *
	Either (push) send store's fixity report to depositor or (pull) save fixity report
Either receive (push) store's fixity report or (pull) request store's fixity report	
Process fixity reports, for each file in the store's fixity report compare the store's fixity with the expected fixity using at least two algorithms.	

* The fixity report mentioned here is an xml document conforming to the PREMIS schema. An example is given later.

If fixity values agree then proceed
Include fixity values in the depositor's catalog of expected values
Delete temporary copy of source file

Optionally report success to store

Receive success message

Curation

This fixity management protocol for the curation of files in a storage system assumes a secure communication channel from the depositor to the store and vice-versa.

Action on failure is not described. It is likely that this will involve manual intervention.

Depositor:

Store:

Compute the fixity of stored files
Create the store's fixity report
Either (push) send store's fixity report to depositor or (pull) save fixity report

Either receive (push) store's fixity report or (pull) request store's fixity report
Process fixity reports, for each file in the depositor's fixity catalog compare the store's fixity with the expected fixity using at least two algorithms.

If fixity values agree then proceed

Optionally report success to store

Receive success message

Production

This fixity management protocol for the production of files from a storage system assumes a secure communication channel from the depositor to the store and vice-versa.

Action on failure is not described. It is likely that this will involve manual intervention.

Depositor:	Store:
Identify the store *source* file and request from the store	
	Receive the source file request Download the source file to the depositor's *target* file
Receive the target file Compute the target fixity Create depositor's fixity report Process fixity report, compare the target fixity with the expected source fixity using at least two algorithms. If fixity values agree then proceed	
Optionally report success to store	
	Receive success message

Fixity report

Please note: for illustration only, this is NOT to be regarded as a template.

```
<?xml version='1.0'?>
<premis xmlns="info:lc/xmlns/premis-v2"
        xmlns:xlink="http://www.w3.org/1999/xlink"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="info:lc/xmlns/premis-v2
        http://www.loc.gov/standards/premis/v2/premis-v2-3.xsd"
        version="2.3">
  <object xsi:type='file'>
    <objectIdentifier>
      <objectIdentifierType>UUID</objectIdentifierType>
      <objectIdentifierValue>990a4dce-beea-5e04-8110-2a18b4c4f7ca
      </objectIdentifierValue>
    </objectIdentifier>
    <objectCharacteristics>
      <compositionLevel>0</compositionLevel>
      <fixity>
        <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
        <messageDigest>1234</messageDigest>
        <messageDigestOriginator>SCAT</messageDigestOriginator>
      </fixity>
      <fixity>
        <messageDigestAlgorithm>SHA-1</messageDigestAlgorithm>
        <messageDigest>77a3c7f76206e34913829b3002b00a464ce3db4c</messageDigest>
        <messageDigestOriginator>SCAT</messageDigestOriginator>
      </fixity>
      <size>0</size>
      <format>
        <formatDesignation>
          <formatName>xxxxx</formatName>
        </formatDesignation>
      </format>
    </objectCharacteristics>
  </object>
</premis>
```