DELIVERABLE

Project Acronym: **E-ARK**

Grant Agreement Number: **620998**

Project Title: European Archival Records and Knowledge Preservation

DELIVERABLE DETAILS

DELIVERABLE REFERENCE NO.	D2.4
DELIVERABLE TITLE	Pilot Documentation
REVISION	1.0

PRINCIPAL AUTHOR(S) Name(s)	Organisation(s)
István Alföldi István Réthy	National Archives of Hungary (NAH
REVIEWER(S) **Name(s)**	**Organisation(s)**
Clive Billenness Janet Delve Kuldar Aas	University of Brighton (UoB) The National Archives of Estonia (NAE)

Project co-funded by the European Commission within the ICT Policy Support Programme Dissemination Level		
P	Public	X
C	Confidential, only for members of the Consortium and the Commission Services	

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Submitted Revisions History

Revision No.	Date	Authors(s)	Organisation	Description
0.1	10/10/16	István Alföldi	National Archives of Hungary (NAH)	Draft document structure
0.2	24/10/16	István Rethy	National Archives of Hungary (NAH)	First version
0.3	27/10/16	István Alföldi	National Archives of Hungary (NAH)	Draft
0.4	04/11/16	István Alföldi	National Archives of Hungary (NAH)	Final version
1.0	7/11/16	Clive Billenness	University of Brighton (UoB)	English language check and formatting

Acknowledgements of additional material

Contributions are acknowledged from

- Anders Bo Nielsen (Danish National Archives)
- Phillip Mike Tømmerholt (Danish National Archives)
- Alex Thirifays (Danish National Archives)
- Hans Fredrik Berg (National Archive of Norway)
- Terje Pettersen-Dahl (National Archive of Norway)
- Arne-Kristian Groven (National Archive of Norway)
- Tarvo Kärberg (National Archive of Estonia)
- Karin Oolu (National Archive of Estonia)
- Raivo Ruusalepp (Estonian Business Archive)
- Ats Rand (Estonian Business Archive)
- Gregor Završnik (National Archive of Slovenia)
- Boris Domajnko (National Archive of Slovenia)
- Joze Skofljanec (National Archive of Slovenia)
- Miguel Ferreira (KEEPS)
- Zoltán Lux (National Archives of Hungary)
- Mezei József (National Archives of Hungary)
- Clive Billenness (University of Brighton)

Statement of originality:
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

1. Executive Summary

E-ARK project

The goal of the European Archival Records and Knowledge Preservation (E-ARK) Project is to pilot archival services to keep 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. E-ARK will demonstrate the potential benefits for public administrations, public agencies, public services, citizens and business by providing easy and efficient access to the archived records.

The project brings together a core group of European national archives, four leading research institutions, three providers of archiving software solutions and services, two government agencies, and two international membership organisations that represent the communities who stand to benefit from the project: data owners/providers, archives, software vendors and solution providers.

E-ARK will, over a three year period, harmonise archival processes at a pan-European level 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.

Work Package 2 (description from DoW)

The E-ARK General Model definition is a public deliverable of Work Package 2.

The overall objective of this work package is to ensure that the scenarios implemented at 7 identified pilot sites are both realistic and relevant, that they bring together a meaningful subset at each site of the use cases in order to establish a general model of the E-ARK service.

WP2 will

- Identify specific use cases that will each be implemented in at least one pilot scenario, covering:
  - Export from business systems
  - Creation of SIPs from unstructured and structured data
  - Execution of the complete SIP -> AIP -> DIP data-flow to support migration and submission/access scenarios
  - Existing use cases for access to content in physical and virtual reading rooms (with appropriate access controls) and as web-applications
  - Additional use cases that augment the main pilot programme including short “stretch tests” and 3rd party validation
- Identify and mitigate legal and regulatory constraints.
- Provide support and advice about the operational environment of the pilot sites to the teams in WP3-6 during the planning phase (which corresponds to their main cycles of iterative (agile) design and development.
- Support the teams working at the pilot site in the planning and deployment phase
- Ensure smooth execution of the pilots.
- Document the recommended practices and lessons learned in the project knowledge base.

T2.4 Future pilot deployment (M25-M27)
The objective of this task is to finalize the pilots in harmony with the D2.1.

The Electronic Archiving Service consists of a series of activities covered by software tools and manual workflow steps. These tools are currently partly in existence, some are being developed by E-ARK project, many more are to be added by developments of the digital preservation community in the future. The role of this task is to identify the most relevant scenarios for the E-ARK Service, define which scenario which level of activity is needed in order to bridge the gap of the currently existing solutions (e.g. integration, software development, interface definition).

In order to make the E-ARK service as widely as possible to demonstrate the functionality of the service built on D2.1 from the pilot will be finalized around the pilot sites. In order to plan ahead for a pilot project previously identified three levels:

1. Full scale project pilot activities – implementation, by consortium members, of one or more scenarios at one or more locations for a period of six months or longer. Members of DLM forum and DPC will receive details of the pilot implementation and be invited to participate as observers. There are seven full scale pilots.

2. Additional project pilot activities – implementation, by consortium members of shorter 'stretch' pilots that extend the scenarios or apply them in different contexts. This may include the participation of members of DLM Forum and DPC who are directly not members of the E-ARK consortium

3. External validation activities – implementation of project results by members of DLM Forum and DPC as part of an extended 'Beta' program with limited involvement from consortium members. Outcome of this task is the high-level requirement specification of the full scale pilots and also scenarios, sites and requirements of the 2nd and 3rd level pilots.

T2.5 Support and execution of pilots. (M7-M33)
The task is concerned with the implementation of the pilots defined in D2.3. The Task Leader contributes to providing an appropriate methodological framework for all pilot for specifying the input/output points and the uniform principles applied in the different areas like source data management, user training, user documentation,
interim reports and the final reports. This way the results of the pilot sites are comparable and can be reliably proven in this E-ARK-service pilot. There are seven 6-month pilot sites running concurrently and these are defined in Section B3.2a, Approach.

This document corresponds to the deliverable:

D2.4) Pilot documentation

Pilot documentation: This package of documentation will provide technical and end-user guidance to support not only the pilot sites but also possible future deployments thereafter. [Project Month 33]

Structure of this deliverable

The deliverable is a package of linked documents.

This **Summary** contains the common information and short overview of the pilots, along with references to the final version of the **Pilot Definition** Excel files and Pilot Documentation Packages. The **Pilot Definition** files provide detailed information about scenarios, data sets and step-by-step preparation and process step instructions. The **Pilot Documentation Package** is created by the pilot staff responsible for the pilot execution. This package contains additional information along with screenshots (and videos in some cases) of the tools during the execution of the pilot.

Summary (this document) – Created by WP2

Pilot Package – Pilot 1

- Pilot Definition (Final version) – Created by WP2 and Pilot 1 Lead

- Pilot Documentation files – Created by Pilot 1

Pilot Package – Pilot 2

- Pilot Definition (Final version) – Created by WP2 and Pilot 2 Lead

- Pilot Documentation files – Created by Pilot 2

Pilot Package – Pilot 3

- Pilot Definition (Final version) – Created by WP2 and Pilot 3 Lead

- Pilot Documentation files – Created by Pilot 3

Pilot Package – Pilot 4

- Pilot Definition (Final version) – Created by WP2 and Pilot 4 Lead

- Pilot Documentation files – Created by Pilot 4

Pilot Package – Pilot 5

- Pilot Definition (Final version) – Created by WP2 and Pilot 5 Lead

- Pilot Documentation files – Created by Pilot 5

Pilot Package – Pilot 6

- Pilot Definition (Final version) – Created by WP2 and Pilot 1 Lead

- Pilot Documentation files – Created by Pilot 6

Pilot Package – Pilot 7

- Pilot Definition (Final version) – Created by WP2 and Pilot 7 Lead

- Pilot Documentation files – Created by Pilot 7

Note that this document covers pilot activities belonging to the *E-ARK Full-scale pilots (1)*. The *Additional project pilots (2)* and *External validation activities (3)* will be defined in a separate document.

2. Pilot planning

E-ARK will pilot an end-to-end OAIS-compliant e-archival service covering ingest and reuse of structured and unstructured data addressing the needs of data subjects, data owners and data users. It will integrate tools currently in use in partner organisations, and provide a framework for providers of these, and similar tools, to ensure compatibility and interoperability. The project has three phases resulting in a set of tool instantiations, a validated pilot platform and a set of recommended practices based on evaluation of the pilot. This approach supports the planned three-tier piloting strategy (full-scale pilot, shorter 'stretch' pilots and external validation)

The work has been organised into six work packages, as shown in the diagram below. Specialist skills are associated with each WP and this grouping of activities also reduces inter-dependences between work packages and localises risk. The detailed definition of the work required in each work package includes a diagrammatic 'product flow' diagram. These express the flows and dependences within and between work packages.

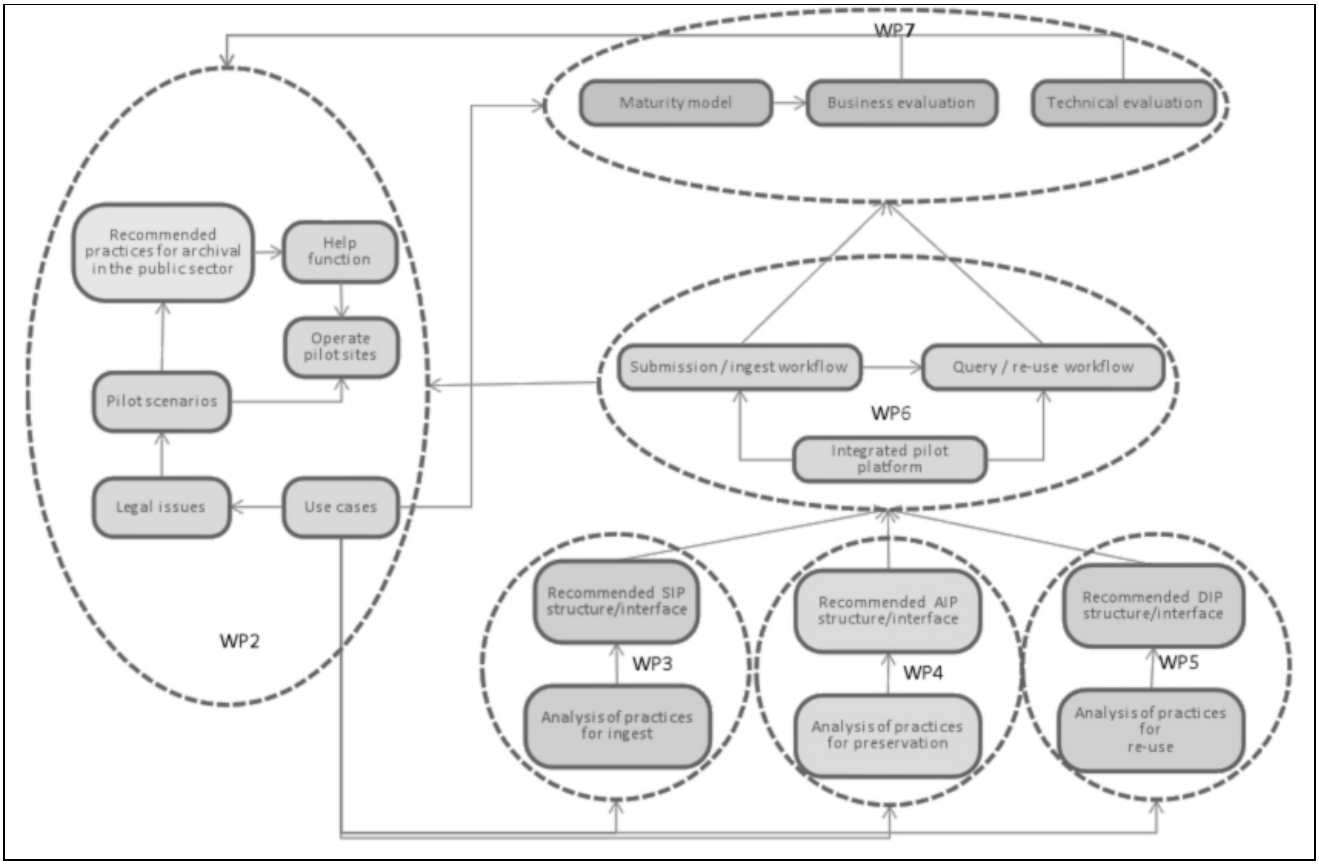


Figure 1: E-ARK – Overall Approach

WP2 is concerned with ensuring that the needs of each pilot site are addressed in the work packages that actually deploy the tools, and that the pilot scenarios are achievable and reflect any legal and logistical constraints. It also supervises the acquisition of appropriate data from the data-owners working with each pilot site and, finally, documents the knowledge gained from the pilot in the form of recommended practices.

WP3, WP4 and WP5 are responsible for the information packages that encapsulate the content and related metadata that is being archived, respectively during the workflows for **submission** (SIP - the data structures used by the data owner to enable ingestion of the content), **archival** (AIP - the data structures used by the repository operator to enable preservation functions) and **dissemination** (DIP – the data structures used for extraction and re-use of content). The mapping of SIP to AIP and AIP to DIP provide the mechanism for integration of tools/services in the pilot and compliance with these three data-structures provides the mechanism for interoperability between tools/services.

WP6 provides access to ingest and re-use tools/services to be deployed in the pilot, based on the implementation of a repository supporting the open source AIP schema from WP4. Pilot sites can either use this open-source solution or work with their platform-providers to implement SIP/AIP and AIP/DIP mappings of their own, supported through their community of interest within the project.

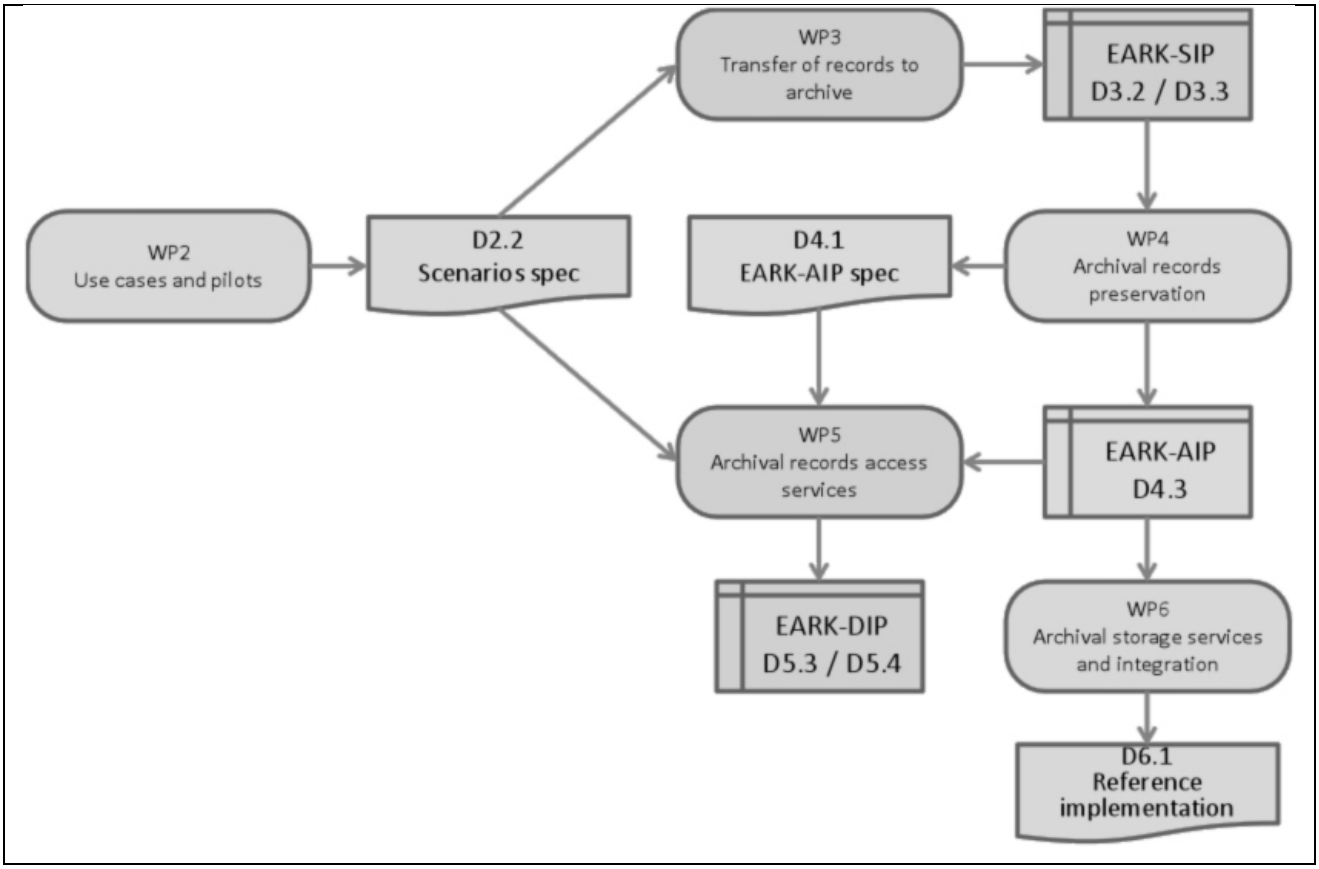


Figure 2: E-ARK Technical Integration

WP7 is responsible for evaluating the pilot service from technical and commercial perspectives based on criteria established for each scenario by WP2 and will utilise a maturity model developed in the TIMBUS project. Following the pilot deployments, both technical and business evaluations will be carried out and stored in a knowledge base, based on the indicators created for each pilot component. For example, a formal specification of the pilot ingest workflow will include information about how it has been developed and tested.

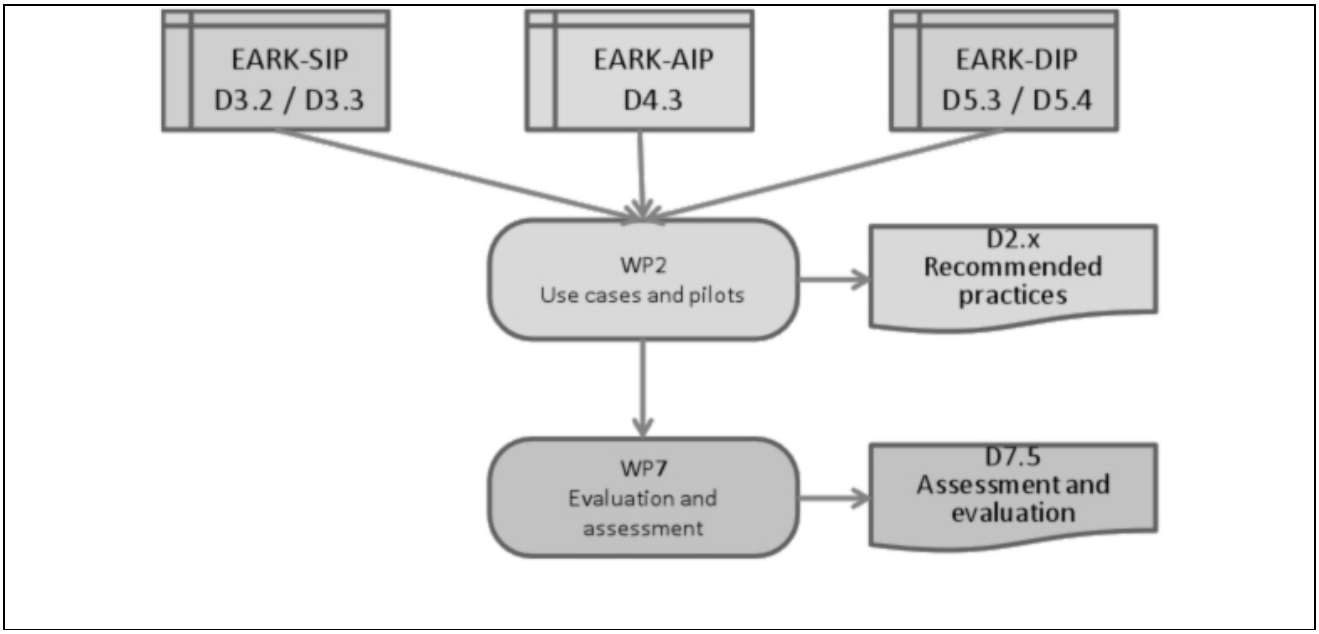


Figure 3: Pilot Workflows

More specifically, there are two distinct work-streams orchestrating the work required to integrate the pilot service and the work required to deploy, support and evaluate the pilot. This is summarised above, one leading to the WP6 deliverable for an *"Integrated Platform Reference Implementation"* (M24) and the other leading to the WP7 deliverable *"Pilots Assessment – Final"* (M36).

Piloting, which is the responsibility of WP2, consists of seven instances of parts of the E-ARK service.

2.1 The full scale pilots planned in the E-ARK Description of Work (DoW)

T2.5.1 Full scale pilot no. 1. – SIP creation of relational databases

Task leader: Danish National Archives.

Supported by: Magenta

- **Scope:** Not less than 4 databases of different sizes and complexities (one contains several million records)
- **Object:** Creating SIPs for relational databases using the tool created in WP3, T3.3: SIP Creation Tools, for further evaluation.
- **Participants:** Danish National Archives (digital archive), Magenta, the data provider institution creating the archival records.
- **Resource plan:** 8 person months for setting up the pilot (assisting the archivists and data provider in preparing the transfer), carrying out the pilot (transfer, quality checking, metadata amendments), testing the results and reporting.
- **Timeframe:** M28-M33
- **Preconditions:** M03.3 and M03.4
- **Position in the project:** DNA will pilot SIP creation and ingest specified by WP3
- **Contribution to the project outcome:** the pilot demonstrates the applicability of the project outcomes in creating SIPs from relational databases

T2.5.2 Full scale pilot no. 2. – SIP creation and ingest of records

Task leader: National Archives of Norway

The main part of the pilot includes the export of electronic records and their metadata from EDRM systems and databases of Norwegian public sector institutions, transfer and ingest them to the NAN digital repository.

- **Scope:** Not less than 2 transfers of unstructured records with mixed restricted and unrestricted material, and not less than 1 transfer of structured records.
- **Object:** Extract data from EDRMS and databases, create SIPs for structured and unstructured records using ESSArch Tools, ingest the SIPs to the repository using ESSArch Preservation Platform, for further evaluation.
- **Participants:** National Archives of Norway (digital archive), data provider
- **Resource plan:** 6 person months for setting up the pilot (assisting the archivists and data provider in preparing the transfer), carrying out the pilot (transfer, quality checking, metadata amendments), testing the results and reporting
- **Position in the project:** NAN will pilot SIP creation and ingest specified by WP3
- **Timeframe:** M28-M33
- **Preconditions:** M03.3 and M03.4
- **Contribution to the project outcome:** the pilot demonstrates the applicability of ESSArch Tools and the ingest functions of ESSArch Preservation Platform.
- **Data owners:** to be defined at the time of the pilot.
- **Platform:** ESSArch Tools will be used to create the SIPs, and ESSArch Preservation Platform will be used to create and manage the AIPs, both delivered by ES Solutions. NAN IT-department is responsible for the systems operation.

T2.5.3 Full scale pilot no. 3. – Ingest from government agencies

Task leader: National Archives of Estonia

The main part of the proposed pilot includes the export of electronic records and their metadata from EDRM systems of Estonian public sector institutions, transfer and ingest to the NAE digital repository.

In addition Estonian agencies have the responsibility to make public electronic records with no access restrictions available on their web sites, which means that the pilot will also enable this through standardised linking/access methods that are implemented in the agencies' digital infrastructure / web site.

- **Scope:** export public records from an EDRM system 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 (the original EDRMS at the agency); The whole set will include about 5000 records (but depends on the exact agency of course).
- **Objects:** EDRMS at a governmental agency (Alfresco), records preparation tool (UAM), digital preservation and access systems (SDB, AIS);
- **Participants:** National Archives of Estonia (digital archive), one governmental agency (data provider), general public (access to records);
- **Number of users:** Archivists at NAE (dealing with the ingest and preservation, about 3 persons); archivists at the agency (about 2-3 persons preparing the export/transfer and providing means for continuous in-house usage), general public - we have around 1000 daily users at the archives virtual reading room / AIS but obviously we are not able to predict how many of these will actually access and use the information ingested through the pilot;
- **Resource plan:** about 4 person months (includes updates to the EDRMS installation at the agency, to UAM and SDB/AIS, setting up and running the pilot).

- **Position in the project:** NAE will implement and pilot the records export requirements, SIP format and transfer-ingest workflow specified by WP3 and the access services specified by WP5;
- **Timeframe:** setting up pilot sites through M25 – M27, running the pilot for six months through M28 – M33, which means that the records are available for the general public for at least three months;
- **Preconditions:** M03.3, M03.4, M04.2, M05.4, M05.6.Records are available at the agency in digital form and enriched with metadata; it is possible to export the records; records export, preparation, transfer, ingest and access functionalities have been updated according to project deliverables in Alfresco, UAM, SDB and AIS;
- **Contribution to the project outcome:** the pilot demonstrates the applicability of the project outcomes inside the framework of Estonian public sector legislation and the tools applied at NAE.
- **Platform and data owners:** a specific data provider has not been selected for NAE, NAE notified the Ministry of Economics and Communication (in charge for co-ordinating e-Gov and electronic records management in Estonia) and they have promised their full support when it comes to actually selecting the specific agency. We are aiming to use Alfresco as the commercial system which we ingest data FROM (there are about 10-20 agencies in Estonia who use it – so quite a few possibilities). SDB is the preservation platform which we employ to ingest data.

T2.5.4 Full scale pilot no. 4. – Business Archives

Task leader: National Archives of Estonia

Supported by: Estonian Business Archives

Estonian Business Archives, Llc. is a privately owned archiving services provider. The main client base of the company is comprised of private businesses in Estonia for archiving and preservation of both paper and digital records. The business archives pilot in the E-ARK project will focus on transfer of electronic records from private companies to the digital archive solution of the Estonian Business Archives and their subsequent description required for archiving and preservation.

- **Scope:** Transfer of business records to a digital archive solution in a business archive, quality control, enhancement of description and AIP creation.
- **Object:** bespoke business system that contains records (pilot will test an annual batch of ca 4,500 records); financial and CRM systems that contain records (pilot will test an annual batch of ca 15,000 records).
- **Participants:** Estonian Business Archives, Llc (digital archive), two private companies (data providers).
- Number of users: The archived business records are for the sole use of their owner-company only.
- **Resource plan:** 4 person months for setting up the pilot (assisting the companies' archivists in preparing the transfer; setting up and configuring the IT infrastructure at EBA), carrying out the pilot (transfer, quality checking, metadata amendments, AIP creation), testing the results and reporting.
- **Position in the project:** The pilot will report on the suitability of the ES Tools and ES Preservation Platform for processing electronic records from business systems.
- **Timeframe:** M25-M27: setting up the pilot sites; M28-M31: running the pilots; M32-M33: testing and reporting.
- **Preconditions:** M03.3, M03.4, M04.2, M05.4, M05.6.
- **Contribution to the project outcome:** The business archives pilot will provide a view how the tools developed by the project can be implemented in the private sector setting. The pilot will assess to what extent these tools add value to the existing archiving services and workflows established in the corporate sector. The nature of objects used in the pilot – business information systems that contain or manage records – is slightly different from the public sector use cases that mostly rely on EDRM systems or databases of records.
- **Platform and data owners:** The systems that records will be transferred from and the current digital archive solution at the EBA are all bespoke solutions.

T2.5.5 Full scale pilot no. 5. – Preservation and access to records with geodata

Task leader: National Archives of Slovenia.

Supported by: Danish National Archives

During the e-ARK project the standardised method for ingesting geo data will be developed. This will allow the archives to offer geodata as a selection and display criteria of records by means of integration of current state of the art tools.

- **Scope:** Pilot will prove that the SIP and DIP implementations fulfil specific requirements for the records containing GIS data, test the instructions (for the producer and for the archive) regarding all phases of ingest, to prove that the archival use of GIS data is possible (via open data method, direct access in the archives and use GIS data as search criteria in the DIP contents).
- **Object:** pilot report with recommendations about urgent improvements and possible future improvements support for WP6 & WP7 setting up the work environment of selected E-ARK archival tools provide real life examples how the project deliverables can be used
- **Position in the project:** Pilot will prove usability of specification and tools for supporting ingest (WP3 D03.3) and access (WP5 D5.3, D5.4) of archival records with specific data. Uses specifications and tools for supporting ingest (WP3 D03.2, D03.3) and access (WP5 D5.2, D5.3, D5.4)
- **Participants:** National Archives of Slovenia (digital archives), Danish National Archives (best practice exchange)
- **Resource plan:** 7 person months (6 pm for National Archives of Slovenia 1 pm for DNA)
- **Preconditions:** M03.3, M03.4, M04.2, M05.4, M05.6.
- **Timeframe:** M25-M27: setting up the pilot sites; M28-M31: running the pilots; M32-M33: testing and reporting.
- **Platform:** DBExport Tool

T2.5.6 Full scale pilot no. 6. – Seamless integration between a live document management system and a long-term digital archiving and preservation service

Task leader: KEEP SOLUTIONS

RODA (Repository of Authentic Digital Records) is a long-term digital repository system that implements an ingest workflow that not only validates SIPs, but also checks its contents for virus, does format identification, extracts technical metadata, and migrates file formats to more “preservable” surrogates. RODA also provides access to digital information in several forms such as search/navigate over available metadata as well as online visualisation and download of originals, preservation formats and dissemination derivatives. Administration interfaces allow back-office users to manage fonds/collections and define rules for preservation actions. All interactions between users (human and machines) and the repository are logged for security and accountability reasons. RODA ensures that ingested data is authentic by recording PREMIS metadata on all actions performed by the repository, records provenance in archival metadata standards such as ISAD(g), and ensured integrity and availability by frequently monitoring data and making sure that it has not been tampered with. More recently, RODA has been enhanced to support preservation plans developed in Plato, thus proving a full-cycle preservation environment for digital objects ensuring usability and readability of ingested data.

RODA currently supports the Digital Archiving and Preservation Service at the Portuguese National Archives. This service allows public bodies to submit digital content to the archiving service for long-term preservation. The Digital Archiving and Preservation Service takes care of the necessary procedures to keep data accessible for long periods of time (in the scale of decades). Producers have special privileges in the system, allowing them to manage their data and change the structure of their fonds/collections. Data is submitted via SIP files that need to be manually prepared by producers using an offline tool called RODA-in.

- **Scope and objectives:** The goal of this pilot is two-fold. On one hand, Keep Solutions demonstrates that the pan-European SIP structure designed in the WP3 is adequate to support the media types currently supported by RODA (i.e. relational databases, text documents, video, audio and images) and, on the other hand, that the most adequate and scalable form of ingest is to automate the SIP creation process. In order to achieve this, we will tap into a running Document Management System and, based on appraisal and selection strategy installed, we will extract, transform, aggregate and create Submission Information Packages that conform to the pan-European SIP format defined in WP3 that are ready to be ingested in RODA.
- **Participants:** In this pilot we will make use of data produced by several bodies of the Portuguese public administration. One already confirmed is a project partner, the IST. The IST is a Portuguese public university that delivers top quality higher education and engages in research, development and innovation activities. In its activities, several forms of content with high *administrative*, *legal*, *financial* and *informational* value are produced every day. During the project lifetime the IST will engage in a parallel project to re-engineer a large part of the technology that supports its administrative services, which will include the acquisition and deployment of an integrated archival system. This makes this pilot an excellent example as information assets to be ingested from the actual production systems are expected to be highly unstructured and in desperate need of preservation. Besides the IST, the consortium will also take advantage of the role that AMA plays in the structure of the Portuguese Public Administration to complement this case with more data providers.
- **Resource plan:** 7 person months. 6 PM for KEEPS for development, testing and integration and 1 PM for IST for consulting and liaison with the departments that will provide data to the pilot.
- **Position in the project:** RODA already supports preservation actions and dissemination interfaces for 5 media types. This pilot will focus on enhancing the ingest process by connecting the long-term repository to the Document Management Systems active at the data producer's location this way demonstrating SIP suitability for packaging various content types and scalability by providing a seamless ingest process that requires little or no human intervention.
- **Timeframe:** Between M25–M27 the pilot will be deployed. Between M28–M33 the ingest process will run in parallel with the SIP creation process.
- **Preconditions:** pan-European SIP format defined (WP3). RODA must be enhanced to support the new SIP format (WP3). Automatic SIP creation tool/middleware must be developed to integrate the data provider DMS with the long-term repository.
- **Contribution to the project outcome:** The pilot will demonstrate that the pan-European SIP structure designed in the WP3 is adequate to support the content types currently supported by RODA (i.e. relational databases, text documents, video, audio and images) and, on the other hand. The pilot will also demonstrate and provide a framework for automatic SIP creation and DMS-Repository interoperability showing the scalability of whole ingest process.
- **Platform and data owners:** The owner of the data in this pilot will be the IST. Multiple systems are currently in place to support document management processes, e.g. an internally developed records management system called "DOT", a commercial workflow software called eDocLink, and an archival management system called ICA-Atom. In this pilot a prioritization of existing platforms will be made to choose the ones that will be included in the pilot.

T2.5.7 Full scale pilot no. 7. – Access to databases

Task leader: National Archives of Hungary.

Supported by: Danish National Archives

NAH will extract structured content from an Oracle database with the tools developed by WP3. The pilot will examine the applicability of data-warehouse concepts in an archival environment in order to maintain both the

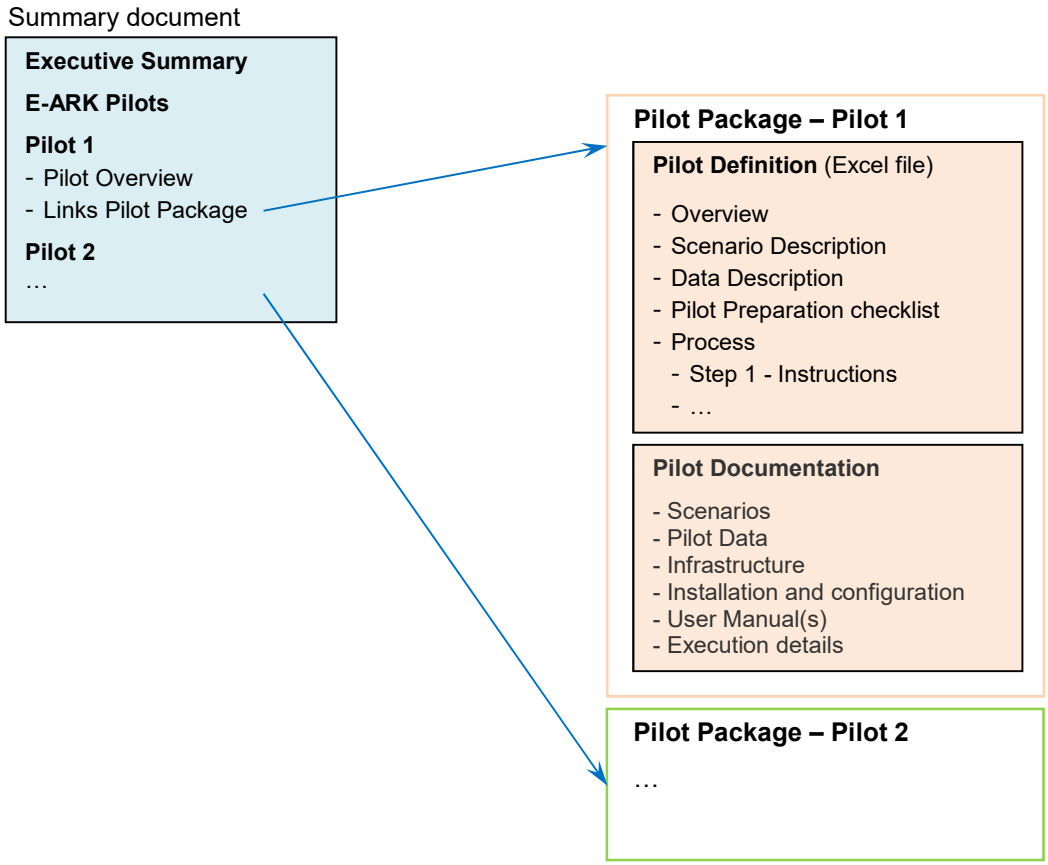
original structure and intellectual interpretability of ingested data. The working prototype for access will be a user-friendly web-based application based on the DIP specification of WP5.

- **Scope:** Representation of not less than 2 databases of different sizes and complexities with restricted and open content.
- **Objects:** Extract data from the EDRMS and the databases, create SIPs for structured and unstructured records using the ESSArch Tools, ingest the SIPs to the repository using the ESSArch Preservation Platform, for further evaluation.
- **Participants:** National Archives of Hungary (digital archives), data provider
- **Resource plan:** 6 person months for setting up the pilot (assisting the archivists and the data provider in preparing the transfer; setting up and configuring the IT infrastructure at NAH), carrying out the pilot (transfer, quality checking, metadata amendments, AIP creation), testing the results and reporting.
- **Position in the project:** NAH will primarily implement and pilot the applicability of specifications and tools related to access (WP5 D5.3, D5.4). The pilot will also prove usability of specifications and tools for supporting ingest (WP3 D03.3) of archival records.
- **Resource plan:** 7 person months (6 pm for National Archives of Slovenia 1 pm for DNA)
- **Preconditions:** M03.3, M03.4, M04.2, M05.4, M05.6.
- **Timeframe:** M25-M27: setting up the pilot sites; M28-M31: running the pilot; M32-M33: testing and reporting.
- Contribution to the project outcome
- **Data owner:** Prosecution Service of Hungary
- **Platform:** DBExport Tool, Oracle APEX, development in Java

3. Structure of this Document

The documentation corresponding to the deliverable ***D2.4 Pilot Documentation*** consists of a summary document (this document) and pilot specific documents.

- **Summary** (this document)
  The Summary contains the common information and short overview of the pilots, along with links to the final version of the Pilot Definition excel files and Pilot Documentation Packages.
- **Pilot Definition** (Final version)
  The Pilot Definition Excel spreadsheet provides detailed information about scenarios, data sets and step-by-step preparation and process step instructions.
- **Pilot Documentation Package**
  The Pilot Documentation Package is created by the pilot staff responsible for the pilot execution. This package contains additional information along with screenshots (and videos in some cases) of the tools during the execution of the pilot.



The whole D2.4 documentation consists of the following documents and folders:

Summary (this document) – Created by WP2

Pilot Package – Pilot 1
- Pilot Definition (Final version) – Created by WP2 and Pilot 1 responsible
- Pilot Documentation files – Created by Pilot 1

Pilot Package – Pilot 2
- Pilot Definition (Final version) – Created by WP2 and Pilot 2 responsible
- Pilot Documentation files – Created by Pilot 2

Pilot Package – Pilot 3
- Pilot Definition (Final version) – Created by WP2 and Pilot 3 responsible
- Pilot Documentation files – Created by Pilot 3

Pilot Package – Pilot 4
- Pilot Definition (Final version) – Created by WP2 and Pilot 4 responsible
- Pilot Documentation files – Created by Pilot 4

Pilot Package – Pilot 5
- Pilot Definition (Final version) – Created by WP2 and Pilot 5 responsible
- Pilot Documentation files – Created by Pilot 5

Pilot Package – Pilot 6
- Pilot Definition (Final version) – Created by WP2 and Pilot 1 responsible
- Pilot Documentation files – Created by Pilot 6

Pilot Package – Pilot 7
- Pilot Definition (Final version) – Created by WP2 and Pilot 7 responsible
- Pilot Documentation files – Created by Pilot 7

4. Pilot Overview

4.1 Pilots and the E-ARK General Model

OAIS Relevance

The following table provides OAIS process – Pilot cross-reference information. This table is part of the EARK General Model v2.0.

Pilot – OAIS Process cross reference table	
E-ARK	General Model v2.0

Full-scale Pilot		Pre-Ingest	Ingest	Archival Storage Preservation	Data Management	Access
Pilot 1	SIP creation of relational databases (Danish National Archives)	■				
Pilot 2	SIP creation and ingest of records (National Archives of Norway)	■	■	■		
Pilot 3	Ingest from government agencies (National Archives of Estonia)	■	■	■	■	■
Pilot 4	Business archives (National Archives of Estonia, Estonian Business Archives)	■	■	■		
Pilot 5	Preservation and access to records with geodata (National Archives of Slovenia)	■	■	■	■	■
Pilot 6	Seamless integration between a live document management system and a long-term digital archiving and preservation service (KEEP SOLUTIONS)	■	■	■	■	■
Pilot 7	Access to databases (National Archives of Hungary)	■	■	■	■	■

■ Focus of the pilot

■ Elements also used/tried within the pilot

Pilot Scenarios and E-ARK Use cases

WP3-4 (Pre-Ingest/Ingest) and WP5 (Access) have been focusing on the following business use-case scenarios in the scope of the E-ARK project:

Pre-Ingest/Ingest use-cases

- Extract and Ingest relational database based on SIARD 2.0
- Extract and Ingest ERMS records based on MoReq2010

- Extract and Ingest computer files from simple file-system – GML
- Extract and Ingest computer files from simple file-system

Access use-cases

- Access databases via Sofia (SQL)
- Access databases via SOLR (not SQL)
- Access single ERMS records via Alfresco CMS
- Access geodata via QGIS
- Access data with OLAP via oracle

The following table provides cross-reference information about scenarios defined in the full-scale pilots and the above listed E-ARK use-cases. This table is part of the EARK General Model v2.0.

Pilot Scenario – E-ARK Use-cases cross reference table	
E-ARK	**General Model v2.0**

	Pre-Ingest / Ingest					Access					
Pilot Scenario (Pilot #. Scenario #)	**Extract and Ingest relational database based on SIARD 2.0**	**Extract and Ingest ERMS records based on MoReq2010**	**Extract and Ingest computer files from simple file-system – GML**	**Extract and Ingest computer files from simple file-system**	**Other**	**Access databases via DBVTK (SQL)**	**Access databases via SOLR (not SQL)**	**Access single ERMS records via Alfresco CMS**	**Access geodata via QGIS**	**Access data with OLAP via Oracle**	**Other**
1.1 Extracting records from database	■										
1.2 Extracting records from database	■										
1.3 Extracting records from database	■										
1.4 Extracting records from database	■										
2.1 SIP Creation and Ingest of unstructured records		■									
2.2 SIP Creation and Ingest of unstructured records		■									
2.3 SIP Creation and Ingest of structured records					■						
3.1 Extract records from EDRM and ingest into Preservica		■									
3.3 Provide access to records from governmental institution through CMIS interface								■			■

4.1 SIP Creation and Ingest of business records from bespoke business system	■										
4.2 Import business records from SIARD packages to bespoke business system						■					
4.3 SIP Creation and Ingest of business records from bespoke business system	■										
4.4 Import business records from SIARD packages to bespoke business system						■					
5.1 SIP Creation and Ingest of records with Geodata			■		■						
5.2 SIP Creation and Ingest of records with Geodata									■		■
5.3 Search and Access information using Geadota			■		■						
5.4 Search and Access information using Geadota									■		■
6.1 Automatic ingest of records from a semi-active archival management system					■						
7.1 SIP Creation and Ingest of old (not normalized) database in SIARD 2.0 format	■										
7.2 SIP Creation and Ingest of unstructured files				■							
7.3 Extract SIARD Package from Preservica / E-ARK AIP (APEX/OWB access)										■	
7.4 Search and present SIARD based information with E-ARK access tools (HADOOP, HIVE Presentation)						■					
7.5 Access information from unstructured files							■				

Pilots and E-ARK Tools

The following table provides cross-reference information about scenarios defined in the full-scale pilots and E-ARK tools developed by WP3-WP5 or in WP6 as part of the integrated prototype. This table is part of the EARK General Model v2.0.

Scenarios - Tools

E-ARK **General Model v2.0**

Pilot	Scenario	Pre-Ingest							Ingest - Storage				Storage - Access																				
		Database Preservation Toolkit	Alfrsco Export Module	RODA-in	ESSArch Tool Producer (ETP)	ESSArch Tools Archive (ETA)	UAM	SIP creator (E-ARK Web)	SIP2AIP (E-ARK Web)	RODA Repository	ESSArch Preservation Platform	HDFS-Storage	Catalogue	OMT - Search and Dsiplay GUI	Order Submission Service	OMT - Order Management Tool	Lily - Ingest	ESSArch Preservation Platform	E-ARK Web (Search)	AIP2DIP (E-ARK Web)	DBPTK	IP Viewer	DB Viewer (Sofia)	ERMS Viewer (Alfresco)	Single file Viewr	QGIS	Geoserver	Peripleo	Orade (OLAP Viwer)	CMIS portal/viewer
Pilot 1 (DNA)	1. Extracting records from database	■																														
	2. Extracting records from database	■																														
	3. Extracting records from database	■																														
	4. Extracting records from database	■																														
Pilot 2 (NAN)	1. SIP Creation and Ingest of unstructured records				■	■					■																					
	2. SIP Creation and Ingest of unstructured records				■	■					■																					
	3. SIP Creation and Ingest of structured records				■	■					■																					
Pilot 3 (NAE)	3.1 Extract records from EDRM and ingest into Preservica		■				■																									
	3.2 Extract records from EDRM and ingest into Preservica		■				■																									
	3.3 Provide access to records from governmental institution through CMIS interface																							■								■
	3.4 Provide access to records from governmental institution through CMIS interface																							■								■
Pilot 4 (EBA)	4.1 SIP Creation and Ingest of business records from bespoke business system	■		■																												
Pilot 5 (SNA)	5.1 SIP Creation and Ingest of records with Geodata				■	■					■																					
	5.2 SIP Creation and Ingest of records with Geodata				■	■					■																					
	5.3 Search and Access information using Geadota												■	■	■	■	■	■	■			■				■	■	■				
	5.4 Search and Access information using Geadota									■			■	■	■	■	■	■	■			■				■	■	■				
Pilot 6 (KEEPS)	6.1 Automatic ingest of records from a semi-active archival management system									■																						
	6.2 Automatic ingest of records from a semi-active archival management system									■																						
Pilot 7 (NAH)	7.1 SIP Creation and Ingest of old (not normalized) database in SIARD 2.0 format	■		■					■			■																				
	7.2 SIP Creation and Ingest of unstructured files			■					■			■																				
	7.3 Extract SIARD Package from Preservica / E-ARK AIP (APEX/OWB access)											■					■		■	■	■		■									
	7.4 Search and present SIARD based information with E-ARK access tools											■					■		■	■	■									■		
	7.5 Access information from unstructured files											■					■		■	■		■			■							

Pilots Definition

The scenarios, data and tool usage along with pilot preparation and step-by-step process activities are defined in detail in the Pilot Definition excel documents. The final version of the Pilot Definition excel file of each pilot is part of this deliverable and can be found in the Pilot Package folders.

The Pilot Definition description follows this structure

- Pilot
  - **Scenario**
    - Business use-case (from General Model)
    - Used Information package types
    - Used E-ARK tools
    - Data Set description
      - Content description
      - Metadata description
    - Pilot preparation description and status information
    - Process description
      - Process step and low-level use-case (from General Model)
        - Used E-ARK and local tools
        - Preliminaries and start condition
        - Input/Output
        - E-ARK (and local) tools usage details

4.2 Pilots information summary

Pilot 1

Pilot 1	**SIP Creation on relational databases**		
Task leader	Danish National Archives		
Supported by	Magenta		
Scope	The scope of this Pilot is to test the E-ARK SIP Creation tool with not less than 4 databases of different sizes and complexities (one contains several million records)		
Object	Creating SIPs for relational databases using the tool created in WP3, T3.3: SIP Creation Tools, for further evaluation		
Short description	The goal of the pilot is to make four successful data extractions from live authentic databases into the SIARD 2.0 format.		
Contacts	Name (Title)	E-mail	Skype
Contact Person	Anders Bo Nielsen	abn@sa.dk	
Pilot staff member	Phillip Mike Tømmerholt	pmt@sa.dk	philliptommerholt_rigsarkivet
Pilot staff member			
Scenarios			
Scenario 1	Extracting records from database (Data Set 1)		
Description	Extracting records from database containing no documents.		
OIAS relevance	Pre-Ingest		
Use-case	Extract and Ingest relational database based on SIARD 2.0		
E-ARK specifications	SIARD 2.0		
E-ARK Tools	Database Preservation Toolkit		
Data	Health system from The Danish National Serum Institute		
Description	Database containing information from reported infectious diseases at a national level. 50-60 tables and about 90.000 records in the main table.		
Data type	Microsoft SQL Server 2008		
Metadata format	Not relevant		
Quantity	small		
Scenario 2	Extracting records from database (Data Set 2)		
Description	Extracting records from database containing no documents.		
OIAS relevance	Pre-Ingest		
Use-case	Extract and Ingest relational database based on SIARD 2.0		
E-ARK specifications	SIARD 2.0		
E-ARK Tools	Database Preservation Toolkit		
Data	Registry of Cultural Events from Kultunaut Aps		
Description	Database from the commercial company Kultunat Aps, which holds information about cultural events at a national level, from events arranged by local communities to cultural events from the Danish cultural institutions. The database contains more than 5 million records.		
Data type	MySQL		
Metadata format	Not relevant		
Quantity	large		
Scenario 3	Extracting records from database (Data Set 3)		
Description	Extracting records from database containing documents. The DNA will go to the producers site with the tool on a USB. The DNA will together with the producer use the tool and make extractions into two formats: SIARDDK and SIARD2.0.		
OIAS relevance	Pre-Ingest		
Use-case	Extract and Ingest relational database based on SIARD 2.0		
E-ARK specifications	SIARD 2.0		
E-ARK Tools	Database Preservation Toolkit		
Data	Administrative system from The Danish National Archives		
Description	Database containing information about all incoming scientific research data, and public deliveries of research data. Database containing BLOBs/documents. Size 131 gigabyte.		

Data type	Microsoft SQL Server 2008
Metadata format	Not relevant
Quantity	small
Scenario 4	Extracting records from database (Data Set 4)
Description	Extracting records from database containing documents. The DNA will go to the producers site with the tool on a USB. The DNA will together with the producer use the tool and make extractions into two formats: SIARDDK and SIARD2.0.
OIAS relevance	Pre-Ingest
Use-case	Extract and Ingest relational database based on SIARD 2.0
E-ARK specifications	SIARD 2.0
E-ARK Tools	Database Preservation Toolkit
Data	Administrative and health records system from Ministry of Higher Education and Science.
Description	Studenterrådgivningen is an institution under Ministry of Higher Education and Science, whose purpose is to provide social, psychological, and psychiatric counseling, and treatment to students in their educational situation. The database contains about 100.000 BLOBS/documents.
Data type	MS SQL Server 2008
Metadata format	Not relevant
Quantity	large

Pilot 2

Pilot 2	**SIP creation and ingest of records**		
Task leader	National Archives of Norway		
Supported by			
Scope	Not less than 2 transfers of unstructured records with mixed restricted and unrestricted material, and not less than 1 transfer of structured records.		
Object	Extract data from EDRMS and databases, create SIPs for structured and unstructured records using ESSArch Tools, ingest the SIPs to the repository using ESSArch Preservation Platform, for further evaluation		
Short description	The main part of the pilot includes the export of electronic records and their metadata from EDRM systems and databases of Norwegian public sector institutions, transfer and ingest them to the NAN digital repository.		
Contacts	Name (Title)	E-mail	Skype
Contact Person	Arne-Kristian Groven	arngro@arkivverket.no	
Pilot staff member	Terje Pettersen-Dahl	geihau@arkivverket.no	
Pilot staff member	Geir Haug	tepe@arkivverket.no	
Pilot staff member	Jørgen Ø. Vik-Strandli	jorvik@arkivverket.no	
Scenarios			
Scenario 1	SIP Creation and Ingest of unstructured records (Data Set 1)		
Description	Extract unstructured records from EDRMS based on the Norwegian NOARK 4 standard. Create SIP using ESSArch Tools. Ingest the SIP to the repository using ESSArch Preservation Platform, for further evaluation.		
OIAS relevance	Pre-Ingest, Ingest		
Use-case	Extract and Ingest ERMS records (similar to MoReq2010)		
E-ARK specifications	E-ARK-SIP		
E-ARK Tools	ESSArch Tool Producer (ETP), ESSArch Tool Archive (ETA), ESSArch Preservation Platform		
Data	Noark 4 output from EDRMS		
Description	EDRMS data from public producer converted into Noark 4 output (real production data)		
Data type	Noark 5 XML file, documents in PDF/A (or a few other specified formats), in TAR file		
Metadata format	XML: METS, PREMIS, ADDML (local)		
Quantity	20GB		
Scenario 2	SIP Creation and Ingest of unstructured records (Data Set 2)		

Description	Extract unstructured records from EDRMS based on the Norwegian NOARK 5 standard. Create SIP using ESSArch Tools. Ingest the SIP to the repository using ESSArch Preservation Platform, for further evaluation
OIAS relevance	Pre-Ingest, Ingest
Use-case	Extract and Ingest ERMS records (similar to MoReq2010)
E-ARK specifications	E-ARK-SIP
E-ARK Tools	ESSArch Tool Producer (ETP), ESSArch Tool Archive (ETA), ESSArch Preservation Platform
Data	Noark 5 output from EDRMS
Description	EDRMS data public producer converted into Noark 5 output (real production data)
Data type	Noark 5 XML file, documents in PDF/A (or a few other specified formats), in TAR file
Metadata format	XML: METS, PREMIS, ADDML (local)
Quantity	5 GB
Scenario 3	SIP Creation and Ingest of structured records (Data Set 3)
Description	Extract data from old database output, create SIPs for structured records using ESSArch Tools, ingest the SIPs to the repository using ESSArch Preservation Platform, for further evaluation.
OIAS relevance	Pre-Ingest, Ingest
Use-case	Extract and Ingest ERMS records (similar to MoReq2010)
E-ARK specifications	E-ARK-SIP
E-ARK Tools	ESSArch Tool Producer (ETP), ESSArch Tool Archive (ETA), ESSArch Preservation Platform
Data	Old database (CSV)
Description	The data set here is the national registry of licenced hunters containing data from the period 1985-1999.
Data type	CSV format (input), tar file
Metadata format	XML: METS, PREMIS, ADDML (local)
Quantity	Containing 338.500 registred persons. 105 MB

Pilot 3

Pilot 3	**Ingest from government agencies**		
Task leader	National Archives of Estonia		
Supported by			
Scope	Export public records from an EDRM system 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 (the original EDRMS at the agency); The whole set will include about 5000 records (but depends on the exact agency of course).		
Object	Native EDRMS at a governmental agency (Alfresco DELTA), records preparation tool (UAM), digital preservation and access systems (Preservica, AIS)		
Short description	The main part of the proposed pilot includes the export of electronic records and their metadata from EDRM systems of Estonian public sector institutions, transfer and ingest to the NAE digital repository. In addition Estonian agencies have the responsibility to make public electronic records with no access restrictions available on their web sites, which means that the pilot will also enable this through standardized linking/access methods that are implemented in the agencies' digital infrastructure / web site		
Contacts	Name (Title)	E-mail	Skype
Contact Person	Karin Oolu	karin.oolu@ttu.ee	karinoolu
Pilot staff member	Tarvo Kärberg	tarvo.karberg@ra.ee	tarvo.karberg
Scenarios			
Scenario 1	Extract records from EDRM (of a governmental institution), create SIP and ingest to Preservica		
Description	Export public records from an EDRM system of a governmental agency, create SIP, and ingest to the Preservica system at the National Archives of Estonia.		
OIAS relevance	Pre-Ingest, Ingest		

Use-case	Extract and Ingest ERMS records based on MoReq2010 (Alfresco is not Moreq-compliant system)
E-ARK specifications	E-ARK-SIP, SMURF
E-ARK Tools	Universal Archiving Module (UAM)
Data	Records and metadata exported from native ERMS (DELTA) Export Module at Ministry of Justice of Estonia
Description	Data set consists of different documents of Ministry of Justice from 6 series with different retention period.
Data type	ddoc, docx, PDF, TIFF
Metadata format	SMURF ERMS
Quantity	15 files
Scenario 2	Provide access to records from governmental institution through RESTful services
Description	Estonian agencies have the responsibility to make public electronic records with no access restrictions available on their web sites, which means that the pilot will also enable this through standardized linking/access methods that are implemented in the agencies' digital infrastructure / web site.
OIAS relevance	Access
Use-case	Access single ERMS records via CMIS Browser (To be consolidated with a CMIS interface access solution)
E-ARK specifications	SMURF
E-ARK Tools	CMIS Browser
Data	Records and metadata exported from native ERMS (DELTA) Export Module at Ministry of Justice of Estonia
Description	Data set consists of different documents of Ministry of Justice from 6 series with different retention period.
Data type	ddoc, docx, PDF, TIFF
Metadata format	SMURF ERMS
Quantity	15 files

Pilot 4

Pilot 4	**Business Archives**		
Task leader	National Archives of Estonia		
Supported by	Estonian Business Archives		
Scope	Pre-ingest preparation and transfer of business records to a digital archive solution in a business archive		
Object	bespoke business system that contains database records		
Short description	Estonian Business Archives, Llc. is a privately owned archiving services provider. The main client base of the company is comprised of private businesses in Estonia for archiving and preservation of both paper and digital records. The business archives pilot in the E-ARK project will focus on transfer of database records from a private company to the digital archive solution of the Estonian Business Archives.		
Contacts	Name (Title)	E-mail	Skype
Contact Person	Raivo Ruusalepp	raivo@eba.ee	raivoruu
Pilot staff member	Ats Rand	ats.rand@eba.ee	atsrand
Pilot staff member			
Scenarios			
Scenario 1	Migration and Ingest of business records from bespoke business system		
Description	Export business records from bespoke business system. Ingest to local archival system of EBA.		
OIAS relevance	Pre-Ingest, Ingest		
Use-case	Extract and Ingest relational database based on SIARD 2.0		
E-ARK specifications	E-ARK SIP, SIARD 2.0		
E-ARK Tools	Database Preservation Toolkit		

Data	Records from bespoke business system
Description	Business system with 14 tables. The database contains approximately 12 000 records.
Data type	MS-SQL as mdf
Metadata format	none
Quantity	more than 12 000 rows
Scenario 2	Extracting records from database (Data Set 2)
Description	Extracting records from database containing no documents.
OIAS relevance	Access (not DIPs involved only restoring data from SIARD packages)
Use-case	Access databases via DBVTK (SQL)
E-ARK specifications	SIARD 2.0
E-ARK Tools	Database Preservation Toolkit
Data	Records from bespoke business system
Description	Business system with 14 tables. The database contains approximately 12 000 records.
Data type	MS-SQL as mdf
Metadata format	none
Quantity	more than 12 000 rows
Scenario 3	Migration and Ingest of business records from bespoke business system
Description	Export business records from bespoke business system. Ingest to local archival system of EBA.
OIAS relevance	Pre-Ingest, Ingest
Use-case	Extract and Ingest relational database based on SIARD 2.0
E-ARK specifications	E-ARK SIP, SIARD 2.0
E-ARK Tools	Database Preservation Toolkit
Data	Records from bespoke business system
Description	Business system with 63 tables (+several history and support tables that are not needed for a complete structure of the working database). The database contains approximately 200 000 records.
Data type	MS-SQL as mdf
Metadata format	none
Quantity	more than 200 000 rows
Scenario 4	Extracting records from database (Data Set 2)
Description	Extracting records from database containing no documents.
OIAS relevance	Access (not DIPs involved only restoring data from SIARD packages)
Use-case	Access databases via DBVTK (SQL)
E-ARK specifications	SIARD 2.0
E-ARK Tools	Database Preservation Toolkit
Data	Records from bespoke business system
Description	Business system with 63 tables (+several history and support tables that are not needed for a complete structure of the working database). The database contains approximately 200 000 records.
Data type	MS-SQL as mdf
Metadata format	none
Quantity	more than 200 000 rows

Pilot 5

Pilot 5	**Preservation and access to records with geodata**
Task leader	National Archives of Slovenia
Supported by	Danish National Archives
Scope	Pilot will prove that the SIP and DIP implementations fulfill specific requirements for the records containing GIS data, test the instructions (for the producer and for the archive) regarding all phases of ingest, to prove that the archival use of GIS data is possible (via open data method,

	direct access in the archives and use GIS data as search criteria in the DIP contents).		
Object	Pilot report with recommendations about urgent improvements and possible future improvements support for WP6 & WP7 setting up the work environment of selected E-ARK archival tools provide real life examples how the project deliverables can be used		
Short description	During the e-ARK project the standardized method for ingesting geo data will be developed. This will allow the archives to offer geodata as a selection and display criteria of records by means of integration of current state of the art tools. This pilot is also supported by a video on YouTube showing the step-by-step use of the RODA tools. The Video is available at https://youtu.be/2uxD7Zn9l7M		
Contacts	Name (Title)	E-mail	Skype
Contact Person	Gregor Završnik ()	gregor.zavrsnik@gov.si	gregor.zavrsnik
Pilot staff member	Alenka Starman ()	alenka.starman@gov.si	
Pilot staff member	Anja Paulič ()	Anja.Paulic@gov.si	
Pilot staff member	Joze Skofljanec ()	joze.skofljanec@gov.si	
Scenarios			
Scenario 1	SIP Creation and Ingest of records with Geodata		
Description	Create SIP from records and metadata exported from GURS (The Surveying and Mapping Authority of the Republic of Slovenia). SIP creation and ingest of at least one small vector geodataset with less than 100 records and one with more than 1000 records. Archivist creates a Submission agreement for SIP creation, according to E-ARK guidelines for geodata SIP creation. Producer creates a SIP containing geodata, acording to Submission agreement, based on EARK SIP specifications for geodata. Archivist technicaly validates the submited SIP package, according to E-ARK guidelines for geodata SIP creation. Archivist confirms, that content validation of the submited SIP package was performed. An AIP is generated from the SIP and gets ingested into the archival repository.		
OIAS relevance	Pre-Ingest, Ingest		
Use-case	Other (SIP Creation and Ingest of records with Geodata)		
E-ARK specifications	E-ARK SIP, E-ARK AIP (with GeoData)		
E-ARK Tools	RODA-In, ESSArch Tools Archive (ETA), SIP2AIP (E-ARK Web), ESSArch Preservation Platform, EAD Editor, QGIS		
Data	Two sets from the Surveying and Mapping Authority of the Republic of Slovenia: 1.) Records and metadata of municipalities as valid until 1994, exported from GURS, database 2.) Records and metadata of administrative units until 1994, exported from GURS		
Description	Records and metadata of maps with Geodata		
Data type	GML document with metadata in XML format, ESRI Shapefile, csv		
Metadata format	ISO 19115 (INSPIRE)		
Quantity	62 records (cca. 3MB) + 1204 records (cca. 12,4 MB)		
Scenario 2	Search and Access information using Geadota		
Description	Create DIP from AIP containing record with Geodata. Present Geodata information with QGIS along with content and metadata from DIP. A data object containing geodata can be identified by using search criteria as specified by E-ARK Tool requirement specification after search index was updated from an AIP. Selected data objects are selected and order is issued. DIP is prepared according to order specification and end user credentials. DIP file structure with file descriptions (mime type, short description) is presented to the enduser. Geodata from the order can be accessed in the designated viewer (QGIS). The user checks authenticity of the DIP by accessing PREMIS documentation. Access to DIP is documented and captured metadata can be exported.		
OIAS relevance	Access		
Use-case	Other (Access of records with Geodata)		
E-ARK specifications	E-ARK AIP, E-ARK DIP (with GeoData)		
E-ARK Tools	OMT - Search and Display GUI, Order Submission Service, OMT - Order Management Tool, Lily – Ingest, ESSArch Preservation Platform, E-ARK Web (Search), AIP2DIP (E-ARK Web), IP Viewer, QGIS, Geoserver, Peripleo		

Data	Two sets from the Surveying and Mapping Authority of the Republic of Slovenia: 3.) Records and metadata of municipalities as valid until 1994, exported from GURS, database 4.) Records and metadata of administrative units until 1994, exported from GURS
Description	Records and metadata of maps with Geodata
Data type	GML document with metadata in XML format, ESRI Shapefile, csv
Metadata format	ISO 19115 (INSPIRE)
Quantity	62 records (cca. 3MB) + 1204 records (cca. 12,4 MB)
Scenario 3	SIP Creation and Ingest of records with Geodata
Description	Create SIP from records and metadata exported from ARSO (Environmental Agency of Republic of Slovenia). SIP creation and ingest of at least one vector geodata with at least 250 records. Data iz exported directly from their own system into GML format. And their sistem also exports INSPIRE metadata. Archivist creates a Submission agreement for SIP creation, according to E-ARK guidelines for geodata SIP creation. Producer creates a SIP containing geodata, acording to Submission agreement, based on EARK SIP specifications for geodata. Archivist technicaly validates the submited SIP package, according to E-ARK guidelines for geodata SIP creation. Archivist confirms, that content validation of the submited SIP package was performed. An AIP is generated from the SIP and gets ingested into the archival repository.
OIAS relevance	Pre-Ingest, Ingest
Use-case	Other (SIP Creation and Ingest of records with Geodata)
E-ARK specifications	E-ARK SIP, E-ARK AIP (with GeoData)
E-ARK Tools	ESSArch Tools Producer (ETP), ESSArch Tools Archive (ETA), ESSArch Preservation Platform, EAD Editor, QGIS
Data	Records and metadata of Natura 2000 areas created in 2004, exported from ARSO database
Description	Records and metadata of maps with Geodata
Data type	GML document with metadata in XML format, ESRI Shapefile
Metadata format	INSPIRE
Quantity	286 records (cca. 9,6 MB)
Scenario 4	Search and Access information using Geadota
Description	Create DIP from AIP containing record with Geodata. Present Geodata information with QGIS along with content and metadata from DIP. A data object containing geodata can be identified by using search criteria as specified by E-ARK Tool requirement specification after search index was updated from an AIP. Selected data objects are selected and order is issued. DIP is prepared according to order specification and end user credentials. DIP file structure with file descriptions (mime type, short description) is presented to the enduser. Geodata from the order can be accessed in the designated viewer (QGIS). The user checks authenticity of the DIP by accessing PREMIS documentation. Access to DIP is documented and captured metadata can be exported.
OIAS relevance	Access
Use-case	Other (Access of records with Geodata)
E-ARK specifications	E-ARK AIP, E-ARK DIP (with GeoData)
E-ARK Tools	OMT - Search and Display GUI, Order Submission Service, OMT - Order Management Tool, Lily – Ingest, ESSArch Preservation Platform, E-ARK Web (Search), AIP2DIP (E-ARK Web), IP Viewer, QGIS, Geoserver, Peripleo
Data	Records and metadata of Natura 2000 areas created in 2004, exported from ARSO database
Description	Records and metadata of maps with Geodata
Data type	GML document with metadata in XML format, ESRI Shapefile
Metadata format	INSPIRE
Quantity	286 records (cca. 9,6 MB)

Pilot 6

Pilot 6	**Integration between a live document management system and digital archiving and preservation service**		
Task leader	KEEP SOLUTIONS (KEEPS)		
Supported by	Instituto Superior Técnico (IST)		
Scope	The goal of this pilot is two-fold. On one hand, KEEP SOLUTIONS will demonstrate that the pan-European SIP structure designed in the WP3 is adequate to support the media types found in today's Electronic Records Management Systems (e.g. text documents, video, audio, images, etc) and, on the other hand, that the most adequate and scalable form of ingest is to automate the SIP creation and delivery process to the preservation service.		
Object	In order to achieve the goals of this pilot we will tap into two live Electronic Records Management Systems (ERMS) and, based on the appraisal and selection strategies installed, extract, transform, aggregate and create Submission Information Packages (SIP) that conform to the A1:R21-European SIP format defined in WP3. The pilot will also demonstrate the capabilities of the preservation services that follow the transfer of data to repository, namely, ingest and access by providing means to access Dissemination Information Packages from the producers Electronic Records Management Systems served by the preservation service.		
Short description	The aim of pilot 6 is to assess the efficacy of the E-ARK Information Package Specifications which defines how metadata and data should be packaged in order to move records between the three stages of records keeping - active, semi-active and inactive. On a typical setting, a record that needs to be archived usually falls into one these three “ages”: - Active - when the metadata and data are “live” being used and modified regularly. - Semi-active - when the metadata and data are archived for a short period – say up to 5 years. - Inactive - when the metadata and data are moved to a long-term repository for permanent conservation. The pilot aims to do ensure the seamless transference of information between the semi-active and the inactive stages in a way that no relevant data or metadata is lost in the process. To accomplish this goal, a special integration tool has been developed that implements the package specifications and orchestrates the entire transfer process. The pilot worked with data from a public institution whose “active” records have been initially produced and managed in an electronic records management system and then transferred to the archival service of that same institution for temporary conservation - semi-active stage. The archival service is, however, not prepared to face the challenges of long-term digital preservation, so the records that have been selected for permanent conservation need to be transferred to a long-term digital repository (the third “age”). This is where this pilot comes in. The whole goal of the pilot is to ensure that the information package specifications developed in E-ARK and the integration procedures developed are appropriate to support the transference of records between a active or semi-active archival system and a long-term preservation repository.		
Contacts	Name (Title)	E-mail	Skype
Contact Person	Miguel Ferreira	mferreira@keep.pt	jmaferreira
Pilot staff member	Luís Faria	lfaria@keep.pt	luis100
Pilot staff member	Hélder Silva	hsilva@keep.pt	hsilva_keep
Pilot staff member	Sebastien Leroux	sleroux@keep.pt	slerouxatkeep
Pilot staff member	Rui Rodrigues	rrodrigues@keep.pt	rui.tiago.mr
Pilot staff member	Ricardo Vieira	rjcv@tecnico.ulisboa.pt	ricardojoao.vieira
Pilot staff member	João Cardoso	joao.m.f.cardoso@tecnico.ulisboa.pt	joao.m.f.cardoso
Scenarios			
Scenario 1	Automatic ingest of records from a semi-active archival management system		
Description	This scenario aims to demonstrate the ability to seamlessly transfer data from a semi-active records management system to a long-term preservation repository with little or no human intervention. The scenario is based on real-world operations already in place at a public organization since mid-2015. The scenario enhances the established practice by adding an additional component to its		

	architecture that will be responsible for the long-term preservation of historical records once they reach their inactive age. The long-term preservation repository runs as a back-end service of the Archival Management System and aims to support its data curation activities.
OIAS relevance	Ingest
Use-case	Other (Ingest of Archival Management Records using the SMURF profile.)
E-ARK specifications	E-ARK SIP, E-ARK AIP
E-ARK Tools	Repository Integration Pipeline (RIP), RODA Repository
Data	Historical records
Description	Data used in this pilot scenarion was comprised of a collection of digitised books related to the Peninsular War dating from 1778 to 1834. The collection is composed of 964 records stored in a relational database following the semantic elements of EAD. The dataset also contains a total of 34.600 pages of documentation in uncompressed TIFF files at 300 dpi. The total amount of data is around 1.2 TB. This collection can be inspected at its original location at http://arquivo.cm-mafra.pt/details?id=173037.
Data type	300 dpi uncompressed TIFF files
Metadata format	EAD
Quantity	964 records described in EAD containing a total of 34.600 pages of 300 dpi uncompressed TIFF files. The total amount of data is around 1.19 TB.

Zoltan Lux	lux.zoltan@mnl.gov.hu	lux.zoltan1
József Mezei	mezei.jozsef@mnl.gov.hu	jmezei_92

Pilot 7

Pilot 7	**Access to Databases**		
Task leader	National Archives of Hungary		
Supported by	Danish National Archives		
Scope	Representation of not less than 2 databases of different sizes and complexities with restricted and open content.		
Object	Extract data from the EDRMS and the databases, create SIPs for structured and unstructured records using the ESSArch Tools, ingest the SIPs to the repository using the ESSArch Preservation Platform, for further evaluation		
Short description	NAH will extract structured content from an Oracle database with the tools developed by WP3. The pilot will 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. The working prototype for access will be a user-friendly web-based application based on the DIP specification of WP5		
Contacts	Name (Title)	E-mail	Skype
Contact Person	Zoltan Lux	lux.zoltan@mnl.gov.hu	lux.zoltan1
Pilot staff member	József Mezei	mezei.jozsef@mnl.gov.hu	jmezei_92
Scenarios			
Scenario 1	SIP Creation and Ingest of old (not normalized) database in SIARD 2.0 format		
Description	Create SIP from old (not normalized) database B25. The data is in CSV exports of DBASE files. Create both E-ARK and local SIPs and ingest them into E-ARK Web HDFS storage and Preservica archival repository. Both E-ARK and local AIPs are generated during the ingest.		
OIAS relevance	Pre-Ingest, Ingest		
Use-case	Relational database based on SIARD 2.0		
E-ARK specifications	E-ARK SIP, E-ARK AIP		
E-ARK Tools	DBPTK, RODA-In, SIP2AIP (E-ARK Web), HDFS-Storage		
Data	Hungarian Prosecution Office database		
Description	Old (not normalized) database in CSV exports of DBASE files.		
Data type	CSV files		
Metadata format	none		

Quantity	more then 300.000 cases and 500.000 name. (1,6 GB)
Scenario 2	SIP Creation and Ingest of unstructured files
Description	Create SIP from scanned documents of the Meeting minutes of the Central Coimmettee of the Hungarian Socialist Party. The image files are in PDF format with EAD metadata. Create both E-ARK and local SIPs and ingest them into B27and Preservica archival repository. Both E-ARK and local AIPs are generated during the ingest.
OIAS relevance	Pre-Ingest, Ingest
Use-case	Other (Extract and Ingest computer files from simple file-system)
E-ARK specifications	E-ARK SIP, E-ARK AIP
E-ARK Tools	RODA-In, SIP2AIP (E-ARK Web), HDFS-Storage
Data	Scanned meeting minutes of the Central Coimmettee of the Hungarian Socialist Party
Description	Scanned documents in file systems in PDF file and corresponding metadata (EAD)
Data type	PDF/JPG files (representations)
Metadata format	EAD
Quantity	123.225 files. (101 GB)
Scenario 3	Extract SIARD Package from Preservica/E-ARK AIP
Description	Access database information of the Hungarian Prosecution Office in SIARD format using APEX and OWB access. Both E-ARK and local DIPs are generated during access.
OIAS relevance	Access
Use-case	Other (Access database via APEX and Oracle BI)
E-ARK specifications	E-ARK AIP, E-ARK DIP
E-ARK Tools	HDFS-Storage , Lily – Ingest, E-ARK Web (Search), AIP2DIP (E-ARK Web), DBVTK
Data	Hungarian Prosecution Office database
Description	Old (not normalized) database in CSV exports of DBASE files.
Data type	CSV files
Metadata format	none
Quantity	more then 300.000 cases and 500.000 name. (1,6 GB)
Scenario 4	Search and present SIARD based information with E-ARK access tools
Description	Access database information of the Hungarian Prosecution Office in SIARD format using HADOOP based search and access with HIVE Lily Presentation in local environment.
OIAS relevance	Access
Use-case	Access data with OLAP via oracle
E-ARK specifications	E-ARK AIP, E-ARK DIP
E-ARK Tools	HDFS-Storage , Lily – Ingest, E-ARK Web (Search), AIP2DIP (E-ARK Web), DBVTK
Data	Hungarian Prosecution Office database
Description	Old (not normalized) database in CSV exports of DBASE files.
Data type	CSV files
Metadata format	none
Quantity	more then 300.000 cases and 500.000 name. (1,6 GB)
Scenario 5	Access information from unstructured files
Description	Create DIP from scanned documents of the Meeting minutes of the Central Coimmettee of the Hungarian Socialist Party. The image files are in PDF format with EAD metadata in E-ARK Web HDFS storage and Preservica. Create both E-ARK and local DIPs.
OIAS relevance	Access
Use-case	Access databases via SOLR (no-sql) Access data from E-ARK web / HDFS storage and from locals system. SOLR is used for search the full text index generated of the documents.
E-ARK specifications	E-ARK AIP, E-ARK DIP
E-ARK Tools	HDFS-Storage, AIP2DIP (E-ARK Web), , Lily – Ingest, E-ARK Web (Search), Single file Viewr
Data	Scanned meeting minutes of the Central Coimmettee of the Hungarian Socialist Party
Description	Scanned documents in file systems in PDF file and corresponding metadata (EAD)
Data type	PDF/JPG files (representations)
Metadata format	EAD
Quantity	123.225 files. (101 GB)