D.4.1 Archival Information Package (AIP) Formats and Restrictions

DOI: 10.5281/zenodo.1172649

Grant Agreement Number:	620998
Project Title:	European Archival Records and Knowledge Preservation
Release Date:	13th February 2018
Contributors	
Name	Affiliation
Jan Rörden	National Archives of Estonia
Piret Randmäe	National Archives of Estonia
Rainer Schmidt	Austrian Institute of Technology
Manfred Thaller	University of Cologne
Clive Billenness	University of Brighton
David Anderson	University of Brighton
Janet Anderson	University of Brighton

STATEMENT OF ORIGINALITY

Statement of originality:
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

TABLE OF CONTENTS

List of figures

Executive Summary

Within the overall strategy of E-ARK the AIPs have a small but highly significant role. While SIPs have to be created by all archival systems to be used within individual solutions, and DIPs have to be deployed in all environments, the conversion between the three pan-European formats for information packages (SIP ==> AIP ==> DIP), has to be realized only once, leading to significant savings.

The structure of the E-ARK AIP has therefore not to be derived from individual implementations of existing technology providers, but from the major abstract concepts currently existing. This report examines and describes these concepts and derives recommendations from them, which will be used as a guide line in the next step, the design of the pan-European AIP format and the conversion tools supporting it.

Introduction

The present work – D4.1: Report on available formats and restrictions – is part of the Task 4.1: E-ARK-AIP specification within Work Package 4: Archival Records Preservation of the E-ARK Project. It is based primarily upon research in the literature about Archival Information Package (AIP)[1] concepts, as well as on the results of a survey and a series of qualitative interviews that were conducted in cooperation with Work Packages 3 (Transfer of Records to Archives) and 5 (Archival Records Access Services).

It is recommended that this report be read in conjunction with E-ARK Deliverables D3.1 – "Report on available best practices" and D5.1 – "GAP report between requirements for access and current access solutions" (both available on-line on the project website http://eark-project.eu)

Within E-ARK, this deliverable serves as a starting point for the definition not only of the pan-European AIP format itself, but also as a starting point for the design of the software components which, according to the description of work, shall convert the SIPs to AIPs and the AIPs to DIPs. Only if the requirements for all three information package formats are known is it possible to start defining the interfaces. Without the information collected here, no other work related to the AIPs would be possible within E-ARK.

A pan-European AIP format: The vision.

In the OAIS progression "*SIP to AIP to DIP*" (SIP – Submission Information Package, DIP – Dissemination Information Package), the AIP can easily be overlooked, as all it does is to reside somewhere in a repository between creation and usage. Its neglect creates serious pitfalls for preservation, however:

- It is widely agreed that when a SIP is turned into an AIP, the formats of data files it contains shall be checked for fitness for preservation and potentially migrated into more suitable file formats. Such migrated files may then be packaged into proprietary container formats, which are specific to a given repository. As installations of repository systems are less numerous by several orders of magnitude than readers for popular file formats, it is, however, much more likely that a proprietary container format becomes unreadable than a proprietary file format, for which millions of copies of software able to read it have been distributed. This might be seen as too grim a picture. Indeed, many providers or repository system do announce that they follow open (XML based) structures. But using an XML based format, which is used only in a small number of instances, and where the documentation may be not widely distributed (intuitive elements being less than intuitive in the long run) is only a small improvement over a proprietary format, which on top of being proprietary is also binary. A pan-European AIP format has to be based completely upon using widely accepted description standards and packaging methods, using constructions going beyond existing ones only if absolutely necessary.

- Independent from the first issue is the problem that an institution installing a preservation system needs an exit strategy in the same way as for any other information system. Hardware and software *will* change, so AIPs residing within a given repository *will* have to be transferred to another repository eventually. While it is in principle feasible to convert the complete content of a petabyte-range repository into DIPs, which are then re-ingested into the AIP format for a new repository, very

[1] OAIS: AIP definition

simple reflection will make it clear that this creates a very serious financial risk for the future. If two repository systems share the AIP format, we are essentially discussing a copy operation, which is restricted mainly by the bandwidth by which the two systems are connected. If they do not share such a common AIP format, the costliest operation which exists in a repository - ingest - has to be executed again for all SIPs, which will than undergo massive consistency and integrity checks of the SIPs, before they are re-converted into an AIP. This is assuming that the DIP format of the first repository generation is identical to the SIP format of the second generation repository. If this assumption is wrong, what has been a simple copy operation is replaced by a small software project, which creates a DIP to SIP conversion tool on top of all other costs.

- Asynchronous to systematic operations upon the contents of a repository triggered by changes in the technical infrastructure of the repository, operations related to all – or a significant portion of – the AIPs within a repository will also be triggered by other events. The most obvious occurs when formats contained within the AIPs become obsolete, forcing mass migration. In this situation, the usage of a standard AIP format across Europe generates extremely significant synergies for the archives and a huge market potential for service providers. Any migration tool or service offered for an AIP format used throughout Europe can compete in a market defined by a very large number of customers. The same is true for all tools which may be needed to augment AIPs as a result of stricter requirements for security that may arise in the future.
- While cost considerations may prevent it currently, and there is no clear definition of requirements yet, we consider it obvious that multiple copies of AIPs, stored at sufficiently distant locations to make the destruction of all repository sites by the same catastrophic event unlikely, will have to become a standard requirement in the near future. Managing these repositories with intentionally different systems has obvious advantages, as this means, that they cannot be endangered by an undetected malfunction in one of the repository systems. To avoid the necessity of multiple ingest processes into different repository systems, the AIP format should be transparent to the individual repository system. This allows copying for AIPs easily between repositories, increasing their security.
- A further problem is that current repository systems typically rely on the integrity of a few critical system components. Individual files in a repository are recoverable only if these system components survive. This is a big disadvantage compared with the analogue case. The catastrophic events at the city archive of Cologne have demonstrated that a very great portion of the content has been recovered, even after the archive – as an intact storage system – has been destroyed completely. If the Cologne city archive had held a digital repository, where just a single critical hardware component was physically destroyed, the remaining components of the system might have been totally beyond recovery – individual storage units being readable, but not interpretable, when a critical, central catalogue is destroyed.

The vision of a pan European AIP format is therefore simple: Have a digital container, which allows one to keep digital content safe over long periods, be vendor independent, enable multiple copies on different repository systems, and allow for recovery even from truly catastrophic events.

This vision implies that the AIP layer is as independent as possible from the archiving system as a whole.

Within the E-ARK architecture, we see the pan-European AIP format as a possibility for significant synergies and savings. Figure 1 visualizes this. Many vendors currently provide solutions to transfer the content of living systems into consistent information packages. Preservica's "SIP Creator"[2] is a good example of that. In our interpretation, this is actually outside of the OAIS model, or at least outside the diagram, with which it is frequently identified, as the OAIS concept of "Ingest" assumes that a consistent SIP exists. The creation of such a SIP – which could be defined as "Pre-Ingest" – currently results in either a vendor-specific SIP, or it is mixed into the actual Ingest. If a software provider or service provider replaces the proprietary SIP with the pan-European solution, the whole costly processes of ingest – consistency checks, format verifications, etc. - can be left for a vendor-independent conversion unit for which an exemplary implementation will be provided by work package 4. A vendor of archival solutions can therefore concentrate on the interface, which can be optimized according to the requirements of individual customers, and rely for the backend on reusable components. (And the same holds true for the later conversion of an AIP into a DIP.)

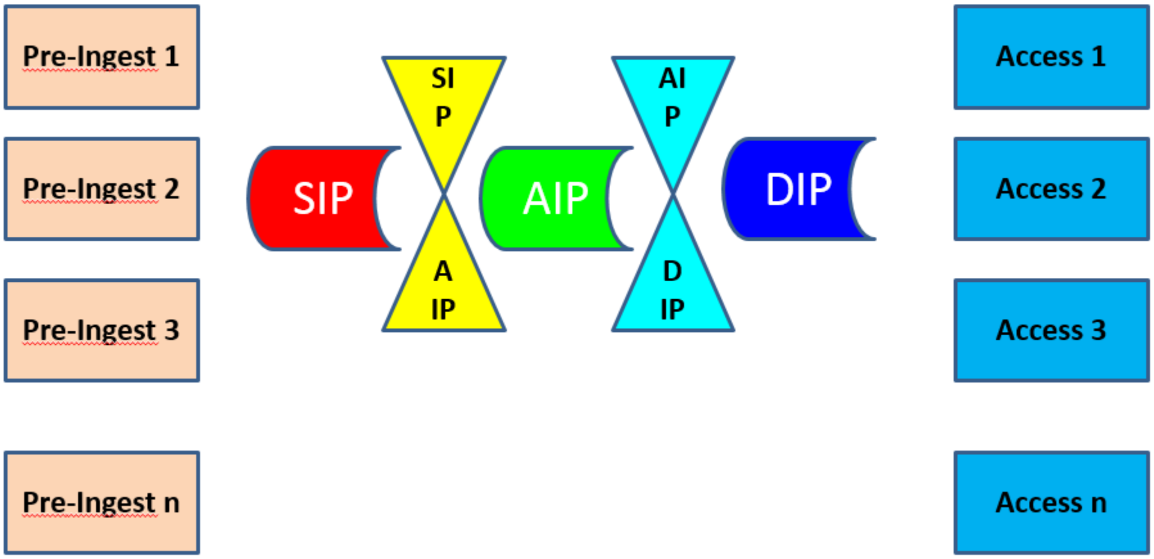

Figure 1: AIPs within the preservation workflow

Consequences of the vision for the structure of this report

To allow this vision, a pan-European AIP format has to have the following abstract characteristics:

(a) It should be *physically autonomous*.

That is, it must be possible to transfer the AIP from one repository into another one transparently, without changing it. While it is not necessary, for some repository systems it may be useful to convert the AIP into the DIP format of the new repository during transfer. This allows the new repository to extract the metadata required to implement its own access functionalities (the presentation of the content of the stored IP within

[2] Preservica Standard Workflows 5.0, 22-May-2013. P. 2-8.

a life system, that is). However, this is a concern of the target repository, not a requirement on the AIP format. If only a single AIP, without any other component of the repository system in which it is residing, survives whatever is responsible for the loss of those components, it must still be possible to extract its content and prepare it as a DIP for processing by current software. "Any other component of the repository system" in this definition includes all software components which are either proprietary to a specific vendor of repository systems or open source but only used by small communities, since in either case their future support is not assured.

(b) It should be *logically autonomous.*

That is, if only a single AIP, without any other component of the repository system in which it is currently residing, survives, it must still be possible to interpret its content for inclusion into current information systems. Each AIP must therefore connect the data it encapsulates inseparably with the all metadata which are needed to interpret these data. Interpreting data in this sense means that a human operator needs to be able to fully understand the context and interpret the content accordingly.

The following examination of candidate formats for a pan-European AIP format addresses therefore only solutions which are independent from any vendor specific system. The focus is mainly on the logical structure of the AIP. The decision that the individual files have to be packed into a physical container which can be processed by widely available software components, independent of any vendor, radically reduces the number of physical archiving formats we have to look at.

(c) Side effects.

As we are discussing an AIP which conceptually covers all sorts of digital data, we discuss here the format of an AIP primarily independent of the data and metadata formats, which are contained within it. (Data and metadata formats may be referred to in order to clarify some point. Data and metadata may also be mentioned, when they occur within a third-party document, which is summarized here.) As a side effect, the SIARD format, which is a clear candidate for the encoding of data from relational databases within the AIP, is not discussed in this document. It is a data format, which we expect to use within the data section of an AIP, it is not an AIP format itself.

(d) Specific criteria.

Section 2 - *AIP: concepts and implementations* - starts with a list of the categories, which have been used to compare the existing proposals for AIP formats, which we have looked at in detail. These categories are specified there as functional areas, which are used to structure the comparative descriptions. Below we list the requirements which an AIP must fulfil in these categories, derived from the vision described above and the abstract characteristics in items (a) and (b) of this section.

A pan-European AIP format (and the tools creating an object in this format) *must* provide:

- Technical controls to check for the integrity and authenticity of the IP.
- A general mechanism for storing metadata which is generic enough to be extended easily by additional metadata schemas as they become defined within existing or additional user communities of the AIPs. "Storing" metadata implies the capability of checking the consistency of

data/metadata relationships within an SIP.

- A general mechanism for storing data, which is generic enough to be extended easily to new data formats. “Storing” data implies the capability to identify and validate these additional formats.
- A mechanism to create an identity for an AIP, which allows it to be addressed unambiguously over long periods.
- A mechanism to administer relationships between different AIPs.
- A protocol covering all processing that occurred to the AIP since its creation, including a mechanism to extend such a protocol when further processing occurs during the lifetime of an AIP.
- A versioning mechanism, which combines the requirement of an AIP being immutable with the requirement to add additional information about its content after creation.

These criteria are explicitly used to describe existing solutions. Beyond them, the following criteria have been used tacitly, as reflecting the state of the discussion:

- All sub-formats used have to be openly accessible and publicly maintained.
- It must be possible to recover the content of an AIP after disasters relying exclusively upon open source, community supported tools which are widely used and fully documented.

1. Approach: Quantitative survey and qualitative interviews

This report on the state of the art on AIP formats is submitted at the same time as the deliverables D3.1 and D5.1, which report on the same for SIP and DIP formats. The work on SIP and DIP formats has been directed by an attempt to get an overview of individual technical solutions, collected by very intensive collections of information from individual institutions. For SIPs and DIPs this cannot be avoided, as these form the “interface” between currently existing technical solutions and the formats to be proposed by E-ARK. The vision described above shows the AIPs in a completely different position: the software tools needed to work on AIPs interface not with any of the existing SIP and DIP formats, but with the pan-European SIP and DIP formats, to be designed by work packages 3 and 5.

The purpose of this deliverable has therefore not been to look at individual solutions of technology providers, but at the general principles for AIP construction currently available, a study which has been primarily based upon descriptions and discussions of AIP formats provided outside of individual projects.

Nevertheless, work package 4 has been participating at a very low level of intensity in a best practice study on all types of formats and tools, which has been primarily driven by work packages 3 and 5, across the structure of tasks within E-ARK’s description of work. For the present report, the information derived from this survey has been ancillary. Readers interested in the survey are strongly encouraged, to read the deliverables D3.1 and D5.1. This report is *not* sufficient to appreciate the valuable work provided by work packages 3 and 5 within this cross task survey. Abbreviated information on this survey is, nevertheless, attached to this deliverable in the form of three appendices.

The systematic reasons, why this survey has been less important for this deliverable, have been given above. Unfortunately, there are also empirical reasons, why we have found it less useful, than it has been for the deliverables D3.1 and D5.1. As mentioned, what we have really needed for this report has been concrete and detailed technical information on the *structure* of AIPs. The cross task best practice survey is a quantitative online survey with more than one hundred participants and a series of extended individual interviews, mainly via Skype, have been implemented. As we needed general structural information, only two questions regarding AIPs have been inserted into the survey and general questions into the interview. (Cf. The three appendices on the cross task survey to this document.) However, the feedback received has given us valuable access to information on some of the AIP formats used, frequently by URLs of technical documents. The discussion in section 2.7 below rests heavily on this feedback.

But such precise information has been provided only rarely. To illustrate the problem: Q.18 for archives/ Q.14 for private organizations/ Q.14 for government organizations/ Q.14 for private companies/ Q.14 for "others" reads: *Please, briefly describe the submission and archival information packages formats used in your organisation or supported by your solution(s).* This question was answered 34 times and skipped 159 times. This information may have been considered confidential in some cases, e.g., when a service provider developed the format used under some contract which assured that it could be re-used for other projects later and was intellectual property of the service provider.

It seems to be more plausible, however, that this information was not, or not easily, accessible to the person completing the survey. The 34 answers received – there was a text field where participants were asked to provide their answers – supports this interpretation. Some answers are: *"XML and pdf-a", „Not defined in detail as yet. But for next two years incoming data will be stored in a MS Windows folder structure."*

A related question – *"Please, briefly describe the submission and archival information packages formats used in your organisation or supported by your solution(s) and provide a URL link."* – resulted in answers like *"METS", "PREMIS", "XML", "PDF/A"* as well.

These respondents were apparently unaware of the difference between an individual data or metadata format and an information package. We assume this indicates that they are only vaguely aware of the underlying technical problems.

2. AIP: concepts and implementations

In the following, various AIP formats will be described. Those formats cover on the one hand purely theoretical AIP designs, and on the other hand actual implementations in live systems. We have drawn particularly valuable insights about the latter ones from the qualitative interviews. This report does of course not cover every single implementation of the AIP concept, nor does it cover every theoretical design: the ones described here were picked for their influence on other systems, because they are widely known and used or because of how detailed the available format descriptions are.

To allow for a more qualified comparison between the various formats, the description of every AIP concept will be structured in a way that it is clear whether certain requirements for the "perfect AIP format" are met, and if so how they are fulfilled. The aspects covered are:

- Security. How is authenticity and integrity guaranteed?
- Metadata. Which metadata formats are used and how are they stored?
- Data. How is data stored within the AIP?
- Identity. How is an AIP identified?
- Relationships. How are relationships between AIPs described?
- History. What happened to the AIP after its creation?
- Versioning. Is it possible to have different versions of the same AIP (object)? If yes, how is it realised?

Note that the available AIP descriptions vary very much in terms of how much detail they provide and how they weight the importance of certain aspects.

Problems of the physical storage layer – properties of different storage devices, physical distribution across multiple locations and the like – are not considered in this report.

2.1 US Patent 13/219,630 Method And System For Preparing Digital Information For Long-Term Preservation[3]

This US patent describes a system and a method for

> *"creating or extending a preservation-ready digital document. This document is represented so as to be durably intelligible and reliably trustworthy. It includes within itself standardized metadata, provenance information, and reliable links to chosen documents within a World-Wide network of digital repositories. These links and the documents' own identifier(s) are chosen to uniquely, unambiguously, and forever identify what they refer to. This system provides a robustly durable method of preserving an unbounded number of digital objects for as long as their representing bit-strings are kept in existence and findable by now-conventional digital library technology […]. The overall system herein described provides this service without requiring that pre-existing software be modified, and without requiring that any information object that it is intended to protect be modified from what its declared authors, editors, and producers created and provided as input for this preservation packaging service."*[4]

It is worth noting that the concept presented here is not meant to replace existing formats or approaches for long-term preservation, but to be compatible with them and provide valuable enhancement.[5] Furthermore, the author does not want his work to be thought of being applicable for archiving only, as he states that "LDP and archiving [are] distinct topics that overlap, that can be coupled to please clients, and that can be combined to provide services that transcend what is possible with either alone."[6]

The introductory paragraph is a relatively accurate description of what an ideal AIP format (the author uses the term PIP, Preserved Information Package; the TDO – Trustworthy Digital Object – is a particular PIP

[3] Available here: http://www.google.com/patents/US20130054607 22/05/2014. Further referred to as "Gladney".

[4] Gladney, p. 1.

[5] Gladney, paragraph 0031

[6] Gladney, paragraph 0040

design)[7] should be capable of, namely storing digital objects and related metadata for an unlimited amount of time (limited only by physical storage). "Storing" does not exclusively mean to ensure an intact bitstream, but to ensure that the stored objects are reliably useful in the future. This includes authenticity, integrity and that the stored object can be accessed, despite the fact that the future systems used for this purpose are not known today, as well as someone in the future might not have any prior knowledge about said object. As stated by the author: "Each TDO is reliably interpretable and verifiably authentic whether or not any authority for its content is available to answer questions and/ or testify about the information conveyed within this package."[8]

To be more detailed on the challenges mentioned above, Gladney lists what a TDO should be capable of:

"[Each TDO provides essential parts of #2, #3, #4, #5. #6, #7 and #8.]

1. *Ensuring that a copy of every preserved document survives as long as it might interest someone;*
2. *Ensuring that authorized customers can find and use any preserved document as its producers intended, avoiding errors introduced by third parties that include archivists, editors and programmers;*
3. *Ensuring that any customer has ready access to means for deciding whether information received is sufficiently trustworthy for his intended application;*
4. *Hiding information technology complexity from end users (producers, archivists, and consumers);*
5. *Replacing human effort by automatic procedures whenever doing so is feasible;*
6. *Empowering authors, editors, and other information producers to package information so as to relieve overloading of professional cataloguers;*
7. *Enabling gradual migration of huge numbers of current digital holdings to preservation-worthy representations and packaging; and*
8. *Enabling effortless interchange of packaged information among institutions and individuals."* [9]

In the following paragraphs, Gladney continues with the description of an editor that should be used for the creation of a TDO. Since this is not really needed for the understanding of the TDO concept and structure, this report will continue with describing how AIP structures are resembled and requirements are met by the author's invention.

Every TDO conforms to the following data schema: Fig. 2 depicts a (simple) TDO, Fig. 3 a TDO Protection Block and Fig. 4 describes the nesting of TDOs, which can be compared to the concept of AICs (AIC – Archival Information Collection). Also, a TDO does not have to have payload data, it could "be useful to describe pending payload, or to relate information in other digital objects, or to describe information that might never be contained in digital objects."[10]

[7] Gladney, paragraph 0011
[8] Gladney, paragraph 0013
[9] Gladney, paragraph 0022-0030
[10] Gladney, paragraph 0077

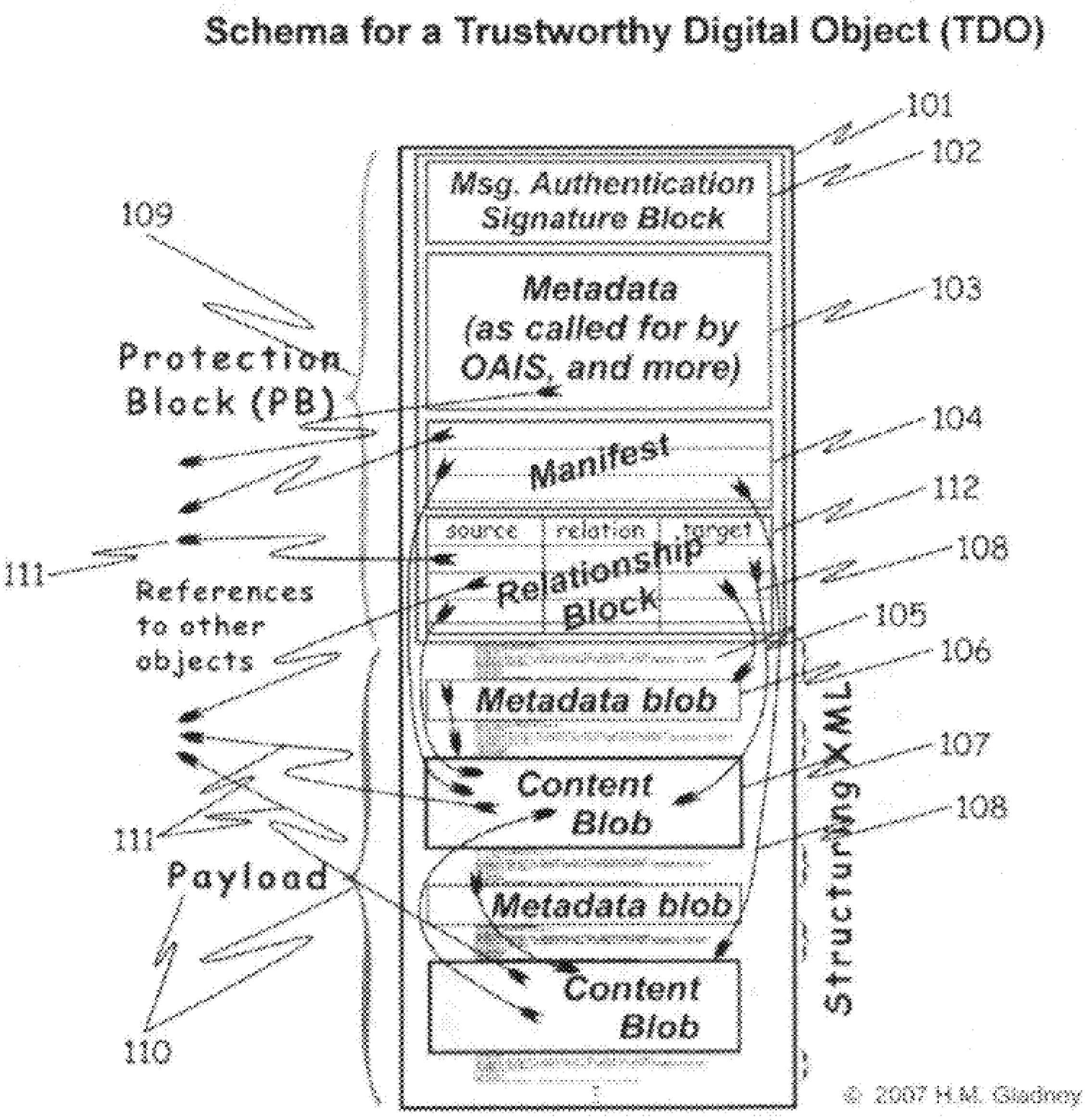

Figure 2: Schema of a TDO

Figure 2 depicts the following: 102 Message Authentication Signature Block, 103 metadata, 105 XML as "glue language" for TDO components, 106 Metadata blob (content depends on type of data being described), 107 Content Blob (might be accompanied by 106), 108 (current TDO) and 111: arrows indicate a link (program pointer, bibliographic citation, WWW link, ambiguous literal descriptor); could also refer to real-world objects (geographic location, ISBN number...), 109 Protection Block (more detailed in Fig. 3; every other block may be missing or empty), 110 Payload (any number of 107), 112 Relationship Block (table, representing a mathematical relation; used to add new bookmarks to a blob without destroying authenticity).

Security

The Protection Block (PB) is mandatory for a proper TDO[11] and provides metadata and various other information needed to identify the TDO and possible relations with other TDOs, as well as a description of the payload content and more optional content (201-209)[12]:

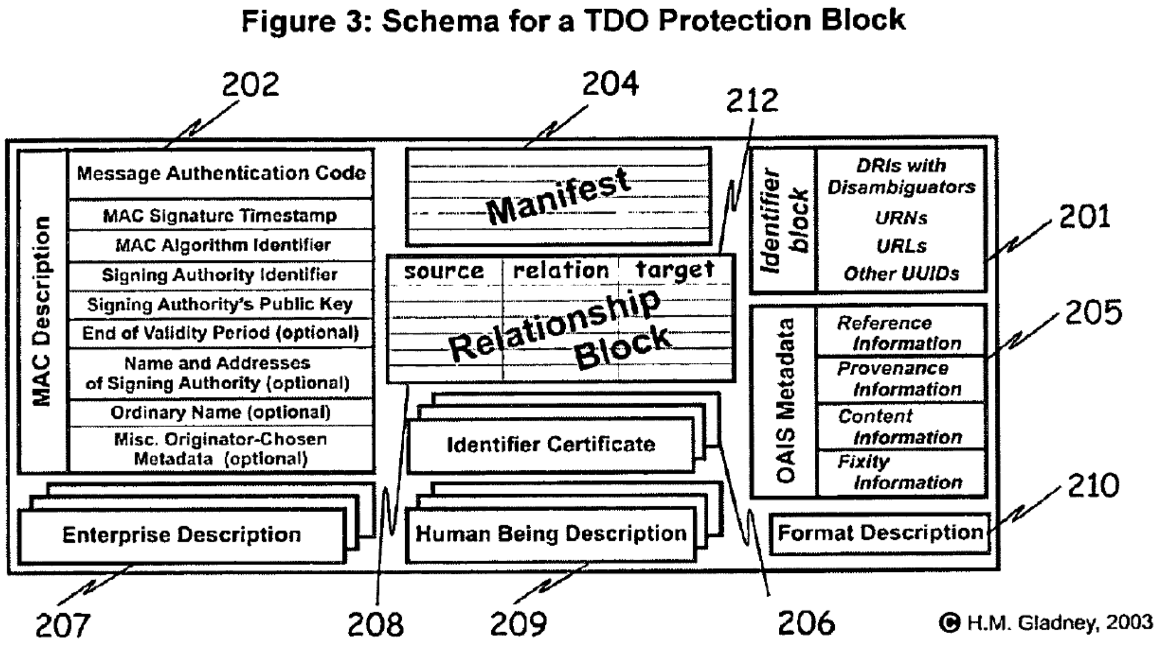

Figure 3: TDO Protection Block

Figure 3 shows a schema of a Protection Block: 201 Identifier block (at least one UUID, any number of additional identifiers), 202 MAC Description (digital signature information for cryptographic authenticity protection), 204 Manifest (table identifying every content and metadata blob within TDO payload), 205 OAIS Metadata block, 206 Identifier Certificate, 207 Enterprise Description, 208 Relationship Block, 209 Human Being Description, 210 Format Description.

What is depicted here as MAC Description is essentially used to "seal" the content with cryptographic certificates that are recursively chain rooted in a trustworthy institution[13]. There is no specific information on how this should be achieved, only that it "holds digital signature information required for cryptographic authenticity protection."[14]

Metadata

The whole Protection Block can be viewed as an aggregation of metadata, describing every aspect of the TDO. This ranges from OAIS metadata that is needed for preservation, linking with other TDOs, description of the payload content, and so on. Additionally, the Payload contains descriptive metadata linked to the

[11] Gladney, paragraph 0089
[12] Gladney, paragraph 0088
[13] Gladney, paragraph 0047-0048
[14] Gladney, paragraph 0091

data. There are no limitations on metadata formats used.

Data

Data is stored inside the Payload in an unlimited number of content blobs (bitstrings); this number can be zero. Gladney does not iterate on recommended file formats, only that each of those content blobs "might be accompanied by a Metadata blob (106)."[15]

Identity

Within the Protection Block, Gladney puts an Identifier Block (201). At least one UUID has to be assigned, and he suggests more options like URN, URL, DRIs.

Relationships

Inside a TDOs Protection Block, the Relationship Block is – as suggested by the name – responsible for keeping track of the relations of the current TDO. This includes on the one hand relations between two or more TDOs, and on the other hand a relation between the current TDO and a real-world object. This real-world object could be identified via geographical coordinates, an ISBN number if it is a book and likewise for other object types.

History

No clear information regarding how to handle information about the history of a given TDO is presented by the author; however, this information would be included in the Protection Block.

Versioning

Gladney describes a scenario where nested TDOs will be useful: "Some book author creates his masterwork and, wanting to preserve this so that his eventual biographers would know his words precisely, creates an Author's TDO, sending this object to a copy editor. [...]"[16] Nesting of TDOs is thus just versioning, where the most recent TDO version always contains the predecessors or the original one.

[15] Gladney, paragraph 0076

[16] Gladney, paragraph 0098

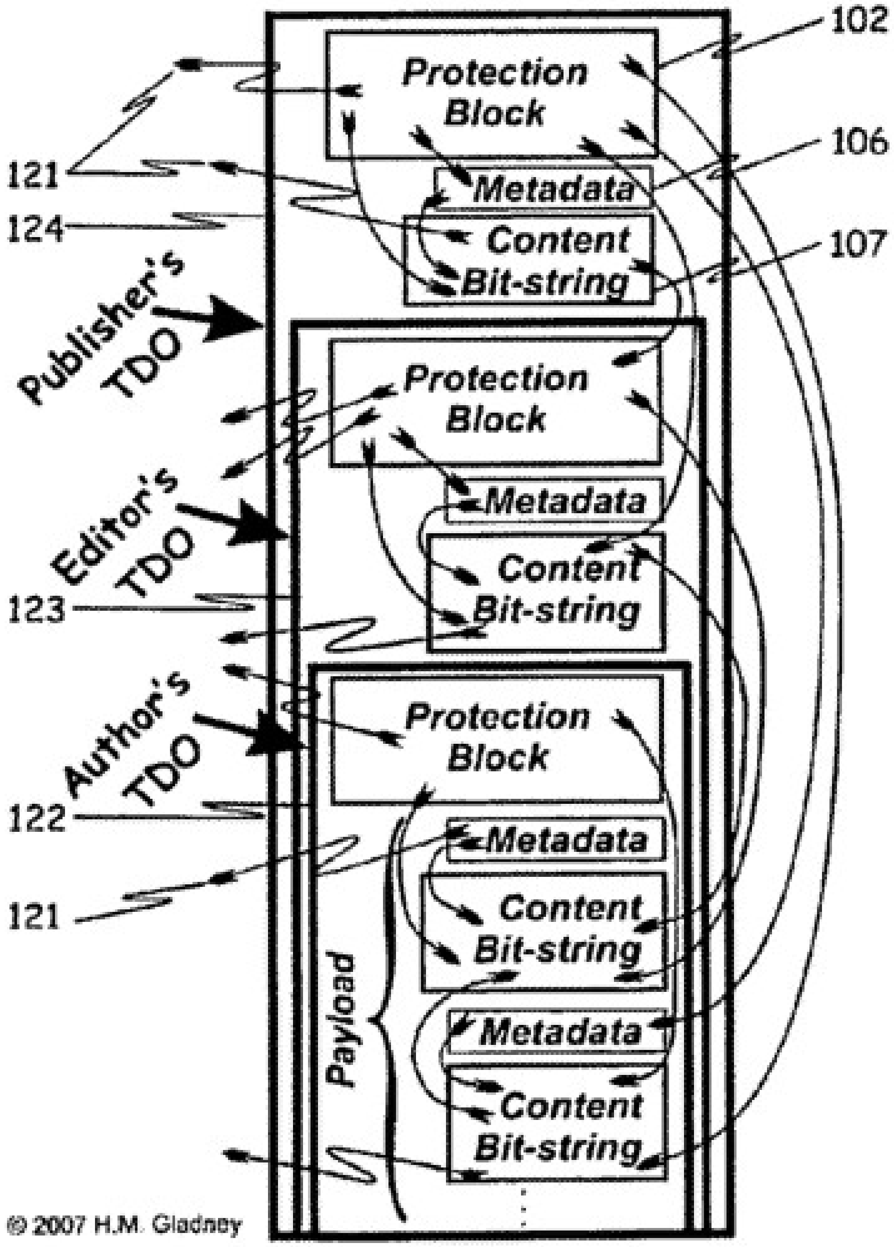

Figure 4: Schema of nested TDOs

Figure 4 shows how TDOs should be nested: 121 linking of external objects (can be an alternative to nesting TDOs, just link the edited versions with each other), 122 Author's TDO, 123 human copy editor (made improvements to certain paragraphs; those are linked to the original version), 124 publisher TDO version (changes to layout etc.).

Gladney also describes another concept of extending a TDO, but this does not suit the requirements for an AIP format as envisioned by the E-ARK project since extending the TDO would require some changes to the data inside the TDO.

2.2 BSI TR-03125 Preservation of Evidence of Cryptographically Signed Documents, Annex TR-ESOR-F: Formats and Protocols[17]

The BSI ([German] Federal Office for Information Security) covers several tasks: "[It] investigates security risks associated with the use of IT and develops preventive security measures. It provides information on risks and threats relating to the use of information technology and seeks out appropriate solutions."[18] The guidelines published by the BSI do not equate to federal law; however, if they are recommended to other federal or state agencies (like TR-03125 is), it is mandatory to comply with them unless there are important reasons for not doing so.

However, it is not mandatory to implement this guideline (in this case, the XSD provided) exactly as it is published. Any implementation is permitted, as long as it fulfils the requirements described in the guideline.

"With the Technical Guideline BSI-TR 03125 "Preservation of Evidence of Cryptographically Signed Documents", the BSI is providing a guide that describes how electronically signed data and documents can be stored in a trustworthy manner in the sense of legally valid preservation of evidence over long periods of time - until the end of the retention periods." The guideline does not intend to replace requirements and definitions in the field of digital archiving but extend those for electronically signed documents. Thus, the BSI TR-93125 is a description of middleware that can be included into digital archives to fulfil those additional requirements.

> "*Concretely, this Technical Guideline describes a differentiated catalogue of obligatory (shall), recommended (should), and optional (can) requirements with regard to all elements and areas in which there is a need to design in order for agencies and institutions to develop effective, sustainable, and economical technical scenarios for the storage of electronically signed documents and data with the preservation of evidence. These are primarily*
>
> - *Recommended data and document formats*
> - *A recommended storage format for archival information packages*
> - *Recommendations for a reference architecture or alternative architectures*
> - *Requirements for components (upstream application systems) and modules (Cryptographic module) as well as their dependencies.*"[19]

This is by far the most detailed concept for an AIP, which we have encountered. The following chapter follows the structure of the above quoted report, but abbreviates very heavily. For an in-depth description of the XML structures which are briefly explained below, the reader is referred to the original document.

A further note: Being a very detailed report on a format which is actually being used, the BSI TR-03125 describes very detailed solutions for the handling of items of information. This might lead to the assumption

[17] https://www.bsi.bund.de/EN/Publications/TechnicalGuidelines/TR03125/BSITR03125.html, 22/05/2014. Annex F: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/TechGuidelines/TG03125/TG-03125AnnexTR-ESOR-F.pdf?__blob=publicationFile, 22/05/2014.

[18] https://www.bsi.bund.de/EN/TheBSI/Functions/functions_node.html, 24/06/2014.

[19] https://www.bsi.bund.de/EN/Publications/TechnicalGuidelines/TR03125/BSITR03125.html, 26/05/2014.

that the described solutions can only be accepted "as is". We have clarified with the authors, however, that other technical bindings between the underlying concepts and a concrete XML-format specification are definitely possible. Following the requirements presented here, which are considered necessary to make a stored document court-worthy, does *not* imply the endorsing of individual coding decisions.

Similarly, we quote the formats recommended in the report to indicate their scope. We do not imply by that, that an AIP endorsing the logical structure recommended here would not be able to include other formats as well.

2.2.1 Definition of the Archival Information Package (XAIP)[20]

"An Archival Information Package, i.e. an electronic document in the sense of this document intended for long-term storage in an electronic archive system, is a self-explanatory and well-formed XML document that can be verified against a valid and authorised XML schema (also called XML formatted Archival Information Package or XAIP for short in the following)."[21] This AIP contains the content data and metadata that is required to track administrative and transaction procedures throughout the whole retention period. Describing the AIP in a valid XML schema ensures that:

- *"The Archival Information Packages can be evaluated for syntactic correctness before submission to the electronic long term storage,*
- *Necessary additions or augmentations to the meta data can be made by expanding or augmenting existing meta data structures and/or including additional XML schema, and*
- *The cryptographic security measures needed for proving the authenticity and integrity of data subject to the duty of retention on account of legal requirements, such as electronic signatures or electronic time stamps can be permanently and reliably linked to the securing data."*

Furthermore, the AIP structure is based on the OAIS reference model, METS, VERS (Victorian Electronic Records Strategy) and XFDU (XML Formatted Data Unit). *"An Archival Information Package […] should have the following data structure:*

- *An archive package header (*packageHeader*) with information about the logical structure(s) of the XAIP document and the sender,*
- *a data section for meta information for description of the transactional and archiving context of the content data (*metaDataSection*),*
- *a data section for the content data (*dataObjectsSection*), and*
- *in the event of the storage of electronically signed documents, a data section for the storage of signatures, certificates, signature verification information, and electronic time stamps (*credentialsSection*)."*

[20] Annex TR-ESOR-F, p.9.
[21] TR-ESOR-F, p. 9.

Security

All files regarding Security are stored inside the credentialsSection; this covers electronic signatures and their associated documents as well as for example files concerning copyright.

> *"The credentials section serves to accommodate* EvidenceRecord *elements pursuant to [RFC4998] / [XMLERS] and other supplemental evidence data such as signatures, time stamps, certificates, and associated status information for each of the content data objects stored in the XAIP document. The relationship between the Evidence Records or the supplemental evidence data and the corresponding payload or meta data is realised with the* relatedObjects *attribute of the Credential element explained below. If no evidence record objects are needed in the entire XAIP, the* CredentialsSection *can be omitted completely."*[22]

If the corresponding AIP is subject to migration, "the CredentialData structure can also be used to store the Evidence Records that belong to the XAIP document. For doing so, the EvidenceRecord data element pursuant to [RFC4998] / [XMLERS] that can accommodate multiple reduced archive time stamps for one payload data object is stipulated." As AIPs "must also be able beyond the original (first) archiving to depict different versions and migrations from one archive to another in a correct and traceable manner", the evidence records "also have to be taken into account along with the XAIP container." Additionally, the old XAIP container can be seen as a payload data object and thus included into the new XAIP container.

Signatures

Commonly used signature formats are mostly ASN-1 or XML based. The following formats are recommended, although it might be required to use other formats that are more application-specific.[23]

- PKCS#7 / CMS / CadES (ASN-1 based):

  "The Cryptographic Message Syntax (CMS) signature format going back to [PKCS#7] pursuant to [RFC5652] is the most commonly used ASN.1 based signature format in practice. Because the data to be signed in this signature format are treated merely as binary objects – without consideration for an internal structure – any data can be signed, but the signatures cannot be embedded into the payload without further ado. [….] CMS based signatures should be verified before the generation of the initial archive time stamp. In doing so, a CMS signature should be supplemented by a so-called CAdES-X Long (CadES-X-L) signature pursuant to [RFC5126, section 4.4.3.1] so that the renewed signature verification with any CAdES conformant signature application component is possible."[24]

- XML Signatures / XadES:

  In contrast to the ones based on CMS, it is possible to sign only certain defined parts of a document when using XML based signatures. *"XML based signatures should be verified before the generation of the initial archive time stamp. In doing so, a XML signature should be supplemented by a so-called XAdES-X Long signature pursuant to [ETSI 101903, Annex B.2] so that the renewed signature verification with any XAdES conformant signature application*

[22] TR-ESOR-F, p. 16.
[23] TR-ESOR-F, p. 55
[24] TR-ESOR-F, p. 55.

> *component is possible.“*[25]

Additionally, the signatures are also verified through certificates, hashtrees and timestamps. The general idea is to combine as many trustworthy mechanisms as possible to guarantee that the linked object stays court-worthy.

Metadata

> *“The meta data section (*metaDataSection*) of an XAIP document is a complex XML data type and shall contain all meta information that is needed for transparent and permanent interpretation of the transaction and archiving context for all content information packages summarized (subsumed) in the Archival Information Package. Multiple meta data packages can be stored in this section, each of which contains the meta information for a separate content data object stored in the Archival Information Package (XAIP document).”*[26]

The recommended file formats are XML and XSD to formalise and/or validate the XML structures.

Data

Data is to be stored inside the DataObjectsSection, either XML or Base64 encoded. The structure of this element is described as following:

> *“The data object section of an XAIP document,* dataObjectsSection*, should have the content data to be archived in the* dataObject *data structure.“ The number of those objects is not limited, and could be used to store data in different formats or entire folders with different files. Each content data object “should be described by the following data elements: Obligatory, a unique identification characteristic (*dataObjectID*) for this section, obligatory, a* contentData *data element for inclusion of the content data, optional, a cryptographic checksum of the content data, and optional, information about any transformations of the payload data object that have been carried out.“*[27]

> *“It shall be ensured for a long-term legally compliant preservation of electronic primary information (content data) that the negotiability and (machine) readability of the stored electronic information can be guaranteed by suitable measures for the duration of the statutory preservation periods at a minimum. In order to ensure the long-term negotiability, solely standardised data formats that are stable over the long term with a public description should be used.”*[28]

The recommended file formats that comply with this requirements are listed below; they are grouped into document (records) files and multimedia files (audio and video). The multimedia file formats contain content data (audio, video, picture, text) and (not necessarily) the relation between those objects.[29]

[25] TR-ESOR-F, p. 56.
[26] TR-ESOR-F, p. 14.
[27] TR-ESOR-F, p. 15.
[28] TR-ESOR-F, p. 44.
[29] TR-ESOR-F, p. 50.

Documents (Records)

1. Text (ASCII): ASCII is a character set and a text format; it does not include layout information. It should be used for simple text like metadata, where the layout is not important. For special characters, different Unicode standards (further developed ASCII) must be used.[30]

2. PDF/A: PDF/A is an established ISO standard for electronic documents and suitable for long-term archiving. It is recommended for static documents.[31] Further information on how to transform documents into PDF/A and related formats is available in the guideline the guideline (p.46 and following).

3. ODF (Open Document Format): The ODF is standardised by OASIS and is an XML based document format for text, spreadsheet calculations, presentations and other office documents. The layout information is separated from the content, so both can be processed individually. It can be used for all (primarily) character orientated documents.[32]

4. TIFF: TIFF allows for lossless saving of graphic information and is therefore recommended, when the graphic information is mandatory for a document. It can be combined with LZW compression and CCITT bitlevel coding.[33]

5. JPEG: JPEG is a compression procedure and a graphic format, most commonly used on the Internet. It is advised to use JPEG only when compromises have to be made between picture quality and file size.[34]

6. PNG: PNG allows for loss-less compression and incremental display of graphics and is therefore very suitable for use on the Internet. It is recommended to be used for figures.[35]

Audio formats

1. Ogg Encapsulation Format: Ogg is a container format for multimedia files that is “free from software patents and unlimited and therefore it is suitable for long-term electronic storage, in particular on account of the detailed format specifications that are available.”[36] With another codec, it can also be used as a recommended video standard.

2. MP4 / MPEG-4 Part 14: MP4 is another container format for multimedia files. As with Ogg, it can also be used as one of the recommended video standards; it is of course an open and manufacturer independent standard.[37]

3. AAC (Advanced Audio Coding): AAC is a standardised format, but with lossy compression. “It is treated as an improved successor to MP3 because the quality is usually higher with the same bit-

[30] TR-ESOR-F, p. 45.
[31] TR-ESOR-F, p. 46.
[32] TR-ESOR-F, p. 48/49.
[33] TR-ESOR-F, p. 49.
[34] TR-ESOR-F, p. 49.
[35] TR-ESOR-F, p. 50.
[36] TR-ESOR-F, p. 51.
[37] TR-ESOR-F, p. 51/52.

rate."[38]

4. EBU Broadcast Wave Format (BWF): This is a format of the European Broadcasting Union (EBU).

Video formats

1. Ogg: Ogg is a container format for multimedia files that is "free from software patents and unlimited and therefore it is suitable for long-term electronic storage, in particular on account of the detailed format specifications that are available."[39] When used for videos, Theora is recommended as the video codec.[40]

2. MP4 / MPEG-4 Part 14: As said above, MP4 is also recommended for video files in SAGA V 4.0.[41]

Identity

The identity of an AIP – as well as its history – is implemented within the packageHeader section of the XAIP. It should contain information about the whole XAIP, but the one of the two mandatory elements is "an AOID element for storing a unique Archival Information Package ID (AOID) generated by the TR-ESOR-Middleware or the ECM/long-term storage".[42] The full list of elements the packageHeader contains:

1. *"An optional package identifier (*packageID*) for the Archival Information Package that is generated by the archiving business application and can be used to identify the Archival Information Package (e.g. within the business application),*

2. *an AOID element for storing a unique Archival Information Package ID (AOID) generated by the TR-ESOR-Middleware or the ECM/long-term storage,*

3. *an optional* schemaLocation *(such as a URL) with information about the version and the name of the syntactic definition of the XML schema at the basis of the Archival Information Package,*

4. *an optional* packageInfo *field with basic information about the Archival Information Package in text format that makes a future user able to understand the format of the XAIP document and interpret the contents,*

5. *a sequence of* versionManifest *elements in which the contents of the various versions of the Archival Information Package are specified,*

6. *an optional canonicalization element that refers with an URI to the canonicalization algorithm with which the XAIP at hand was normalised and finally*

[38] TR-ESOR-F, p. 52.
[39] TR-ESOR-F, p. 52.
[40] TR-ESOR-F, p.52. See http://www.theora.org, 21/07/2014.
[41] TR-ESOR-F, p. 52.
[42] TR-ESOR-F, p. 10.

7. *an optional extension element that contains additional information if necessary."*[43]

Relationships

In case an XAIP gets updated or modified, the outdated or original version can be included as a payload data object.

History

While a possibility to describe versions of an archival object is provided, we find it difficult to recognize features for a history in the sense PREMIS describes it.

Versioning

As mentioned in the paragraph about identity, the packageHeader section contains another mandatory element. It is "a sequence of versionManifest elements in which the contents of the various versions of the Archival Information Package are specified".[44]

2.3 Archivematica

Archivematica is a free and open-source digital preservation system. There are many more requirements the Archivematica AIP (and the system as a whole) could fulfil, compared to today; since it is an open source project, it has to rely on different partners to provide, add or maintain certain functionalities. A list of features that will be added as well as a wish list can be found on the Archivematica wiki.[45] By design, Archivematica it is able to provide long-term access to digital objects while complying with best practice standards. However, it does not provide an access system itself (AtoM[46] is used for that) or a storage system (although the integration of Arkivum[47] and vice versa was mentioned, as they are only providing bit-stream preservation[48]), which means Archivematica has to be combined with solutions covering those aspects. Here, interoperability for a long time span has to be ensured, otherwise it would render Archivematica's efforts nearly useless. The software itself is built upon a series of micro-services which process the information packages. They can be connected into custom workflows and "distributed to processing clusters for highly scalable configurations."[49] There is an in-depth description of the whole workflow to be found on the website. Here, we will focus on the aspects regarding T4.1, namely the AIP concept and the Archivematica format policy.

General AIP structure

The first thing to notice is the name of the AIP. It consists of two parts, one either being the name of the SIP or a newly assigned one, and the second one being a UUID that was assigned during SIP formation. This UUID is a direct representation of the file structure in the storage system: it is made of so-called "UUID

[43] TR-ESOR-F, p. 10.
[44] TR-ESOR-F, p. 10.
[45] https://www.archivematica.org/wiki/Development_roadmap:_Archivematica#Archivematica_1.4, 27/06/2014.
[46] https://www.accesstomemory.org/en/, 26/05/2014.
[47] http://arkivum.com/, 26/05/2014.
[48] Interviews: Archivematica & Arkivum
[49] https://www.archivematica.org/wiki/Overview, 26/05/2014.

quads", each quad representing a directory, forming a path inside the storage system. The AIP is stored at the end of this path.

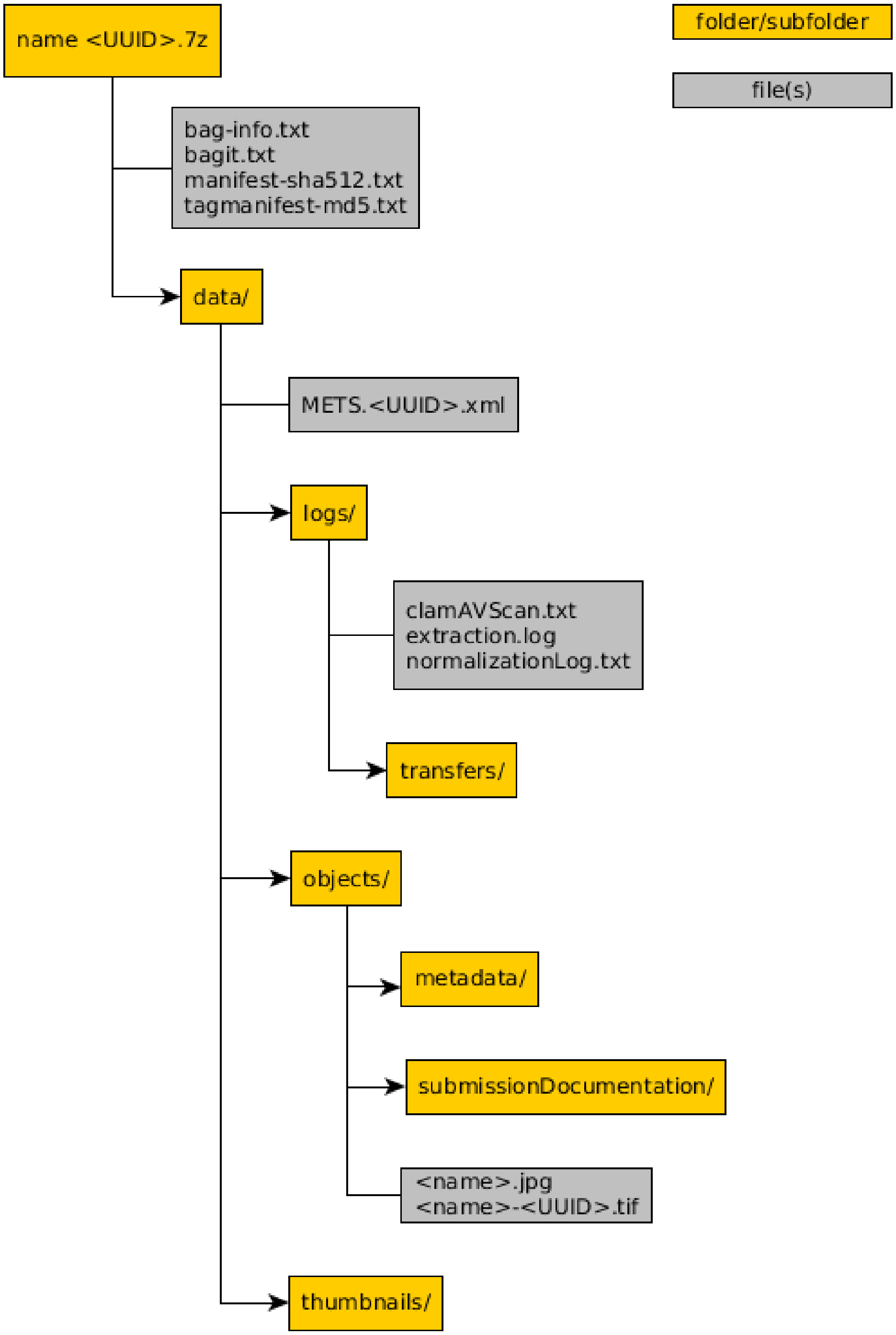

Figure 5: Archivematica AIP structure

On the first level inside the AIP container (format can be decided during the ingest process) the BagIt[50] files can be found as well as a folder called "data". Inside this folder there are three folders – /logs, /objects, /thumbnails – and a METS file. This file contains the full PREMIS implementation; its role is to link original objects to their preservation copies, their descriptions and submission documentation and to the PREMIS metadata to the objects in the AIP.

Inside /log, the log files for normalization, malware scan and extraction (from unpacking packages) are placed. An additional subfolder /transfers contains logs from transferring the SIP from the producer to the archive during the SIP workflow.

The /objects folder contains both original and normalized files from the SIP; the original SIP structure is preserved, i.e. the existing folder structure within the AIP is kept. "/metadata contains metadata that may have been imported with the transfers. /submissionDocumentation contains submission documentation for each transfer which is part of the SIP and each transfer's METS.xml file. The structmap for the transfer is the closest approximation of original order for the transfer."

Additionally, Archivematica also supports creation and management of AICs. An AIC in this system consists of "any number of related AIPs and a METS file containing a fileSec and a logical structMap listing all the related AIPs. In storage, a pointer.xml file gives storage and compression information for each AIC METS file and each AIP."

Security

Archivematica packs their AIPs according to the BagIt specification, and therefore includes the obligatory checksums. Additionally, logs created during AV scan, normalization, extraction and transfers are stored within the AIP.[51]

Metadata

As already mentioned, metadata according to BagIt is used. Further, METS is used to link original files with preservation copies, file descriptions and submission documentation. PREMIS metadata is also linked to the corresponding object. Additional metadata can be imported and stored as well.

Data

As already mentioned, Archivematica keeps the original files, but the standard workflow is to normalize them to preservation- and access-ready formats. This is handled by the before-mentioned micro-services, which can be adapted as needed. The preservation formats are all open source; the choice for a specific file format is based upon the availability of open-source normalization tools, best practices and "significant characteristics"[52] for each media type. Every normalized file is saved as a combination of the original files'

[50] BagIt http://www.digitalpreservation.gov/documents/bagitspec.pdf (Library of Congress)
[51] https://www.archivematica.org/wiki/AIP_structure, 27/06/2014.
[52] https://www.archivematica.org/wiki/Significant_characteristics, 26/05/2014.

name and the UUID of the AIP: <original-name>-<UUID>.format.[53] The preservation strategy not only includes normalizing to a preservation format upon ingest, but also into an access format.[54]

As already mentioned, Archivematica has a so-called format policy, that is a guideline for which formats are suitable for long-term storage and which tools are used to normalize. In the following, we will list which formats are used for preservation as well as the tools used for normalization. Some formats are not normalized; also, for some normalization processes the required tool is not yet identified, these are marked here with an asterix(*). Since Archivematica will update this list as needed, the reader is referred to their website for up-to-date information.[55]

Audio. File formats: AC3, AIFF, MP3, WAV, WMA. Preservation format: WAVE (LPCM), tool used: FFmpeg.

Email. File formats: PST. Preservation format: MBOX, tool used: readpst. In development: Maildir (stored in original format).

Office Open XML. File formats: DOCX, PPTX, XLSX. Preservation format: original.

Plain text. File formats: TXT. Preservation format: original.

Portable Document Format. File formats: PDF. Preservation format: PDF/A, tool used: Ghostscript.

Presentation files. File formats: PPT. Preservation format: original.

Raster images. File formats: BMP, GIF, JPG, PCT, PSD, TIFF, TGA. Preservation format: Uncompressed TIFF, tool used: ImageMagick. Kept as original: JP2, PNG.

Raw camera files/Digital Negative format (in development). File formats: 3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F. Preservation format: original.

Spreadsheets. File formats: XLS. Preservation format: original.

Vector images. File formats: AI, EPS, SVG. Preservation format: SVG, tool used: Inkscape.

Video. File formats: AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, SWF, WMV. Preservation format: FFV1/LPCM in MKV, tool used: FFmpeg.

Word processing files. File formats: DOC, WPD, RTF. Preservation format: ODF* (for WPD and RTF), original (DOC).

Identity

The identity of an AIP is assigned during SIP formation and consists of the name of the SIP(s) and a UUID (<name>-<UUID>, i.e. Pictures_of_my_cat-aebbfc44-9f2e-4351-bcfb-bb80d4914112). This UUID serves – besides the identification – a second purpose:

> *"The AIP directories are broken down into UUID quad directories* for efficient storage and retrieval. (*UUID quad directories: Some file systems limit the number of items allowed in a directory, Archivematica uses a directory tree structure to store AIPs. The tree is based on the*

[53] https://www.archivematica.org/wiki/AIP_structure 27/06/2014.
[54] https://www.archivematica.org/wiki/Media_type_preservation_plans, 03/07/2014.
[55] https://www.archivematica.org/wiki/Media_type_preservation_plans, 03/07/2014.

> *AIP UUIDs. The UUID is broken down into manageable 4 character pieces, or "UUID quads", each quad representing a directory. The first four characters (UUID quad) of the AIP UUID will compose a sub directory of the AIP storage. The second UUID quad will be the name of a sub directory of the first, and so on and so forth, until the last four characters (last UUID Quad) create the leaf of the AIP store directory tree, and the AIP with that UUID resides in that directory.)"*[56]

Relationships

Within Archivematica, it is possible to combine multiple AIPs through an AIC. Such an AIC consists of "any number of related AIPs and a METS file containing a fileSec and a logical structMap listing all related AIPs."[57]

History

The history of an AIP can be viewed by checking the metadata stored within the AIP (normalization log, transformation log...).

Versioning

Versioning is not implemented in the current Archivematica release (1.1, released 10/04/2014). However, it is part of the Archivematica roadmap and scheduled for the 1.4 release.[58]

2.4 Norway: Riksarkivet

The Riksarkivet in Norway implemented ESSArch[59] with the AIC (Archival Information Collection) concept suggested in the OAIS model. They also have an approach in the SIP-AIP conversion that is slightly different from most solutions.

First, the one thing worth mentioning in the AIP generation process is that all data the archive receives from the producer is classified as content, this applies even to the metadata the producer might produce; additionally, a checksum for the SIP is generated. This way, the data and metadata are easier to store and the metadata of the SIP will not be changed due to administrative operations during the SIP-AIP conversion. Preserving the original submission as well is part of ensuring authenticity of the stored data.

[56] https://www.archivematica.org/wiki/AIP_structure 27/06/2014.

[57] https://www.archivematica.org/wiki/AIC, 27/06/2014.

[58] https://www.archivematica.org/wiki/Development_roadmap:_Archivematica#Archivematica_1.4, 27/06/2014.

[59] Cf, http://www.essarch.org and

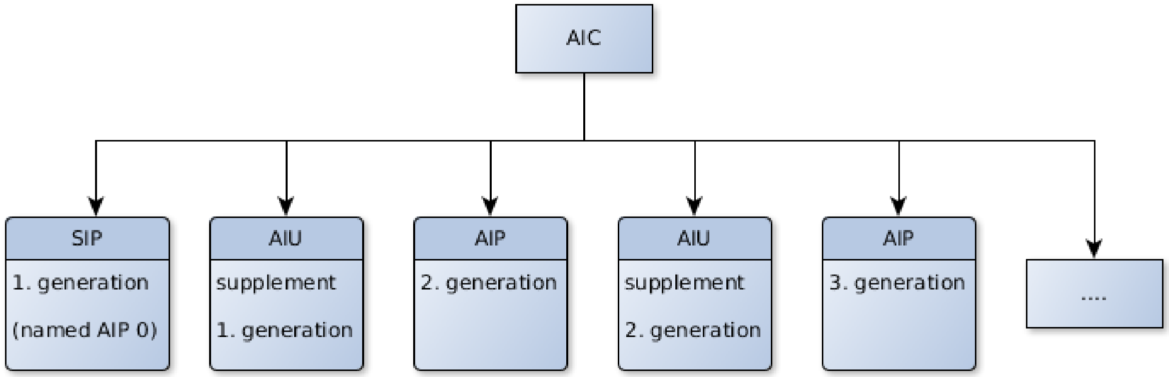

Figure 6: Riksarkivet AIC structure

Second, the implementation of the AIC concept also plays a major role in terms of authenticity. Once an AIP is created it will, of course, never be changed again, not even for metadata editing. Instead, a new AIP will be created that is linked to the AIC; for changes to the metadata, it is planned to create a new AIU. Additionally, the SIP itself (without any changes) will be stored as an AIP (AIP zero); migration to file formats suitable for LTS, preservation metadata etc. will be added in the AIP of the so-called "first generation".

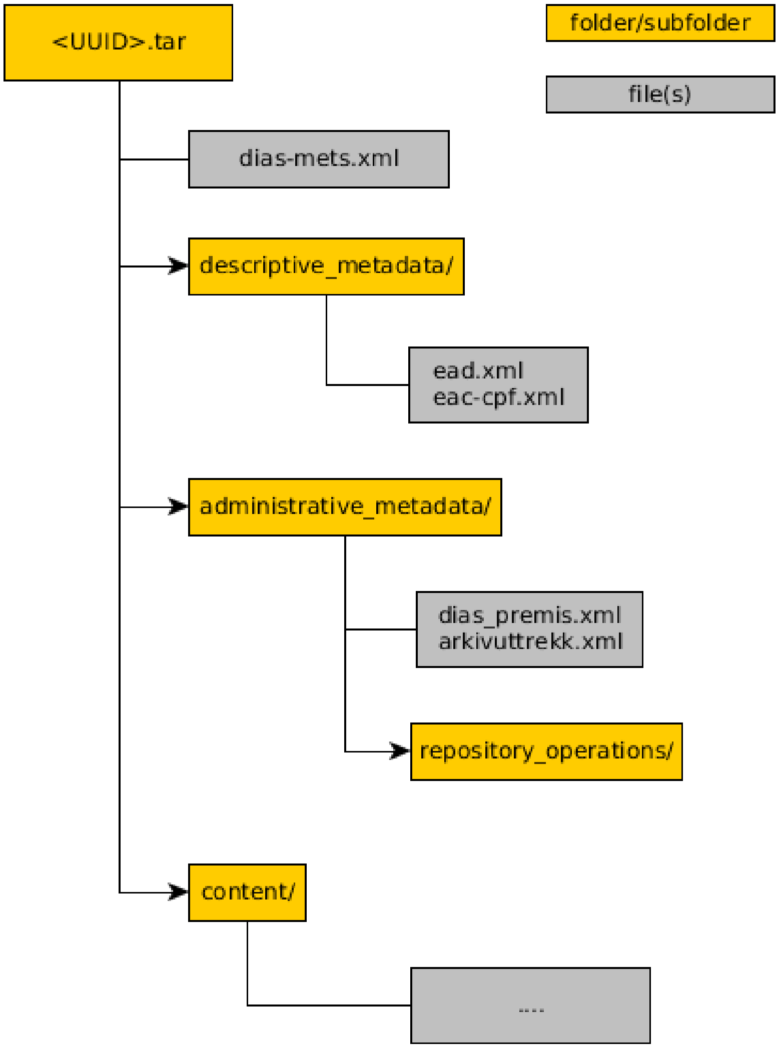

Figure 7: Riksarkivet AIP structure

Security

The integrity of the AIP content is provided through the use of checksums, which are first created upon SIP delivery. Second, the whole SIP is stored within the AIP.

Metadata

There are various types of metadata that are used: METS, PREMIS, EAD, EAC-CPF and ADDML. They are stored as XML files.

Data

Data is stored inside the content/ subfolder.

Identity

Every AIP is identified through the UUID, which is essentially the name of the .tar file.

Relationships

The AIC contains information about all AIP and AIU generations that belong to one object.

History

The files that give information about the history of a specific AIP are stored within the administrative_metadata/ subfolder and provide log files for technical operations performed on the AIP.

Versioning

Versioning is included in the AIC concept, respectively implemented through the AIP/AIU generations.

2.5 DNS

The "Digital Archive of NRW" (DA-NRW) is a project that was "started to develop a system that is able to preserve the cultural heritage data of the German State North Rhine-Westphalia for the long term." It is funded by the State's Ministry for Family, Children, Youth, Culture and Sports. To support it, a functionally complete preservation system, known as "DNS" (DA-NRW software Suite) has been implemented by the chair for Applied Computer Science in the Humanities (HKI) at the University of Cologne. Part of its definition[60] is an AIP format, which we discuss below.

[60] *Das Digitale Archiv NRW in der Praxis. Eine Softwarelösung zur digitalen Langzeitarchivierung,* ed. by M. Thaller, Hamburg, 2013 (= *Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik*, vol. 5)

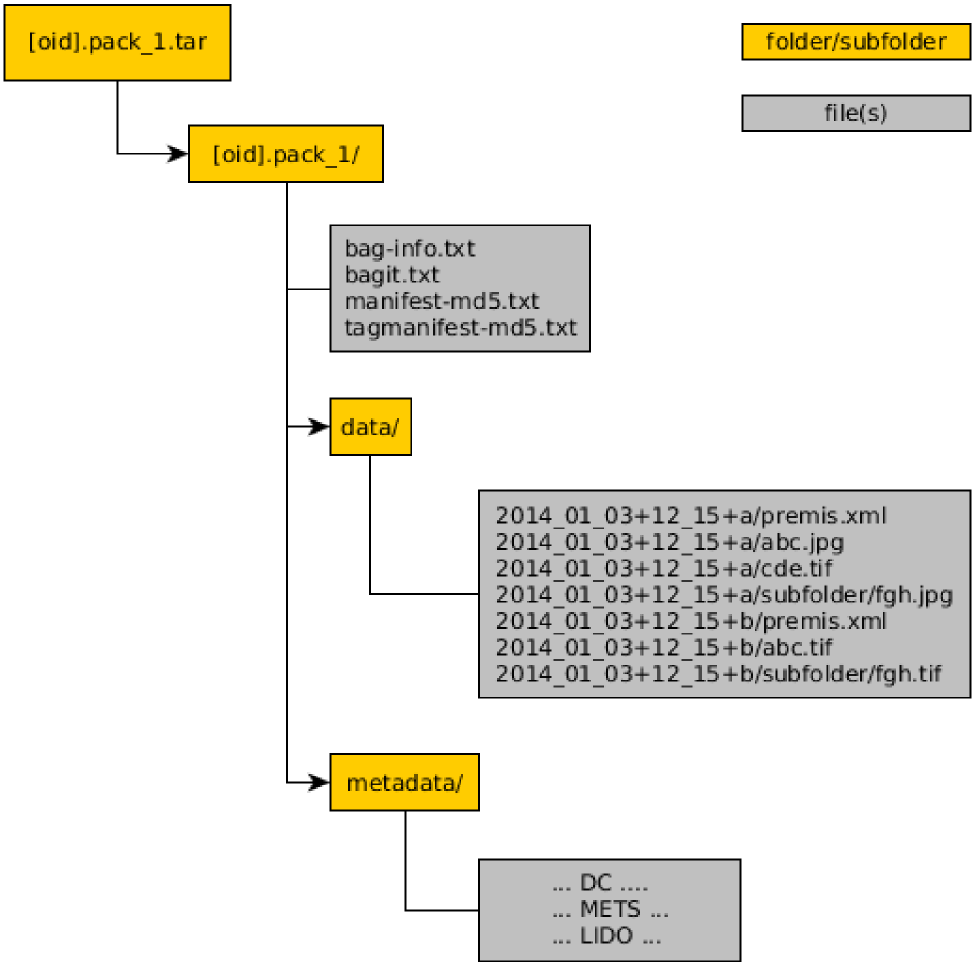
[oid].pack_1.tar
folder/subfolder
file(s)
[oid].pack_1/
bag-info.txt
bagit.txt
manifest-md5.txt
tagmanifest-md5.txt
data/
2014_01_03+12_15+a/premis.xml
2014_01_03+12_15+a/abc.jpg
2014_01_03+12_15+a/cde.tif
2014_01_03+12_15+a/subfolder/fgh.jpg
2014_01_03+12_15+b/premis.xml
2014_01_03+12_15+b/abc.tif
2014_01_03+12_15+b/subfolder/fgh.tif
metadata/
... DC ....
... METS ...
... LIDO ...

Figure 8: DNS AIP structure

Security

The DNS AIP format has no concept regarding security mechanisms besides the checksums required by the BagIt standard. There is no option for digital signatures.

Metadata

Descriptive metadata is stored in a separate subfolder of the AIP; there is no limitation to certain metadata schemes, the one that fits the content best should be used.

Data

The data stored within the AIP is always preserved in its original format, including any hierarchy represented by subfolders. This is called alpha-representation, or a-rep (*<...>+a*). Files that are not already in a format suitable for long-term storage are normalized to suitable formats; again, the original hierarchy will be kept. This is called beta-representation, or b-rep (*<...>+b*) Files that have been ingested in a suitable format are not normalized.

Identity

Every AIP has its own unique identifier, resembled here through [oid] and the .pack_1...n addition.

Relationships

In the version currently implemented, the software does not provide support for the consistent preservation of relationships between AIPs (as might, e.g., be needed to protect against deletion anomalies). The great strength of the immutable object identity in the concept has, however, been derived from the need to support such structures in the future.

History

The premis.xml provides information about the history of the related AIP.

Versioning

The version is visible through the name of the AIP itself, *pack_1* means that this is the first version of the AIP specified through the afore mentioned [oid]. Such a pack is called a "delta"; an AIP can consist of an unlimited amount of deltas (pack_1 … pack_n). A delta contains one or more files that are either an addition of the pack_1 AIP, or if needed edited metadata files. It is not yet implemented, but planned to merge deltas during routine processes. As it is of now, during the DIP creation process all deltas are checked, and if there is more than one version of a single file, only the newest version is extracted from the storage unit.

2.6 Preservica

Preservica, a software vendor that offers solutions for digital preservation,[61] is quite popular among archiving institutions.[62] The software offers functionalities for every archiving step, ranging from ingest and storage to access solutions and including "Active Preservation technology" as well as a security approach.[63]

Since the software is proprietary, it was not possible to obtain the very details of Preservica's approach to archiving; nonetheless, we were provided with some information regarding the logical structures of the AIP and the workflows used, so giving an overview on how Preservica works is possible. Since we do not have access to the whole documentation, some information provided here might be vague and/or outdated. As the related documents are not available to the public, the reader cannot access them for further information and they are referenced for completeness only.

First, there is no clear differentiation between SIP, AIP and DIP. The specification of a XML-bound IP, or XIP, is used for all three of them, with some minor changes depending on context.[64] Considerations on AIPs happen at the conceptual level only, reflecting needs of individual customers. We found no directly derived specifications for an actual AIP format focusing on its long term survival. Additionally, an AIP can be adapted to the users' needs; which makes it even more difficult to describe a "Preservica AIP", since it does not exist

[61] http://preservica.com/, 02/07/2014.

[62] http://preservica.com/customers/, 02/07/2014.

[63] http://preservica.com/preservica-works/, 02/07/2014.

[64] Preservica: Information Package Structure Definition 5.0, 31-Oct.2012. p. 9.

in a definite structure. In the following, we will thus describe how the general structure of an AIP within Preservica looks like and how it complies with the requirements described above.

Security

The only security measures that are in-built are checksums.[65] Adding for example digital signatures and related mechanics might be possible, but this is unknown to the author of this report. Besides that, integrity checks are performed by background jobs within the storage system that replace damaged files.

Metadata

A Preservica AIP has, according to the available documentation, one metadata.xml file that is stored on the top-level directory of an AIP. This XML file contains information about various events (virus checking etc.), technical metadata about the files within the AIP.[66] Furthermore it contains a complete description of the AIP. This covers (the list should by no means viewed as complete)[67]:

- Is the AIP part of a collection?
- A description of the content.
- Information about the manifestation of the described object (presentation, preservation or container).
- Technical information about the file(s).
- Fixity information: checksum and the algorithm used to create it.

This means that Preservica already stores a lot of information about the AIP and its content; as already mentioned, an AIP can be customized according to the needs of individual archives, institutions or to fulfil other requirements due to circumstances that might be important.

Data

Data is stored within the /content subfolder and follows the file hierarchy within the SIP. Files can be described by additional metadata.

Identity

Every AIP is identified through a UUID that is represented through the name of the root directory.

Relationships

Relationships between AIPs are saved inside the metada.xml file in the top level directory inside the AIP.[68]

History

Preservica tracks every operation that has been applied to an AIP. This information is stored within the metadata.xml on the first level of an AIP container.

[65] Preservica, p. 7.
[66] Preservica, p. 9.
[67] Preservica, p. 3-8.
[68] Preservica, p. 3-4.

Versioning

Preservica stores information about AIP versions inside the metadata.xml file; furthermore, outdated versions of a file are marked, so that certain operations (like normalization checks, export as a DIP) are only performed on the newest version.[69]

2.7 Other approaches

The six approaches towards AIP formats described above, are definitely not the only attempts at defining AIP formats which exist.

Possibly the oldest approach, which at first look could satisfy all or most of the requirements assumed above would be the "XML encapsulation" approach, which embeds binary data and (potentially also binary) metadata into a general XML based structure. This approach has probably been most generically described already in 2005 by Filip Boudrez in *Digital containers for shipment into the future*[70]. This approach has from our point of view *one* shortcoming, which reduces its applicability so seriously, that we have not considered it in detail.

In the encapsulation approach, as defined in the paper of Boudrez, the AIP as a whole is structured by a topmost XML layer, into which binary objects can be embedded. In the approaches we have discussed at greater length, the topmost layer consists of a simple directory tree. That is, as soon as the physical container is opened, all metadata and data components can be processed separately of each other; even if one component – as e.g. an obsolete version of an auxiliary metadata format – cannot be processed, as its definition has been lost, all other meta data components and the data components can still be handled independently. If the physical container has as its outermost layer an XML structure, this XML structure *must* be understood and processed, before any other content is retrieved. It introduces a single point of failure, therefore, which should definitely be avoided.

To clarify a potential source of confusion: "(XML) encapsulation[71]", on the other hand, is also used by the industry to describe the integration of (XML encoded) metadata into a binary object[72]. (And Boutrez himself (Digital Containers, p.5) also discusses under the general heading the encapsulation of metadata into binary data files.)

As we consider the "single point of failure" problem prohibitive, we have not discussed further the following examples:

- The German Bundesarchiv provides a "Datenschema zur Archivierung" under the URL https://www.bundesarchiv.de/imperia/md/content/abteilungen/abtb/bbea/xbarch_1_4_2.xsd. This format seems to allow the inclusion of the data stream into the top level XML description of the AIP

[69] Preservica, p. 3-4.

[70] http://www.expertisecentrumdavid.be/docs/digital_containers.pdf

[71] The term as such is *highly* generic: One may observe that in an early paper of the National Library of Australia OAIS' IP concept itself appears already as the top level example for an encapsulation strategy: http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/20.html

[72] E.g. Oracle with regard to image files: http://docs.oracle.com/cd/B19306_01/appdev.102/b14302/ch_metadata.htm

(in which case it does not meet the “single point of failure” criterion) or the reference to a data stream outside of the AIP (in which case it creates a vulnerability in the concept of “data and metadata inseparably connected within one physical container). The format also defines a specific encoding for technical metadata and metadata for the rendering environment. These we would strongly recommend, to encode in a standard independent from the AIP structure as such – within E-ARK the previous work of UPHEC[73] would be appropriate.

- An Estonian specification of an encapsulation format exists under http://rahvusarhiiv.ra.ee/en/universal-archiving-module. The same problems as in the German case apply.
- The Catalunyan https://www.aoc.cat/content/download/6657/24722/file/estructuraPitMets.pdf, besides raising similar problems as the ones discussed, is so closely connected to METS, that it can from our point of view not be used as a structure encapsulating different metadata concepts together with data.

3. Container formats

So far we have described the structure of the AIPs. There seems to be almost universal agreement that these structures should be stored as individual digital files, which contain all their components.

This chapter will give an overview about the most commonly used formats for containers that are used to store and archive a set of files together. This is to give the reader an idea of what our future recommendations regarding the archiving format will be based on. The three candidate formats have been picked due to their high usage and/or because they are standardised and described in publicly available documents. They will be briefly described here, with a focus on the properties that make them useful (or not useful) for a pan-European AIP.

Besides the implicit requirement of being ubiquitous and well documented, which implies tools exist that support the format for a very long time, we have looked at two characteristics, which are currently supported only by a minority of container formats, though they could be considered of high strategic value for long term preservation.

Integrity check

Integrity checking is often used when storing data to magnetic storage devices, although today a much smaller number of errors resulting from storage or transmission are expected. Some archiving formats contain data to flag these errors which are detected by software used to read those files.

Recovery record

This refers to redundant data embedded in files that are used to detect and correct storage/transmission errors through software.

[73] *The Trustworthy Online Technical Environment Metadata Database: TOTEM*, Delve, J. and Anderson, D. (eds.), 2012, Hamburg: Verlag Dr. Kovac, pp137-164

Looking at different container formats with these requirements as a background, some important observations on important container formats are:

3.1 ACE (.ace)

ACE is a proprietary file format used for data compression and archiving. For Linux and Mac OS X a freeware command-line decompression tool is available; additionally, a freeware decompression DLL is available for MS-Windows based systems. However, this software is freeware and not open-source. Furthermore, there does not seem to be publicly available documentation available for this software. Also worth noticing is that the concrete formats used by distributions of ACE provided by different suppliers (or even different versions provided by one supplier) are not fully compatible. The important attributes of ACE are:[74]

- multi-volume archives;
- integrity check, repair functionality;
- authenticity verification.

3.2 RAR (.rar)

RAR is a proprietary archive file format supporting data compression and error recovery. The decompression algorithm is available under open source conditions.[75] Important characteristics of the .rar archive format are:[76]

- files/archives up to 8,589 billion gigabytes in size are supported; number of files is unlimited;
- recovery record and recovery volumes “allow to reconstruct even physically damaged archives”;
- multi-volume archives are supported.

This format is still under active development; it is available for Linux, Windows, Mac OS X, FreeBSD and even Android.[77] In addition to the decompression algorithm, basic data structures are available; more detailed information on the structures should be available through the UnRAR source code.[78]

3.3 Tape Archive (.tar)

The file format .tar (derived from tape archive) is an early Unix format and was standardised by POSIX.1-1988 and POSIX.1-2001. It is often used for (uncompressed) archiving, as it allows for the storing of many files into one large file while preserving various file system information. It writes data in blocks of 512 bytes to the storage medium; because of the way file size information is stored, archived files can only have a maximum size of eight gigabytes. Since 2001, different implementations of the tar format allow bigger files through the use of base-256 encoding for the file information.

There are several different tar formats, but gnu (and oldgnu, its predecessor) and posix (POSIX.1-2001 specification, most flexible and feature-rich format) are the only ones that allow an unlimited file size and

[74] http://www.winace.com/, 02/07/2014.
[75] http://www.rarlab.com/rar_add.htm, UnRAR for various OS available. 02/07/2014.
[76] http://www.rarlab.com/rar_archiver.htm, 02/07/2014.
[77] http://www.rarlab.com/download.htm, 02/07/2014.
[78] http://www.rarlab.com/technote.htm, 02/07/2014.

do not restrict file names to a certain number of characters.[79]

Operating systems that are based on Unix usually include tools that support the usage of tar files (GNU tar being the default one for most Linux distributions); for Windows OS, third party programs exist that offer these functionalities.

3.4 Preliminary recommendation

The two properties discussed would be very welcome in a physical container format for long term preservation. Due to the deficiencies quoted for *.ace and *.rar, we would recommend still to consider *.tar the primary candidate for physical storage format. (Which in fact seems to be consensus among the majority of existing repositories.) We have not discussed *.zip, as this format is less flexible when it comes to recovery from reading errors or similar problems.[80] More specifically: Programs reading zip files must not scan top-down, but check the central file directory on where a file chunk starts. Since there can be other data between those chunks, scanning is problematic; however, every file inside the zip container should be preceded by a local header describing that file.

It has been recommended to us by the authors of BSI TR-03125 to consider PDF A/381 as a candidate format for a physical AIP container. (PDF A/3 is a container which is used to encapsulate files, in contrast to the “PDF document format”). Highly attractive, as the modelling of an AIP format along the lines of one of the main standards would be, this would contravene the principles which in section 2.7 above lead to the exclusion of XML encapsulated formats. We will revisit this format later in the project, however.

.

4. Legal requirements and restrictions

Although the “documentation of national requirements for authentication of stored documents for legal purposes“ was explicitly mentioned in the DOW as a subject of this deliverable, it is not possible to provide a comprehensive overview. We know that this is not the outcome that was desired when formulating D4.1, but unfortunately it is the case.

The underlying problem here is that the legislation on this aspect of digital archiving in some European countries is either very detailed (see the BSI guideline), very general or simply non-existent. One can however access the different answers to the corresponding survey questions online.[82] Considering this and the fact that technical requirements and possibilities are subject to change, we recommend a different approach on this matter. From the point of view of WP4, it would be best to provide as many security measures as possible, and include them in the pan-European AIP. Additionally, this should be realized by a modular concept so that users can either add more or different security measures, or exclude those that they do not need. Certain baseline requirements should be met regardless, to ensure that even with a lack

[79] http://www.gnu.org/software/tar/manual/tar.html#SEC132, 02/07/2014.

[80] http://www.pkware.com/documents/casestudies/APPNOTE.TXT, 04/07/2014.

[81] ISO 19005-3:2012 Document management -- Electronic document file format for long-term preservation -- Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3)

[82] http://www.cdpa.co.uk/EARK/showquestions.php?Group=All

of official requirements, a certain level of authenticity will be achieved. This should include (by no means excluding other options):

- unique IDs for each AIP
- technical/administrative metadata storage operations that the related AIP was subject to (file migration, change of physical storage unit, adding new content, …) or that restrict certain operations that an AIP can be subject to (because of retention dates, access restrictions, …)
- checksums to guarantee data integrity - the first should be created upon SIP ingest
- a concept to include digital signatures and certificates

How each of these requirements are implemented will be determined during the creation of the actual E-ARK AIP specification; this will require coordination between the WP3, WP4 and WP5. This is mandatory, because some security requirements have to be fulfilled by different parts of the actual software implementation. For example, access restrictions that are written in a metadata file inside the AIP are worthless when they are not considered during DIP creation or can by bypassed in some way. The same is true for administrative operations that need to be documented.

However, when talking about these requirements it is important not to neglect an important point: Considering the timeframe that an AIP will be stored, and that a lot of operations will be supervised and conducted by human operators, all technical methods used today to ensure the authenticity of digital data will be obsolete if someone chooses to bypass them with criminal intent. It is more likely than not that this will be possible no matter how much effort is put into protection today. There is no such thing as safe data.

5. Conclusion: What we have not yet done

We have concentrated completely on the properties of AIPs in general, without considering the peculiarities of AIPs which shall encapsulate databases or content from records management systems. This becomes particularly notable in section 2.2.1 above, when the data formats discussed do not consider databases. There are two reasons for that: The discussions within E-ARK on data formats to be supported, have not yet been concluded. The need to archive multimedia streams, a requirement which is not normally discussed within database archiving projects, has been mentioned repeatedly. Therefore we have looked at a general AIP solution, which can host *all* types of data. On the other hand, SIARD[83] is a *de facto* standard starting point for any discussion of the appropriate format for preserving databases; E-ARK acknowledges that by referring to it in the Description of Work, as a format used for that purpose. As it is a data format, not a container format for multiple object types, it is nevertheless not discussed here. The ideas of Lukas Rosenthaler from Bale that RDF triples are not only useful for an integrated metadata repository for long

[83] "http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en&download=NHzLpZeg7t,lnp6I0NTU042l2Z6ln1ad1IZn4Z2qZpnO2Yuq2Z6gpJCDdIR8fmym162epYbg2c_JjKbNoKSn6A-- " provides a short version of the format definition in English. The full version has originally been written in German, accessible as "http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=fr&download=NHzLpZeg7t,lnp6I0NTU042l2Z6ln1ad1IZn4Z2qZpnO2Yuq2Z6gpJCDeIN,gGym162epYbg2c_JjKbNoKSn6A--" has not been translated into English, though a German version is available.

term preservation[84] but also for archiving the contents of databases, which have been presented at recent conferences, are however not yet published in quotable form. This approach will be considered but cannot yet be discussed in detail.

For the same reason, we have not discussed individual metadata formats here. It is clear that, among the participating archives, many will use METS to store metadata about the objects they intend to store; we have been concerned in this document, however, not with this or any other individual metadata standards, but with a structure into which as many different ones as possible may fit.

[84] Ivan Subotic, Lukas Rosenthaler, Heiko Schuldt, "A Benchmark for RDF-based Metadata Management in Distributed Long-Term Digital Preservation,", in: *IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 2013, pp. 11-16.

Appendix A: Joint Methodology (written by Piret Randmäe)

Note: As mentioned initially, the survey has been much less central for this deliverable, than for the state of the art reports if work packages 3 and 5. As it *has* influenced the findings described above, however, we have included it in an abbreviated form.

Tasks T3.1, T4.1 and T5.1 formed a Cross-task collaborating to analyse current solutions and best practices for Ingest, Archival Storage and Access respectively. This was done to align work, be effective and avoid redundancy but also to ensure that stakeholders were not approached several times by different tasks from the E-ARK project asking for details about their digital arching practices.

The objective of this cross-task work was to build a knowledge base about best practices, tools, requirements and restrictions relevant to archiving solutions. The collected information will feed into the onward work within E-ARK to specify common formats for OAIS Information Packages (SIP, AIP and DIP) and in the work of developing common tools for archival services. We conducted our work through desktop research, an online survey sent to a wide range of stakeholders and a series of qualitative interviews with selected stakeholders. We included following groups into this work:

- Archives (national, municipal and private archives. This group was also intended to include libraries),
- Private companies that have developed archiving services,
- Private organisations that have developed archiving services,
- Projects that have developed archiving services,
- Public organisations (creators of digital content (Producers) and regulatory bodies).

We contacted Organisations throughout Europe, as well as in North America, Australia and New Zealand Our findings gave a unified view of three areas of research, each specified to support work in one of our reports:

- **Ingest.** Best practices for pre-ingest, ingest and ingest tools
- **Archival Storage.** Available formats and restrictions for storage and different national requirements for authentication for legal purposes.
- **Access.** GAPs between requirements for access and current access solutions.

Desktop research

The purpose of the desk top research was to get overall knowledge of current practices and access solutions.

Method

We began with desktop research as an initial stage of our task followed by primary research – quantitative online survey that was afterwards ensued by qualitative interviews. Our desktop research comprised of data collation, gathering overall knowledge from available published resources. That information, reports and publications on similar matters, were then analysed and cross referenced.

Online survey

The purpose of the survey was to gather overall information about the practices for digital archiving from a broad range of stakeholders.

Method

- **Survey type.** Quantitative survey via an online questionnaire with a mix of question types
  - Yes/No questions
  - Multiple choice and comment
  - Choose from list (drop-down)
  - Essay box questions
- **Media.** Online survey using SurveyMonkey. Survey invitation sent out to numerous stakeholders via e-mail.
- **Period.** The initial survey period was from 02-20 April 2014, which was later extended to the beginning of May.
- **Documentation.** Survey question sets and survey results can be found online: http://www.cdpa.co.uk/EARK/showquestions.php?Group=All.

Qualitative research is good at providing information in breadth, from a larger number of units, but if wanted to explore a topic in depth, quantitative methods can be too shallow. Hence we decided to use quantitative data collection on our survey in order to collect data from as many respondents as possible so we could achieve broad-based results. To distinguish best practice used worldwide we then needed to go for in-depth qualitative techniques, in our case qualitative interviews. From collected answers we then chose eleven organisations based on their given answers that interested us especially.

Stakeholders

The survey was sent out to broad range of stakeholders from all five stakeholder groups.

Questions for the survey

The questions for the survey were created considering the needs of each task. We used two level internal quality assurance to ensure that the questions were appropriate, understandable and covered all relevant topics for better end results. Each set of questions was reviewed by other task members in the cross-task group and finally all questions went through quality assurance by E-ARK partners outside our cross-task group.

The questions from the survey can be divided into four categories:

1. General questions about background, legislation and contact information.
2. Questions concerning pre-ingest, ingest and ingest tools.

3. Questions about preserving archival information packages and file formats.
4. Questions about requirements for access and current access solutions.

Construction of survey

There were 94 questions in total in the survey. However not all questions were asked of every respondent. We created targeted questions depending on which stakeholder group the respondent belonged to. There was also a dynamic logic on given answers. For example if *(Q.19) Does your Organisation provide access to digital material?* was answered "Yes" then the survey logic skipped *(Q.20) Why do you not provide access to assets?* and went straight to *(Q.21) Which specific content types do you currently provide access to?*. This was done to ensure that respondents only were asked relevant questions.

Note about libraries in the survey

The intention was to include libraries in the stakeholder group "Archives". However, libraries responding to the survey generally identified themselves as belonging to the category "Other". Since the survey was constructed with individual sets of questions targeted at each stakeholder group, the consequence was that libraries were given a set of questions which was meant for group "Other". Fortunately, as the group "Other" contained all "Archives" questions, no relevant questions (from ingest or storage point of view) were lost. In fact, questions about SIP and AIP formats were asked from both Archives and Libraries.

Qualitative interviews

The purpose of the qualitative in-depth interviews was to gather details about selected (interesting/significant) solutions used worldwide. The answers collected are used for description of best practices and as input for the onward work of WP3, WP4 and accordingly WP5 to create a common SIP and DIP specifications and ingest/access tool(s).

Method

Our method used in qualitative interviews comprised elements from structured as well as semi structured interviews.

- **Interview type.** structured/semi-structured interview
- **Platform.** Media used for conducting the interviews was Skype.
  - and face-to-face in the very few cases when it was possible
  - 4 persons (institutions) answered in writing to our qualitative interview questions
- **Interview period.** Interviews were held throughout May 2014. Interviews lasted on average one hour. The shortest interview was 45 minutes while the longest was about 1h 15 minutes.
- Questions were sent to interviewees beforehand to enable them to familiarise and think about questions before interviews.
- Interviews held on Skype were recorded using an MP3 Skype Recorder. A summary of the interview was written and sent to interviewees for verification afterwards. There were 3 interviewers' roles in our interviews:

- Person who asked questions. Interviewer's mission was to have a conversation with the respondent by asking key questions and other related questions. The exact set of questions depended on the responses of the. The interviewer played a neutral role and didn't give his or her opinion in the interview process.
- Person who took notes. The notes in written form were the primary source for the later analysis. The voice recordings were used for making sense in complicated answers if needed. It was allowed to ask additional questions if the answer was unclear or not detailed enough by the person taking notes.
- Person who monitored and controlled the process. That person started, observed and closed the interview. They were encouraged to interrupt the interview whenever needed to gain and maintain the control over process. This person could also ask follow-up questions if something was left unclear or of particular interest, but the interrupting should not be consistent.
- After a few interviews conducted with the three interviewer's roles it was discovered, that the same work can be done just as efficiently by two interviewers. So the tasks of a person monitoring the overall process of an interview were then divided by person taking notes and person asking most of questions.

In qualitative interviews the interviewees are given space and time to expand and elaborate their answers and experiences that it was not possible to do in survey. Moreover, their answers are not pre-categorised in the interview.

Semi-structured interviewing is more flexible than standardised methods such as the structured survey. Although the interviewer in this technique will have some established topics for investigation, this method allows for the exploration of emergent themes and ideas rather than relying only on concepts and questions defined in advance of the interview. The interviewer would use a standardised interview guide with set questions which will be asked of all respondents. The questions tend to be asked in a similar order and format to make a form of comparison between answers possible. However, there is also scope for pursuing and probing for novel, relevant information, through additional questions often noted as prompts on the schedule. The interviewer frequently has to formulate impromptu questions in order to follow up leads that emerge during the interview.

We created internal and external interview guides to ensure that all relevant topics would be covered and to allow clarification and discussion about interesting aspects. We chose to make detailed internal interview guides with comprehensive questions. Because interviews are carried out in collaboration with T5.1, T3.1 and T4.1 and by making detailed interview guides we ensured that all relevant questions are asked even when persons from that task are not present. In external interview guides we explained shortly the process of the interview and added also questions asked in the interview so that the interviewee can think about the answers and be prepared if needed.

Stakeholders for interview

We used representation and back-tracking for identifying of stakeholders with best/good practices for the interviews.

- Representation: we chose a representative cross section of stakeholders that:
    - Come from different Organisation types (i.e. Archives, Vendors),
    - Hold different data types (both format types and structured/unstructured data),
    - Are subject to different legal requirements (e.g. retention periods, dispensations, confidentiality),
    - Use different strategies/methods (e.g. normalization of data on Ingest, on demand access, offline/online storage, emulation/migration),
    - Use different systems.
- Back-tracking: We identified the stakeholders who provided us the most interesting answers in the quantitative survey and then chose them as interviewees for the qualitative interview. Each task have different interest and criteria for selection of stakeholders and as such not all interviews will be equally relevant for all tasks.

Interview questions

The questions for the interview were created also considering the needs of each task. We used two levels of internal quality assurance just like we did on creating survey questions for better results. Each set of questions was reviewed by other task members in our cross-task group and finally all questions were gone through by members outside our cross-task group.

We carried out pilot interviews with National Archives of Hungary, The Archives of the Republic of Slovenia, National Archives of Norway and the Danish National Archives prior to other interviews to detect any possible problems that might occur; to see if we would fit in the desired one hour time-frame and make sure that all questions are well and universally understood. Also the questions were amended based on feedback from the pilot interviews and they were further refined iteratively throughout the whole interview process based on feedback from interviewees. The full list of questions can be found in Appendix C.

Interview guidelines

The following guidelines were developed to give the best possible conditions for interviews and ensure consistency.

General principles

- All potential respondents should be contacted prior interviews.
- All terms and rules should be introduced during the contact taking process.
- All key questions should be sent beforehand.
- All privacy concerns should be regulated with the legal agreement.
- All prior information about the respondents and their current situation should be clear to all interviewers beforehand.

Questions

- The questions will be created prior to the interview.
- Open ended questions will be allowed. But when open ended questions are used it is a good idea to have a list of topics that should be covered in the question to ensure that the needed information is obtained.
- Questions will be grouped by respondent's type.
- The interviewer will ask each respondent's group the same set of key* questions.
- Ordering and phrasing of the key* questions will be kept consistent from interview to interview.
- All key questions should be easily identified in the questions list.

Establishing the connection and recording the interviews

- Interviewers use Skype even if the respondents use telephone because of the agreed recording functionality and constant quality.
- All conversations will be recorded with the MP3 Skype Recorder tool. If the respondent rejects the recording agreement then the recording should not take a place.
- Recordings will not be shared with third parties.
- All recordings will be deleted latest by the end of 2014.
- Interviewers are aware of possible technical issues with the sound quality, microphone malfunctions, and a lag in the Internet connection speed and have a backup plan prepared in advance.

Things which should be avoided (based on QDATRAINING guidelines)

- Talking over participant
- Interrupting participant (not allowing participant time to finish talking before asking the next question)
- Finishing sentences for participant (putting words in their mouths)
- Asking more than one question at a time (very often, you will only get a response to the last one the participant heard)
- Asking narrow questions (framing the question too narrowly)
- Asking leading questions
- Filling up silences (not giving the participant time to think or expand) which is very common amongst less experienced (and also some very experienced) qualitative interviewers
- Not following the topic guide (not to be confused with not allowing emergent topics) or being consistent across and between interviews in relation to key topics from the topic guide which should have been drawn from the research question itself

- Not allowing interesting and emergent topics to be developed because of a rush to get to the next question or prompt
- Not being courteous enough
- Not having due cognisance where a power relationship exists between the interviewer and participant.
- Arguing with the participant (yes we are serious and have an excellent example in the workshop)
- Being judgemental (we have a wonderful example in the workshop)
- Not signalling when the end of the interview is approaching allowing the participant to say anything they may have on their mind
- Fumbling with equipment and being unfamiliar with the equipment being used
- Failing to record the interview altogether
- Recording in a noisy and distracting environment (only limited control available to the researcher on this one but cognisance is important nevertheless where choices do exist)

Things do before the interview starts

- The leader will state that "With the permission of the interviewee, this interview is being recorded for accuracy purposes only".
- State that that interviewee will receive the written summary from the interview for reference and to correct any mistakes before it is used in the reports
- The leader will introduce the participants.

Appendix B: Quantitative Survey Questions

Only very little information derived from these questions entered this deliverable. The reader is referred to the online source with the combined questions and answers for completeness, however.[85]

Appendix C: Qualitative Interviews Questions

We tried to structure the interview questions according to the OAIS model: ingest – storage – access and modified them in regards to the stakeholder group the interviewee belonged.

Interview questions for archives

The (pre-) ingest of digital objects

1. Steps in pre-ingest process
   - Please describe the usual negotiation process between producer and archive.
   - Please describe the usual records export process and procedures at agencies of what your

[85] http://www.cdpa.co.uk/EARK/showquestions.php?Group=All

archive is aware of.

1. Steps in ingest process
    - Could you briefly describe your usual workflow for digital archiving (including pre-ingest steps)?
    - Could you briefly describe any other more complicated workflows you use in your institution?

The processing and storage of digital objects

1. Maintenance of AIP
    - Please explain how your AIPs are stored. What kind of logical and physical containers do you use?
    - How are your AIPs preserved over time, which strategies do you apply?
    - How do you ensure authenticity (in a legal context) for your stored data?
1. Access to AIP
    - Do you keep track of every access that has been made to a specific AIP while it is in storage (i.e. who access it, when, etc.)?
    - How do you handle restricted access to certain data (and thus to AIPs)?

The accessing of digital objects

1. Data and creation of DIPs
    - What are the typical steps in your workflow when providing access to data?
    - What happens to the DIPs after use?
    - Could you briefly describe the information packages you use in your institution?
1. Dissemination and access
    - Which tools do you use for providing access to your collections?
    - How can users search your collections and find out what data he/she needs? (In other words: how can users find the correct DIP(s))
    - How can the content of one or more DIPs be searched?
    - How can disseminated data be used by users?
    - What access restrictions and requirements must your access service comply with?
    - How does your system handle confidentiality, retention dates, dispensations, user identification/authorization etc.?
1. Users
    - What are the most typical use-cases for your access services?

- What do you know about your end-users' needs?
- How user friendly is your access system in your opinion?
- General
- What would you say are the biggest advantages/weaknesses of your access service?
- What kind of access would you like to offer but are not capable of offering currently?

Interview questions for service providers

The (pre-) ingest of digital objects

- How does your solution support negotiation process between producer and archives?
- Could you briefly describe your customers usual workflow for digital archiving (including supported pre-ingest steps)?
- Could you briefly describe any other more complicated workflows what are supported by your solution?

The processing and storage of digital objects

- Please explain how your AIPs are stored. What kind of physical containers do you recommend?
- Please explain the logical structure of data stored by your system.
- How are your AIPs preserved over time, which strategies can be applied by your solution?
- How do you ensure authenticity in your system?
- Please explain how and on what circumstances your system creates DIPs from AIPs?
- Does your solution keep track of every access that has been made to a specific AIP while it is in storage (e.g. who accessed it, when etc.)?
- How does your solution handle restricted access to certain data (and thus to AIPs)?

The accessing of digital objects

- What are the typical steps in the workflow when providing access to data using your system?
- What typically happens to DIPs after use?
- Are your access service adjusted to your clients' local conditions?
- What functionalities does your access system have? (if possible you are very welcome to support your answer with snapshots of the interfaces in your access system?)
- How users (e.g. a researcher) search collections for the purpose of identifying which IPs contain the specific information he/she wants?
- How can content in one or more DIPs be searched?
- How does your system handle confidentiality, retention dates, dispensations, user

identification/authorization etc.?

- Do you have any knowledge of how end-users typically use your access services?
- What do you know about the needs of the end-users of the access service?
- How user friendly is your access system to end-users in your opinion?

General

- What would you say are the biggest advantages/weaknesses of your access system?