E-ARK Progress Update March 2015

www.eark-project.eu

Prepared by: Kuldar Aas (National Archives of Estonia),
Andrew Wilson (University of Portsmouth)

Content: This report summarises E-ARK actions between December 2014 and March 2015, and provides a brief overview of work planned for the following few months.

Introduction

The primary scope of E-ARK is to provide standardisation in four key areas:

- delivering electronic data to long-term preservation (i.e. preparation and transfer of data to archives),
- open formats to support submission, preservation and access to archived data (i.e. what kind of metadata would be needed to store next to the actual data, and how the data and metadata should be formatted),
- access to data stored in long-term repositories (i.e. which methods and tools could be used to advance the level of access which currently is rather lacking for born-digital records),
- using data mining and big data technologies to facilitate large scale research on archived data.

The Project started on 1 February 2014 and has at the end of January 2015 finished its first year of activity.

Overview of actions from December 2014 to March 2015

The main effort of E-ARK in months 11 – 14 has been continuation of the development of E-ARK standards used in data preparation, transfer, preservation and access. A major outcome was the forwarding of three deliverables to the European Commission in January 2015. Two of these deliverables describe the principles of E-ARK Submission and Archival Information Packages (accordingly deliverables D3.2 and D4.2). The third presents a first draft of the E-ARK Maturity Model for digital preservation systems.

The project continues with refining the details of the information packages and thereafter shifts focus towards the development of the E-ARK toolset. One important result in

preparation for the development phase has been the delivery of a Requirements Specification template. The template allows all work packages to specify tool and information package requirements in a common way throughout the whole E-ARK project. The objectives of the template are two-fold. Firstly, it is considered important to streamline the way archivists communicate with developers; developers would be ill-served with different methodologies and formats for requirements. Secondly, acknowledging that the information gathering in the first tasks represented a bottom-up approach to identifying requirements, the template establishes a top-down approach.

A project all staff meeting, anticipated in the previous newsletter was held at the University of Portsmouth on 16-18 February. A separate report on the meeting is at the end of the work package report.

As for the other actions in the Work Packages the main effort has been as follows:

- WP1 (Project Coordination) has carried out its usual range of project management activities;

- WP2 (Use Cases & Pilots) has concentrated on supporting WP3 – WP5 in the development of Information Package and tool requirements by providing input about the exact needs of E-ARK pilot sites;

- WP3 (Transfer of Records to Archives) has finalised and delivered to the European Commission the draft SIP specification, deliverable D3.2.

  This SIP specification is in conformance with the common specification for E-ARK IPs. We have also aligned the specification with other known SIP formats which are known to the project from the best practice report (D3.1 "Report on Available Best Practices"). Despite that we would like to encourage readers to inspect deliverable D3.2 closely and share their opinion on it. We would especially like to ask Advisory Board members to give feedback to the project on two questions:
  - "Is the E-ARK SIP Draft Specification useful and why (or why not)?"
  - "Would I use it in my solution(s) and why (or why not)?"

  In addition WP3 continued working on pre-ingest and ingest workflows and has updated the draft diagrams first created in D2.1.

- WP4 (Archival Records Preservation) has mainly been working on deliverable D4.2 which was successfully published and forwarded to the European Commission in January 2015. The Work Package has also continued with tool development and has added support for relational database archiving (SIARD format) into their SIP2AIP tool. In collaboration with WP6, work has also started on gathering further requirements for updating the SIP2AIP tool to a more mature state and integrating with other E-ARK tools;

- WP5 (Archival Records Access Services) has primarily been working on the requirements specification template mentioned above. The template has also been put into use by creating high-level illustrations and descriptions of the entire access process to ensure that project participants have a common understanding of the archival access process. Based on the high-level descriptions, work has started on analysing metadata requirements and discussing a proposed structure for the E-ARK DIP;

- WP6 (Archival Storage, Services and Integration) has continued working on the E-ARK reference implementation. Three main areas of focus have been:
  - evaluation of design issues for implementing and integrating outcomes of WP4 with the integrated prototype developed in WP6 and prototyping of an initial approach for integrating EPP, SIP-to-AIP conversion, and a scalable search facility;
  - continued contributions to the technical evaluation of the WP4 SIP2AIP converter with respect to its integration and deployment on the AIT infrastructure;
  - development of an initial version of a customizable graphical search interface for demonstration and prototype development purposes.

- WP7 (Evaluation & Assessment) has focused on developing the deliverable D7.1, the first version of the information governance maturity model (see below). Other activities included a draft design for a 'knowledge centre' information system to support the harmonisation of best practices, standards, and other appropriate references relating to information governance;

- WP8 (Dissemination & Exploitation) has continuously updated the project website as well as maintained our newsfeed. WP8 was also the main organiser of the February project meetings and the database preservation workshop described below.

Deliverables produced in the period

There were three deliverables produced for the European Commission in the December 2014 to March 2015 period, and we would very much appreciate it if you could take the time to read these closely and send us feedback by the end of April 2015. The documents are available on our public web site (http://www.eark-project.com/):

- *D3.2 E-ARK SIP Draft Specification:* This document describes a draft SIP specification for the E-ARK project. It provides an overview of the SIP structure and sets out the main metadata elements for E-ARK SIP. This document will provide the initial input for the technical implementations of E-ARK ingest tools.

- *D4.2 E-ARK AIP draft specification*: This deliverable describes a blueprint for the structure of an E-ARK AIP. Since an AIP has a potentially unlimited life span, the document also includes a discussion of how such an AIP may keep its identity

unchanged, while its physical representation may change over time. Finally, the document contains a chapter on the way in which the AIP format is to be embedded into the technical workflow within which an AIP exists.

- *D7.1 A Maturity Model for Information Governance – An Initial Version:* This deliverable sets out a draft maturity model for information governance which will be used to assess the E-ARK Project use cases. The model has an intentional focus on the processes being harmonized in the project, i.e. ingest, archival preservation, and dissemination (access). The model dimensions are based on SEI CMMI. The maturity levels are based on ISO9001. The definition of the maturity model is based on a development method of J. Becker (for more information please consult the deliverable). The deliverable describes the outputs of the first four phases of the method, i.e. it (1) defines the problem, (2) reviews existing maturity models, (3) defines the strategy used for the development, and (4) develops the model. The remaining steps of the method, which include application, validation and review of the maturity model, are due to be completed by the end of June 2015.

Next steps (March – June 2015)

During the spring period in 2015, work packages 3 – 6 will continue to develop the draft E-ARK standards for data preparation, transfer, preservation and access. Most crucially the work on IP specifications will continue and broaden. Two additional subgroups have been established to work more explicitly on:

- E-ARK requirements for exporting records and metadata from electronic records management systems (ERMS): this subgroup will develop the principles and formats for exporting archival records from business systems that manage documents and records, as a pre-ingest process. It will also focus on developing metadata requirements based on the MoReq standard. The primary goals for early 2015 are:
  - the definition of records system specific export requirements to be implemented by a records system that exports to an archive;
  - the specification of a reference data model and metadata elements. These are to be used for the definition of the core E-ARK SIP profile for ERMS;

- E-ARK SIARD-E format: this subgroup will develop a specification of an E-ARK archival relational database format (working name SIARD-E), based on the best practices from the existing SIARD, SIARDK and DBML formats. The subgroup will also gather additional requirements for updating appropriate tools with support for the validation, quality and reuse of archived databases.

In more detail the work in individual Work Packages will be as follows:

- WP1 (Project Coordination) will continue the management of the project.

- WP2 (Use Cases & Pilots) will continue supporting WP3 – WP5 in the current requirements gathering phase from the point of view of the E-ARK pilots. It will also start gathering information around legal issues in E-ARK pilot implementations. The aim is to produce the appropriate deliverable – D2.2 – by month 17 (June 2015).

- WP3 (Transfer of Records to Archives) will expand and continue to develop the SIP draft specification to include support for MoReq 2010 based metadata and the proposed SIARD-E format. By May 2015, the next version of the SIP specifications as well as tool requirements will be available and large-scale development of specific E-ARK tools (records management export module for Alfresco, database preservation toolkit and SIP creation tools) will start.

- WP4 (Archival Records Preservation) will continue working on the SIP2AIP tool and the AIP specification, particularly focussing on the needs and possibilities for enriching and de-normalising structured data during ingest. A conceptual solution for storing parallel representations of structured data in an AIP (original database structure and the de-normalised presentation) will be available by summer 2015. Work will also continue around the integration of the SIP2AIP tool with the EPP digital preservation platform.

- WP5 (Archival Records Access Services) will finalise the access use cases and the functional requirements for the E-ARK access scenarios. A draft DIP format is scheduled for release at the end of April and we will be seeking Advisory Board comments at that time. In addition the use cases and requirements will be used to start the development of the E-ARK access toolkit.

- WP6 (Archival Storage, Services and Integration) will focus on the upcoming deliverable *D6.1 Faceted Query Interface and API*, which is due to be delivered at the end of April. Technical work will include developing support for full text indexing of ingested office documents. In this context, it is planned to deploy a publicly accessible demonstrator which allows other work packages to evaluate the WP6 search service API. The WP6 search interfaces will be evaluated with respect to supporting the DIP creation tool together with WP5.

- WP7 (Evaluation & Assessment) continues to work on a first version of the E-ARK knowledge base. The knowledge base will comprise:
  - A **Requirements Sources Manager** to manage the standards and reference documents that contain Information Governance requirements. The component will allow semi-automatic extraction of requirements from different sources;
  - A **Vocabulary Manager** to manage the terms and definitions in the existing sources;
  - A **Requirement Manager** that will manage requirements from existing sources;
  - **Assessment Services** that will provide a set of services that will allow verification of compliance with existing requirements. An **On-Line Self-Assessment** tool will provide an online assessment based on the Information Governance Maturity Model while the **Schemas Validator** will allow the verification of XML Schemas against defined XSD or requirements (e.g. validation of XML Schemas based on the SIP, AIP and DIP schemas);

- The **Data Store** will manage and store all the information from the previous components. The **View Manager** will be responsible for consolidating, personalising and displaying the information stored in the Data Store.

WP7 is also currently gathering feedback on the draft E-ARK maturity model which will be used to measure the success of the pilots in year three of the project.

- WP8 (Dissemination & Exploitation) will continue the update of communication channels and carry out dissemination.

E-ARK Project All Staff Meetings, 16-18 February 2015

From 16 – 19 February, E-ARK Project all staff meetings, as well as a meeting of the Project Board and the Executive Steering Committee, were held at the University of Portsmouth.

The program started with an all staff meeting on the afternoon of 16 February which discussed a number of issues, including:

- Common Specification
- Requirements Specification template
- Specification of tools do be developed in E-ARK
- Use of MoReq metadata in archiving
- Extending SIARD

The all staff meeting continued on the morning of 17 February with an introduction by the Project Coordinator, Janet Delve, followed by a discussion of a number of administrative issues, including end of year reporting. The next session of the meeting consisted of Advisory Board and Work Package updates from the respective WP leads.

On Tuesday afternoon a developers meeting was held in parallel with sequential sessions on editing the E-ARK website using the JOOMLA software, and using the EC NEF system. These sessions were followed by a project board meeting in the late afternoon.

Wednesday morning saw a series of Work Package meetings including a WP2 & 7 joint meeting and a parallel joint meeting of WPS 4 & 6. In the afternoon a WP5 meeting was held, with an Executive Steering Committee meeting taking place in parallel.

We would like to highlight that these meetings helped essentially to visualise the connections between individual work packages and groups. In practical terms the meetings helped especially to transfer knowledge between development, research and archival partners in preparation of the large-scale development phase (to be started in late spring 2015). As the outcome we were ensured that we can be confident in our roadmap towards a set of interoperable E-ARK specifications and tools.

Database/Data Mining Workshop: *Is a Data Warehouse a Data Archive and Why Does It Matter?* **19-20 February 2015**

One of the items in the core of the E-ARK work programme is the archiving and reuse of structured data derived from relational database management systems. These systems are one of the essential building blocks of information technology. Ubiquitous but often obscured behind layers of scripting, processing, forms and queries, they are arguably the most important invention of the Twentieth Century. There's no question that databases present a complex challenge to preservation. They can be difficult to document and difficult to understand even when they are documented. The complex interdependencies of data, query and scripting make migration problematic and highly specialised.

Relational databases and to some extent data warehousing approaches, which favour structure and homogeneity, are sometimes contrasted with 'big data' approaches that tend to favour heterogeneity and de-normalisation. It could be suggested that a concern with relational databases is outmoded and that the preservation community could simply adopt big data approaches. But the contrast can be overstated, especially when preservation issues are discussed. In practice 'big data' tools seem to offer improved workflows that complement rather than replace existing data warehouse tools. And even if 'big data' tools are the solution for access they still need to integrate with fundamental preservation processes and standards.

Better technical guidance and organisational know-how are needed if the digital preservation community is to offer confident and consistent solutions to long-term access for relational databases.

This workshop, held over one and a half days at the University of Portsmouth, was jointly sponsored by the E-ARK Project and the Digital Preservation Coalition (DPC). The workshop was split over two days, with day one aiming to:

- Clarify where databases, data warehouses and big data complement / overlap each other
- Review the state of the art in the preservation of databases
- Present case studies of current tools and practices around the preservation of relational databases
- Introduce commercial approaches to 'data warehousing' and explore the relationship with preservation
- Introduce big data approaches for database preservation

Day two's aims were to:

- Review the state of the art in the use of 'big data' and its implications for preservation
- Examine and debate the use cases for archived databases / big data
- Identify recommendations for further research and guidance in the preservation of 'big data'

Copies of the presentations are available at: http://www.eark-project.com/resources/conference-presentations/dpc-eark-event. Videos of the event will be available soon, also from the E-ARK website.