Database Preservation Toolkit

State-of-the-art and future plans in the context of the E-ARK project

Is a Data Warehouse a Data Archive and Why Does It Matter?

Hélder Silva hsilva@keep.pt
KEEP SOLUTIONS

February 19-20, 2015
Open source
Vendor freedom, no licensing costs

Ready to use
By archivists, producers and consumers

Authenticity
Preservation metadata, ISO 16363 compliant
Database Preservation Toolkit

Developed within RODA project (2006)

Now stand-alone open-source project (2014)

http://keeps.github.io/db-preservation-toolkit/

Imports and exports between DBMS and DB formats

Supports preservation formats: DBML, SIARD
In the Archive

Diagram of OAIS (Open Archival Information System) components:
- Preservation Planning
- Data Management (DM)
- Ingest
- Access
- Archival Storage
- Administration
- Management

Interactions:
- Producer to Ingest: SIP
- Consumer to Access: DIP
- Data Management (DM) to Ingest and Access
- Archival Storage to Ingest and Access
In the Archive
Pre-ingest / Production / SIP creation

Origin formats:
- MySQL
- Oracle'12
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC

db-preservation-toolkit

Preservation formats:
- DBML
- SIARD
Back to the Archive
Back to the Archive
Back to the Archive
Access

Preservation formats

- DBML
- SIARD

db-preservation-toolkit

Access format / viewer

- Viewer
- ...

...
Preservation formats: DBML

First preservation format

Plain XML

  Associated XML Schema
  Software/hardware independent
  Easy to process
  Descriptive
  Extensible

Stores the data and database structure in a single XML file

Never meant to be a standard but no “standard” was available at the time (2006)
Preservation formats: SIARD

Developed by the Swiss Federal Archives

SIARD

Suite: tools to create/view
Format: open file format specification

Standards based (XML, SQL:1999, Unicode)

Stores everything in an uncompressed ZIP file

Swiss standard since 2013 (eCH-0165)
Preservation formats: SIARD

Added in version 2.0

Both input & output module

Generated output was validated using KOST-Val

https://github.com/KOST-CECO/KOST-Val
Validating SIARD
Validating SIARD
Validating SIARD

Import modules
- MySQL
- Oracle'12
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC
- DBML
- SIARD

Streaming data model

Export modules
- MySQL
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC
- DBML
- SIARD
- Viewer

DBMS Preservation format
- db-preservation-toolkit
- SIARD

Importing and exporting databases through different modules.
Validating SIARD

DBMS

PostgreSQL → db-preservation-toolkit → SIARD

Preservation format

Import modules
MySQL
Oracle 12
SQL Server
PostgreSQL
DB2
MS Access
ODBC
DBML
SIARD

Streaming data model

Export modules
MySQL
SQL Server
PostgreSQL
DB2
MS Access
ODBC
DBML
SIARD
Viewer
Validating SIARD

DBMS

PostgreSQL

Preservation format

Import modules

MySQL
Oracle'12
SQL Server
PostgreSQL
DB2
MS Access
ODBC
DBML
SIARD

Streaming data model

Export modules

MySQL
SQL Server
PostgreSQL
DB2
MS Access
ODBC
DBML
SIARD
Viewer

db-preservation-toolkit

Validating SIARD

db-preservation-toolkit

db-preservation-toolkit

SIARD
Validating SIARD

**DBMS**

- PostgreSQL

**Preservation format**

- DB-preservation-toolkit
- SIARD

**Import modules**

- MySQL
- Oracle’12
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC
- DBML
- SIARD

**Export modules**

- MySQL
- SQL Server
- PostgreSQL
- DB2
- MS Access
- ODBC
- DBML
- SIARD
- Viewer

**Script**

- Database initialization (create database, insert data with all datatypes available)
- Textual dump + DBMS to SIARD conversion (white path)
- Execute SIARD to DBMS conversion (red path) + Textual dump
- Do textual comparison of both dumps
- Database cleanup (delete both databases created)
db-preservation-toolkit vs. SIARD Suite

Throughput

- db-preservation-toolkit
  - 1 GB: 5 MB/s
  - 5 GB: 14 MB/s
  - 25 GB: 27 MB/s

- SIARD Suite
  - 1 GB: 1 MB/s
  - 5 GB: 2 MB/s
  - 25 GB: Crashed !!!

**Note:** The SIARD Suite crashed at 25 GB.
db-preservation-toolkit vs. SIARD Suite

<table>
<thead>
<tr>
<th>Data Size</th>
<th>db-preservation-toolkit</th>
<th>SIARD Suite</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GB</td>
<td>3 min</td>
<td>16 min</td>
</tr>
<tr>
<td>5 GB</td>
<td>6 min</td>
<td>47 min</td>
</tr>
<tr>
<td>25 GB</td>
<td>12.5 min</td>
<td>Crashed !!!</td>
</tr>
</tbody>
</table>

Execution time
Future plans:
Fast database viewer
Preservation formats

- DBML
- SIARD

Fast viewer resources

- db-preservation-toolkit

Fast viewer application

- Lucene / Solr index
- Web interface & REST API

Future plans:
- Fast database viewer
Future plans: Fast database viewer

Preservation formats
- DBML
- SIARD

Fast viewer resources
- db-preservation-toolkit
- Lucene / Solr index

Fast viewer application
- Web interface & REST API

SIARD “flavours”
- Danish (SIARD-DK)
- E-ARK (SIARD-E/SIARD 2.0)
Database Preservation Toolkit

State-of-the-art and future plans in the context of the E-ARK project

Is a Data Warehouse a Data Archive and Why Does It Matter?

Hélder Silva hsilva@keep.pt
KEEP SOLUTIONS

February 19-20, 2015