Retaining digital information over the long-term

Viv Cothey
viv.cothey@{wlv.ac.uk, gloucestershire.gov.uk}

5 March 2019

Abstract:

Long-term business processes are needed in order to achieve long-term information retention. In this paper we introduce the "authentic preservation" business process in order to achieve the *known authentic survival* of digital information. Since we are constrained by having access to only a succession of short-term contributions from supporting technological systems, we need to introduce a long-term "authenticate" process as a constituent of authentic preservation. Lastly we need to provide for "disorderly exit" plans. We present a co-operative model for an authentic preservation business process supported using technological systems generally available to local authorities.

Introduction:

This paper is part of an Archives First project undertaken by a consortium of local authority archives or similar "memory" organisations (see appendix 1) following an earlier project investigating digital preservation within a local authority context (Cothey and Pickavance, 2017). The current project goals include specifying appropriate solutions to meet local authority so-called digital preservation needs. In this paper we set out one of the solutions that has been identified.

At an early stage in the project's investigation a "100 year use case" was documented. This was in order to understand and communicate the local authority digital preservation need. The use case follows from the Statutory Guidance on Adoption (2013) in respect of a so-called "adoption case records", see appendix 2. An important qualification is that although information must be retained for at least 100 years, it can be disclosed only under quite restrictive conditions.

It was discovered that the digital preservation market is not well informed in respect of local authority retention requirements such as the 100 year use case. In consequence this requirement is not yet fully served. The basics are simply stated, they are; get the “stuff” from the producer, package it, store it, and then after 100 years, locate it, retrieve it, authenticate it and present it to the consumer.

More formally the business processes for long-term retention are as illustrated in figure 1.

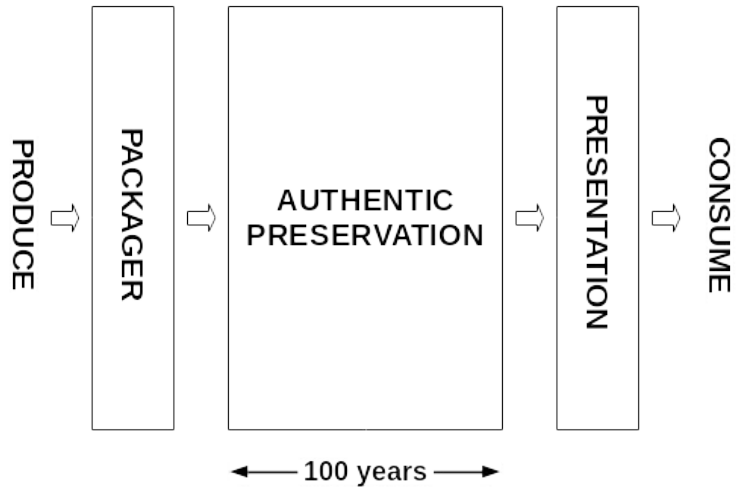

Figure 1: Long-term retention business processes

The defining characteristic is that in 100 years time the Archival Information Package (AIP) that is retrieved (and is transformed to a Dissemination Information Package (DIP) by the presentation process) must be demonstrated to be the same as the AIP that was created (by the packager process) and was stored 100 years previously.

The business process that achieves this is here called “authentic preservation”. Its purpose is to achieve the known authentic survival of the retained information. That is, as just described, any particular AIP when retrieved is known to be the same as it was when it was created. Furthermore this is true for all AIPs that have been created.

In order to understand authentic preservation we decompose it to reveal its three constituent business processes, *store*, *discover* and *authenticate*. This is shown diagrammatically in figure 2.

Figure 2: Authentic preservation business processes

Collectively these processes facilitate the known authentic survival of retained information. They have the following outline purposes;

store	receives AIPs from the packager process and responds to requests from *discover* by passing a copy of the AIP to the presentation process
discover	receives AIP metadata from the *packager* process, provides a search service to identify relevant packages and passes requests for AIPs to *store* and *authenticate*
authenticate	receives AIP message digest values from the *packager* process in order to maintain a database of all AIP message digest values and responds to requests from *discover* by passing authenticating message digest values to the *presentation* process.

Each of the processes, *store*, *discover*, and *authenticate* need to be supported by a succession of temporary or short-term technological systems. The *packager* process operates only once in respect of any one AIP. The format of the process output, that is the AIP, needs to be defined but its technological support can be easily changed. Similarly the *presentation* process operates only once in respect of any one DIP. The DIP contains information created from an AIP and intended to be consumed in response to a particular access request. Although its content will be derived from the AIP payload, the DIP content cannot be anticipated since the scale of redaction needed as well as the “format du jour” (Rusbridge, 2006) is not known until the access request is received.

Sophisticated storage management systems and catalogue systems for discovery are generally available. Fixity management systems which support the long-term authenticate process are a novelty and are described below. Each system can be expected to be operational for a limited lifetime, say ten years. This is because of, for example, system obsolescence or local authority re-procurement rules. Replacing a legacy system at the end of its operational life entails migrating data from the legacy system to the replacement system.

Figure 3 illustrates this life-cycle mismatch between the long-term business process and the supporting temporary technology systems.

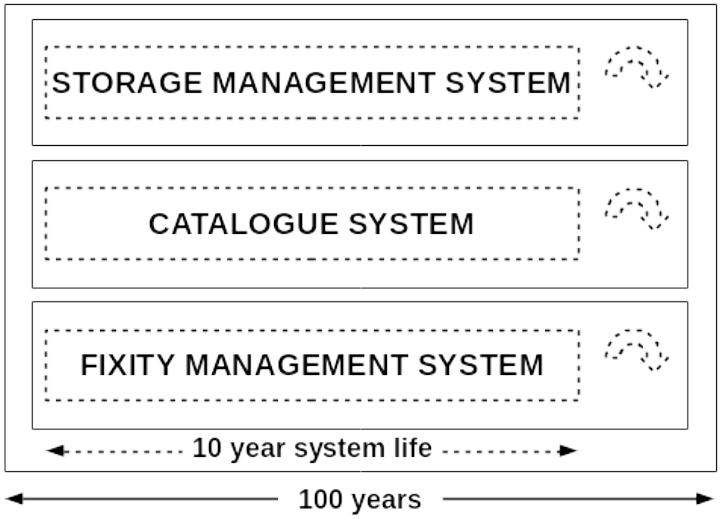

Figure 3: Successive temporary technological systems supporting authentic preservation processes

The long-term characteristic of the authentic preservation business process can only be achieved by using a sequence of technological systems each being essentially short-term or temporary. It is worth quoting Abrams et al (2009) when explaining their change of focus;

> "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."
>
> (Introduction)

The fixity management system (see Cothey, 2010, p. 214) maintains an independent database of the cryptographic hashes or message digests for each package.[1] This database can be just a simple collection of text files. Note that this fixity management process is additional to the intrinsic fixity mechanism that is present in procedures such as BagIt (Kunze et al., 2018) and which is used to verify the internal integrity of a “bag”.

The authentic preservation process is vulnerable to threats to the temporary technological systems. During its operational lifetime standard disaster recovery procedures can be expected to mitigate most risks. But authentic preservation is especially vulnerable to failures when migrating a temporary system to its replacement at the end of an operational life. This end of life risk can only be mitigated by an effective “exit plan”.

Local authorities rely on external technology suppliers for both products and services. Institutional or organisational failure such as the “compulsory liquidation with immediate effect" of Carillion (Wikipaedia, 2019) shows the need to plan for a “disorderly” exit from the technology systems that support authentic preservation. It must be anticipated that there will be no access to any asset or resource, for example hardware, software or services, that was controlled or provided by the failed supplier. The local authority itself must be included as a potential candidate for failure.

The planning tool developed by DigitalPreservationEurope (2008) is helpful when thinking about (disorderly) exit planning.

Appropriate solutions to meet local authority long-term information retention needs must include disorderly exit plans. A straightforward approach to achieving this is to concurrently operate at least two sufficiently independent combinations of the technology systems used in support of authentic preservation. Ideally this includes an overlap of any contractual dependencies. These requirements were identified by Fryer (2015) and in part by Brown and Fryer (2014). The arrangement is illustrated diagrammatically in figure 4.

1 This is a example of a possible “blockchain” application.

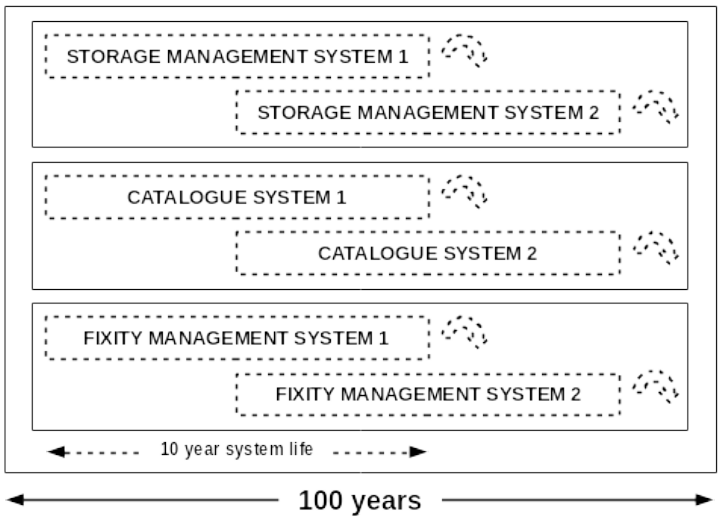

Figure 4: Overlapping concurrent short term system support for authentic preservation

Note the contrast in time-scales. For example, the storage management technology systems that support the store constituent of the authentic preservation business process will be renewed many times within the 100 year period. It can be expected that (orderly) exit plans will perform. Under normal conditions, even including the need to invoke disaster recovery procedures, the systems will perform reliably.

However one can have no such reassurances in respect of a disorderly exit. The vulnerability of an authentic preservation process to an institutional failure will depend upon many details, for example the extent of coupling or modular separation between the supporting technological systems. At one extreme a tightly coupled monolithic arrangement from a single supplier will present an all-or-nothing threat compared to a loosely coupled arrangement with several suppliers for each interchangeable module.

At the present time it is considered not economically feasible for a local authority to develop valid plans for restoration that extend to the disorderly failure of any of the current suppliers in the digital preservation market. This is an area of current interest and investigation. In consequence disorderly exit planning must focus on alternative provision unaffected by the supplier failure.

For example, there are several catalogue systems that can address the discover constituent of the authentic preservation process. If the cataloguing metadata for the retained information is captured by more than one catalogue system and is stored in

different storage management systems then the discover process is better protected from the disorderly failure of either the cataloguing system supplier or the metadata storage system supplier. In principle this union approach to aggregating catalogues is already an accepted practice.

The remainder of the paper sets out a co-operative model for an authentic preservation business process followed by the paper's conclusions and recommendations.

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

A co-operative model for an authentic preservation business process:

The model is illustrated in figure 5 which can be seen to be similar to figure 1 above. However in figure 5 the temporary technological support for authentic preservation is replicated by a second local authority.

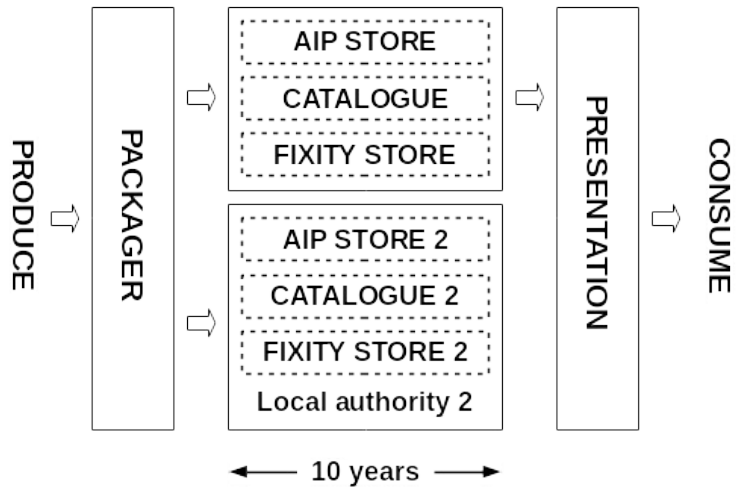

Figure 5: Co-operative model for authentic preservation

In the model the authentic preservation component supports the store/discover/authenticate business processes shown in figure 2 above.

At their most basic,

AIP STORE	is a simple file store of AIPs
CATALOGUE	is the existing collection catalogue which cross references the location of both physical and digital material
FIXITY STORE	is a simple file store of AIP fixity information.

All the data is managed as part of the local authority's corporate data and thereby benefits from corporate information security, data restoration and disaster recovery procedures.

The model assumes that effective exit plans are in place so that system migrations are reliable.

Normally therefore just the upper authentic preservation component will be sufficient to achieve the known authentic survival of digital information. However the purpose of the model is to provide for reliable recovery from the disorderly failure of a supplier, provider or institution when orderly exit plans fail.

The model achieves this by replicating the storage and catalogue functions so that the packager process passes information additionally to "Local authority 2". The model presumes that the suppliers, providers and institutions are sufficiently independent and not shared by both local authorities. Not only should there be geographic separation, the local authorities should not, for example, share the same IT outsource supplier (cf Carillion).

In the event that it is necessary to recover from a disorderly failure, then the AIP store, catalogue or fixity store can be repopulated from "Local authority 2".

Implementation of the model could achieve level 4 compliance across all five categories[2] in the current (version 1) levels of digital preservation scheme of the National Digital Stewardship Alliance (NDSA, 2019).

2 Storage and geographic location, file fixity and data integrity, information security, metadata, and file formats.

Conclusion:

Authentic preservation processes are necessary in order to achieve the known survival of digital information. There is no digital preservation “silver bullet”. Retaining digital information over the long-term is a curatorial process that depends upon well managed procedures and not on any particular technological product.

A continuing constraint is the life-cycle mismatch between the long-term authentic preservation process and the temporary nature of supporting technology systems. Disorderly exit plans must be created that can ensure the survival of information in the face of the disorderly failure of suppliers, providers or institutions.

A co-operative model to provide authentic preservation can avoid single points of disorderly failure.

Recommendation:

Consortium members should develop their understanding of authentic preservation and investigate opportunities for co-operation in order to create effective disorderly exit plans.

Acknowledgements:

The Archives First follow-up project is partly funded by The National Archives, UK. Additional funding is provided by the Archives First membership.

The author thanks Claire Collins and Steve Askew for their critical contributions.

References:

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

Brown A and Fryer C (2014). Achieving sustainable digital preservation in the cloud. In *Proceedings of 2nd annual conference of the International Council on Archives, Girona, 11-15 October 2014*. Available from <http://www.girona.cat/web/ica2014/eng/comunicacions.php>.

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 and Pickavance C (2017). Archives First: digital preservation project. Available from <https://gloucestershire.gov.uk/archives/digital-curation/digital-curation-projects/archives-first/201709-archivesfirst-digital-preservation-final-report.pdf>.

DigitalPreservationEurope (2008). DPE repository planning checklist and guidance DPE-D3.2. Available from <https://digital.library.unt.edu/ark:/67531/metadc799759/m1/1/>.

Fryer C (2015). Project to production: digital preservation at the Houses of Parliament, 2010-2020. *International Journal of Digital Curation* 10(2).

Kunze J, Littman J, Madden E, Scancella J and Adams C (2018). The BagIt file packaging format (v1.0). *Internet Requests for Comments 8493*. Available from <https://tools.ietf.org/id/draft-kunze-bagit-17.txt>.

NDSA (2019). Available from <https://ndsa.org/documents/Levels_v1.pdf>.

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

Statutory Guidance on Adoption (2013). Available from <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/270100/adoption_statutory_guidance_2013.pdf>.

Wikipedia (2019). Carillion. Available from <https://en.wikipedia.org/wiki/Carillion>.

Appendix 1

Archives First consortium members

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, 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

Digital preservation for local authorities

"The 100 year use case"

25 June 2018
[adjusted to be appendix 2 of "Retaining digital information over the long-term"]

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 3 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 3).

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 3).

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 3).

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.

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.