# Addressing Major Digital Archiving Challenges

Dr Janet Delve
University of Portsmouth
Eldon Building, Winston Churchill Avenue
Portsmouth, PO1 2DJ, UK
Janet.Delve@port.ac.uk

Professor David Anderson
University of Portsmouth
Eldon Building, Winston Churchill Avenue
Portsmouth, PO1 2DJ, UK
David.Anderson@port.ac.uk

Dr Andrew Wilson
University of Portsmouth
Eldon Building, Winston Churchill Avenue
Portsmouth, PO1 2DJ, UK
Andrew.Wilson@port.ac.uk

## ABSTRACT

The E-ARK project (E-ARK is funded by the European Commission's FP7 PSP) is addressing several major challenges faced by archives and institutions/researchers preparing data to send to archives. With the recent emphasis on open access, there has been a sea-change regarding discovery and archival material, so that citizens, businesses and academic researchers as well as the archives and data providers themselves can look forward to novel ways of analyzing archival data. E-ARK is half way through its three-year timespan, and has already produced some concrete solutions to real challenges in this problem space. This poster will graphically demonstrate the various challenges and show how E-ARK is meeting them now, or plans to in the future.

## General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation.

## Keywords

Digital Archives, User Survey, E-ARK, EC, ICT-PSP, Pilot, e-infrastructure, data mining, OAIS, Big Data, born-digital records, ingest, access, EDMRS, database preservation, open access, MoReq.

## 1. INTRODUCTION

By the time of iPres 2015, the E-ARK project will have gone past the halfway point, and have started producing key results and drafts that are relevant to the digital archiving problem space. It is timely to share these results with the wider digital archiving community.

Our focus in essence: in the first year we carried out a suite of best practice surveys to see how various communities carry out basic digital archival tasks: preparing their archival material, storing it, and subsequently accessing it. We also set up a knowledge centre proof of concept to house the expertise gathered over the life of the project. In the second year, we are developing draft standards and open source tools to perform these tasks, underpinned by a comprehensive legal study, which provides a European legislative backdrop for the work we are doing. In the third year, we will have an integrated framework with modular components that will be deployed in seven pilot instances. These will cover a range of data types, archival institutions and discovery methods. Included in this will also be a data mining showcase based on Big Data methods that have been used to help develop the framework. The pilots will be based on real use cases that can serve as exemplars for many other user communities.

## 2. ABOUT E-ARK

European Archival Records and Knowledge preservation (E-ARK) was launched in February 2014 and is a 3-year pilot project within the European Commission's ICT Policy Support Programme (PSP) Competitive and Innovation Framework (CIP) Pilot B Programme under Grant Agreement no. 620998. With 16 partners in 11 EC countries comprising end users, research institutions and systems suppliers, its objective is to provide a single, scalable, robust approach capable of meeting the needs of diverse organisations, public and private, large and small, and able to support complex data types. E-ARK will demonstrate the potential benefits for public administrations, public agencies, public services, citizens and business by providing simple, efficient access to the workflows for the three main activities of an archive - acquiring, preserving and enabling re-use of information.

E-ARK will implement a number of pilot systems in different countries addressing challenges which differ in content and scale in order to create, by the end of the project, in 2017, a suite of openly-accessible end-to-end solutions capable of integration into third-party products and which will be sustained into the future.

Our work is worldwide: the first attempt to bring together working elements of archival systems. As such it is an ambitious project which has several key features: creating standardized pre-ingest formats / specifications; expanding MoReq modules to be used as a key element of the infrastructure; using CMIS and Big Data techniques to promote new ways of access to digital archives, etc. It also addresses a wide range of users: public bodies, commercial institutions, individual citizens and researchers.

Our project will also provide a Digital Preservation Maturity Model which will enable organizations to not only assess their current performance, but also to measure improvement. More information about the project is available from our website
at www.eark-project.eu.

## 3. WHAT ARE THE KEY CHALLENGES?

Here is a list of major challenges in this domain, from a variety of perspectives and communities, with descriptions of the E-ARK approach to addressing these challenges:

- *How do I get my data out of my electronic records system (e.g. Sharepoint) and into an archive?*
  This is a major priority for E-ARK, as it can be a real headache for e.g. government departments to package their data for transfer into an archive in the manner that the archives require. Practical issues include the fact that records systems use their own hierarchical classifications, which do

not necessarily match those used by an archive. How can such compatibility problems be overcome? We have studied current best practice, and are now working on data export specifications that data producers can use to get their data out of their source systems and into the archives in an "archive-acceptable" format. These open specifications are based on the MoReq schema, and cater for a wide range of systems, including Electronic Records Management Systems (ERMSs) as well as simple file based systems. Our Work Package 3 (WP3), led by the National Archives of Estonia, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information.

- *How do I archive databases?*
  E-ARK is producing everything you need for each stage of digital archiving, and we are covering database archiving as well as the archiving of digital records. We have studied current best practice, and based on this we have produced draft specifications showing how to put data (including databases and their contents) into an archive, store them there, and then access them later for discovery and re-use. We have been working closely with the Swiss Federal Archives and the Swiss Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST), and you will find the latest version 2.0 of the Software Independent Archival of Relational Databases (SIARD) format on our website for your feedback. Our Work Package 4 (WP4), led by the Austrian Institute of Technology, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information.

- *Are there any general models or schemas that show the digital archival processes step by step?*
  We have produced a comprehensive general model that is fundamental to our entire project: it covers all the tools, processes, workflows, users etc. and specifically includes the pilot implementations (various parts of our final E-ARK system will be piloted by 7 national archives). Our Work Package 2 (WP2), led by the National Archives of Hungary, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information

- *Are there any new ways to discover archival data? Can we do complex searches or just google type searches?*
  We are looking at new ways of discovery for a wide range of data and many types of users: businesses, researchers, citizens, government departments etc. Whilst sensitive data has to be protected, we are looking for the best tools and techniques for accessing and analyzing any data that is open for discovery. We have studied current best practice in this area, and have used our findings as a basis for our developments which include data mining, Online Analytical processing (OLAP) and other advanced searching techniques. Our Work Package 5 (WP5), led by the Danish National Archives, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information. Our Work Package 6, led by the Austrian Institute of Technology, is also contributing to the advanced searches effort with a report on faceted searches.

- *What has Big data got to do with digital archiving? What is Hadoop and can we use it?*
  Big Data is a broad church, but can be said to involve powerful (fast) analysis of large volumes of varied data to produce valuable new insights with greater accuracy. Big Data has associations with open data, and cloud computing, with an emphasis on large-scale accessibility. The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing. We have developed an integrated system using a Hadoop cluster, running software such as Solr, Hive, Pig, Mamout etc. Big Data also leans heavily on previous architectures such as multi-dimensional databases and data warehousing, and we are using Big Data techniques such as data mining, data warehousing, dimensional modelling and Online Analytical Processing (OLAP) to carry out large-scale analysis of e.g. geographical data (geo-data) using Oracle Warehouse Builder and Oracle OLAP. Our Work Package 6 (WP6), led by the Austrian Institute of Technology, is spearheading the Big Data work, and they have already produced reports, conference presentations etc. with key practical information. Work Package 5 (WP5), led by the Danish National Archives, is using Big Data techniques for discovery, and they have already produced reports, conference presentations etc. with key practical information. Work Package 4 (WP4), led by the Austrian Institute of Technology, is using Big Data techniques such as dimensional modelling to archive databases, and they have already produced reports, conference presentations etc. with key practical information.

- *Are there any standards to help me archive my data properly?*
  Producing specifications and schemas forms a vital part of our work, alongside developing open source tools / workflows and frameworks. Standardizing the digital archival process across Europe and beyond should be a real help to institutions large and small, governmental, commercial or academic. We are creating our schemas to be as flexible and useful as possible – with mandatory elements that are essential to comply with best practice, and plenty of flexible options so that institutions / individuals can customize their archives in myriad different ways. Our work is not just for national archives – we do everything with regional and local archives in mind too. We have several reports dealing with standardizing issues

- *How does digital archiving vary from country to country in Europe?*
  We have a broad range of national archives taking part in E-ARK, with many more countries represented in our Archival Advisory Board. This enables us to take account of many different types of archival practice: some archives currently have no digital archives, some archives deal with everything as a database, some archives deal only with records etc. We covered current archival practice across Europe in our best practice reports in the first year of the project.

- *How does the law affect digital archiving in each European country? Are there any EC laws / directives that affect all digital archiving, and what is on the horizon in this respect?*
  These are vital considerations for E-ARK as each country needs to be able to use our outputs within their own legal framework. For this reason, we have undertaken comprehensive research to determine upcoming legislation that will affect practical digital archiving. We have a dedicated, extensive legal study which we will keep updated throughout the project.

- *Do the archives have any examples or use cases to inspire me?*

We are developing pilot cards to show how our archival partners will actually be using E-ARK in their pilot implementations. These cards will highlight the use cases for each national archive, showing why they joined the project and what benefits they expected to gain.

- *Are there any Open Source digital tools I can use? Can they be integrated? Will they fit with commercial tools / systems and existing Open Source tools / systems?*
  Our tools and platforms are all designed to be scalable and open source, so they will be suitable for your archiving needs. We have leading open source and proprietary commercial partners both in the project consortium, and on our Commercial / Technical Advisory Board, in order to ensure integration and a good fit with existing archival systems. Our aims with respect to scalability are covered in Work Package 6 (WP6).
- *Can I use something developed for a national archive in my regional archive/ local archive/ research data center?*
  Yes, this is our plan: our designs are for all archival shapes and sizes. We have representation from national and regional archives, and would also welcome input from local archives / research data centres.
- *How can I measure how well my organization is performing in terms of digital archiving? Are we beginners, or a bit further along the road?*
  We have been working on a specialized business maturity model to enable institutions to gauge their progress in this regard. All the information necessary for digital archiving, including vocabulary management, is going into a dedicated, long-term Knowledge Base, to be hosted by the DLM Forum on their website. Our Work Package 7 (WP7), led by the Instituto Superior Técnico, Lisbon, Portugal, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information.
- *I am unsure about or don't like something that E-ARK is doing. How can I make my thoughts known to them?*
  Please send us your feedback – email info@eark-project.eu!
- *Does it matter which preservation strategy I use with digital archiving? For example, can I use migration or emulation or a hybrid?*
  E-ARK is preservation-strategy neutral, and we are consciously identifying metadata (data about data) elements catering for both migration and emulation.
- *Do you have any questions for us? If so please get in touch. You can join our mailing list (http://eepurl.com/M35bH), and we are looking for more members on our Data Provider Advisory Board; and local archive members for our Archival Advisory Board (contact Andrew.Wilson@port.ac.uk).*

## 4. ABOUT THE AUTHORS

Dr Janet Delve, E-ARK Co-ordinator, University of Portsmouth.

Dr Andrew Wilson, E-ARK Senior Research Fellow and Advisory Board Co-ordinator, University of Portsmouth.

David Anderson, Professor of Digital Humanities, University of Portsmouth.