Re-purposing the E-ARK AIP format

Kuldar Aas
kuldar.aas@ra.ee
Estonian National Archives

Karin Bredenberg
karin.bredenberg@sydarkivera.se
Kommunalförbundet Sydarkivera

Sven Schlarb
sven.schlarb@ait.ac.at
AIT Austrian Institute of Technology

Carl Wilson
carl@openpreservation.org
Open Preservation Foundation

Introduction

Digital preservation necessitates mitigating very long-term risks. It is almost inevitable that archival content, stored as archival information packages (AIPs), will have to be transferred between repository systems in time. Current practise when replacing a system starts with exporting information packages from the old repository system. The exported information packages are then converted to satisfy the new repository's ingest requirements. Finally, the converted packages are ingested into the new repository.

This transfer process brings considerable - and sometimes hidden - hardware and personnel costs. Tasks such as integrity checks, metadata mapping, file format validation and conversions, packaging, and storage, are demanding and expensive. This can be particularly true at large scale, for example the entire digital holdings of national archives.

E-ARK deliverable D4.1 (Rörden & Randmäe, 2014) introduced the concept of a pan-european AIP format. This aimed to avoid transfer costs by use of a standardised package format, enabling systems to ingest AIPs directly from storage systems without copying or restructuring. The pan-european AIP format would define standards for building modular and reusable components. These could be shared by the digital preservation community and memory organisations.

The AIP format should rely on standards which are ubiquitous in the digital preservation community and be complemented by conventions that bring improved interoperability and tool support.

Outline

The white paper is structured as follows:

- We begin with an overview of current practices for the ingest and creation of AIPs in repository systems and will outline related issues.

- Next we review the purpose of the AIP format, before presenting an overview of the pan-European AIP format and the benefits it might bring to memory organisations, digital preservation practitioners, and solution providers.
- The next section considers the needs of solution providers, whose task is to develop AIP implementations. We recognise the need for a pragmatic specification and that the adoption of common standards and practices is encouraged by an approach that brings the benefit of interoperability without disrupting existing practices.
- Finally, we present a concrete suggestion of changes needed to the current E-ARK AIP format with an overview of the design decisions that would shape the new AIP format.

An Overview of Current Practice and Related Issues

Digital archiving has become increasingly important in the past decade, with new vendors entering the market specifically targeting long-term digital preservation. Memory organisations can now choose between a variety of repository systems. Each solution has its own standards and protocols for receiving, managing, and storing digital objects. The transfer of digital objects to the repository is usually defined by a solution-specific SIP format that is validated against a set of rules. These rules verify that the SIP fulfils the repository's requirements and ensure that it can be ingested successfully.

Content and metadata to be ingested into a repository must be arranged in the SIP structure specified by the vendor. For data not held in an existing repository, some solutions offer software tools to create SIPs from existing metadata and digital files. For material already in a repository, existing metadata must be mapped to the new repository's SIP standard; then AIPs are either converted directly to the new SIP format, or via export as DIPs before conversion to SIPs. Finally the prepared SIPs are validated and ingested into the new system. The results of ingest are AIPs which conform to the organisation's specification or a custom vendor format.

Some memory organisations formulate specific requirements, determining how AIPs should be stored in the long term. Others let the provider determine how the IPs are securely and efficiently transmitted and safeguarded in archival storage. Some customers require that additional replicas of the AIPs are created in a vendor-independent format, for example, to safeguard them on long-term storage devices, such as tape drives. For this purpose some solutions offer the possibility to export AIPs. However, these AIPs are often snapshots which don't retain a complete record of the AIP's provenance. If they do, there is generally no consensus on how this information should be recorded.

Why is this the case? First, the adoption of flexible XML schemas means there are many ways to record any particular feature. This is partly addressed by defining and publishing profiles, as in the case of METS structural metadata, and by providing recommendations for the use of PREMIS with METS (The Library of Congress, 2017), for example. Second, repository vendors look to differentiate themselves by developing efficient custom AIP implementations. Third, while vendors address efficient ingest of compatible SIP formats, system migration isn't seen as a priority. Meanwhile, there are no incentives or benefits for solution providers to develop or adopt common standards.

As a consequence there is insufficient compatibility between AIP implementations. There are a lack of conventions as to how to record specific metadata elements, such as provenance and rights, and the way in which the AIPs are stored. This means less tool support for provenance functions, complex logical AIP structures, and compliance with legal and data protection regulations. Further, solution providers do not have a reliable specification to implement these functions and validate AIPs.

Purpose and Requirements of an AIP Format

An AIP is comprised of content objects together with all metadata necessary to ensure preservation of the content. The OAIS Reference Model states:

“The AIP is defined to provide a concise way of referring to a set of information that has, in principle, all the qualities needed for permanent, or indefinite, Long Term Preservation of a designated Information Object.” (OAIS, 2012, p. 4-36)

The OAIS Reference Model distinguishes Content Information, i.e. the actual content objects to be preserved, and the Preservation Description Information (PDI), i.e. the contextual information which is required to ensure the content objects can be trusted, accessed and interpreted over a long time period (OAIS, 2012, p. 4-37). It is explicitly mentioned that the “contents of each type of PDI are left to the discretion of the individual Archive” (OAIS, 2012, p. 4-37), i.e. what needs to be part of the PDI and how the information is provided the implementers’ decision.

Apart from encapsulating content and metadata, additional core AIP functions are:

- to ensure its authenticity and integrity;
- provide an identifier which denotes the AIP during the whole life-cycle; and
- record provenance, change history, and information related to access, retention periods, etc.

The main difference between an AIP compared to a SIP or a DIP is that the latter are ephemeral. An AIP is permanent and must record the history of changes, i.e. it should be possible to document and trace the creation, update, and deletion of metadata, representations or digital objects.

The OAIS Reference Model defines functional areas with a high level of abstraction to provide a generally applicable framework for repository systems. However, this means it lacks the level of detail needed to achieve interoperability and tool support for exchanging AIPs between repositories. The AIP format should recommend the use of structural, descriptive, and preservation metadata that facilitate interoperability across different repository systems.

Ideally, the AIP contains all the information needed to interpret the content objects or metadata. Reference to external information always carries the risk that this information is no longer available. For this reason, the physical and logical autonomy was postulated in E-ARK Deliverable D4.2 as follows:

“Physical autonomy requires an AIP to be an integrated byte stream, which can be handled independently from any repository system and be processed by ubiquitously available standard tools. Logical autonomy is reached when an AIP contains all the metadata which are necessary to interpret the content of the AIP, even if no additional information about it is available outside the AIP.” (E-ARK D4.2, p. 7)

Related to this is the requirement to store the AIP as a single container file in the form of a continuous byte-stream so that it can be sequentially written to storage devices where E-ARK deliverable D4.2 states:

“On the physical level, that implies that an AIP is represented by a robust container, which is separated into different segments. A robust container is a continuous byte stream, which can be handled by widely available tools, independent of any proprietary software component of a repository. It contains the parts of an AIP integrated into such a byte stream in a widely supported format, which allows the extraction of parts of its content, even if the byte stream has become corrupted. For the time being uncompressed tarballs will be used.” (E-ARK D4.2, p. 7)

However, this requirement is based on the assumption that it is possible to bundle digital objects in individual physical AIP container files. While this might still be applicable for large audio-visual files, it may not be practical for very large content objects, such as databases, for example.

Finally, the AIP format should be vendor independent, i.e. it can be interpreted and maintained by different repository systems. Ideally, repository system migration would be a simple copy operation or might not require AIP transfers at all.

Aim and benefits of a pan-European AIP format

Exchanging AIPs between repositories is challenging, especially if package standards differ or organisations use different repository systems. Larger organizations can achieve a degree of vendor-specific standardisation through building networks (e.g. user groups) or through joint procurements. While using the same repository provides system-level compatibility, differences in AIP implementations can still prevent a successful exchange of the AIPs. The ability to freely exchange AIPs between different repository implementations undoubtedly remains a fundamental requirement in many cases.

More than 10 years ago, several research projects in the US attempted the exchange of content between preservation repositories (Shirky, 2005).This was later put up for discussion by (Caplan et al., 2010) and investigated in an experimental setup as part of the TIPR project (“Towards Interoperable Preservation Repositories”). They concluded that alongside transfer formats, protocols and other technical aspects, semantics are also important for AIP transfers, as they ensure repository systems can trust and “understand” each other.

There is no lack of initiatives working to align metadata standards and best practices. Many such initiatives are driven by the international digital preservation community. The PREMIS Data Dictionary (The Library of Congress, 2020), the ongoing work on METS profiles (The Library of Congress, 2018), the Open Archives Initiative Protocol for Metadata Harvesting (Open Archives Initiative, 2015), highlight just a few examples.

A pan-European AIP format should not aim to introduce new standards. We propose instead to leverage existing standards, already widely adopted by the digital preservation community, while facilitating interoperability through the adoption of sensible conventions. We believe that a higher degree of interoperability can be achieved by these conventions in the long run, through convergence of the repository implementations.

Regarding the establishment of a standard AIP format, there is no practical need to impose it as a format that solution providers and organisations must implement. The ability to export AIPs in a format that is understood by the target system can already avoid the costly procedure of producing system-specific SIPs and thereby simplify ingest into the new repository system.

One strength of a shared AIP format is that it encourages the development of software to support it. Tools to handle more complex use cases e.g. the segmentation of large AIPs into smaller related parts or serialising deltas AIPs for small changes should find a larger potential market.

The Need for Pragmatism

The original AIP format proposed by E-ARK got a mixed reception from vendors. They understandably felt that the AIP structure was an implementation detail of their solution. Some even see AIP feature support and processing efficiency as competitive advantages.

To address the concern, we clearly state that we believe that a repository’s internal logical and physical AIP structure is a decision for the vendor/developer. A diversity of storage and management implementation is the sign of a healthy marketplace and a valuable source of digital preservation innovation. It allows memory institutions a choice of alternative approaches when looking for the right solution to safeguard their digital collections.

That said, we feel that these organisations' long-term preservation aspirations are hindered by the lack of a standard long-term storage format. Complex and staff/resource intensive transfers between repository systems are expensive and expose preserved content and metadata to unnecessary risks. We don’t anticipate that vendors will implement “first-class” back-end support for a standard AIP format, but other approaches are possible. Many repository systems offer dark-archive or escrow storage nodes, which hold an offline copy of the AIPs as a preservation measure. This non-operational node could persist the AIPs in the standard format providing a backup of last resort, and a ready source of data for transfer to an updated repository system. It also mitigates the risk of a vendor going out of business. An even simpler approach might be to offer a “common” AIP export format for transfer between repository systems.

Addressing the Change

The intention is to start with an initial set of AIP extensions to the CSIP. The format needs to be extensible, to cope with future requirements. The AIP specification is simplified by defining the AIP as an instance of the Common Specification for Information Packages (CSIP) without exceptions.

Instead of suggesting a custom structure to incorporate the change history of the AIP in a single container, the AIP format provides recommendations for persisting the logical AIP as a single or several physical containers, with protocols to record the change history. The AIP specification covers more complex use cases, such as information package segmentation and incremental updates to AIPs, avoiding storage redundancy when updating AIPs. We anticipate that repository providers will support such cases in future releases.

The simplest, valid AIP instance is a CSIP of OAIS Package Type[1] “AIP”however, it lacks features to support complex AIP functions. Specific use cases require dedicated agreements as to how they can be fulfilled across solutions. These may be use cases where digital objects or representations are excluded from full-text indexing or certain access tools, or displaying a history of migration actions that contributed to the current state of an AIP.

The AIP only provides additions when there is a clear benefit, for example better tool support or improved functionality for long-term preservation of AIPs. It consists of extensions and recommendations to record the provenance, change history and rights or data protection metadata, and describes how the AIPs are stored.

The question remains as to how organisations and vendors might be encouraged to adopt the AIP specification. Firstly, the AIP is based upon the CSIP. The AIP format then applies to the extent in which the use cases of advanced AIP handling need to be addressed in an organisation, in customer installations, or in a vendor’s solution.

Aside from the benefits of compatibility and interoperability between repository systems across Europe, what might motivate repository vendors to participate? Given many years of experience and investment in digital preservation, it is understandable that vendors are reluctant to implement a new AIP format. The AIP specification is conceived as a vendor independent export/dark archive format.

Concrete benefits for vendors include:

- Reduced system migration costs making the decision to purchase a new system cheaper and easier. The high cost of system migration costs might deter organizations from purchasing a new repository solution.
- The additional effort required to transform and migrate existing archival data into SIPs isn’t always born by the customer. Large scale, complex and long running system

[1] https://earkcsip.dilcis.eu/schema/CSIPVocabularyOAISPackageType.xml

installations and migrations may occupy vendor staff and resources as well as the customers.

- While installing a system, vendors and customers may not think too hard about end of life provision for new software. Given digital preservation's long-term aspirations this isn't a credible position. Providing a documented, vendor independent format that allows for transfer of content and metadata to a subsequent replacement system should reassure the customer at purchase that the vendor takes the issue seriously.

We stated earlier that the AIP specification is derived from the CSIP. Work on the specification will focus on three areas:

- Structure: The structure of the AIP is prescribed by the CSIP. The AIP provides recommendations for additional attributes required for archiving, e.g. a property indicating which representation is the archival master copy. The AIP format also allows the relationships between AIPs to be described. These are not the thematic relationships addressed by descriptive metadata, but relationships concerning the derivation and division of physical AIP containers.
- Provenance, change history, and data protection/privacy: The AIP format makes provisions for recording provenance, an audit history of changes, and legal restrictions or measures to protect sensitive information.
- Storage: The AIP is stored as a physical file container which is a continuous byte stream. The AIP format recommends storing the AIP as a series of physical container packages. This allows other repository systems to incorporate them immediately as a new source of information, ideally without even having to prepare SIPs at all. It also addresses more complex use-cases, such as delta-AIP or information package segmentation.

The new version of the specification will propose recommendations for these areas with a focus on the initial set of use cases to be addressed by the AIP.

Bibliography

Caplan, P. (2008, 12). Repository to Repository Transfer of Enriched Archival Information Packages. *D-Lib Magazine*, *14*(11/12). http://www.dlib.org/dlib/november08/caplan/11caplan.html#McDonough

Caplan, P., Kehoe, W., & Pawletko, J. (2010, 03). Towards Interoperable Preservation Repositories (TIPR). *International Journal of Digital Curation*, *5*. 10.1145/2039274.2039290

The Library of Congress. (2017). *Guidelines for using PREMIS with METS for exchange*. Retrieved 4 24, 2021, from https://www.loc.gov/standards/premis/guidelines2017-premismets.pdf

The Library of Congress. (2018, 06 24). *METS Profiles*. METS Profiles: Metadata Encoding and Transmission Standard (METS) Official Web Site. Retrieved 4 9, 2021, from http://www.loc.gov/standards/mets/mets-profiles.html

The Library of Congress. (2020, 4 28). *PREMIS Data Dictionary*. PREMIS Data Dictionary for Preservation Metadata, Version 3.0. Retrieved 4 9, 2021, from https://www.loc.gov/standards/premis/v3/

OAIS. (2012, 06 01). *OAIS (ISO 14721:2012)*. Reference Model for an Open Archival Information System. Retrieved 01 10, 2021, from https://public.ccsds.org/Pubs/650x0m2.pdf

Open Archives Initiative. (2015, 01 08). *The Open Archives Initiative Protocol*. The Open Archives Initiative Protocol for Metadata Harvesting. Retrieved 4 9, 2021, from https://www.openarchives.org/OAI/openarchivesprotocol.html

Rörden, J., & Randmäe, P. (2014, 9 8). *pdf D4.1 E-ARK Report on Available Formats and Restrictions*. Retrieved 4 26, 2021, from http://www.eark-project.com/resources/project-deliverables/7-e-ark-d41-report-on-available-formats-and-restrictions.html

Shirky, C. (2005, 06 01). *Library of Congress Archive Ingest and Handling Test (AIHT) Final Report*. https://www.digitalpreservation.gov/partners/documents/ndiipp_aiht_final_report.pdf