E-ARK Project
Summary of Activities
Year 3
1 February 2016 - 31 January 2017

EC Project Number 620998

cip competitiveness and innovation
framework programme
2007–2013

Overview and objectives

2016 - 2017 was the final year of E-ARK, and we have achieved the overall project goal of piloting archival services to keep digital records authentic and usable, based on current best practices which address the three main endeavours of an archive: acquiring, preserving and enabling re-use of information. The benefits for public agencies, citizens and business have been demonstrated by providing easy and efficient access to the archived records. Archival processes at a pan-European level have been harmonized, supported by guidelines and recommended practices that cater for a range of data from different types of source including record management systems and databases. The E-ARK approach is public-facing, open source, robust, replicable and scalable, and addresses a wide range of organisations, taking full account of legal constraints.

The following specific objectives for this year have been achieved. The pilot versions of the E-ARK formats, the Submission Information Package (SIP), the Archival IP (AIP), and the Dissemination IP (DIP), were updated to incorporate stakeholder feedback. The open source tools / platforms, which support the formats and accompanying archival processes, plus the Integrated Platform Reference Implementation, were similarly developed in response to user feedback, and were then acceptance tested by the Open Preservation Foundation (OPF).

The General Model, with its overarching workflows, methodology and use cases, was updated and published on the website. With the model, formats and tools all in place, the main objective of this final year, and indeed the whole project, was conducted: running the seven pilots in six countries, taking account of any issues brought up in the legal study.

The pilots started in May and successfully concluded in November. There were also additional pilot activities and international external pilots. The impact at the seven pilot sites was assessed. The Final Conference was held at the National Archives of Hungary in Budapest in early December, and included the Data Mining Showcase, tool workshops, keynote presentations from experts, and detailed explanations of the pilots. Our dissemination strategy for the year has focussed on encouraging testing and take-up of deployed tools, plus expanded awareness.

Two dedicated E-ARK journal special editions are in preparation, and planning undertaken for a Facet E-ARK book. Detailed sustainability arrangements have been put in place, such as setting up the DLM Archival Standards (DAS) Board and non-exclusive, permanent licences issued to the DLM Forum, the Digital Preservation Coalition and the Open Preservation Foundation to enable the project's work to continue. The Knowledge Centre and Maturity Model were both finalised, and made available via the DLM Forum website. All deliverables have been submitted on time, and all milestones met.

Co-Ordination (Work Package 1 - WP1) University of Brighton, UK

Project Management

E-ARK has been successfully managed throughout the year in accordance with the principles of the Prince 2009, MSP (Managing Successful Programmes) and M_o_R (Management of Risk) methodologies. The Project Management Team have maintained regular contact with all work packages, through fortnightly teleconferences, and occasional face-to-race meetings when the opportunity arose through other events having brought project partners together. Once each month, these meetings are extended to include representatives of all Consortium Members, thereby enabling any issue affecting any part of the project to be identified and addressed without delay. Regular Project Board meetings have been held as scheduled. A variety of media have been used for intra-project management during the year: Cisco Webex, Sharepoint, Redmine, Github and Google Drive.

Technical Coordination, National Archives Estonia

The technical focus of the third project year was on supporting the E-ARK pilot sites, gathering pilot feedback and delivering it appropriately to developers and specification owners. Furthermore, the latter months of the project saw the emergence of actions ensuring the sustainability of the technical outcomes. As a result, the project was able to continuously improve all tools and specifications throughout the third year, deliver all the results promised in the Description of Work, and even exceed these expectations by the delivery of additional pilots, specifications and tools. In order to ensure these outcomes, the daily technical management and coordination of E-ARK continued as established during the first two years of the project. Virtual technical meetings were organised on a monthly basis for all technical staff. In addition, one cross work package face-to-face meeting was held on 1 March 2016 during the E-ARK All-Staff Conference.

Multiple cross work package groups continued to be active, most notably the WP3 – WP5 group for maintaining the Common Specification for Information Packages and synchronising the SIP, AIP and DIP specifications; and the WP4 – WP6 group working on data mining issues. Further collaboration was established between WP2 and WPs 3-5 in regard to managing the feedback from pilots. As well, in summer 2016 a specific work group was established to manage the creation of sustainability regimes (DAS Board, relationship to OPF etc.).

Work Package 2 (WP2) - Use Cases and Pilots, National Archives Hungary

WP2 pilot related activities were divided into three phases: to plan, to prepare for and to execute the pilots. In year 3 the focus was on the execution, and WP2 has built on the preparation from year 2, such as the pilot cards, to successfully manage the execution and evaluation of the seven full-scale pilots in six countries. Extensive pilot documentation was drawn up for each pilot instance, delineating and illustrating the working of each pilot instance. In addition to the full-scale pilots, there were also several additional pilot scenarios, as well as an external pilot at the National Archives and Records Administration (NARA) in the U.S.A., plus an extra pilot at MINHAP. A semi-structured pilot impact questionnaire was developed, and representatives from each pilot site were interviewed to evaluate the impact and outcomes from their pilot. The deliverables D2.3, D2.4 and D2.5 were submitted as scheduled. The legal study D2.2 was also updated and an additional D2.2.1 version was also produced, which provided a standalone set of guidelines to archives on the legal issues which affect them.

Work Package 3 (WP3) - Transfer of Records to Archives, National Archives Estonia

For WP3, the main effort in year 3 of the project, was to support pilots, update and fine tune the following tools: the Electronic Record Management System (ERMS) Export Module (EEM), a tool that helps to export records and their metadata from ERMSs in a controlled manner, by being capable of connecting to and exporting content from any repository that supports the CMIS (Content Management Interoperability Services) protocol version 1.0 or higher; the Database Preservation Toolkit (DBTK), a tool for exporting relational databases as SIARD 2.0 and other formats; the ESSArch Tools for Producer (ETP), a tool for SIP creation which supports the E-ARK general model but can easily be configured to support any other pre-ingest, ingest processes and workflows; the ESSArch Tools for Archive (ETA), a tool which facilitates receiving SIPs, performing quality controls and preparing SIPs for ingest into the preservation platform; RODA-in, a tool which is specially designed for producers and archivists to create SIPs from files and folders available on the local file system. R

ODA-in satisfies the need for mass processing data by providing a quick and simple way for creating thousands of valid SIPs (in E-ARK format or other pre-configured formats) with just a few clicks, complete with data and metadata; the Universal Archiving Module (UAM), a tool which allows users to rearrange, classify and further describe the contents of the transfer; validate the transfer according to the rules established by the archives, and finally create SIP packages to be transferred to the digital archives. It is fully compliant with the E-ARK pre-ingest and ingest workflows and the SIP specification. Also, the SIP specification, together with the SMURF (Semantically Marked Up Records Format) for ERMS and the SMURF for SFSB (Simple File-System Based

Records) were updated based on the feedback mainly received from the pilots and the Advisory Board. The updated documentation was handed over to the DAS Board for further management.

Work Package 4 (WP4) – Archival Records Preservation, Austrian Institute of Technology

The work was focused on the implementation of the AIP format specification as defined by deliverable D4.3 "E -ARK AIP pilot specification" in the form of the *earkweb* application framework. During the development phase in year 3, the discussion and refinement of the E-ARK AIP specification was ongoing. For this reason, the AIP specification was still being adapted to the needs of the pilots. The regular Common Specification working group discussions affected the E-ARK AIP format, and most decisions regarding structural requirements and the concrete metadata implementation details were considered for implementation in *earkweb*.

The main task was to implement the SIP-AIP conversion component as part of the reference implementation E -ARK Web (in short: *earkweb*.) A set of tasks was implemented to convert an E-ARK SIP to an E-ARK AIP using the web application frontend of *earkweb* to control the execution of the conversion. Special attention was paid to the scalability of the information package processing by making sure that E-ARK AIPs can be created from E-ARK SIPs in parallel. Furthermore, it was required to actively consider new requirements being presented during the initial testing of use cases defined by the various pilots, particularly to ensure the *earkweb* archiving solution covers the requirements of small archives as well as very big archives with very large data collections. Together with WPs 5 & 6, work on database archiving involved further looking into methods to prepare de-normalized AIPs ready for dimensional analysis via OLAP.

Work Package 5 (WP5) – Archival Records Access Services, Danish National Archives / Magenta

WP5 released the DIP pilot and final format specifications, and prepared Access tools for the Access pilots in Estonia, Slovenia, and Hungary. In Estonia, the CMIS Viewer enabled government agencies to access their own archived material (SMURF format) in the Preservica repository of the Estonian National Archives.

In Hungary, the Database Preservation Toolkit and the Database Visualization Toolkit were piloted. The former loads SIARD database files ready for the latter to render these in a user-friendly format. Also, the data mining showcase, which was created in a collaboration between WP4, 5 and 6, was piloted. In Slovenia, the geodata format specification was tested using the Access Software Platform. The Access Software Platform is a platform that allows the end-user to search for archival material and to order it from the archive; it then allows the archivists to process the ordered archival material and to prepare it for dissemination; lastly it makes the material available for the end-user in an appropriate DIP format. In the Slovenian case, the archival material was accessed using specialised geodata tools, namely Peripleo for search, and QGIS and geoserver services for access.

Work Package 6 (WP6) – Archival Storage, Services and Integration, Austrian Institute Technology

The main WP6 focus was on supporting the deployment of the Integrated Platform Reference Implementation in different configurations at E-ARK stakeholder sites. A flexible packaging mechanism combined with a standalone backend implementation enables custom single-server deployments on demand. The scalable, Hadoop-based backend implementation has been ported to the latest CDH (Cloudera Distribution Including Apache Hadoop) distribution in order to support a recent technology stack, advanced data mining concepts, and enterprise demands. Showcases that dealt with the application of text mining approaches, the extraction and visualization of geographical information, the implementation of a database archiving workflow, and mass document ingest have been implemented and deployed at stakeholder sites in Hungary and Slovenia. The technical details have been documented and summarized in deliverable D6.3. Data mining workflows have been demonstrated at IPRES 2016 in Bern, Switzerland on 6 October 2016; in the context of an E-ARK webinar that took place on 2 November 2016; the E-ARK demo day on 8 November 2016 (Brussels); the DLM Forum on 17 November 2016 (Oslo); and the E-ARK final conference in Budapest on 7 December 2016.

Work Package 7 (WP7) – Evaluation and Assessment Instituto Tecnico Lisbon, Portugal

WP7 improved the existing services of the Knowledge Centre. In particular, the Reference Requirements Management Service (entitled REQs) was improved through the development of change management and feedback functionalities, and the Maturity Assessment was improved to include the changes on the Maturity Model and to provide immediate feedback to the users. Additionally, a new service "MoReq Assessment" was included in the Knowledge Centre to support the process of assessing if records management systems are MoReq2010 compliant. A final Maturity Model was developed that includes a new approach where levels are defined by the set of capabilities the organization should have. Consequently, the self-assessment questionnaire was changed to incorporate the new approach. Additionally, the dimensions of management and infrastructure were added to the questionnaire. The final version of the Maturity Model was used to perform a final pilots' evaluation. In this final evaluation, E-ARK tools were also assessed to verify whether best practices were followed during software development to ensure its sustainability and improve code quality.

Work Package 8 (WP8) – Project Dissemination, University of Brighton, UK

Website: www.eark-project.eu

We have continued to expand the project website which contains information about the project as well as providing access to public deliverables and other resources, such as presentations given at conferences. We achieved an average of 2,750 page views per month during the year, (monitored continuously using Google Analytics) from 136 countries including all EC Member States. We have also continued to make extensive use of our Twitter and LinkedIn feeds, as well as regularly publishing our online Newsletter, successfully expanding our communities of interest on Social Media. We have exceeded all our performance targets for dissemination activity.

We have given many presentations at both national and international events, reaching audiences totalling many thousands.

E-ARK Advisory Boards

At the start of the project E-ARK established three stakeholder Advisory Boards as an integral component of project governance, as well as to enhance project communication and dissemination activities.

The Advisory Boards' main contribution is to assess contributions to and from the project and to help resolve conflicting community views if these arise. In general terms, the Advisory Boards also:

- represent a range of stakeholder interests broader than those represented by project partners
- ensure E-ARK project outputs are compatible with relevant national and international standards and legislation
- keep the project work connected to best practice in digital preservation approaches, tools and services
- help disseminate information about and outputs of the E-ARK project to their stakeholder communities.

The Boards are open to all interested parties and represent a wide range of stakeholder interests. The current membership numbers for the three E-ARK Advisory Boards are:

the Commercial / Technical Advisory Board	10 organisations (+1)
the Archival Advisory Board	31 members representing 20 (+2) organisations
the Data Provider Advisory Board	4 institutions, 1 individual (no change)

All completed project deliverables are circulated to the Boards following their submission to the EC. This practice was continued in Year 3 .

In addition, the revised internal deliverable, Introduction to the Common Specification for Information Packages, was circulated to Advisory Board members in July 2016. It is pleasing to be able to report that the amount of feedback received has increased during the third year of the project. All feedback received is acknowledged, and individual responses prepared by Work Package leads and the Technical Coordinator are sent to respondents.

At least one face to face meeting of the Advisory Boards is held every year. The purpose of the meetings is to inform Board members of progress of various work packages, to discuss issues and concerns, and to receive additional feedback on project deliverables. In year three of the project, two face to face meetings were held. The first of these was held in conjunction with the International Conference on Digital Preservation (iPres) in Bern, Switzerland on 3 October.

The second face to face meeting of the Boards was held in conjunction with the DLM Forum Member Meeting in Oslo, Norway on 17 November. Both meetings were very successful, with 12 attendees at the iPres meeting and 14 at the meeting held in Oslo. Advisory Board members are sent an overview report approximately every quarter that details Work Package progress in the previous three months and foreshadows major project activities in the upcoming three months. In year three, however, only two reports were circulated, in April and July.

The third report, nominally scheduled for late October 2016 was omitted because of its proximity to the two Advisory Board face to face meetings, and the fourth report for year 3 was circulated in early March 2017 as the final newsletter of the E-ARK Project to the Advisory Boards.

Expected Results / Impact

The E-ARK open source, digital archiving framework, complete with accompanying metadata and other standards, has been thoroughly tested and has made a significant impact on the institutions who carried out the pilots: this has been assessed by carrying out semi-structured interviews at the pilot sites based on a detailed impact questionnaire. Highlights from this impact analysis include: major savings in the cost providing pre-ingest tools due to increased competition; harmonisation of geo-spatial data archiving practices which facilitate comparison of Natura 2000 sites across Europe; the benefits of using the *earkweb* tool, together with advanced search and data mining facilities, in national, regional and local archives; and last but not least, robust, common standards and tools that can truly be used interchangeably across Europe.

The project stakeholders have also been surveyed, and the responses indicate that the pilots have led to the development of new skills, and it is predicted that future training opportunities will arise as will new archiving and digital strategies. The results will improve public awareness and allow web-based access to tools. Archival processes will be more open to re-evaluation, and the preservation and visualisation of archives will be enhanced.

The piloting of the tools will lead to increased speed of use; the possibility of new access; improved tools and improved knowledge about big data issues. Technical benefits will include increased efficiency in bringing archive data closer to the producer and the end user; access to ‘real life’ help and greater reuse of data.

Overall tangible benefits include increased competition; new procurement processes; potential reductions in IT budgets; opportunities to keep processes in-house; improved service delivery and critically, new job opportunities. However, consideration needs to be given to the long-term sustainability of the outcomes and impacts of the project. This will require a proactive stance and continuity of the E-ARK brand.