D.2.4.P4 Full-scale Pilot 4 – Business Archives

DOI: 10.5281/zenodo.1171546

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

Contributors	
Name	Affiliation
Raivo Ruusalepp	Estonian Business Archive
Ats Rand	Estonian Business Archive
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 Scenarios

Scenario 1, 2: **Exporting records from local database and importing back to database from SIARD 2.0. In total more than 12 000 business records from bespoke business system**

Scenario 3, 4: **Exporting records from client database and also importing back to database from SIARD2.0. In total more than 200 000 business records from bespoke business system**

2.2 Introduction

The purpose of Pilot 4 was to test the export functionality of business records from a bespoke business system to the digital archive system of the Estonian Business Archives (EBA). The scenario tested the Database Preservation Toolkit on legacy information from a database platform that was to be upgraded and archived its contents to the EBA.

2.3 Organisations involved

The organisations involved in pilot 4 were: Estonian Business Archives and one of the largest cell factory and microbiology laboratories in Estonia (Note: confidentiality agreement with the client is that pilot participants will remain anonymous during the pilot period, after the product testing is successfully confirmed and published, participant names are made public).

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 data provider was Cell Factory who are a one of the largest cell factory and microbiology laboratories in Estonia. Cell Factory was founded in 2005, as a spin-off of Tartu University. The company has been a client of EBA since 2005.

2.4 Software components

The main software products involved in this pilot are:

- Client side database system;
- Database preservation toolkit;
- EBA digital archive solution.

Client side database system

The system generates reports and statistical information based on analyses and medical observations. The System contains 63 tables with approximately 200 000 records with 33 491 files. The size of the database is approximately 160 GB.

Database Preservation Toolkit

Application that transforms most database files into SIARD 2.0 format which acts as a long term preservation standard for archiving databases.

EBA digital archive solution

EBA digital archive solution now has the capability of archiving databases. New services were developed using the Database Preservation Toolkit engine to import and export data from most of the widely used database solutions.

2.5 Data characterization

The data-set used in pilot 4 was:

Description: Business system with 63 tables, plus 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

Metadata format: none

Quantity: 200 000 rows

2.6 Pilot workflow

Preparation

Database preparation is needed in order to ensure archiving of the correct data. For example, that there are no ongoing procedures or daily back-up routines or other actions which can be interrupted or would corrupt data turning the export process.

Client provides access to database - this includes making the database accessible to the Database Preservation Toolkit.

Extraction

EBA IT staff and client's IT staff work together to set up proper configuration to start export data from database with Database Preservation Toolkit.

[bashCommand]

```
java
      -Dfile.encoding=UTF-8
      -jar
dbptk-app-2.0.0-beta5.jar
      --import microsoft-sql-server
      --import-server-name=localhost
      --import-database="ICO_MSSQLPROD_01"
      --import-username=sa
      --import-password="strong_long_password"
      --export siard-2
      --export-file=ico_mssqlprod_20161020.siard
```

[/bachCommand]

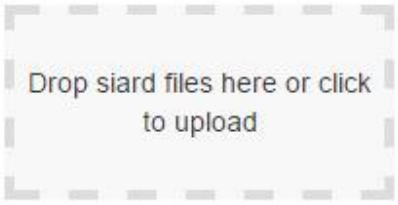
The same configuration should be used for every export instance. Any modification should be documented and made public for both sides.

Transfer

Data is transferred over HTTPS using the EBA digital archive solutions' RESTful API and web user interface to submit SIARD2.0 files to `https://api.eba.ee:8080/v1/files` endpoint:

Id	Name
1	"test_siard.siard"
2	"ico_mssqlprod_20161020.siard"
3	"ico_mssqlprod_20161020_1.siard"
4	"ico_mssqlprod_20161020_2.siard"
5	"ico_mssqlprod_20161020_3.siard"
6	"ico_mssqlprod_20161021.siard"
7	"ico_mssqlprod_20161022.siard"
8	"ico_mssqlprod_20161023.siard"
9	"ico_mssqlprod_20161024.siard"
10	"ico_mssqlprod_20161024_1.siard"
11	"ico_mssqlprod_20161024_3.siard"

Drop File:



Files:

Upload Log:

Access

Import process is started with Database Preservation Toolkit to any available database platform. Preferred database will always be initial database where data was exported from.

[bashCommand]

```
java
      -Dfile.encoding=UTF-8
      -jar
dbptk-app-2.0.0-beta5.jar
      --export microsoft-sql-server
      --export-server-name=localhost
      --export-database="ICO_MSSQLPROD_01"
      --export-username=sa
      --export-password="strong_long_password "
      --import siard-2
      --import-file=ico_mssqlprod_20161020.siard
```

[/bachCommand]

When import is successful, the data is obtained and used through Microsoft SQL Server Management studio or other management studios PL/SQL Developer, phpMyAdmin web interface, depending on database platform.

Estimated time for the pilot:

The entire process for pre-ingest with files piloted took roughly 6 hours. Time expenditure will increase significantly if a larger volume of files is processed since most of the time was spent on preparation and description activities.

2.7 Installation instructions

Since there is no installation need for Database Preservation Toolkit, there are also no specific further instructions to provide. The business sector is oriented towards simplest and most effective solutions, often just the minimal viable product. The lack of GUI made the process somewhat slower.

A note for languages that make use of umlauts, for example, in Estonian letters “ÕÄÖÜ”. In the Database Preservation Toolkit, it is prudent to use the attribute “`-Dfile.encoding=UTF-8`“ to ensure that umlauts do not go missing.