Pilot Execution Instructions and Documentation

Full-scale Pilot 1 - SIP Creation on relational databases

Version 1.0 – November 7, 2016

Final

Created by:

- Anders Bo Nielsen (DNA)
- Phillip Mike Tømmerholt (DNA)
- István Alföldi

Change Log		
Date	**Editor**	**Comments**
28/10-16	Anders Bo Nielsen	Pilot contribution version
07/11-16	István Alföldi	Final version

Table of Contents

1. EXECUTIVE SUMMARY 1

2. PILOT DOCUMENTATION 2

2.1 INTRODUCTION 2

2.2 THE DATABASES 2

2.2.1 Health system from the Danish Serum Institute 2

2.2.2 System from a private sector business Kultunaut Aps 2

2.2.3 Administrative system from The Danish National Archives 2

2.2.4 Administrative and health records system from Ministry of Higher Education and Science 2

2.3 DATABASE PRESERVATION TOOLKIT 3

2.3.1 Database Preservation Toolkit (DBPTK) modules 3

2.3.2 System requirements 3

2.3.3 Pilot workflow 3

2.3.4 Sample commands 4

2.3.5 Screen output 4

2.4 RESULTS OF THE PILOT 4

2.5 SCENARIO OVERVIEW 5

2.6 IMPACT 6

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 Introduction

The aim of pilot 1 is to extract relational databases from four real life relational database management systems and transform these extracted databases from their proprietory format into the system independent archiving formats SIARDDK and SIARD 2.0. The SIARD 2.0 format is developed by the E-ARK project and the Swiss Federal Archives.

This extraction and transformation is performed using the open source Database Preservation Toolkit (DBPTK) made by project partner KEEP Solutions.

2.2 The databases

2.2.1 Health system from the Danish Serum Institute

This is a database containing information from reported infectious diseases at the national level - all Danish citizens from 2010 to 2015.

This database was on a MS SQL Server, it had 90 tables worthy of archiving, with 90,000 records in the main table, and around half a million records in total.

2.2.2 System from a private sector business Kultunaut Aps

This company is harvesting and selling information to different media about all cultural events from the smallest local gatherings to international exhibitions and events in Denmark. They have kept these informations since 1990's, and the Danish National Archives and Kultunaut made an agreement this year, that the data can be preserved at the Danish National Archives. Maybe 5% of our archived databases stem from private donors.

The database was on a MySQL server, and due to known errors in the SIARDDK module at the time, we exported only to SIARD2 format when we visited this company. We then in our own environment, used the export module to export from SIARD2 to SQL Server and then to SIARDDK. The great thing about this database was that it had more than 33 million records, and 28 millions of these were in one table alone.

2.2.3 Administrative system from The Danish National Archives

This and the next database were chosen because they have embedded large objects (LOBs). LOBs will be exported to external files that are placed in folders and subfolders. This is the new feature in SIARD 2.0 and that is why we chose this database since we knew it had more then 1,3 million LOBs. So this was definitely a scalability test.

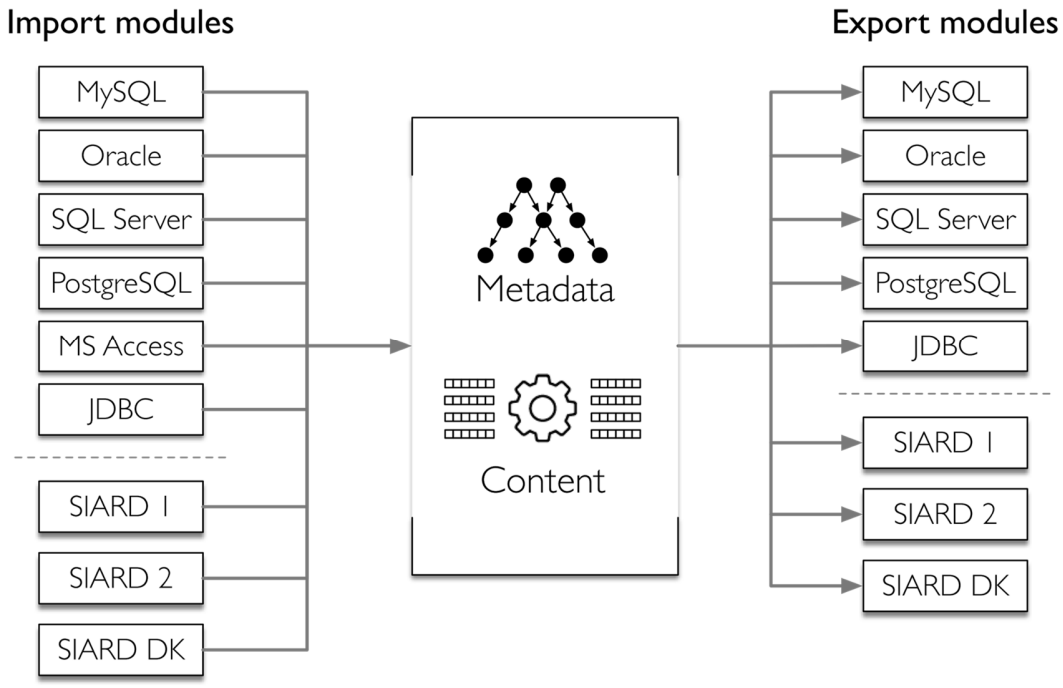
2.2.4 Administrative and health records system from Ministry of Higher Education and Science

We also exported data from an Administrative and health records system from the Ministry of Higher Education and Science. It is a Database containing information about social, psychological, and psychiatric counseling to students in their educational situation. This system had around 100,000 LOBs and was on a SQL Server.

2.3 Database Preservation Toolkit

The open source Database Preservation Toolkit (DBPTK) made by project partner KEEP Solutions is a java program with a CLI (command line interface). The program can connect to a number of different RDBMS products using a modular approach.

2.3.1 Database Preservation Toolkit (DBPTK) modules



2.3.2 System requirements

The DBPTK has few system requirements. It runs on any OS supporting the latest versions of Java.

2.3.3 Pilot workflow

Having succesfully tested the DBPTK on a few small sample databases we contacted the producers (the owners of the databases) and made agreements on when we could arrive at their site and try to extract the databases.

We connected to their RDBMS using a laptop and an external drive.

The most challenging task was to ensure that we had been provided with a user having sufficent rigths on the database system to export the database we desired.

The DBPTK and our equipment was sufficiently fast enough that we could perform the export within normal working hours.

In a few cases we not only exported to SIARD at the producer site but also received a backup of the database in its proprietory RDBMS format, giving us the possibility to redo the export at the premises of the Danish National Archives.

2.3.4 Sample commands

The DBPTK using CLI (a GUI is on its way) has a simple parameter list.

In the sample below a list of the tables in the database is created in order to filter out undesired tables (at the archivist's risk of break referential integrity.):

```
java "-Dfile.encoding=UTF-8" -jar "C:\Program
Files\DatabasePreservationToolkit\dbptk-app-2.0.0-SNAPSHOT-233-3.jar" -i microsoft-
sql-server -is localhost -idb Kingo_Drift -iu earktest -ip istvan2 -ide -e list-
tables -ef "E:\listtablesKingoall.txt"
```

Here is a sample of exporting to SIARD 2.0:

```
java "-Djava.io.tmpdir=E:\Temp" "-Dfile.encoding=UTF-8" -jar "C:\Program
Files\DatabasePreservationToolkit\dbptk-app-2.0.0-SNAPSHOT-233-3.jar" -i microsoft-
sql-server -is localhost -idb Kingo_Drift -iu earktest -ip istvan2  -ide -e siard-2
-ef "E:\4 Kingo_SIARD2.0_external_LOBs\kingo.siard" -ep -eel -eelpf 9999
```

2.3.5 Screen output

During the export the DBPTK shows the processing of tables. Below is an excerpt:

```
Getting contents from table 'dbo.sagskort'
Total of 29261 row(s) processed
Obtained contents from table 'dbo.sagskort'
Getting contents from table 'dbo.sagskort_henvisttiltype_rel'
Total of 662 row(s) processed
Obtained contents from table 'dbo.sagskort_henvisttiltype_rel'
Getting contents from table 'dbo.sagskortansvarlige'
Total of 32666 row(s) processed
CLOB(s) found in table 131. Archived as string
Obtained contents from table 'dbo.sagskortansvarlige'
Getting contents from table 'dbo.sagskortdatoer'
Total of 29505 row(s) processed
CLOB(s) found in table 132. Archived as string
```

2.4 Results of the pilot

All four databases have been succesfully exported to both SIARD 2.0 and SIARDDK format, and since we have our own validation tool for SIARDDK, we can say that the exports comply with SIARDDK. We have also manually validated against the SIARD 2.0 format and found no errors.

We still find that in the future we need a validation tool that systematically validates the SIARD 2.0 format.

2.5 Scenario overview

Scenario	1: Extracting records from database with no documents	2: Extracting records from database (large) with no documents	3: Extracting records from database with documents	4: Extracting records from database (large) with documents
Start	17-5-2016	23-8-2016	12-7-2016	13-6-2016
Status	100%	100%	100%	100%
Database	MS SQL Server 90 tables 90,000 records	MySQL 5 tables 33,000,000 records.	MS SQL Server 289 tables 1,300,000 LOBs	MS SQL Server 180 tables 100,000 LOBs
Producer	Health system from the Danish Serum Institute	System from a private sector business Kultunaut Aps	Administrative system from The Danish National Archives	Administrative and health records system from Ministry of Higher Education and Science
Descrip-tion	Database containing information from reported infectious diseases at the national level. Infectious diseases for all Danish citizens.	Harvesting and selling information about cultural events at national level from the smallest local gatherings to international exhibitions and events in Denmark.	Database containing information about all incoming scientific research data, and public deliveries of research data	Database containing information about social, psychological, and psychiatric counseling to students in their educational situation
Comment **SIARD 2.0**	*SIARD2.0:* 100% extraction of all tables and their data. The DNA has manually validated the SIARD-package up against the *"eCH-0165 SIARD Format Specification 2.0"*. There is no automatic tool for this yet, so there might be small errors in the package, where it does not comply with *"eCH-0165 SIARD Format Specification 2.0"*.	*SIARD2.0:* 100% extraction of all tables and their data. The DNA has manually validated the SIARD-package up against the *"eCH-0165 SIARD Format Specification 2.0"*. There is no automatic tool for this yet, so there might be small errors in the package, where it does not comply with *"eCH-0165 SIARD Format Specification 2.0"*.	*SIARD2.0:* 100% extraction of all tables and their data in one single SIARD-file. The DNA still has to export with a split to a SIARD-file and an external LOB-folder. The DNA also need to validate the SIARD-package up against the *"eCH-0165 SIARD Format Specification 2.0"*	*SIARD2.0:* 100% extraction of all tables and their data. The DNA has manually validated the SIARD-package up against the *"eCH-0165 SIARD Format Specification 2.0"*. There is no automatic tool for this yet, so there might be small errors in the package, where it does not comply with *"eCH-0165 SIARD Format Specification 2.0"*.
Comment **SIARD 2.0**	100% extraction of all tables and their data. The DNA has validated against *"Executive Order on Submission Information Packages"* and found no errors in the product.	100% extraction of all tables and their data. The DNA has validated against *"Executive Order on Submission Information Packages"* and found no errors in the product.	100% extraction of all tables and their data. The DNA has validated against *"Executive Order on Submission Information Packages"* and found no errors in the end product	100% extraction of all tables and their data.

2.6 Impact

The results from the pilot show that the E-ARK project open source Database Preservation Toolkit (DBPTK) made by project partner KEEP Solutions can export live databases in proprietary formats to system independent archiving formats SIARD 2.0 and SIARDDK (the latter provided by E-ARK project partner Magenta).

This open source tool can ease the export of databases by decreasing the cost and time spent for producers of SIPs as well fo