E-ARK: Harmonising pan-European archival processes to ensure continuous access to e-government records and information

Andrew WILSON [a], Miguel FERREIRA [b], Kuldar AAS [c], Anders Bo NIELSEN [d], Phillip Mike TØMMERHOLT [d],

[a] University of Brighton, UK
[b] KEEP SOLUTIONS, Portugal
[c] National Archives of Estonia
[d] National Archives of Denmark

Abstract. There has been a widespread shift to electronic ways of conducting business that has transformed existing relationships between governments, governments and citizens, and governments and business. This move to electronic interactions is supported by new business systems that streamline and automate transactions, enable integration of information and service delivery and enhance collaboration between participants. Such changes in the way government business is carried out have significant implications for how public administrations document their activities and make that information available to both government and citizens to aid future decision making and accountability. Because digital records are particularly vulnerable to technological obsolescence and media decay, ensuring future access to the information created by government is a challenging issue for all jurisdictions. This paper focus on the E-ARK project, a European endeavour to standardise and create tools for consistently transferring digital records between business systems and digital archives. The E-ARK approach has the potential to simplify and make consistent diverse approaches to solving the issue of how to transfer information between the ICT systems in use in government, and the archives charged with the responsibility for ongoing and management of the information considered to be of long-term significance.

General e-Government track (Ongoing research)

Keywords
Digital archiving, access, digital preservation, information management.

1. Introduction

The adoption by governments of increasingly sophisticated ICT has led to the widespread introduction of electronic business systems. A consequence of this shift to electronic ways of conducting government processes is the need to ensure that information managed in these systems is available and accessible for as long as necessary, often across multiple generations of technology. In many cases such information is of long-term value and is transferred to dedicated digital archives which continue offering relevant access services to citizens and government branches.

However, there is no single, widely understood and accepted approach on how valuable digital information should be transferred to digital archives, preserved and accessed for the long-term [1]. In practice, existing approaches to digital archiving mimic the traditional archiving processes designed for paper-based materials, which do not take advantage of the our current ability to mass process large volumes of digital records.

The European Commission has acknowledged the need for more standardized solutions in the area of long-term preservation and access, and has funded the E-ARK project [2] to address the problem. In co-operation with academia, national archival services and commercial systems providers, E-ARK is creating and piloting a pan-European methodology for electronic document archiving, synthesising existing national and international best practices, that will keep digital information authentic and usable over time. The methodology is being implemented in open pilots in various national contexts, using existing, near-to-market tools and services developed by project partners. This approach allows memory institutions and their clients (public- and private-sector) to assess, in an operational context, the suitability of those state-of-the-art technologies.

The objective is to provide a single approach capable of meeting the needs of diverse organisations, public and private, large and small, and able to support various forms of complex data types. E-ARK aims to demonstrate the potential benefits for public administrations, public agencies, public services, citizens and business by providing simple, efficient access to workflows for the main activities of an archive, including export from source business systems, transfer of records to the archives, preservation and enabling access and re-use. The workflows and services being developed by E-ARK will be robust and scalable.

The range of work being undertaken by E-ARK to achieve this objective is wide-ranging and ambitious, and more extensive than can be adequately described here. Accordingly, this paper focusses only on the project's approach to achieving interoperability between source systems and digital archives, specifically, to allow the extraction of information from government agency business systems and its formatting as an information package (IP) for transfer to a digital archive.

2. Background

An early phase of E-ARK involved a study of the current state of digital archiving procedures around the world with an emphasis on activities in Europe, in order to es-

tablish implementation gaps, as well needs for new tools and services. Initial desk research identified a number of reports that clarified and compared how record keeping institutions have approached the issues of digital archiving and digital preservation. The most recent and relevant studies (as at 2014 when the project started) are summarised here:

1. *Digital Preservation Services: State of the Art Analysis* [3] (2012).

   This overview of the state of the art in service provision for digital preservation and curation was carried out as part of the DC-NET project. [4] Its focus is on the areas where bridging the gap between e-Infrastructures and efficient and forward-looking digital preservation services is needed. Based on a desktop study and analysis of some 190 currently available tools and services for digital preservation, the study shows that the majority of tools are small individual tools adapted for local needs. Furthermore, the study finds that there is a lack of services which orchestrate tools and services into holistic preservation solutions. The study is a central contribution to understanding the differences in digital preservation solutions and illustrates the lack of collaboration among different tools available for solving the same tasks.

2. *Common challenges, different strategie*s [5] (2012).

   This high level study was presented as a keynote speech at the 2012 DLM Forum general meeting in Copenhagen. The paper compares strategies and approaches to digital archiving at national archives in Europe. It shows that there are significant differences in the regulatory mandate of national archives as well as vast differences in how much experience national archives have in relation to handling and preserving born-digital material. It also shows that the quantity, types, complexities and the age of digital material vary greatly between national archives. The study has played an important role in raising awareness about the differences in strategies and approaches to digital archiving in Europe.

3. *Database Archiving* [6] (2012).

   This study by the Danish National Archives, investigated and compared approaches to database archiving in Europe. The study outlines the common challenges and problem areas related to database archiving and highlights the fact that even though the majority of archives expect to preserve databases in the future, the current experience is limited.

4. *Analysis of Current Digital Preservation Policies: Archives, Libraries and Museums* [7] (2013).

   This study was undertaken by Library of Congress Junior Fellow Madeline Sheldon and examined cultural heritage institutions in both Europe and the US. The study searched for digital preservation policies, strategies or plans published on the Internet by cultural heritage institutions. The analysis identified a list of policies and bulk of the analysis focused on developing and applying a taxonomy to describe the topics covered by the documents. The identified policies were grouped into a number of categories indicating the institutional source: archives, libraries, and museums. The study does little more than identify the policies and does not go actually analyse the content of the policies.

5. *Standards and Interoperability Best Practices Report* [8] (2013)

The survey was undertaken for the Digital Cultural Heritage Roadmap for Preservation project (DHC-RP) and investigates standards, best practices, and identifiers that are in use by the Digital Cultural Heritage (DCH) sector. The report provides short descriptions and links to various types of important standards and discusses issues and challenges regarding use of these standards. The report suggests that practical tests made within the DCH-RP project have shown that previously developed e-infrastructures must be modified and/or improved in order to provide a "pan-European" solution for the DCH community.

6. *Survey on Digital Preservation* (2013) [9]

This study investigated digital preservation practices and how they are implemented at libraries and archives. The main focus was on North America, but the study included survey respondents from all over the world. The study found, amongst other things, that most organisations do digital preservation locally, but that some participate in collaborative efforts, especially related to digital repositories. The study confirms what has been concluded in other studies, i.e. that the approaches taken to digital archiving differ greatly around the word, even though the challenges are the same.

7. *SCAPE survey on preservation monitoring* [10] (2014).

In 2014 the Scalable Preservation Environments (SCAPE) Project carried out a web-based survey seeking respondents' help on understanding digital preservation incidents, threats and opportunities which are relevant to organisations, and the ways they would like to detect these threats, opportunities and incidents. The survey focus is on monitoring systems for early detection of incidents and threats and is not a general survey about digital archiving practice.

The general conclusion the project drew from these reports and surveys is that harmonising currently fragmented archival approaches across Europe is required to provide the economies of scale necessary for general adoption of end-to-end digital archiving solutions. There is a critical need for an overarching methodology addressing business and operational issues, and technical solutions for ingest, preservation and reuse.

3. Information Interoperability

Achieving interoperability between source and archival systems requires that:

- Data and metadata are in standardised formats so their subsequent use is not inhibited by system differences;
- The data and metadata, and any other information required to use the data, are combined in a single conceptual package;
- The package contains enough information to allow validation both before and after transfer to a digital archive;
- The package is constructed in such a way that its information content can be understood in the long term without reference to external systems or standards.

Ensuring that information can be easily and consistently transferred between systems with all characteristics and components intact requires a coordinated approach and agreement on standardised methods for packaging and sending information. E-ARK has approached this issue by developing a generalised E-ARK Common Specification (see 3.1 below) for how information being transferred and managed over time should be packaged to support interoperability and long-term access.

The E-ARK approach to digital archiving is based on the widely recognised OAIS Reference Model [11]. Consequently, the project follows the definitions of the main archival processes and associated conceptual information package definitions articulated in OAIS:

- Ingest - the Submission Information Package (SIP);
- Archiving - the Archival Information Package (AIP);
- Dissemination - the Dissemination Information Package (DIP).

All of these conceptual *Information Package* types have been tackled in the project and detailed technical specifications have been created for each. All of the specifications are based on the E-ARK Common Specification but extend it with specifics of the relevant processes. In this paper (see 3.2 below) we concentrate on the SIP format, which sets out the requirements for packaging information for transfer between producer systems and the archives. The interoperability encouraged by the specification also allows archives to replace repository systems as needed while remaining compatible with established ingest workflows. Secondly, vendors will be able to adapt their electronic records management systems to be compatible with the specification, allowing the creation of integrated workflows between producers and archives.

Finally, to guarantee that the integrity and authenticity of transferred information is not compromised, we need to go beyond the actual data and also consider system-specific aspects. For example, a typical real world records management system contains records arranged into aggregations, metadata relating to records and their relationships to other entities, a business classification scheme, a set of retention and disposal schedules, user access controls and definitions, information to support the retrieval of a record by a search engine and so on. All these components, which make up a specific and complete information package, must be transferred together with the data in a way that ensures the integrity, authenticity and understandability of the whole package are maintained. This need has been addressed in E-ARK by the concept of Content Types (see 3.3 below), which allow for the definition of relevant system-specific elements which need to be archived along with the data, and which ultimately are used to extend the scope of the common specification itself.

4. E-ARK Specifications

In this section we explain some of the details of the E-ARK Information Package specifications which are mentioned above, specifically the Common Specification, the Submission Information Package specification, and Content Type profiles.

4.1. The Common Specification for IPs

The backbone of archival interoperability in E-ARK is provided by the so-called Common Specification for Information Packages [12]. The OAIS compliant specifica-

tion is built on the requirements presented above and provides a unified set of rules for packaging any data and metadata into a single conceptual package which can be seamlessly transferred between systems, preserved and reused in long term. The core of the common specification is a definition of an Information Package structure (Figure 1).

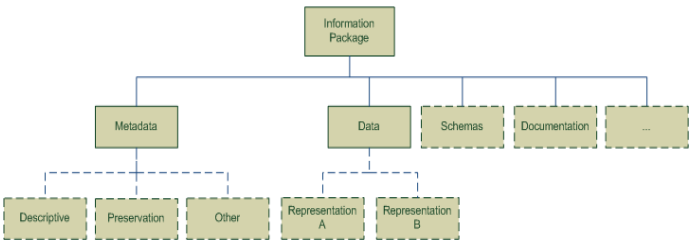

Figure 1: E-ARK Information Package structure

The structure allows for the separated inclusion of any metadata, data, relevant schemas and documentation. Further the metadata in the package can be divided into descriptive (metadata needed to find and understand the data), preservation (metadata needed to ensure the integrity and authenticity of data, metadata and the whole package) and other (any other metadata which is deemed relevant by the source system or the archives).

A specific feature of the data component is that it can be further divided into multiple representations. We cannot expect the file formats and data structures which have been originally used to last for the long term and we have to foresee the need for (potentially multiple) data migration cycles during the preservation of the information.

Lastly, to ensure that the whole package can be understood and reused in the long term users have the possibility of making the package self-sustainable by including any relevant schemas and documentation which might not be available externally in the future. Documentation may take many forms and constitutes what the OAIS calls *Representation Information,* i.e. information which cannot easily be classified as semantic or structural, e.g. software, algorithms, encryption, written instructions and many other things may be needed to understand the data.

In addition to the structure, the Common Specification details the use of core structural and packaging metadata. Essentially each package includes a core XML metadata file which follows the widely recognized METS standard [13]. The core METS metadata serves the main purpose of:

- Identifying the package and its components in a persistent and unique way;
- Providing a standardized overview of all components of the package;
- Connecting relevant pieces of data, metadata and other components to each other.

Ultimately, the METS metadata ensures that everything inside the information package can be validated according to commonly accepted rules.

4.2. Submission Information Packages (SIPs)

The first stage in the digital archiving workflow is extracting information from the producer's business system and packaging it for transfer to the archive's system. The OAIS Reference Model [11] conceptualises information submitted to an Archive as

one or more discrete transmissions of Submission Information Packages (SIPs). The E-ARK SIP specification provides a detailed description of the structure and main metadata elements that should be part of an E-ARK SIP and also functions as initial input for the technical implementations of pre-ingest and ingest tools that automate the creation and transformation of SIPs.

In its simplest form, an E-ARK SIP is a packaged set of files and folders inside a ZIP or TAR container (Figure 2). A SIP can contain one or more representations of a single intellectual entity (e.g. Rep-001 and Rep-002 under the "representations" folder in the diagram). The SIP can hold metadata that is related to the intellectual entity as a whole; at the same time each representation may also contain its own specific metadata, although separation of metadata in this way is purely optional.

In addition, as provided for in the E-ARK Common Specification, the information package folder must include a mandatory core metadata file named "METS.xml", which includes the information needed to identify and describe the structure of the package itself and the rest of its constituent components. One vital requirement for the E-ARK SIP specification is that it is able to be extended to support any content type a digital repository needs to ingest. Accordingly, the specification allows for the development of additional separate content type descriptions for different types of information being submitted to the archives.

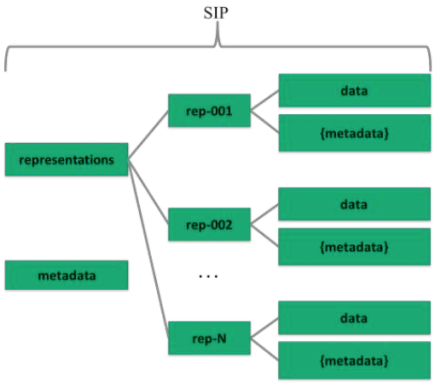

Figure 2 - E-ARK SIP structure.

4.3. Content Type-Specific SIP Profiles

As discussed above, a SIP can contain content type specific data and metadata. Types of data files and their structural relationships, and metadata elements vary for different content types. Metadata is submitted to an archive so that it can support functions in the archive. Metadata produced by a content type specific business system will variously be intended to support descriptive, structural, administrative, technical, preservation, provenance (relating to authenticity) and rights (relating to IP, retention and access) functions.

The METS standard used in the E-ARK SIP specification does not offer one single structure in which content type specific metadata could be stored as a whole. In order to efficiently use metadata to support archival functions, the SIP defines separate METS sections as containers for the various metadata functions, such as the METS header for package management, the <dmdSec> for Encoded Archival Descriptions

(EAD) [14] and other descriptive metadata standards, and the <amdSec> for preservation (PREMIS) [15], technical and other functions. In order to use the submitted metadata, it has to be extracted into the standards used in the SIP METS sections. To do this the content type specific metadata elements need to be mapped to those METS sections and implemented using the agreed standards. Therefore, complementary SIP metadata profiles are needed for the key E-ARK content types to define how the submitted content-specific metadata should be mapped to the E-ARK SIP structure. E-ARK has developed two specific content type SIP profiles:

- *SMURF Profile* [16] (Semantically Marked-Up Record Format), which contains mappings for both electronic records management systems (based on MoReq2010 [17]) and for simple file-system based records. The SMURF profile specifies further how to archive the necessary elements of an ERMS system, including the classification scheme, aggregations and classes, disposal schedules, and user access controls.
- *Relational Database Profile* which is based on the SIARD format [18]. SIARD is an open format developed by the Swiss Federal Archives. The format is designed for archiving relational databases in a vendor-neutral form. The format proposes a common standard for describing core elements of the live DBMS: data; structure; stored procedures; triggers; views; and queries. A new version of SIARD (i.e. SIARD 2.0) has been developed by E-ARK in straight collaboration with the Swiss Federal Archives.

5. Content Type-Specific Tools

Constructing SIPs to conform to the content type specific profiles manually is time consuming and onerous, as well as potentially causing inefficiencies and errors in package construction. To overcome such limitations to interoperability, E-ARK is partnering with a number of developers of archival software (ES Solutions, KEEP SOLUTIONS and Magenta) to enhance the functionality of those relevant tools currently in use in E-ARK partner institutions. In particular, the project is developing open source tools that automate the construction of submission packages (SIPs) as described below.

5.1. SMURF Profile

RODA-in [19] is a specially designed tool to support the creation of E-ARK compatible SIPs ready to be submitted to an OAIS-based archival system. The tool is intended to be used by producers and archivists to create SIPs from files and folders available on the producers' local file systems.

In its newest release (version 2) the tool aims to satisfy the need for mass processing of data and quickly create thousands of SIPs with little human intervention. The tool includes features such as:

- Create, load and edit classification schemes (managed by the OAIS repository)
- automatic aggregation of files/folders into intellectual entities based on aggregation rules;
- Automatic association of metadata to SIPs;
- Drag'n'drop support for quick creation of SIPs from folders

- Support for various descriptive metadata formats (Encoded Archival Description, Dublin Core, as well as custom-tailed metadata schemas);
- Definition of metadata templates to meet the metadata profiles of each producer/Archive
- Creation of SIPs of unlimited size;
- Compatibility with BagIt [20] and E-ARK SIP formats

Furthermore, the tool is multiplatform and open-source and has been tested to work on Windows, Mac OS X and Linux.

5.2. Relational Databases Profile

SIARD, as mentioned above, is an open format for archiving relational databases. A SIARD archive is a ZIP-based package containing mostly files based on XML, SQL:1999 and Unicode. SIARD DK on the other hand is variation of the original SIARD format that has been created by the Danish National Archives to fit their specific requirements. Differences from the original SIARD format include a different hierarchy of folders and file placement specification within the package, normalisation formats, and creation of the SIARD archive as a folder instead of a ZIP file to allow distribution of large databases across multiple storage devices.

E-ARK has collaborated with the Swiss Federal Archives, the Danish National Archives and also with vendors of digital preservation services to create a new version of SIARD that should meet the requirements of all of these stakeholders. SIARD 2 is the most recent update to the SIARD format and is backward-compatible with SIARD 1 [18].

The Database Preservation Toolkit [21] is an open-source tool that allows the extraction of data from various Database Management Systems (DBMS) and the creation of a corresponding SIARD 2 preservation format. The tool also enables the transference of data from the preservation format into an active DBMS (potentially distinct from the one that was originally used to hold the data). The current version of the tool supports conversions from and to Microsoft SQL Server, MySQL, Oracle, PostgreSQL, SIARD 1, SIARD 2 and other JDBC supported systems. It also supports Microsoft Access as an input format and SIARD DK as an output format [22].

6. Conclusion and ongoing work

Ongoing access by citizens to information created by governments is a *sine qua non* of the modern world. But access and re-use of government information of long-term value depends, crucially, on ensuring the reliable and error free movement of records between government business systems and the archives charged with the responsibility of providing ongoing access to those records. Additionally, the movement of records between systems may occur many times during their lifespan and requires robust interoperability between those systems.

This paper has described the E-ARK approach on standardising and creation of freely available tools for consistently transferring digital records between business systems and digital archives. The E-ARK approach described in this paper has the potential to simplify and make consistent currently diverse approaches to solving the issue of how to transfer information between the ICT systems in use in government,

and the archives charged with the responsibility for ongoing and management of the information considered to be of long-term significance. End-users will benefit enormously from the adoption of standardised approaches to information exchange across European record keeping institutions.

The E-ARK is currently being validated by means of 7 full scale pilots that will run for a period of 9 months [23]. These pilots aim to demonstrate the suitability of the E-ARK proposed standards and tools to support current electronic archival needs covering all relevant activities from ingest to data reuse while simultaneously addressing the needs of the stakeholders involved, e.g. data producers, data subjects, data owners, data holders and data users. Pilots will integrate E-ARK tools together with systems in use in partner organisations, and provide a framework to ensure compatibility, interoperability and enhancement of current standards.

Pilots are being conducted in 6 different European organisations and focus distinct aspects of the OAIS life-cycle:

- *Pilot 1*, lead by the Danish National Archives, aims to assess E-ARK SIP creation tools and the Database Preservation Toolkit with not less than 4 databases of different sizes and complexities containing several million records;
- *Pilot 2*, lead by the National Archives of Norway (NAN), aims to demonstrate the ability to export electronic records and their metadata from Electronic document and records management systems (EDRMS) and databases of Norwegian public sector institutions, transfer and ingest them in the NAN digital repository. ESSArch Tools will be used to create E-ARK SIPs, and ESSArch Preservation Platform will be used to create and manage the E-ARK AIPs;
- *Pilot 3*, lead by the National Archives of Estonia (NAE), aims to export public records from an EDRMS of a governmental agency to the National Archives of Estonia and make these available through our own catalogue (i.e. Archival Information System, AIS) as well as provide an API for accessing the records from other systems (e.g. the original EDRMS at the agency); The whole set will include about 5000 records;
- *Pilot 4*, also lead by the National Archives of Estonia (NAE), will focus on the migration and ingest of more than 200.000 business records from bespoke business system from private companies to the digital archive of the Estonian Business Archives and their subsequent description required for archiving and preservation;
- *Pilot 5*, lead by the National Archives of Slovenia, aims to prove that the SIP and DIP implementations fulfill specific requirements of records containing geo-referenced data comprising all phases of ingest. Success criteria include the creation, verification, ingest and access to more than 1000 records with geodata layer;
- *Pilot 6*, lead by KEEP SOLUTIONS, aims to demonstrate that the E-ARK SIP is adequate to support the media types found in today's EDRMS and, that the most adequate and scalable form of transference of data between producers and archives is to automate the SIP creation and delivery process by means of specially developed interoperability components. Success criteria include the ingest of no less that 900 historical records

automatically packaged via a custom developed integration component adding up to 1.2 TB of data.

- *Pilot 7*, lead by the National Archives of Hungary, aims to extract structured content from an Oracle database and examine the applicability of data-warehouse concepts in an archival environment in order to maintain both the original structure and intellectual interpretability of ingested data and to enhance reuse by providing advanced analytics on original data. The resulting prototype will be a user-friendly web-based application.

Finally, pilot organisations will be assessed according to a capability model called Information Governance Maturity Model. The model ensures that the assessment focuses on capabilities that the pilots want to achieve, and allows researchers to compare their status at the beginning and at the end of the project.

References

[1] E-ARK Project. *D3.1 E-ARK Report on Available Best Practices* (2014). Available from http://eark-project.com/resources/project- deliverables/6-d31-e-ark-report-on-available-best-practices

[2] http://www.eark-project.com

[3] Ruusalepp, R. & Dobreva, M. (2012). "Digital Preservation Services: State of the Art Analysis". Available at www.dc-net.org/getFile.php?id=467

[4] See http://www.dc-net.org

[5] Kristmar, K. V.. "Common challenges, different strategies". DLM Forum General Meeting Copenhagen 2012.

[6] Velle, K. "Database Archiving". Available at https://www.sa.dk/media(4588,1033)/EBNA-Minutes,_CPH_29-30_May_2012.pdf. Danish National Archives 2012.

[7] Sheldon, M.. "Analysis of Current Digital Preservation Policies: Archives, Libraries and Museums". Library of Congress 2012. Available at http://www.digitalpreservation.gov/documents/Analysis%20of%20Current%20Digital%20Preservation%20Policies.pdf?loclr=blogsig

[8] Justrell B., Toller E. "Standards and interoperability best practice report". Digital Cultural Heritage Roadmap for Preservation project (DHC-RP) 2013. Available at http://www.dch-rp.eu/getFile.php?id=165

[9] Bergin, M. B. (2013). "Sabbatical Report: Summary of Survey Results on Digital Preservation Practices at 148 Institutions". Available at http://works.bepress.com/cgi/viewcontent.cgi?article=1012&context=meghan_banach

[10] Faria L, Duretec K., Kulmukhametov A., Moldrup-Dalum P., Medjkoune L., Pop R., Barton S., Akbik A. "SCAPE survey on preservation monitoring". SCAPE Project 2014. Available at http://www.scape-project.eu/wp-content/uploads/2014/05/SCAPE_D12.2_KEEPS_V1.0.pdf

[11] CCSDS. *Open Archival Information System Reference Model (OAIS)*, 2012. Available from http://public.ccsds.org/publications/archive/650x0m2.pdf

[12] http://www.eark-project.com/resources/specificationdocs/50-draftcommonspec-1

[13] http://www.loc.gov/standards/mets/

[14] https://www.loc.gov/ead/

[15] http://www.loc.gov/standards/premis/

[16] http://eark-project.com/resources/project-deliverables/52-d33smurf

[17] http://www.moreq.info/index.php/specification

[18] http://www.eark-project.com/resources/specificationdocs/32-specification-for-siard-format-v20

[19] http://rodain.roda-community.org

[20] https://tools.ietf.org/html/draft-kunze-bagit-06

[21] Database Preservation Toolkit: http://www.database-preservation.com/

[22] Ramalho, J. C., Faria, L., Silva, H. Coutada, M. Database Preservation Toolkit: a flexible tool to normalize and give access to databases. *DLM Forum – 7th Triennial Conference: Making the Information Governance Landscape in Europe*. Lisbon Portugal 2014.

[23] Alföldi and I. Réthy, "Detailed Pilots Specification," 2016. Available from http://www.eark-project.com/resources/project-deliverables/60-23pilotsspec/file