A use case approach to archival digital preservation: an analysis

Viv Cothey

Working draft

[Initial 12 July 2018]

27 November 2018

1. Introduction

This working paper is prepared as part of an Archives First project and is preliminary in nature.

The aim of the paper is to develop an understanding of the “system” architectures that could satisfy the need of local authorities to provide for digital preservation. Note that the term “system” used here implies a broad scope rather than technology.

The goal of archival information preservation is to ensure the survival of information, that is, the information continues to be accessible for an arbitrary period of time.

The paper identifies some of the failure events which threaten the survival of “digital information” which must remain accessible for at least 100 years. The “100 year use case” shown in Appendix A is a particular instance. The paper proposes risk avoidance and risk mitigation strategies to manage these potential failure events. A common mitigation strategy is to make use of “redundancy” to avoid a single failure event destroying information.

Anticipated failure events may not materialise, and indeed one hopes that they will not. However responsible information preservation needs to be based on more than good luck!

Identifying potential failure events and risk mitigation planning requires an analysis of the whole digital preservation system. This approach to the analysis is much influenced by the
Planning tool for trusted electronic repositories (DigitalPreservationEurope, 2008).

Preservation systems must be trans-generational, that is, information must survive the replacement of any infrastructure component (organisations, people, buildings, services etc.) by the next generation of that component. Potential failure events that arise here are commonly mitigated by succession plans and exit plans. It will be seen that adequate component exit planning is central to a digital preservation system architecture.

The analysis assumes an OAIS approach to information preservation.

Achieving the digital preservation survival goal entails three activities which can be summarised as,

“get the bits”,
“preserve the bits”, and,
“access the bits”.

The sequence of activities here illustrates the dependency that each activity has on its precursor.

The paper is organised to discuss these activities. In passing the paper touches on several issues that relate to long term information management and which have been under represented in the digital preservation discourse or have been unnecessarily conflated with digital preservation.

These issues include,

- archives versus libraries
- authenticity
- bibliographic catalogues versus archive catalogues
- the Cloud
- digital/electronic libraries, document management systems and digital asset management
- disaster recovery and business continuity
- hardware and software obsolescence
- Open Archive Information System (OAIS) not to be confused with Organization for the Advancement of Structured Information Standards (OASIS) or Open Archives Initiative (OAI)
- provenance, and
- the Web and electronic publishing.

The paper concludes by deducing that a layered modular architecture that supports the portability of preserved information would satisfy the need of local authorities to provide for digital preservation.

2. Getting the bits

This section of the paper identifies that in the example presented by the 100 year use case, “getting the bits” is a non-trivial exercise. Success in initiating a digital preservation process cannot be assumed.

It is generally taken that the starting point for digital preservation is a (digital computer) file. Accordingly there is concern over file formats, software obsolescence and the like. This is explored later in the next section, “Preserving the bits”.

However an earlier investigation (Archives First ########) showed that much, possibly most, digital information that local authorities need to preserve is currently held as data in proprietary databases. The databases form part of computerised transaction processing (TP) systems. Each local authority business area operates its own, usually slightly modified, version of a proprietary line of business TP product.

The 100 year use case derives from work that (English) local authorities do in respect of children in care.

As yet there is no evidence that any supplier of any of the relevant TP products has demonstrated extracting information from its database and has constructed an “adoption record” or similar. Note that “record” is here used in the archival sense and not in the database sense. The lack of attention to this challenge of extracting information is in part explained by the current “legal hold” relating to children in care but this excuse does not apply to the other areas of a local authority’s work.

A second strand of this Archives First project is to work with one of the line of business product suppliers and their TP system in order to be able to demonstrate the required information extraction.

Obtaining information that is independent of its originating proprietary business TP system is a non-trivial task that needs supplier support and intervention.

However, when successful it is anticipated that each “adoption record” will comprise an organised collection of document, image and possibly audio-visual files.

Extracting the bits in the 100 year use case is a significant and materially different activity to copying a computer document or similar, whatever its format, or extracting files from a document management system. By way of comparison, it is anticipated that the democratic services’ document management system will export preservable documents.

3. Preserving the bits

The problem of “preserving the bits” or the long term storage of digital files is sometimes thought of as being the central problem for digital preservation. In this section of the paper several concomitant problems are presented as being of greater importance.

The 100 year use case preservation goal is to ensure the *known* survival of at least one copy of the complete collection of document and image files mentioned above. Achieving this entails that it must always be possible to *discover* the information and to render file content so as to be able to *access* the information.

Hence the “preserving the bits” activity requires not only successfully storing a given bit-string but being able to locate required information, ensure its authenticity and address issues of hardware and software obsolescence. Strategies to manage or mitigate these survival threats will be discussed below.

However from the outset it should be noted that there is a significant institutional and supplier risk. It is reasonable to assume that no organisation contributing to a 100 year preservation task is itself going to survive for 100 years. Therefore the task must be shared over time between a succession of contributors. Each succession will present its own set of threats that affect a successful handover. Exit plans can be used to manage these risks. Two forms of succession can be identified. The first, for example at the expiry of a service agreement, can provide for an “orderly” handover. The other form of succession is “disorderly”. An exit plan for a disorderly handover assumes a “Carillion” like collapse of a contributor. Not only can the failed contributor not provide any assistance, clients of the failed contributor cannot access any of their property.

Hardware obsolescence is not a significant threat since the mechanics of “preserving the bits” is not predicated upon the long term operation of any particular hardware. The collection of hardware components employed constitutes the hardware infrastructure. Upgrades and replacement of hardware infrastructure components is a routine activity within the IT sector.

Software obsolescence is a potential failure event because of the essential dependency that digital information has on using software to transform bits into information. There are two forms of dependency and therefore risk. Firstly there is the intrinsic dependency that the preserved bits has on appropriate rendering software in order to make the information accessible. This *file format* risk is discussed in the next section, “Accessing the bits”.

Secondly, the digital preservation task depends upon may different software (and hardware) products to support preservation functions such as the transmission, storage, management and verification of the preserved bits. Collectively this is a software infrastructure. Software infrastructure components may be very specialized and their use confined to a single contributor.

Seen in the context of 100 year time scale, the dependency on an individual software infrastructure component is ephemeral. That is, a particular software component is unlikely to remain unmodified for more than a few years. As with hardware, the upgrade or replacement of software infrastructure components is a routine activity within the IT sector.

However there must be no interdependency between the software or hardware infrastructure and the preserved bits. That is the activity of preserving the bits is agnostic with respect to the storage software or hardware infrastructure used. Failure within the storage infrastructure can be managed and mitigated by either the disaster recovery and business continuity arrangements of the organisation concerned or by the exit plans mentioned above.

Tactics such as off-site data-backup or even 100% concurrent operational duplication can form part of an organisation’s disaster recovery (DR) and business continuity arrangement. These arrangements should allow an organisation to re-establish an agreed level of *operational* service following a disaster. (Operational services are generally based on data that is modified in order to respond to a service request. DR restores the state of the data to correspond to some short time prior to recovery invocation.)

Disaster recovery arrangements have a necessary but insufficient role in the activity of digital preservation. Generally major organisations have effective disaster recovery arrangements and an excellent record in avoiding data loss. However data loss does occur, for example Kings College, London (PA Consulting, 2017).

Digital preservation does not tolerate any data loss.

Cloud based service providers make use of distributed data centres and remote access via the Internet principally the Web. They quote *durability* statistics such as “11 nines” to imply a 99.999999999 per cent durability of objects over a given year which corresponds to an average annual expected loss of 0.000000001 per cent of objects.

Since the statistical methodology is not made public then it is difficult to understand what this really means. The Cloud providers do not publish any empirical loss data, rather they rely upon a probability model estimating the likelihood of loss. Such estimates are vulnerable to “Black Swan” hazardous events that have not been adequately included in the model.

Cloud providers are claiming that, according to their models, even if users lose data, *on average*, they will not lose much.

One sees a similar statistical presentation from promoters of various lottery products although in this case the argument is the reverse and is based on empirical data; someone each week is guaranteed to become a millionaire! On average however each player might expect to win only, say, £0.000001 per year, that is, you will win, but on average, not much.

The lottery promoters have it right. There are many millions of files stored in the Cloud. Each year some will be lost (that is the data is unrecoverable). For the owner of the information this could be a catastrophe; but on average, across all files, the data loss is minuscule. Furthermore it is not at all obvious that the file's owner would be aware of its loss until they needed the information.

Community memory projects often include a Web component in order to enhance the accessibility of the digital resources that have been developed. It can be an uncomfortable surprise to such project teams that providing Web access to information is not an effective digital preservation tactic. Not only is there frequently no long term funding plan to pay for Web hosting but Web hosting businesses have proved to be ephemeral with data being lost (for example, WPStrands, 2018).

Managing the continued authenticity of the preserved bits is central to the task of ensuring the known survival of the information. Any failure is reported in time to take effective business continuity action.

This gives rise to the notion of a *trusted digital repository*. By this we mean an organisation that has responsibility for storing the preserved bits and which

(a) provides routine fixity reports that demonstrate the continued stability of the preserved bits, and
(b) demonstrates effective exit plans, both orderly and disorderly.

Archivists are able to warrant that a traditional record is *authentic*, that is, the record as presented is unchanged since crossing the archive threshold. (Note that this is not the same as claiming that any information is "true".) This is sometimes known as the custodial model after Jenkinson and which emphases the physical integrity of the record as the basis of his moral defence.

Fixity or cryptographic hash values provide archivists with the tool needed to maintain their moral defence of the digital record even though they no longer retain physical custody. Without this the confidence of users in the authenticity of preserved information cannot be maintained and the evidential value of the record cannot be established.

In *all* threat circumstances *all* of the bit-strings must survive *and* bit-string fixity must be demonstrable. The risk of failure in this task is mitigated by *preservation* redundancy, that is the use of at least two organisationally and technological discrete trusted repositories.

Being organisationally discrete includes ensuring that the repository contract termination/expiry dates are sufficiently staggered.

Technologically discrete means not only that the repositories do not directly share any technological or infrastructure resource, but they are not technologically synchronised. An example of synchronisation here would be two repositories independently using the same version of a third party software product within their storage infrastructure. When the third party issues a maintenance patch for the product both repositories apply it at the same time. If the patch is faulty then both repositories lose data simultaneously (see for example, Speed R, 2018).

This strategy of preservation redundancy also assists developing an effective exit plan in the case of a disorderly failure of a contributor, cf. Carillion.

Note however that preservation redundancy does not excuse any contributor from maintaining its own disaster recovery and business continuity arrangements.

4. Accessing the bits

In this section we consider a range of threats to being able to access preserved information. As mentioned earlier it must be possible to both *discover* the bits and to *render* them.

The challenges of discovery in the context of long term information management has been ignored while attention has been distracted by the supposed threat of format obsolescence to rendering.

"Accessing the bits" is firstly an information retrieval (IR) problem, albeit with some special features. Users engaged in IR have a variety of tools to facilitate *discovery*. These are often call "finding aids". Later in this section we discuss more obviously technical issues.

Finding aids

Information technology has had a significant impact on IR with the "online public access catalog" for libraries being a seminal application. Professionals and end-users now have access to improved searchable catalogues that, it is assumed, provide a sophisticated finding aid for resource discovery.

However effective searching as an IR technique is much helped by having an understanding of how the finding aid (and it's entries) has 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 it does for archive end-users. At the root of this problem are some fundamental differences between libraries and archives. The differences are explored more fully by Schellenberg who devotes an entire chapter of *Modern Archives* to distinguishing between archives and libraries (Schellenberg, 1956).

The differences have not been generally recognised by the IT and IR communities who have applied a simpler bibliographic paradigm to model the management 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 (where similarity is based on what the resource is *about*). There are several internationally recognised schemes all of which facilitate users browsing for information. The classification of knowledge is generally hierarchical becoming ever more specific according to the needs of the user community.

For example, Library of Congress classification, class C,

- C Auxiliary sciences of history
  - CD Diplomatics, Archives, Seals
    - CD921 Archives

For open access libraries the classification assigned to a resource will generally determine where the resource is physically located.

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. 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 by many libraries. The cataloging in publication projects operated by national libraries maintain this standardisation.

A key feature of resource discovery in libraries is that finding, that is, locating a needed resource, is based on what knowledge the resource is about. “Zen and the art of motorcycle maintenance” by Robert Persig is at 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 classification scheme that identifies the bureaucratic context (and therefore origin) of a given resource. In particular this establishes the provenance of the resource and the information that it provides. Although two archives may be similar, for example they may both receive material from a local authority, their finding aids are likely to be different since archive finding aid records are not standardised and are not shared in the sense that bibliographic records are shared.

A key feature of archival discovery is that finding is based on how the resource was created.

An illustration of this contrast in approach occurs when archives include books in their finding aids. For example, the library catalogue entry for Matthews’ book on St Ives is based on 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

Every library that has the book will use the same metadata values for a full catalogue record, since how to build these values has also been standardised. The classification system used here is Dewey Decimal.

By way of comparison, the archive discovery metadata (from The National Archives, Kew) provides provenence but the bibliographic information is barely recognizable and of little use to any other library (or archive).

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

A normal author-title search is here not effective at discovering the resource.

Digital content may present additional 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.

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.

In respect of the 100 year use case it is presumed that resource discovery is mediated by a professional archivist or equivalent who will use the archive finding aid. Some special considerations in support of privacy/confidentiality apply. The finding aid will reveal the unique identifier of an Archival Information Package (AIP). A copy of the AIP can be obtained from the trusted digital repository and its authenticity can be confirmed by inspecting its fixity.

The archive finding aid/discovery tool could be a single point of failure. Not only must the information resource remain accessible for 100 years, so must the discovery tool.

A specific mitigation strategy would be to ensure that the AIP contains sufficient metadata to re-establish a finding aid in the event that this is required.

Unlike bibliographic records, archive finding aid records are bespoke and not portable. Given both the accumulated intellectual value and the operational reliance upon the archive finding aid, appropriate exit planning (and testing) should be undertaken.

More technical issues

Being able to access the AIP is a pre-requisite for accessing the bits. This information package is a container for an organised collection of (document, image and audio-visual) digital files and will itself have a particular file format.

There is therefore a risk that it will not be possible to render the container file format where rendering in this context means successfully extracting the container contents. This would be a particular example of software obsolescence. The risk of file format obsolescence could be mitigated by redundancy – having more than one version using different container file formats. But one still has to select a container file format. Choosing enduring file formats is an endemic challenge for digital preservation.

File formats most at risk are those where there is a dependency on proprietary or otherwise restricted intellectual property. File formats least at risk are those where

- there are multiple independent implementations of software to read the format
- there is non-restrictive publication of all necessary intellectual property to create software to read the format, and
- there is a sufficiently large user population to generate an enduring interest in maintaining the software to read the format.

The “zip” container file format was introduced in 1989 and was associated with a data compression utility. A Zip product was commercially available which can include encryption options. The file format algorithms have been adopted by both proprietary and non-proprietary software manufactures. In consequence there are many independent implementations of software to read a zip container.

A version of zip is also described by the ISO/IEC 21320-1:2015 “Information technology: document container file” standard. This is relevant to the 100 year use case being analysed since it prohibits both compression and encryption.

Lastly the zip format is widely used as a component within office productivity file formats from both Microsoft and the OASIS’s Open Document Format.

An AIP conforming to ISO/IEC 21320-1:2015 thus satisfies all the “least at risk” criteria defined above.

“Tar” (tape archive) introduced in 1979 is a satisfactory alternative container file format but is less well recognised.

Being able to render the AIP zip container file to gain access to its content is a necessary pre-requisite for “accessing the bits”.

The 100 year use case AIP is presumed to include an organised collection of document, image and audio-visual files. OAIS requires that an AIP also includes sufficient additional material to make meaningful the package content as well as, optionally, having a mechanism to verify the content fixity. These properties can be satisfied by including in the AIP a package metadata file based on an enduring schema and a manifest which refers to at least two cryptographic hash fixity values for each of the package content files.

It is convenient to introduce some element of standardisation for the AIP beyond defining the container file format so that automatic tools can be developed to assist in managing and processing them.

The BagIt specification published by the Library of Congress and maintained as an IETF (Internet Engineering Task Force) draft is a good candidate for describing the internal arrangement of the AIP container. The specification and associated software implementations satisfy the “formats least at risk” criteria described above. The specification requires that there be a cryptographic hash based manifest of the BagIt payload. Automated processes can therefore confirm the authenticity of an AIP payload.

Two threats remain that could adversely affect accessing the preserved information. These are that a document, image or audio-visual file format cannot be rendered which here means displayed, and less obviously, that the package metadata fails to make the data meaningful.

Even though a file format currently satisfies the “formats least at risk” criteria described above, it is conceivable that the format becomes inaccessible in the future. A mitigation strategy for this forms part of a so-called “preservation plan”. This plan is based on knowing the range of file formats contained within the AIPs.

There are tools available to detect the format of a file. For example the “file” utility identifies the format of an arbitrary file. “file” is an example of the longevity of many open-source software products. It was originally published in 1973 which predates Microsoft. Since 2002 the PRONOM file format registry has maintained its PUID service which assigns a unique identifier to not only the principal format but possibly more fine grained sub-version formats. Other format registries have briefly offered a similar format identification service.

Should a file format unexpectedly become threatened then the AIP can be reconstituted with an additional format version that complements the “at risk” format. Given an initial judicious choice of file formats then this threat event is considered minimal and less probable than the “Carillion” type event!

The second threat here is that the package metadata fails to make the data meaningful. A mitigation is to base the metadata file on an enduring schema, for example, METS. Gaps in the METS schema can be filled with additional schema, such as Extensible Metadata Platform (XMP), ISO 16684-1:2012 part 1 and ISO 16684-2:2014 part 2. XMP satisfies the formats least at risk” criteria described earlier.

As with all metadata, although the field names or “tags” are highly specific, the values assigned are not always controlled. Canonical, mostly English language, values for bibliographic metadata tags are defined by the Anglo-American Cataloguing Rules (AACR) and more recently Resource Description and Access (RDA). This is not the case for tags used by the bespoke scheme that is the archives’ finding aid that was discussed earlier.

The information’s provenance (as given by the bespoke classification) is a further factor in making the data meaningful. Knowing how the information was created is an essential part of determining what the information means. The importance of accuracy, precision and trans-generational consistency is well recognised by the (archive) professionals involved. The threat of loss of meaning is here mitigated by the survival of this profession.

4 Conclusion

There is no digital preservation “silver bullet” solution that can be acquired. Rather the 100 year use case demands an active curatorial process working with a succession of contributors/partners.

Digital preservation should not be thought of as a technological problem, rather it is an institutional/organisational problem.

No contributor/partner should be relied upon. This includes the host local authority.

Possible “architectural” possibilities are predicated on there being preservation redundancy. That is at least two independent digital preservation storage solutions are operated concurrently.

In order to support orderly/disorderly exit planning a layered modular architecture is preferred rather than a monolithic approach. Greater independence between layers reduces risk.

Effective exit plans are best supported by there being data portability. That is, it should be straightforward to relocate (just the) AIPs between (successive) trusted digital repositories.

Four preservation layers can be conceptualised. These are shown diagrammatically in Figure 1.

The “deposit” layer is responsible for obtaining information from the business system and delivering it in an appropriate format and with the required metadata to the packaging layer.

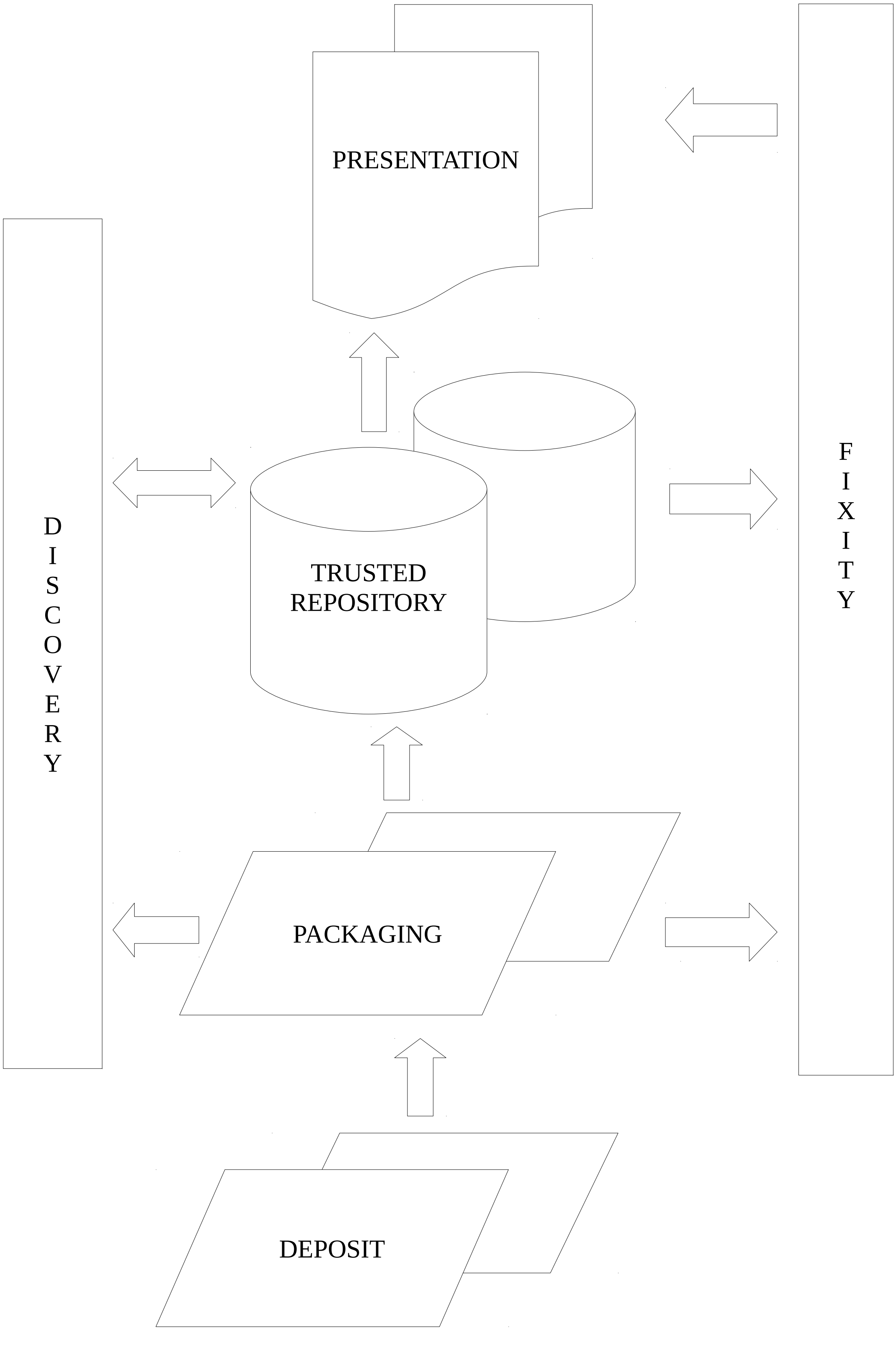
The “packaging” layer is responsible for creating the AIP and informing both the “discovery” system and the “fixity management” system. The “packaging” layer delivers AIPs to the “trusted repository”.

The “trusted repository” supports the “fixity management” system and recovery of the “discovery” system. When required, the “trusted repository” delivers copy AIPs to the “presentation” layer.

The “presentation” layer is responsible for transforming the information content of of AIPs to the *format du jour* (Rusbridge, 2008). The “presentation” layer is informed by the “fixity mananagement” system.

A particular vulnerability especially in the context of this use case is the “discovery” system.

Figure 1.


PRESENTATION
DISCOVERY
TRUSTED REPOSITORY
FIXITY
PACKAGING
DEPOSIT

References

DigitalPreservationEurope, 2008. DPE Repository Planning Checklist and Guidance DPE-D3.2. [Available from: https://digital.library.unt.edu/ark:/67531/metadc799759/m2/1/high_res_d/platter.pdf]

PA Consulting (2017). IT infrastructure resiliency review. [Available from https://regmedia.co.uk/2017/02/23/kcl_external_review.pdf].

Rusbridge C (2006). Excuse me...some digital preservation fallacies? [Available from http://www.ariadne.ac.uk/issue46/rusbridge].

Schellenberg T R (1956). Modern archives: principles and techniques. Chicago, University of Chicago Press.

Speed, R (2018). On the third day of Windows Microsoft gave to me: a file munching run of deltree. [Available from https://www.theregister.co.uk/2018/10/05/windows_10_wipes_files/].

WPStrands (2018). The shocking truth about your backups. [Available from https://wpstrands.com/dont-rely-on-your-hosts-backups/].

Appendix A

Digital preservation for local authorities

“The 100 year use case”

25 June 2018 adjusted for appendix B

1. Introduction

1.1 This is an evolving working document for Archives First: project two. The document will be developed as experience/insights are gained.

1.2 The purpose of the document is to record and communicate the 100 year use case where digital information needs to be preserved for 100 years. This is a requirement of the Statutory Guidance on Adoption 2013 in respect of a so-called “adoption record”.

1.3 These records are currently subject to the GDPR as enacted by the Data Protection Act 2018 (DPA).

1.4 Not all local authority digital preservation use cases are affected by the DPA and not all information has a statutory retention as long as 100 years.

However, the 100 year use case is applicable in all instances where the requirement is to retain information of enduring value in perpetuity.

1.5 An OAIS approach is assumed. In particular the AIP container, by definition, includes all the information that is being preserved together with sufficient material for a user to access the information.

1.6 A feature of the use case is that information access is restricted*;* general access is not permitted for 100 years.

1.7 In order to further generalise the applicability of the use case AIP creation is an automated process. Manual processes would be both error prone and unable to scale to the volumes required.

1.8 Some additional technical comment is given in Appendix B in order to provide clarification.

2. Packaging (AIP creation)

2.1 Information in the AIP is born-digital together with digital attachments. The information will have been created and managed by a business transaction processing system over several years having been migrated from legacy systems.

2.2 The information asset owner is the appointed business manager.

2.3 The Data Controller responsible for compliance is the local authority.

2.4 A trigger event will cause information in the transaction processing system to be collated and exported as a structured collection of simple document and image format computer files.

2.5 Package metadata that is metadata describing the package rather than individual computer file metadata will be compiled and recorded as a computer file using an enduring format and schema (see Appendix B).

2.6 The AIP is created by including the structured collection of simple document and image files and the package metadata file in a single container file which has a UUID name. The AIP creation process includes calculating and recording package fixity values (see Appendix B).

2.7 Packages are created automatically.

3. Storage

3.1 Depositing the AIP in a trusted dark store.

3.1.1 The AIP and its package fixity values are created locally by the depositor.

3.1.2 A copy of the AIP is deposited in a reliable secure long-term digital storage system. It is assumed that this storage system is remote.

3.1.3 The AIP is encrypted whilst in transit and a suitable transmission security protocol is employed. The AIP is decrypted following receipt and is stored as plain text.

3.1.4 Several fixity values for the (plain text) AIP are calculated by the storage system and reported to the depositor for comparison with the locally created fixity values. The deposit is successful only if the fixity values correspond.

3.1.5 There is no local copy of the AIP.

3.2 Maintaining trust

3.2.1 Maintaining trust is an active management process. The trusted store regularly demonstrates the continued authenticity of the AIPs in its custody by recalculating and reporting fixity values.

3.2.2 No AIPs or any package content files are deleted (silently or otherwise).

3.2.3 The storage system conforms to all relevant reliability and security standards.

3.2.4 The storage system reports the results of DR AIP restore testing.

3.2.5 An AIP escrow arrangement is in place. The escrow copies are stored securely by a third-party. AIPs in transit between the storage system and the escrow store are encrypted; escrow copies of the AIPs are plain text. Escrow invocation is tested. Escrow invocation does not require any proprietary software.

3.3 Exit

3.3.1 The termination arrangement provides for an orderly transfer of AIPs to another trusted dark store.

4. Discovery

4.1 AIP discovery is facilitated by a locally maintained searchable catalogue that holds a copy of the AIP package metadata together with the AIP UUID.

5. Presentation (DIP creation)

5.1 A discovery system is maintained locally which provides the UUID name of a required AIP.

5.2 There is a secure user authentication procedure in place which the trusted dark store uses to verify the requester's credentials.

5.3 In response to a valid request from a verified user, the storage system will provide a copy of the AIP. The AIP is encrypted in transit.

5.4 Following decryption, the requester calculates several fixity values for the retrieved AIP which are compared with the locally stored values. The retrieval is successful only if the fixity values correspond.

5.5 Creating the DIP, that is managing the transformation AIP to DIP, is a mediated process (see Appendix B).

5.6 Document and image AIP content file formats are transformed automatically and without being executed.

5.7 The DIP is made available to a qualified end user, that is an end user to whom some or all of the information can be disclosed.

5.8 The end user is advised to employ anti-virus software.

6. Managing risks

This section is much influenced by the Planning Tool for Trusted Electronic Repositories (DigitalPreservationEurope, 2008), (PLATTER).

6.1 Financial

6.1.1 Both the depositor and the trusted store are exposed to existential financial (including organisational) risk.

6.1.2 Suitable escrow arrangements mitigate the effect of failure by the trusted store.

6.1.3 Failure by the depositor is not managed. (Management options could include insuring the credit risk, risk pooling, or last resort arrangements.)

6.2 Key personnel

6.2.1 Both the depositor and the trusted store are vulnerable to the loss of key personnel.

6.2.2 The risk is mitigated by

i. avoiding there being a single key individual
ii. relying only on industry standard practice
iii. maintaining relevant skill-sets
iv. maintaining full documentation

6.2.3 The depositor's authorised users are key personnel.

6.2.4 All staff in the depositor's chain of authority are key personnel.

6.3 Preservation plan

6.3.1 The depositor is vulnerable to the future obsolescence of file formats used in the AIP content which prevent access.

6.3.2 The risk is mitigated by managing the range of file formats used. If a format is deemed to be at risk because, for example, it is proprietary and no open reader exists, then a non-proprietary version is included within the AIP.

6.3.3 Demonstrating authenticity requires local access to a register of AIP fixity values. This is supported by a local operational system which is maintained day to day in the usual way. Fixity value data is retained using a non-proprietary format.

6.3.4 Both the depositor and the trusted store are vulnerable to the adverse effects of technological developments (both hardware and software).

6.3.5 This risk is mitigated by the identification of critical technology and an appropriate “technology watch”.

6.4 Succession plan

6.4.1 The depositor and the trusted store are exposed to succession failure, both technological and human.

6.4.2 The risk is mitigated by relevant transition and handover procedures including testing.

6.4.3 All key personnel and technology are included in the succession plan.

6.5 Discovery system

6.5.1 The discovery system is provided by a locally maintained catalog and is exposed to multiple failure modes (i.e financial, organisational, technological etc.).

6.5.2 Risk is partially mitigated by appropriate discovery metadata (package metadata) being included in each AIP which could be used to re-populate a catalog.

6.6 Disaster plan

6.6.1 A disaster is an unexpected and rapid change event that adversely affects the ability of either the depositor or the trusted store to provide the expected level of service.

6.6.2 The risk of a disaster is mitigated by there being an agreed disaster plan which includes invocation, communication and response.

Appendix B

AIP container file

A popular information package container specification is BagIt created by the Library of Congress. A reference implementation is available.

Historically both tar and zip serialization were supported by the reference implementation but latterly BagIt ignores serialization leaving this to the user.

Zip is now commonly used due to the widespread availability of open source cross-platform tools.

Either tar or zip serialized container files can be compressed. However this should be avoided.

AIP container file names should be unique. A popular way to achieve this is to use a UUID. A file name extension should be optional.

Several fixity values for the AIP are calculated and recorded. A fixity value is a cryptographic hash or message digest obtained by encoding the container file bit string. Message digests are often described ambiguously as checksums.

AIP content is not “virus checked”.

AIP content is not encrypted. The need to maintain decryption keys for a 100 years external to the AIP and for this to be a pre-requisite to accessing preserved information breaks OAIS.

DIP container file

The DIP container file is similar in outline to the AIP container file.

The differences are,

- the DIP is essentially ephemeral and the container file name need not be unique since it will be used by only a single end user,
- there is no requirement to manage DIP fixity,
- in addition to the transformed content files, the DIP also contains relevant intellectual property and terms of use statements.

Package metadata

METS (Metadata Encoding for Transmission Standard) maintained by the Library of Congress is an example of a relevant enduring XML schema.

Any ancillary schema used will also be enduring either because they are open or because schema documentation is included in the AIP.