THE E-ARK PROJECT
SUMMARY OF YEAR 1

European Archival Records and Knowledge Preservation

FP7—PSP GA No 620998

www.eark-project.eu

Overview and objectives

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

The project kicked off at IST Lisbon in February 2014, and has progressed well throughout the year, with regular internal meetings, and many appearances at external events (conferences etc.). External stakeholders have been fully engaged via three Advisory Boards (Archival, Commercial/Technical, Data Provider) as well as at several key conferences throughout the year.

The technical objectives have been reached in terms of:

- gathering an extensive understanding of the best practices in pre-ingest, ingest, preservation and access to archival records;
- the draft specifications of ingest and archival storage packages for records and databases;
- work on the draft specification for the access package is well under way (due month 15);
- developing a Hadoop-based integrated framework;
- preliminary work started on the legal study;
- the proof of concept of the Knowledge Centre, and
- good progress also made against our “stretch targets” of data mining.

Co-Ordination (Work Package 1 - WP1) University of Portsmouth, 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. Quarterly 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 first year of the E-ARK project has been successful from a technical standpoint. The main goal of the project is to improve the level of interoperability and standardisation in archival practices, including the pre-ingest, ingest, preservation and access to archival records. To meet this goal, the main effort has been the communication and harmonisation of approaches between the project partners and beyond. This was initially addressed by two main actions, covered by milestone MS01 Best Practice Overview (month 6):

WP2 provided an internal view of the current status by gathering detailed information about the needs and available best-practices at the E-ARK pilot sites. The outcomes are generalised and documented in the deliverable D2.1;

WP3, 4 and 5 joined forces to provide an external view of the current state of the art by preparing a survey for the archival, service provider and data provider communities to gather additional information about the best practices outside the circle of project partners. The outcomes are documented in respective deliverables D3.1, D4.1 and D5.1.

The milestone was achieved, and the outcomes of the work were also made public, presented (most notably at the DLM Forum meetings in June 2014), and there was also extensive consultation with the E-ARK Advisory Board members as well as other interested stakeholders in the respective communities (archives, data providers, service and technology providers).

Based on the knowledge gathered, work has continued on the provision of the first versions of the E-ARK standards. Most notably, all partners have come together to create an upper layer based on amending the individual deliverables described in the project Description of Work:

In regard to Information Package (IP) specifications (Submission IP (SIP), Archival IP (AIP), Dissemination IP (DIP)) a Common Specification has been developed which outlines the common aspects and requirements for all of these, therefore ensuring that future work will be well aligned in regard to the single specifications. This Common Specification has also been taken into account when developing deliverables D3.2, D4.2 (both delivered month 12) and D5.2 (due in month 15);

WPs 3 – 5 have created a common Requirements Template which facilitates the describing of all aspects of tool standardisation (high level and detail process and information flows, use cases, functional and non-functional requirements) in a common and interoperable way. This Requirements Template is already used for describing the requirements and workflows for the appropriate pre-ingest, ingest and access workflows, which are intended to be implemented within Year Two of the project.

In regard to the core areas of IP standardisation and workflow synchronization, the project has successfully come to an internal common understanding which will support their continued implementation vai actual software solutions, and the dissemination of the E-ARK products outside the project.

Regarding scalable computing and data mining, the project has managed to successfully integrate the intended software of the integrated platform (ESS Preservation Platform) with scalable computing technologies (Lily, Hadoop, HDFS). As such the current state of the integrated platform allows using out-of-the-box big data research methodologies on top of the whole content stored in the digital archives. During the second year the development of the integrated platform will be continued: in particular the components being developed in WP3-5 (E-ARK pre-ingest, ingest and access) will be added to the current storage and analysis capability.

The second project milestone for year 1, MS02 Knowledge Centre Service – Validated Proof of Concept Service was also achieved (M12): the first version of the Information Maturity Model and the Vocabulary Manager (the first component of the Knowledge Centre) were successfully developed, focussing particularly on Ingest, Archival Preservation, and Dissemination.

Concerning external communication and dissemination, the focus of the project from a technical point of view was to establish and widen the knowledge about the project as well as extend the participation of the Advisory Boards. This has also ensured that all E-ARK outcomes have been consulted with a wide representation of the communities. We have raised a lot of interest inside the target communities which is also visible in the number of members who have joined the Advisory Boards. The most notable outcomes are:

E-ARK has been actively pursuing collaboration with the e-government community to ensure that the standardisation towards data providers is widely visible and accepted. Partnership with the EC-funded e-SENS project, which works on establishing EU-wide e-government interoperability, has been established for this purpose. This partnership allows the E-ARK project to get informed in a timely manner about most recent developments in the government IT sector and therefore assure that our outcomes meet the generic technical and organizational requirements of European government agencies;

E-ARK has established a strategic partnership with the Swiss Federal Archives who are now participating in the work around the E-ARK SIP format and database pre-ingest and ingest.

During the second year of the project we will continue the communication effort: however, in contrast to the first year, the commercial archival service providers will be particularly targeted as part of the main scope of activities when E-ARK tools become available (as compared to the archival institutions in the first year).

A summary overview of the technical work packages now follows.

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

The most important action of WP2 was the creation of the General Model in deliverable D2.1 “General pilot model and use case definition” which was submitted to the European Commission in month 6 (M6). The E-ARK framework comprises a set of small blocks: tools, interfaces, data packages each one designed according to the different stages of the archival workflow, the workflow itself interweaving the data into a high level business process in which the actual digital archiving is manifested. The definition of these interconnected components in an overarching model was crucial

for both the pilots and the reference implementation, with resulting cross-references that actually created a backbone for the whole project.

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

WP3 submitted deliverable D3.1 "Report on Available Best Practices" in month 6. One common specification for information packages is under development to help shape a common vision about general principles (e.g. structure, quality criteria) of all the information packages (incl. AIPs and DIPs) in E-ARK. The deliverable D3.2 "E-ARK SIP Draft Specification" was delivered in month 12. Related work defining the requirements for the records export (data selection, extraction etc.) will continue resulting in an input to MoReq2010 export requirements and also a basis for more detailed implementation requirements.

Work Package 4 (WP4) – Archival Records Preservation, University of Cologne, Germany

WP4 produced deliverable D4.1 "Report on available formats and restrictions", submitted month 6 and deliverable D4.2 "E-ARK AIP draft specification". The work on the specification of the SIP to AIP conversion component has been conducted in close cooperation with WP6, connecting the specification with preliminary implementation work, according to AGILE development principles. The emerging reference implementation for a scalable e-Archiving service now already contains a pre-prototype of the SIP to AIP conversion component. While this pre-prototype has been built with static content as test material, parallel to its implementation, work has started on possibilities to connect database / record management system content into the emerging conversion component. Similarly, an examination has started of the current simplified version of PREMIS to support a more advanced rights management.

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

WP5 delivered in month 5 D5.1 "GAP report between requirements for access and current access solutions", describing the landscape of access services today and highlighting user access needs. From the findings – that the main barriers to achieving user needs are about legislative issues and the lack of maturity of existing solutions – WP5 is now preparing a DIP format specification based on the requirements analysis as well as from inputs from other WPs, requirements identified during the pilots, existing access tools and an analysis of relevant metadata standards.

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

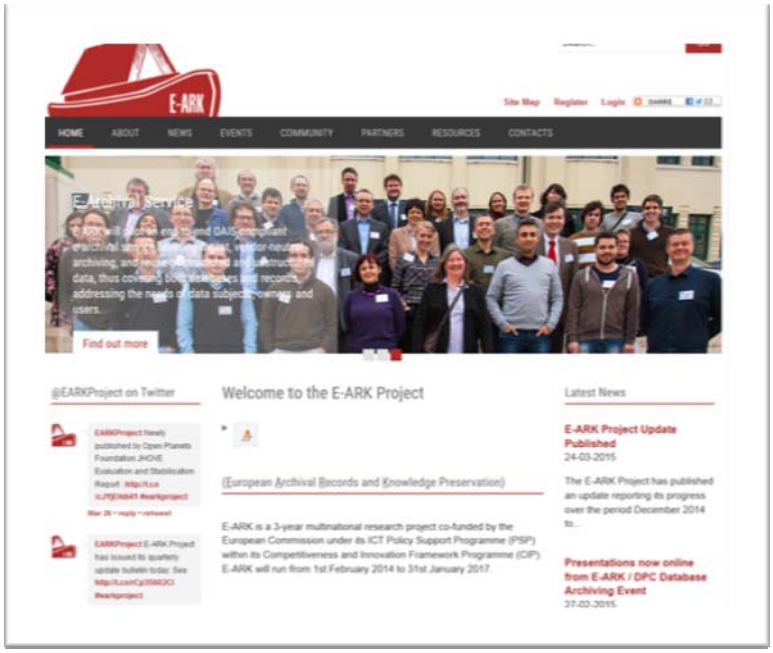
WP6 has contributed to the E-ARK software development and configuration environment. An integrated development environment has been created using a number of software tools, including "Jenkins Continuous Integration", "Maven Parent" and Github. Using this environment, an initial service has been implemented allowing components to upload large files to our Hadoop Distributed File System. A first version of the service has been released, documented, and made available on the AIT infrastructure for integration and testing. With respect to data mining, initial tests which ran a clustering algorithm on the present SolR text index using Apache Mahout have been carried out. Ideas for data mining and an initial software project for de-normalizing existing relational databases were presented in Vienna, and made available as a contribution to a paper at the DLM forum.

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

WP7 has developed the first version of the Information Maturity Model (submitted under Deliverable 7.1) based on Archival best practices namely ISO14721:2012 (OAIS), ISO16363:2012 (TRAC) and ISO20652:2006 (PAIMAS). The model has an intentional focus on the processes being harmonized in the project, i.e. Ingest, Archival Preservation, and Dissemination. Additionally, a Vocabulary Manager was developed to manage the set of terms in use on the project along with the ones defined in existing Information Governance reference documents and standards. The tool allows the definition of relations between terms supporting the harmonization of terms by, for example, identifying identical or similar terms. The Vocabulary Manager is the first component of the Knowledge Centre to be presented at the end of the project.

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

We have given presentations at 10 third-party events, including iPRES, PASIG and ICA, and DLM Forum's Triennial Conference. Our web presence, which was established on Day 1 of the Project, is attracting almost 500 hits per month from more than 350 unique visitors, of whom 70% are first-time visitors. We have created an online monthly newsletter (news.eark-project.eu) which contains a mixture of project-related news and items of general interest to the archiving community and which has attracted 330 readers. Our now-daily Twitter feed has 125 followers while 95 people have signed up on our website to receive regular e-mail notifications about the project. Our tasks in Year 2 will be to expand our engagement activities to reflect the growing availability of project tools and services in order to attract and engage with our wider stakeholder community to examine and test our tools. We will also support WP2 in local publicity with the partners who are running our pilots.



Expected Results / Impact

E-ARK will provide national, regional and local archives with an open source digital archiving framework, complete with accompanying metadata and other appropriate standards. The resulting harmonisation will help communication across archives, but still respect different national legislative backgrounds. Organisations supplying data to archives will be provided with streamlined pre-ingest standards / tools which reduce their costs. A wide range of users (including business, researchers and citizens) of archives will benefit from improved access to digital archives, also covering databases.

Contact

info@earkproject.eu