Applied Interoperability in Digital Preservation: Solutions from the E-ARK Project

Kuldar Aas
National Archives of Estonia
J. Liivi 4
Tartu, 50409, Estonia
+372 7387 543
Kuldar.Aas@ra.ee

Andrew Wilson
University of Brighton
CRD, Grand Parade
Brighton, BN2 4AT, UK
+44 (0)1273 641 643
A.Wilson4@Brighton.ac.uk

Janet Delve
University of Brighton
CRD, Grand Parade
Brighton, BN2 4AT, UK
+44 (0)1273 641 620
J.Delve@Brighton.ac.uk

ABSTRACT
This paper describes the interoperability solutions which have been developed in the context of the E-ARK project. The project has, since February 2014, tackled the need for more interoperability and collaboration between preservation organizations. The solutions being developed include harmonized specifications for Submission, Archival and Dissemination Information Packages; and pre-ingest and access workflows. Furthermore, the specifications have been implemented using a range of software tools and piloted in real-life scenarios in various European archival institutions.

This paper provides a statement on the need for interoperability, and an overview of the necessary specifications and tools, and it calls for preservation organizations to continue collaboration beyond the lifetime of the E-ARK project.

Keywords
E-ARK; Interoperability; OAIS; SIP; AIP; DIP; pre-ingest; ingest; access; long-term preservation; digital archives.

1. INTRODUCTION
The adoption of increasingly sophisticated ICT in information creation and management has led to an exponential increase in the amount of data being created by / in a huge variety of tools / environments. Consequently, preservation organizations across the globe also need to ingest, preserve and offer reuse for growing amounts of data as well. When also taking into account the growing financial pressure which many organizations experience, we can conclude that there is a growing need for more efficiency and scalability in digital preservation. In more technical terms, institutions need to develop efficient guidelines and tools to support the export of data and metadata from source systems, produce or reuse metadata for preservation purposes, deliver information to the digital repository, ingest it, and finally provide relevant access services to appropriate end-users.

However, there is no single, widely understood and accepted approach on how valuable information should be transferred to digital repositories, preserved and accessed for the long-term [1]. In practice, existing approaches to archiving the same kinds of information are national or institutional, and differ in regard to their conceptual, technical and administrative underpinnings.

The European Commission has acknowledged the need for more standardized solutions in the area of long-term preservation and access, and has funded the E-ARK project[1] to address the problem. In co-operation with research institutions, national archival services and commercial systems providers, E-ARK is creating and piloting a pan-European methodology for digital archiving, synthesizing existing national and international best practices that will keep digital information authentic and usable over time. The methodology is being implemented in open pilots in various national contexts, using existing, near-to-market tools and services developed by project partners. This approach allows memory institutions and their clients to assess, in an operational context, the suitability of those state-of-the-art technologies.

The range of work being undertaken by E-ARK to achieve this objective is wide-ranging and ambitious, and more extensive than can be adequately described here. Accordingly, this paper will focus mainly on the Information Package specifications provided by the project, and introduce the range of tools which support these specifications.

2. NEED FOR INTEROPERABILITY
As mentioned above it is crucial to have more scalability and efficiency in archival processes. In particular, preservation organizations need to ensure that the data and metadata they receive and to which they offer access is formatted according to common and standardized principles. More specifically interoperability between source, preservation and reuse systems requires that:

- data and metadata are in standardized formats so their subsequent use is not inhibited by system differences;
- the data and metadata, and any other information required to use the data, are combined in a single conceptual package with all components being uniquely identified;
- the package contains enough information to allow validation both before and after transfer to a digital archive;
- the package is constructed in such a way that its information content can be understood in the long term without reference to external systems or standards.

In digital preservation terms this means that we need to come to a common agreement on the core technical and semantic principles of Information Packages (as defined in the OAIS Reference Model [2]). The main benefit of such standardization is that preservation organizations would be enabled to collaborate across institutional and legislative borders more effectively. Additionally, new opportunities would be opened for the reuse of tools which allow, in a standardized manner, the creation, identification, validation and processing of Information Packages. This, in turn, would reduce the effort

[1] http://www.eark-project.eu/

needed to maintain and develop bespoke local tools and ultimately save costs for any individual organization.

The E-ARK project has defined the standardization of Information Packages as its core activity. At the heart of this effort is the generalized E-ARK Common Specification for Information Packages (see 3.1 below). This specification is based on an extensive best-practice review of the available national and institutional specifications [1] and defines a common set of principles for how information being transferred and managed over time should be packaged to support interoperability and long-term access.

However, the Common Specification itself is not sufficient to achieve an adequate level of interoperability. In addition, the specific needs of pre-ingest, ingest, preservation and access processes need to be tackled. Accordingly, the project has also developed further, more detailed, specifications for Submission, Archival and Dissemination Information Packages. All of the specifications are based on the E-ARK Common Specification, but extend it with the specifics of the relevant processes (see 3.2 below).

The E-ARK Common Specification: SIP, AIP and DIP specifications can be called content agnostic as they allow the packaging of any data and metadata. However, to guarantee that the integrity and authenticity of information is not compromised, we need to also consider specific aspects related to the data in question as well as the environment from which it originates. For example, a typical real world records management system contains records arranged into aggregations, metadata relating to records and their relationships to other entities, a business classification scheme, a set of retention and disposal schedules, user access controls and definitions, a search engine and so on. All these data, metadata and environmental components, which make up a specific and complete information package, must be transferred together with the data in a way that the integrity, authenticity and understandability of the whole package are maintained. To allow for interoperability on such a fine-grained level, E-ARK has implemented the concept of Content Information Types (see 3.3 below). The Content Information Types provide a regime for specifying in detail the precise metadata, data, documentation, and system-level issues relevant for a particular type of Content Information, ultimately extending the scope of the Common Specification itself.

3. E-ARK SPECIFICATIONS

In this section we explain some of the details of the E-ARK Information Package specifications which are mentioned above.

3.1 Common Specification for Information Packages

The backbone of archival interoperability in E-ARK is provided by the so-called Common Specification for Information Packages [3]. The OAIS compliant specification is built on the requirements presented above and provides a unified set of rules for packaging any data and metadata into a single conceptual package which can be seamlessly transferred between systems, preserved and reused in the long term. The core of the common specification is a definition of an Information Package structure (Figure 1).

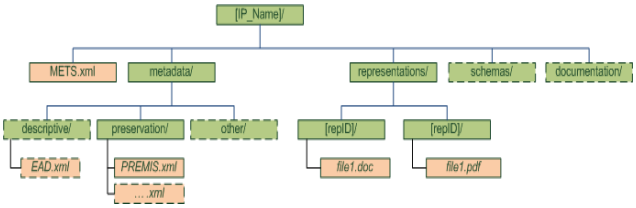

Figure 1: Basic E-ARK Information Package Structure

The structure allows for the separated inclusion of any metadata, data, relevant schemas and documentation into the package. Furthermore the metadata in the package can be divided into descriptive (metadata needed to find and understand the data), preservation (metadata needed to ensure the integrity and authenticity of data, metadata and the whole package) and other (any other metadata which is deemed relevant by the source system or the preservation organization).

A specific feature of the data component is that it can contain one or more representations of a single intellectual entity. The Common Specification allows also a single representation to include only the data of the specific representation or even duplicate the whole structure (Figure 2). Exploiting the last option allows implementers to differentiate between the package as a whole and a specific representation. For example, organizations can include generic descriptive metadata into the root metadata folder and at the same time keep detailed preservation metadata only within respective representations. Also, this offers the possibility of describing new emulation environments as a separate representation, thereby not endangering the integrity and authenticity of the original data and metadata. However, such splitting of metadata between the package and representations is purely optional within the Common Specification.

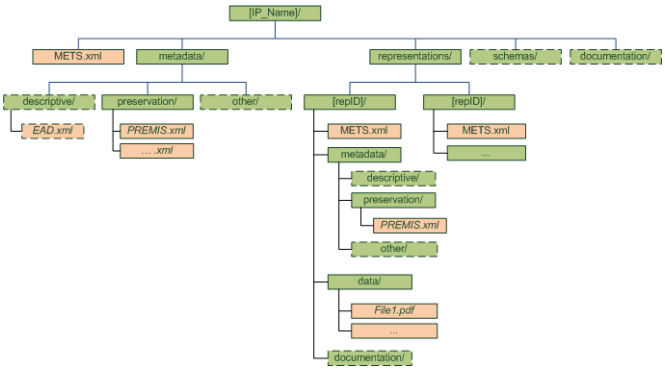

Figure 2: Full E-ARK Information Package Structure

Lastly, to ensure that the whole package can be understood and reused in the long term, users have the possibility of making the package self-sustaining by including any relevant schemas and documentation which might not be available externally in the future.

As well as the mandated folder structure, the information package folder must include a mandatory core metadata file named “METS.xml”, which includes the information needed to identify and describe the structure of the package itself and the rest of its constituent components. As the name indicates the file must follow the widely recognized METS standard[2]. The METS.xml file needs also to be present in all representations in the case where the full folder structure is being used (Figure 2). The METS metadata serves the main purposes of:

[2] http://www.loc.gov/standards/mets/

- identifying the package and its components in a persistent and unique way;
- providing a standardized overview of all components of the package;
- connecting relevant pieces of data and metadata to each other.

In short, the METS metadata is the main tool and driver for interoperability that allows everything inside the information package to be validated according to commonly accepted rules.

In comparison to the METS standard itself, the Common Specification imposes a few additional requirements to be followed. One key requirement is the availability of a specific structural map (METS <structMap> element) which must describe the data, metadata and other components of the package. Again, this requirement is the key towards allowing different developers to create interoperable tools for validating and checking the integrity of Information Packages. Furthermore, the Common Specification provides some additional rules, for example a specific regime for linking to external metadata from the METS file, the availability of IDs etc. For full details of the Common Specification METS profile please consult the full Common Specification document [3].

To support the scalability of Information Packages, the Common Specification allows also for the splitting of intellectual entities across multiple physical packages or for the creation of Archival Information Collections (AICs). This can be achieved by formatting individual representations or parts of representations as a stand-alone Common Specification package and creating a "grandfather" IP which provides references to all of the components. The only requirement for both the components and the grandfather IP is the availability of the METS.xml file, created according to the regime defined in the Common Specification.

3.2 SIP, AIP and DIP Specifications

As mentioned above, the Common Specification provides a set of core requirements which are both process and content agnostic.

To cover for the needs of specific archival processes (pre-ingest, ingest, preservation and access) the E-ARK project has developed separate Submission, Archival and Dissemination Information Package specifications. While all of these specifications follow the rules of the Common Specification, they also widen its scope via the addition of specific details.

The E-ARK Submission Information Package specification [4] concentrates on the details of the pre-ingest and ingest processes. As such it provides additional possibilities for describing a submission agreement in the package, adding further details about the transfer process (i.e. sender and receiver), etc.

The E-ARK Archival Information Package specification [5] concentrates mainly on the need for authenticity. As such it describes multiple possibilities for adding new representations in the original Information Package by either including these in the original package or formatting them as new Common Specification packages. Furthermore, the E-ARK Archival Information Package specification makes special arrangements for keeping the original submitted Information Package intact throughout any preservation actions.

The E-ARK Dissemination Information Package [6] concentrates on the details of access needs. For example, it makes special provisions for the inclusion of specific Representation Information as well as order related information. It can include also an "event log" which can be used for proving the authenticity of the package, even when the original submission itself is not included in the package and is not provided to the user.

3.3 Content Information Type Specifications

As discussed above, an Information Package can contain any type of data and metadata. However, the types of data files, their structural relationships, and metadata elements vary for different Content Information types. For example, metadata produced by a specific business system will variously be intended to support different aspects of descriptive, structural, administrative, technical, preservation, provenance and rights functions.

The METS standard used in the E-ARK Common Specification does not offer one, single structure in which content type specific metadata could be stored as a whole. In order to efficiently use metadata to support archival functions, the Common Specification defines separate METS sections as containers for the various metadata functions, such as the METS header for package management, the <dmdSec> for EAD[3] and other descriptive metadata standards, and the <amdSec> for preservation (PREMIS[4]), technical and other functions. In order to use the submitted metadata, the content type specific metadata elements need to be mapped to those METS sections and implemented using agreed standards. To ensure interoperability on such a detailed content-specific level, complementary metadata profiles are needed for key Content Information types to define how the submitted content-specific metadata should be mapped to the E-ARK Common Specification structure.

To meet this need, the E-ARK Common Specification allows for the creation of additional Content Information Type Specifications. In effect, these detailed specifications can detail the specific requirements for package metadata, data structure, and relations between data and metadata. Essentially anybody is welcome to set up new Content Information Type Specifications as long as these do not conflict with the requirements presented in the Common Specification.

The E-ARK project itself has developed two such specifications:

- SMURF [7] (Semantically Marked Up Record Format) specification, which details the archiving of data and metadata from Electronic Records Management Systems (the specification is semantically based on the MoReq2010[5] standard) or for simple file-system based (SFSB) records (specification based on the EAD standard). The SMURF profile specifies in particular how to archive the necessary elements of an ERMS system, including the classification scheme, aggregations and classes, disposal schedules, and user access controls.
- Relational Database Profile which is based on the SIARD format [8]. SIARD is an open format developed by the Swiss Federal Archives. The format is designed for archiving relational databases in a vendor-neutral form. The format proposes a common standard for describing core elements of the live DBMS: data; structure; stored procedures; triggers; views; and queries.

4. TOOLS

As mentioned above, the E-ARK specifications are primarily intended to lead interoperable tool development. To validate the

[3] https://www.loc.gov/ead/

[4] http://www.loc.gov/standards/premis/

[5] http://www.moreq.info/index.php/specification

applicability of the specifications in real life scenarios the project has committed to providing a set of software tools that automate the creation and processing of Information Packages created according to the specifications. Also, as is a typical convention for EC-funded projects, E-ARK has committed to providing all the tools as open-source, freely available for the whole community to use and participate in developing. The project has not committed itself to developing a single tool for any specification but instead aims to provide a set of tools for the same task within the archival workflow (Figure 3).

For example, the basic SIP creation task can be handled by four quite different tools - ESS Tools for Producers (ETP)[6], RODA-in[7], Universal Archiving Module (UAM)[8] and E-ARK Web[9]. All of these tools implement specific features which make them suitable for different users. The ETP allows for the setup of complex ingest profiles and is therefore suitable for larger organizations; RODA-in excels in the package creation of "loose data" (for example when archiving a whole hard drive at once); UAM is specifically able to deal with data originating from Electronic Records Management Systems and E-ARK Web is a lightweight web-based environment for creating Information Packages manually.

However, all of these tools create Information Packages according to the E-ARK Common Specification. Therefore, they can be used for creating packages not only for transfer to a specific national or institutional repository but to ANY repository supporting the Common Specification as an input format.

This example also illustrates very well the aim of the E-ARK project. Once a common agreement is achieved on core technical principles, organizations will be able to select their tools out of a set of different options, instead of being obliged to use a fixed choice that is then linked to a fixed standard.

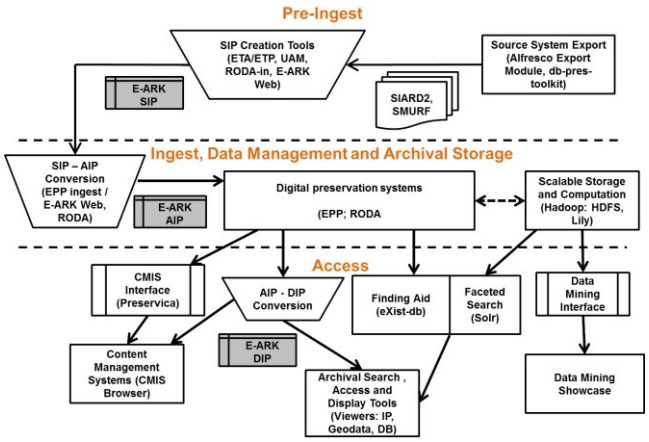

Figure 3: Overview of the E-ARK toolset

To demonstrate this in practice, the E-ARK project is running dedicated pilot installations at seven sites in six European countries where selected E-ARK tools are deployed alongside already available infrastructures. The pilots run between May and October 2016, with results published by early 2017.

[6] http://etp.essarch.org/

[7] http://rodain.roda-community.org/

[8] http://www.arhiiv.ee/en/universal-archiving-module/

[9] https://github.com/eark-project/earkweb

5. CONCLUSIONS

Ongoing access to information is a *sine qua non* of the modern world. But long-term access to and re-use of information depends, crucially, on ensuring the reliable and error free movement of information between their original environments and the digital archives. Additionally, the movement of information between different environments may occur many times during its lifespan and requires robust interoperability between those environments.

Thus, an approach for ensuring that digital information can be easily and consistently transferred between systems with all their characteristics and components intact is an urgent requirement for memory institutions. With its Common Specification, E-ARK has developed a coordinated approach to, and agreement on, standardized methods for packaging and sending information between systems, which is OAIS compliant. With its range of accompanying tools, the E-ARK approach has the potential to simplify and make consistent the currently diverse approaches to solving the issue of information transfer.

However, such standardization needs also to be carried on beyond the lifetime of the E-ARK and we are making every effort to ensure that the work of the project is also acknowledged, continued and broadened by the whole digital preservation community, not only the project partners. This is effectuated by the broad dissemination of the project from the outset via the partners DLM Forum[10] and the DPC[11], and also the practical involvement of three highly-involved advisory boards. The project outputs will be sustained long-term by the DLM Forum.

6. ACKNOWLEDGMENTS

This paper is presenting work which has been undertaken within E-ARK, an EC-funded pilot action project in the Competiveness and Innovation Programme 2007-2013, Grant Agreement no. 620998 under the Policy Support Programme.

We would like to acknowledge the work on the Common Specification by Karin Bredenberg of the Swedish National Archives and seconded for a period to E-ARK partner ES Solutions; to Dr. Angela Dappert (now at the British Library) on the SMURF specification; to Dr. Krystyna Ohnesorge's team at the Swiss Federal Archives on the SIARD 2.0 specification.

7. REFERENCES

[1] E-ARK Project. D3.1 E-ARK Report on Available Best Practices (2014). Available from http://eark-project.com/resources/project-deliverables/6-d31-e-ark-report-on-available-best-practices

[2] CCSDS. Open Archival Information System Reference Model (OAIS), 2012. Available from http://public.ccsds.org/publications/archive/650x0m2.pdf

[3] E-ARK Project. Introduction to the Common Specification for Information Packages in the E-ARK project. Available from http://www.eark-project.com/resources/specificationdocs/50-draftcommonspec-1

[4] E-ARK Project. D3.3 E-ARK SIP Pilot Specification. Available from http://eark-project.com/resources/project-deliverables/51-d33pilotspec

[10] http://www.dlmforum.eu/

[11] http://www.dpconline.org/

[5] E-ARK Project. D4.3 E-ARK AIP Pilot Specification. Available from http://eark-project.com/resources/project-deliverables/53-d43earkaipspec-1

[6] E-ARK Project. D5.3 E-ARK DIP Pilot Specification. Available from http://www.eark-project.com/resources/project-deliverables/61-d53-pilot-dip-specification

[7] E-ARK Project. D3.3 SMURF – the Semantically Marked Up Record Format – Profile. Available from http://eark-project.com/resources/project-deliverables/52-d33smurf

[8] Swiss Federal Archives, E-ARK Project. eCH-0165 SIARD Format Specification 2.0 (Draft). Available from http://eark-project.com/resources/specificationdocs/32-specification-for-siard-format-v20