European Archival Records and Knowledge Preservation (E-ARK) Project. Goals and Achievements: an Overview

Anja PAULIČ, Mrs.
archivist-counsellor, Archives of the Republic of Slovenia, Zvezdarska 1, Ljubljana, Slovenia
e-mail: anja.paulic@gov.com

European Archival Records and Knowledge Preservation (E-ARK) Project. Goals and Achievements: an Overview

ABSTRACT

E-ARK project is co-founded by the European Commission under its ICT Policy Support Programme (PSP) within its Competitiveness and Innovation Framework Programme (CIP). The aim of the E-ARK project is to provide efficient access to the workflows related to the three main activities of an archive - acquiring, preserving and enabling re-use of information. Gathering existing national and international best practices is a key process in creating a usable methodology for electronic document archiving. Within its first year, E-ARK achieved all its planned milestones, submitting all eight contracted public deliverables to the European Commission. These deliverables already dealt with topics such as overview of the current situation of the digital archiving best practices, SIP, AIP and DIP specifications, user requirements, several inputs for technical implementations of E-ARK tools, and many more. At the moment, the methodology of the E-ARK project is being used in several pilots in various national contexts.

Key words: the E-ARK project, digital archiving, archiving best practices

Progetto europeo di conservazione della documentazione archivistica e del sapere (E-ARK). Obiettivi e risultati: una panoramica

SINTESI

Il progetto E-ARK è cofinanziato dalla Commissione Europea nel suo Programma si sostegno alle politiche di ICT (PSP) nell'ambito del Programma quadro per lo spirito di competizione ed innovazione (CIP). Scopo del progetto E-ARK è quello di fornire accesso efficiente ai flussi di lavoro relativi alle tre maggiori attività di un archivio - l'acquisizione, la conservazione e l'abilitazione al riuso dll'informazione. Raccogliere le esistenti buone pratiche nazionali ed internazionali è un processo chiave nella creazione di una metodologia utilizzabile per l'archiviazione elettronica. Nel corso del suo primo anno di vita, E-RK ha raggiunto tutti gli obiettivi prefissatisi, presentando tutti gli otto contratti pubblici finali alla Commissione Europea. Questi hanno trattato di argomenti quali la situazione attuale delle buone pratiche della digitalizzazione archivistica, delle specifiche SP, AIP e DIPdei bisogni dell'utenza, di molti suggerimenti per i miglioramenti tecnici degli strumenti di E-ARK, e parecchi altri. Attualmente la metodologia di E-ARK è utilizzata in molti progetti pilota all'interno di vari contesti nazionali.

Parole chiave: progetto E-ARK, archiviazione digitale, buone pratiche archivistiche

Projekt E-ARK (European Archival Records and Knowledge Preservation): pregled ciljev in dosežkov

IZVLEČEK

Projekt E-ARK je nastal pod okriljem Evropske komisije in je del Okvirnega programa za konkurenčnost, in sicer Programa za podporo politiki informacijskih in komunikacijskih tehnologij. Cilj projekta E-ARK je zagotoviti učinkovit proces dela pri treh poglavitnih arhivskih dejavnostih: pri pridobivanju, ohranjanju in ponovni uporabi arhivskega gradiva. Cilj projekta je na obstoječih nacionalnih in mednarodnih dobrih praksah ustvariti uporabno metodo za arhiviranje elektronskega gradiva. V prvem letu delovanja je projektu E-ARK uspelo doseči vse zastavljene cilje in Evropski komisiji predati osem delovnih poročil. Na projektu so bile tako med drugim obdelane obstoječe dobre prakse elektronskega arhiviranja, specificirane so bile lastnosti informacijskih paketov, ki se pojavljajo v procesu elektronskega arhiviranja, uporabniške zahteve in tehnične zahteve za izdelavo E-ARK orodij za arhiviranje. Metoda za arhiviranje elektronskega gradiva se znotraj projekta E-ARK testira na več pilotnih prevzemih elektronskega gradiva v različnih nacionalnih okoljih.

Ključne besede: projekt E-ARK, digitalno arhiviranje, dobra praksa arhiviranja

1 Introduction: participants and scope of the E-ARK project

E-ARK is a 3-year multinational research project co-funded by the European Commission under its ICT Policy Support Programme (PSP) within its Competitiveness and Innovation Framework Programme (CIP). It started on 1st of February 2014 and will last until 31st January 2017. Broad research aim of the project requires contribution from several different institution from different fields of work with seats across the Europe (Internet 1).

Beside archives (Archives of the Republic of Slovenia, The Danish National Archives, The national Archives of Hungary, The National Archives of Estonia, National Archival Services of Norway) and universities (University of Portsmouth, The University of Cologne, Lisbon Technical University) there are also developers, ministries and foundations collaborating on the project (Austrian Institute of Technology, The DLM Forum Foundation, The Digital Preservation Coalition, ES Solution, Magenta, KEEP Solutions, Agency for Public Services Reform, The Ministry of Finances and Public Administrations). Participants in the E-ARK project are coming from Austria, Denmark, Estonia, Germany, Great Britain, Hungary, Norway, Portugal, Slovenia, Spain and Sweden or are multinational organisations (Internet 2).

The aim of the project is to create a pan European methodology for archiving of electronic documents on the basis of existing national and international best practices for keeping borne digital records authentic and reusable in longer period of time. This methodology is currently being tested in several pilots, which are covering different sets of data to ensure the production of tools that will sufficiently cover three main archival activities: acquiring, preserving and accessing the records. The final goal is to build a pan European infrastructure for the archiving, with different legislations and tradition already been taken into account, which will cover the needs of variety of organisations and will be able to support complex data types (Internet 1).

2 Organisation of work within the E-ARK project

The E-ARK project constitutes of eight work packages, each with its own duties and responsibilities. Work done in one package most often intervenes or it is an input for another package that is why a good coordination among packages is crucial. In this paper, the overview of the E-ARK project will be made from the point of the working entities - work packages. Reason for this is to offer clear insight into a working structure of the project and present progress of the project in the most understandable way.

2.1 Work Package 1: Project Management

Because of technical complexion and geographical and organisational distribution of work project management and technical coordination are crucial in producing sustainable and integrated result. This work package among other things handles planning, quality control, risk management, administrative and financial tasks and also communicates and reports to the European Commission and the General Assembly and oversees quality and consistency of work of rest of the packages (WP01 description v2, 2013).

2.2 Work Package 2: Use cases and Pilots

The objective of this work package is to establish a general model of the E-ARK service and to identify those use cases that are relevant in the terms of service, which is to be developed by technical WPs. The task is working on the scenario in which each pilot site will be implement within their pilot and is supporting the sites during the deployment and operational phases of the project (WP2_description_draft_v09, 2013).

2.3 Work Package 3: Transfer of Records to Archives

Work package 3 covers the ingestion part of working with archival records. In general, the package is to ensure that records and their metadata can be exported from source systems, prepared for transfer and finally transferred and ingested into archival repositories. WP 3 is coming across already

existing practices and workflows for the ingest of records and has to deal with different methodologies and tools applied in different European countries. They are concentrating on collecting best practice and creating standard procedure for three areas: exporting records and metadata from source system, providing pan European SIP[1] format and creating tools that will support newly defined SIP format. The usability of newly defined tools and formats will be tested within E-ARK project pilots (Work package 3, 2013).

2.4 Work Package 4: Archival Records Preservation

This work package is defining an AIP[2] format for long-term preservation, and which also covers requirements for the EARK project pilots. Work package is providing reference models, reference architectures and reference implementations of a tool converting the WP 3's SIPs into AIPs (WP04 description_2013-05-07, 2013).

2.5 Work Package 5: Archival Records Access Service

This work package is defining an E-ARK DIP[3] format, and provides access methods and interfaces for structured and unstructured archival records. The overall objective of WP5 is also to provide standardised access to archival holdings. Work package 5 is examining the user's needs and current standards, on which a DIP format will be produced. Their aim is also a production of a tool that will allow the transformation from AIP to DIP format and tools for the search of and access to archival holdings (WP05 description, 2013).

2.6 Work Package 6: Archival Storage, Services and Integration

This work package is providing a software package and reference implementation for an e-Archiving service. The software packages should be easily put to use at institutions that have no existing archival infrastructure. Its inputs are the results from several other E-ARK work packages, namely WP3's SIPs, WP4's AIPs, and WP5's DIPs. Based on our analysis of the requirements from work packages, we will implement a service, which will allow following steps in archival workflow: data management, access and re-use of archival data and archival data storage (WP06 description_2013-04-27, 2013).

2.7 Work Package 7: Evaluation and Assessment

This work package's goal is to develop a maturity model for information governance that can be proposed to the community as a new tool. That model is also to be used to assess the initial and final levels of the pilots. Work package has to develop and define guidelines for effective information governance, supported by an information system. It also has to evaluate the impact of the solutions developed in the project on the pilots. One of its goals is also to develop a method to apply the maturity model to real life scenarios (WP7 description V4, 2013).

2.8 Work Package 8: Dissemination

The work package's aim is to position the project and explain its scope and potential impact to all identified target groups. It is promoting the use of the E-ARK project infrastructure and reach out to the potential users. The work package is establishing working relations with participants, scientific communities and individual end users, with among other dissemination of result. Their task is also to promote the value and integration of the E-ARK project and its outputs in participating countries. Last but not the least, this work package also promotes the green nature of the E-ARK project with minimising the carbon impact whenever possible (Work package 8 final, 2013).

1. Submission Information Package.

2. Archival Information Package.

3. Dissemination Information Package.

3 Achievements of the E-ARK project in the first 17 months

The main scope of E-ARK is to provide standardisation in next segments: in preparing for and transferring data to long term preservation, in defining open formats to support submission, preservation and access to archived data, access to archival data stored in long-term repositories and in using data mining and big data technologies to facilitate large scale research on archived data. From February 2014, when project kicked off, it had constant progress, with regular internal meetings and also with contributions at external events, which too reached various stakeholders and other interested public. During first 17 months some technical goals have been achieved. The E-Ark project got an extensive knowledge and understanding of best practices in pre-ingest, ingest, preservation and access to archival records, which was successfully converted into specifications for different steps of the archival workflow. The project start to developing a Hadoop-based integrated framework, some work has been done studying the legal issues, progress has been made on creating the Knowledge base, etc. However, the E-ARK project so far achieved all its planned milestones, submitting all contracted public deliverables to the European Commission (Delve, 2015).

On the level of packages work went as follows:

a. WP1, the project managers, have kept in touch with all packages through teleconferences and sometimes in a face-to-face meetings. They have regularly addressed any issues affecting any party of the project so the work could be continued without delays. During the first period several types of media have been used for communication (Cisco Webex, Sharepoint, Redmine, Github and Google Drive) (Delve, 2015).
   Work Package 1 also organised developers meeting in Vienna in May, 2015 (Aas et al., 2015).

b. WP2 defined components of the E- ARK project's framework in its first deliverable. The overarching model defined tools, interfaces, data packages and the workflow of an archival process, connecting all of it into a large scale model in which digital archiving is presented. The complete definition and presentation of these components represents the skeleton of the whole project (Delve, 2015).
   Since the need for the specification implementation is growing, WP2 is also focusing on supporting developers. Basis is a General Model with requirement specifications of WP3, WP4 and WP5, which are based on the E-ARK use cases. The pilot cards concept, with cards that will contain all the important information about the pilot for tool developers, has been introduced in May, 2015. Work package 2 has also been addressing the legal issues arising from EU directives and regulations (Aas et al., 2015).

c. WP3 submitted two deliverables in the first year (Report on Available Best Practice, which dealt with best practice and available tools and was done in collaboration with WP4 and WP5 and E- ARK SIP Draft Specification). They have dealt with requirements for the exporting of the records (data selection, extraction etc.). During the work, the need for one common specification for all information packages arose - the document is still being written. Also, the pre-ingest and ingest workflows have been drawn. A lot of attention has been paid to understanding and specifying the information necessary for submission agreements (agreement between producer and archive) (Delve, 2015).
   Two subgroups have been set: one focusing on requirement for exporting from ERMS, based on MoReq2010 and the other one on requirements for using SIARD for exporting (non-ERMS) databases (Aas et al., 2015).

d. WP4 also submitted two deliverables, one dealt with available AIP formats and restrictions, other one with AIP specification (Report on available formats and restrictions and E-ARK AIP draft specification, respectively). The work on specification of SIP to AIP conversion has been done in collaboration with WP6 and it is ongoing (Delve, 2015).

e. WP5 has submitted deliverable GAP report between requirements for access and current access solutions, which prime focus was on archival access service and user access needs. Finding out that legislative issues and insufficient, non-user-friendly existing solutions lead the work of the package in defining workflow, DIP structure, use cases for the archival access

services in second deliverable E-ARK DIP Draft Specification. This deliverable is an input for technical implementation of E-ARK access tools (Delve, 2015).

f. WP6 has created integrated development environment using a number of software tools such as "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 (Delve, 2015).
Other focuses of work were development of the architecture for an integrated prototype for ingest, data management and access; continuing contributions to the development of a search facility for WP6 have resulted in an updated search prototype which supports now searching within and across information packages as well as across complex objects such as office documents and PDF files. Initial development of a method for interfacing with the access components from WP5 has started. The work package is continuing the development of the Lily ingest workflow to enable ingest of complete E-ARK packages, as compared to the ingest of single documents (e.g. PDF or HTML files) (Aas et al., 2015).

g. 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 focuses on the harmonizing the processes being used in the E-ARK project (ingest, archival preservation and dissemination). Work package also developed a Vocabulary Manager (part of Knowledge centre, which is to be presented at the end of the project) to manage terminology used in the project. The tool supports harmonizing of terms by, for example, identifying identical or similar terms (Delve, 2015).
This package has been working on an assessments process for the pilots in the project, sending out a questionnaire based on the maturity model. They have also been working on the Knowledge Centre, an information system designed to harmonize best practice, standards and other appropriate references for information governance (Aas et al. 2015).

h. WP8 has given ten presentations at third-party events, including iPRES, PASIG and ICA, and DLM Forum's Triennial Conference. Web activity of the project is attracting almost 500 hits per month from more than 350 unique visitors, of whom 70% are first- time visitors. Work package has established online monthly newsletter and daily twitter feed. Twitter feed has 164 (and rising) followers and around 100 subscribers to e-mail notifications about project (Delve, 2015).

The E-ARK project has also turned to the outside partners and activities. In May 2015, the process of including the results and outcomes of the E-ARK project into two strands of ISO standardisation has started. First, ISO JTC1/SC34 and TC46/SC4 have created a common Joint Working Group, which is evaluating long-term aspects of EPUB 3 format. The E-ARK SIP format is being considered as the main candidate for encapsulating data in EPUB format for transfer to appropriate long-term preservation locations. Second, ISO TC46/SC4 has formed a Working Group for the making of an ISO standard on "Data Exchange Protocol for Interoperability and Preservation". The established Working Group includes members from E-ARK. The goal is to synchronise the standardisation effort with the E-ARK project's SIP specification (Aas et al., 2015).

4 The E-ARK project's short term goals (Aas et al., 2015)

Until the autumn of 2015 the main objective of the E-ARK project is going to be the delivery of first versions of E-ARK tools for data preparation, transfer, preservation and access. The information package specifications, workflows and requirements for developing tools will also be updated. By the end of September 2015 the project expects larger review and testing phase. All packages will continue with their activities[4]:

a. WP 1 will continue the management of the project.

b. WP 2 will concentrate on Legal Issues Study, delivery, which has to be submitted in July

4. The paper was submitted on the 20th 2015.

2015. Pilot sites and other work packages are to provide input and comments on pilot cards so WP 2 can improve them. They will also add some more information about pilots on E-ARK website and start working on Detailed Pilot Requirements document, which has to be finished by March 2016. They are also responsible for updating the General Model, especially with data important for tool developers.

c. WP 3 will start with development of following tools: an open source relational database export tool based on the Database Preservation Toolkit, data export tool for the Alfresco Content Management System and E-ARK SIP creation tool based on the ESSArch Tools (ET). The development of data model for SIP and AIP EMRS entities and their relationships will continue, as will specifying details of the metadata elements that will be supported by the E-ARK ERMS SIP profile, including specific decisions on which metadata standards will be used and supported. WP3 will continue with harmonizing SIARD, SIARDDK and DBML techniques into an open archival relational database format for E-ARK SIPs.

d. WP4 will work on the AIP Format Specification. They are to introduce support for segmented AIPs for very large digital objects. A prototype tool for SIP to AIP conversion will be implemented in the context of the integrated prototype described in a previous Newsletter. Results from this implementation will be integrated with the EPP software from ESS in order to make it available for individually customized for the on-site pilots. The OLAP conversion technique will continue to be developed using the capabilities offered by the dp-preservation toolkit created by KEEPS.

e. WP5 will be occupied with two tasks. The first of these is updating the E-ARK Common Specification for Information Packages with additional details for DIPs and access activities. The second task is to review and refine access tool requirements. This work done during the summer is going to be an input for the developers, so they can initiate a development of the access toolkit. A secondary goal in this period is to create an overview of the access component of the reference implementation being maintained and developed by WP6.

f. WP6 will focus on developing a prototype for integrated ingest, data management and access services, including refinement of the technical architecture. They have to finalise deliverable D6.1 Faceted Query Interface and API[5], which has to be delivered by the end of July. Technical work will concentrate on automating the current ingest process and techniques for managing duplicates and versions.

g. WP7 will carry on the work on the E-ARK knowledge base, which has to be delivered in January 2016. WP7 will also finalise the assessment process for the E-ARK pilots. The results of the assessment will be analysed in following months until September.

h. WP8 will continue to update communication channels and carry out dissemination. They will expand activities to reflect the growing availability of project tools and services in order to attract our wider stakeholder community to examine and test our tools. WP8 will also support WP2 in local publicity with the partners who are running our pilots.

5 Conclusion

The first 17 months of the E-ARK project have been a success. The movement towards the main goal of the project, providing standardisation in archival workflow (the pre-ingest, ingest, preservation and access to archival records) started with gathering information about the best practices and needs of relevant stakeholders. Work packages 3, 4, 5 added an input on specific areas that are covered by each WP respectively. Based on the information gathered, work on creating common specification begun. WP3, 4 and 5 have created a common Requirement Template, which is filled up describing all of aspects of tool standardisation (high level and detail process, use cases, requirements etc.) (Delve, 2015).

Another goal of the E-ARK project, to establish Knowledge Centre Service was also achieved. The first versions of Information Maturity Model and the Vocabulary Manager (the component of the

5. Application Programming Interface.

Knowledge Centre) were successfully developed. On the part of external communication the E-ARK project was able to spread the knowledge about the project and to connect with different communities. The important achievement of the E-ARK project is start of the collaboration with e-government community, which allows the project to get information on most recent developments in government IT sector and meet demands and requirements of European government agencies (Delve, 2015).

The E-ARK project goals for the next six to twelve months are threefold. The project has to continue with establishing and implementing an efficient and effective framework for archival workflows covering ingest, preservation and use (within those workflow entities has to comply with existing policies, legislation, tools and standards, and has to assure efficiency of the framework). Secondly, the E-ARK project has to increase awareness of the E-ARK tools and recommendations amongst possible users. This is going to be achieved with the improvement of understanding of legislative and organizational issues, identifying relevant needs for interoperability and access, with development of the E-ARK framework within pilots and with establishing the positive impact of E-ARK outputs. Third goal set of the E-ARK project is to identify and create new opportunities for managing and using archival data content, which basically means the continuance of collaboration with all relevant stakeholders and being prepared to meet their requirement and providing open source tools and services that can be used to support the E-ARK framework (Billenness et al., 2015).

At this moment (June, 2015) work packages are finalising the appropriate specifications of requirements. Specification will provide input for the project developers, allow the creation of tools over summer and the delivery of a first set of E-ARK tools by autumn 2015 (Aas et al, 2015).

Acknowledgments

E-ARK is an EC-funded pilot action project in the Competitiveness and Innovation Programme 2007-2013, Grant Agreement no. 620998 under the Policy Support Programme.

Reference list

Aas, Kuldar, Wilson, Andrew (2015). E-ARK progress update June 2015. Available at: http://www.earkadmin.com/advisory/Data%20Provider/E-ARK%20overview%20Advisory%20Boards%20June%202015.pdf (accessed on 20. 6. 2015)

Billenness, Clive, Delve, Janet, Anderson, David, McMeekin, Sharon (2015). Annual Dissemination Strategy (Year 2). Available at: http://www.earkadmin.com/Deliverables%20submitted%20to%20EC/620998%20D_8_1_2%20eark_annual_comms_strategy%20year%202.pdf (accessed on 20. 6. 2015)

Delve, Janet (2015). Annual report. Available at: http://www.earkadmin.com/e-ark/EC%20Year%201%20Review/E-ARK%20620998%20annual%20report%202015%20integrated%20(3).pdf (accessed on 20. 6. 2015)

http://www.eark-project.com/about (accessed on 20. 6. 2015)

http://www.eark-project.com/partners-1 (accessed on 20. 6. 2015)

WP01 description v2 (2013). Available at: http://www.earkadmin.com/e-ark/Work%20Package%20Descriptions/WP1%20-%20Project%20Management/WP01%20description%20v2.docx (accessed on 20. 6. 2015)

WP2_description_draft_v09 (2013). Available at: http://www.earkadmin.com/e-ark/Work%20Package%20Descriptions/WP2%20-%20Use%20Cases%20and%20Pilots/WP2_description_draft_v09.docx (accessed on 20. 6. 2015)

Work package 3 (2013). Available at: http://www.earkadmin.com/e-ark/Work%20Package%20Descriptions/WP3%20-%20Transfer%20of%20Records%20to%20Archives/Work%20package%203.docx (accessed on 20. 6. 2015)

WP04 description_2013-05-07 (2013). Available at: http://www.earkadmin.com/e-ark/Work%20Package%20Descriptions/WP4%20-%20Archival%20Records%20Preservation/WP04%20description_2013-05-07.docx (accessed on 20. 6. 2015)

WP05 description (2013). Available at: http://www.earkadmin.com/e-ark/Work%20Package%20Descriptions/WP5%20-%20Archival%20Records%20Access%20Services/WP05%20description.docx (accessed on 20. 6. 2015)

WP06 description_2013-04-27 (2013). Available at: http://www.earkadmin.com/e-ark/Work%20Package%20

Descriptions/WP6%20-%20Archival%20Storage,%20Services%20and%20Integration/WP06%20description_2013-04-27.docx (accessed on 20. 6. 2015)

WP7 description V4 (2013). Available at: http://www.earkadmin.com/e-ark/Work%20Package%20Descriptions/WP7%20-%20Evaluation%20and%20Assessment/WP07%20description%20V4.docx (accessed on 20. 6. 2015)

Work package 8 final (2013). Available at: http://www.earkadmin.com/e-ark/Work%20Package%20Descriptions/WP8%20-%20Dissemination/Work%20package%208%20final.docx (accessed on 20. 6. 2015)

SUMMARY

The paper is introducing European Archival Records and Knowledge Preservation (E-ARK) project and presenting its goals and achievements. The aim of the project is to create a pan European methodology for archiving of electronic documents on the basis of existing national and international best practices for keeping borne digital records authentic and reusable in longer period of time. The project focus is on improving the level of interoperability and standardisation in archival practices, including the pre-ingest, ingest, preservation and access to archival records. The first 17 months showed that the work done on the project is in line with work scheduled in description of work and that the members of the eight work packages are doing their best to reach the set goals. The first 17 months of the E-ARK project have been successful, with all milestones reached and deliverables submitted.

Typology: 1.02 Review Article

Submitting date: 11.03.2015

Acceptance date: 09.04.2015