D.2.4.P7 Full-scale Pilot 7 - Access to Databases

DOI: 10.5281/zenodo.1171715

Grant Agreement Number:	620998
Project Title:	European Archival Records and Knowledge Preservation
Release Date:	12th February 2018

Contributors	
Name	Affiliation
Zoltán Lux	National Archives of Slovenia
József Mezei	National Archives of Slovenia
István Alföldi	National Archives of Hungary
David Anderson	University of Brighton
Janet Anderson	University of Brighton

Table of Contents

1. Executive Summary

This document is part of 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. [month 33] (from DoW)

Structure of this deliverable

The deliverable is a package if linked documents.

This **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. The **Pilot Definition** excel provides 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 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

2. Pilot documentation

2.1 Scenariosű

Ingest

Scenario 1: SIP Creation and Ingest of old (not normalized) database in SIARD 2.0 format

Scenario 2: SIP Creation and Ingest of unstructured files

Access

Scenario 3: Extract SIARD Package from Preservica/E-ARK AIP (APEX/Oracle BI access)

Scenario 4: Search and present SIARD based information with E-ARK access tools

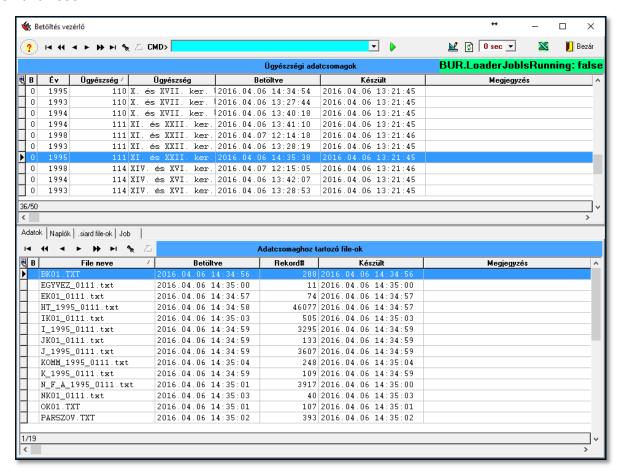**Scenario 5: Access information from unstructured files**

2.2 Datasets

Dataset 1	Hungarian Prosecution Office database
Description	Old (not normalized) database in CSV exports of DBASE files.

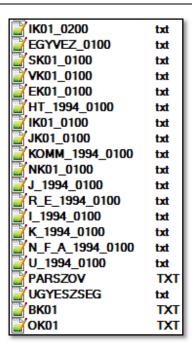|	
Data type	CSV files
Metadata format	none
Quantity	more then 300.000 cases and 500.000 name. (1,6 GB)
Dataset 2	**Scanned meeting minutes of the Central Committee of the Hungarian Socialist Party**
Description	Scanned documents in file systems in PDF file and corresponding metadata (EAD)
Data type	PDF/JPG files (representations)
Metadata format	EAD
Quantity	123.225 files. (101 GB)

2.3 Detailed scenario description

2.3.1 Scenario 1

The Hungarian Prosecution Office delivered a data collection in CSV format from their DBASE database to the Hungarian National Archives. The database contains the criminal records in Hungary between 1994 and 1999.

1 - BÜRke

2 - One Prosecution office's database

3 - A file in csv format

There were no documentation, neither to the original user program (by witch the data were maintained), nor to their DBASE database itself. In order to create an appropriate SIARD 2.0 package, we had loaded the original content into our Oracle database with BÜRke, and we added detailed documentation as comments to the database objects (tables, fields) within the database.

SIARD 2.0 files had been created of the oracle database with DBPTK created by KEEPS SOLUTIONS:

5 - Making and export with DBPTK

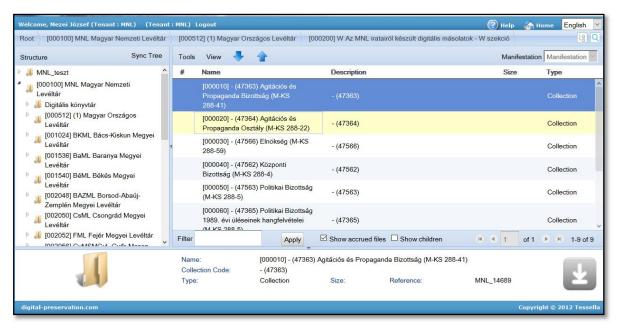E-ARK SIPs had been created, after ingested into our local E-ARK WEB and Preservica.

2.3.2   Scenario 2

3703 E-ARK SIPs had been created of the scanned meeting minutes of the Central Committee of the Hungarian Worker Party (MSZMP). The original creation time interval is 1957 to 1989. SIPs had been loaded into our local E-ARK WEB:

Package name	Process ID	Identifier	Last change	Last task	Process status
HU_MNL_OL_M-KS_288-22_APO_1960_009_13305 - IDd78e6bf1-6ec0-4543-afd1-46d91fe5bbc9	79be1c5a-373a-4374-b03d-f55b4e1bfcf6	urn:uuid:920fffa2-76c2-465b-b8c6-0738e7462775	27.10.2016 15:22:35	LilyHDFSUpload	Warning ⚠
HU_MNL_OL_M-KS_288-22_APO_1960_008_13280 - ID39ea0db0-8d79-4b8e-8916-d3c00d41b675	d8909ae4-cb1f-4a44-b5ec-f01a5c77efc5	urn:uuid:aeeca727-9d81-27e8-5cb1-4501e303211d	22.09.2016 15:09:59	AIPStore	Success ✓
HU_MNL_OL_M-KS_288-22_APO_1960_011_13411 - IDac67b397-ecb0-4b6c-a9e6-7563e4aa3826	bca812a0-4593-49a5-88e0-976de87355ae	urn:uuid:4bcf3131-0a6c-8219-30b6-16da594021af	22.09.2016 15:09:58	AIPStore	Success ✓
HU_MNL_OL_M-KS_288-22_APO_1960_009_13312 - ID743b874d-cf25-4048-8e86-2bdc0a790343	8046b564-3b8a-4e9c-9257-e52e4885f29e	urn:uuid:498cb4ac-8c50-5249-34f6-5d4336c61068	22.09.2016 15:07:42	AIPStore	Success ✓
HU_MNL_OL_M-KS_288-22_APO_1960_015_13633 - ID57251661-1661-4788-a7b4-cb22a62623e8	dec69dc6-a4b2-4a26-8114-99ba59cdbea6	urn:uuid:5674ba12-61b5-b7bd-7d66-96d1bbe2a898	22.09.2016 15:07:37	AIPStore	Success ✓
HU_MNL_OL_M-KS_288-22_APO_1960_008_13300 - ID4f0b3ff5-d7ff-47a3-8848-8300e33a2d87	7fb74ca8-7684-4257-858e-493661262785	urn:uuid:8b621d3b-56f1-2913-4e14-ceecff8c8d63	22.09.2016 15:07:35	AIPStore	Success ✓
HU_MNL_OL_M-KS_288-22_APO_1960_012_13456 - IDf87d4099-82a9-41ab-92f4-76153c895b6b	11c2fda0-332b-4e15-9da3-e69dde98e801	urn:uuid:9d94ad5d-f3f1-95d3-8dd4-d198b2daee31	22.09.2016 15:07:35	AIPStore	Success ✓
HU_MNL_OL_M-KS_288-22_APO_1960_016_13638 - ID3c27e69c-42a8-46a4-bb63-8095f6cc9639	5a97c123-f01e-445e-ad7a-cf74233f2a87	urn:uuid:9e49d4ca-33f1-f37a-e0a4-93906a8d47ef	22.09.2016 15:07:32	AIPStore	Success ✓
HU_MNL_OL_M-KS_288-22_APO_1960_010_13379 - IDd01aa5da-cad3-428d-949b-6155e745cd4c	11468f74-891a-4edc-9862-fbe145d0f230	urn:uuid:e0d04466-f7b1-0fad-6674-6417ba7862d7	22.09.2016 15:07:28	AIPStore	Success ✓
HU_MNL_OL_M-KS_288-22_APO_1960_011_13427 - IDeb8b42e8-34ae-410b-a882-e7fc0290c9ed	243ccdc7-c5ee-44a7-9f0d-489811e805f9	urn:uuid:9f09f095-a190-5b98-eba3-e7d2aada1f70	22.09.2016 15:07:27	AIPStore	Success ✓

6 - E-ARK WEB AIPs
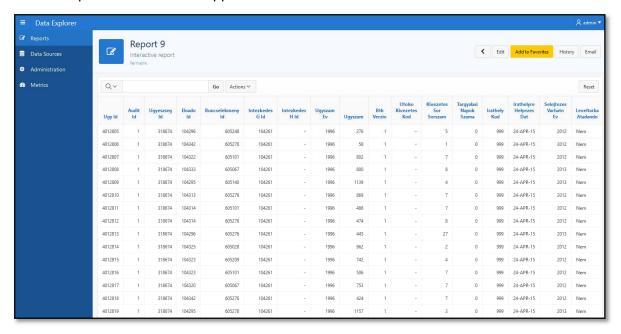
also into Preservica:

7 - **SDB Preservita AIPs**

2.3.3 Scenario 3

Access the database information of the Hungarian Prosecution Office in SIARD 2.0, format using Data Explorer (Oracle APEX). For a more logical presentation of the data if needed, Oracle Warehouse
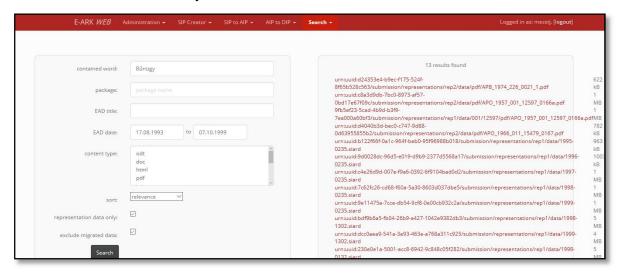
Builder can be used, for transforming the original data model into a data warehouse model, in that case APEX Data Explorer model will be applied to the Data Warehouse schema.

8 - Data Explorer (Oracle APEX)
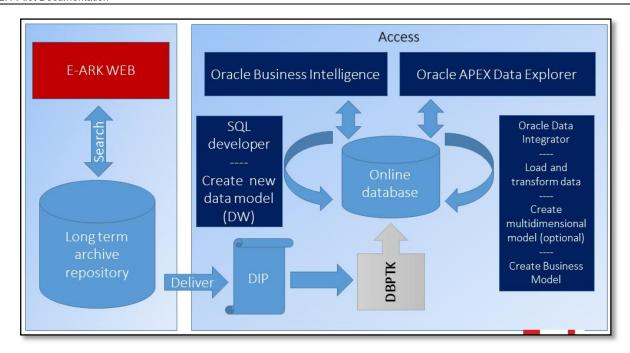
2.3.4   Scenario 4

Access database information of the Hungarian Prosecution Office in SIARD 2.0 format using HADOOP based search and access with Lily Presentation in local environment.

9 - E-ARK WEB

We load the SIARD 2.0 package into a live Oracle database by DBPTK. To presentation the data to the user, Oracle Data Integrator and Oracle Warehouse Builder is used.

By OWB the original model can be converted into a Data Warehouse model. Oracle Data Integrator enables us to design data marts in order to present the data according to specific business requirements. Using ODI, in the presentation layer, we can create a user friendly model which hides the complexity of the data (BI) model from the user and present the data in an understandable form, and also provides a broad opportunity to visualize the data.

10 - ORACLE BI

2.3.5 Scenario 5

Create DIP from scanned documents of the Meeting minutes of the Central Committee of the Hungarian Socialist Party.

11 - EARK WEB Search

The image files are in PDF format with EAD metadata in E-ARK Web HDFS storage and Preservica.

2.4 Organisations

The organizations involved in this pilot are: KEEP SOLUTIONS (creator and developer of RODA-in, DBPTK, DBVTK), AIT (creator and developer of an open source archiving and digital preservation system, E-ARK WEB), XperTeam (developer of Data Explorer, Data Warehouse, and the local Oracle BI infrastructure), Hungarian Prosecution Office

2.5 Tools

The tools involved in this pilot are: E-ARK WEB, RODA-in, DBPTK, Oracle (OLAP Viewer),

2.5.1 E-ARK WEB

E-ARK Web is an open source archiving and digital preservation system. It is OAIS-oriented which means that data ingest, archiving and dissemination functions operate on information packages bundling content and metadata in contiguous containers. The information package format uses METS to represent the structure and PREMIS to record digital provenance information.

E-ARK Web offers functionality for the three types of information packages defined in the OAIS reference model: the Submission Information Package (SIP) which is the information sent from the producer to the archive, the Archival Information Package (AIP) which is the information stored by the archive, and the Dissemination Information Package (DIP) which is the information sent to a user when requested. The system allows executing different types of actions, such as information extraction, validation, or transformation operations, on information packages to support ingesting a SIP, archiving an AIP, and creating a DIP from a set of AIPs.

E-ARK Web consists of a frontend web application together with a task execution system based on Celery which allows synchronous and asynchronous processing of information packages by means of processing units which are called "tasks".

More information: https://github.com/eark-project/earkweb

Installation manual:

2.5.2 RODA-in

RODA-in is a tool specially designed for producers and archivists to create Submission Information Packages (SIP) ready to be submitted to an Open Archival Information System (OAIS). The tool creates SIPs from files and folders available on the local file system.

In version 2 we revolutionized the way SIPs are created to satisfy the need for mass processing of data. In this version you can create thousands of valid SIPs with just a few clicks, complete with data and metadata.

More information: http://rodain.roda-community.org/

2.5.3 DBPTK

The Database Preservation Toolkit allows conversion between Database formats, including connection to live systems, for purposes of digitally preserving databases. The toolkit allows conversion of live or backed-up databases into preservation formats such as SIARD, a XML-based format created for the purpose of database preservation. The toolkit also allows conversion of the preservation formats back into live systems to allow the full functionality of databases.

This toolkit was part of the RODA project and now has been released as a project by its own due to the increasing interest on this particular feature. It is now being further developed in the EARK project together with a new version of the SIARD preservation format.

The toolkit is created as a platform that uses input and output modules. Each module supports read and/or write to a particular database format or live system. New modules can easily be added by implementation of a new interface and adding of new drivers.

More information: http://www.database-preservation.com/

2.5.4 BÜRke

Special local tool, developed to load CSV files into a pre-created Schema of a live Oracle Database (Oracle RDBMS), where the objects of the database are provided with detailed comments.

BÜRke is standalone tool. A configuration file have to be set, which contains the path of the CSV files, and the connection informations to the Oracle database.

2.5.5 Data Explorer (Oracle APEX)

Data Explorer (Oracle APEX) is a local tool developed in Oracle APEX framework. The new version of APEX enables user to create interactive reports by which user can create queries on multiple tables, using simple graphical user interface.

More information: https://apex.oracle.com/en/

Installation manual:

2.5.6 Oracle Warehouse Builder

Oracle Warehouse Builder is a single, comprehensive tool for all aspects of data integration. Warehouse Builder leverages Oracle Database to transform data into high-quality information. It provides data quality, data auditing, fully integrated relational and dimensional modeling, and full lifecycle management of data and metadata. Warehouse Builder enables you to create data warehouses, migrate data from legacy systems, consolidate data from disparate data sources, clean and transform data to provide quality information, and manage corporate metadata.

More information: https://docs.oracle.com/cd/B28359_01/owb.111/b31278/concept_overview.htm

2.5.7 Oracle Data Integrator

A widely used data integration software product, Oracle Data Integrator provides a new declarative design approach to defining data transformation and integration processes, resulting in faster and simpler development and maintenance.

More information:
https://docs.oracle.com/cd/E23943_01/integrate.1111/e12641/overview.htm#ODIGS111

2.5.8 Oracle BI

Oracle BI Enterprise Edition (sometimes simply referred to as Oracle Business Intelligence) provides a full range of business intelligence capabilities that allow you to:

- Collect up-to-date data from your organization
- Present the data in easy-to-understand formats (such as tables and graphs)
- Deliver data in a timely fashion to the employees in your organization

More information: http://docs.oracle.com/cd/E28280_01/bi.1111/e10544/getstart.htm#BIEUG11055

3. Installation manuals

Oracle services by XperTeam:
https://www.dropbox.com/sh/h2wa22gm1abzvn1/AACcq6yyR0JdJgNJylc_31kba?dl=0

E-ARK WEB by AIT: https://www.dropbox.com/s/me211o0duxt6y8p/NAH-Pilot-Installation-AIT.pdf?dl=0

3.1 Infrastructure

The infrastructure we used in this pilot to create SIPs AIPs, DIPs, and visualize data. PCs for the ingest (preparation) part, programs like RODA-in, Server for storage, management, and services, like E-ARK WEB, Oracle, DBPTK.

3.1.1 PCs

- DELL WST3500 – Windows 10 – Intel Xeon W3503 -2 core – 2.4ghz
- Fujitsu MI4W-D3171 – Windows 7 – Intel Core i3 2130 – 4 core – 3.4 ghz

3.1.2 Server

- HS22 in RAC: RedHat Linux, Oracle 12c - Xeon MP 2,64 ghz – 12 core/blade, 96GB memory