Preservation through standardisation with DBPTK

Luís Faria

Research & Innovation Director
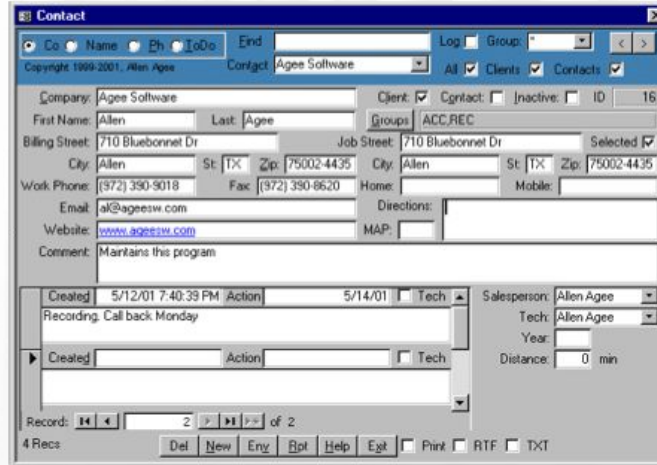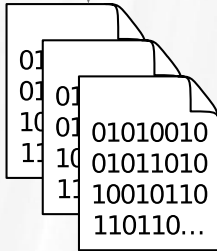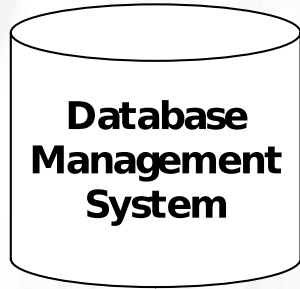
KEEP SOLUTIONS

keep.
Preserving the future

Databases

The **information** that supports institutions and businesses is usually **centralized on databases**.

This information is of **great value** and needs to be **preserved for decades** due to strategic and legal reasons.

The systems that have this information are usually complex with **many software components** playing their part for supporting the **business-logic**, and the **submission** and **presentation** interfaces.

The information is usually laid out in an **organization specifically optimized for the database** and original business objectives, i.e. **not in a user-friendly** organization.

Database Management System

Application

01010010
01011010
10010110
110110…

www.keep.pt

person			
id	**name**	**birth**	**city_id**
1	Mary	1986-03-28	2
2	Phillip	1974-11-08	3
3	Alison	1991-06-10	5
4	Barry	1979-09-14	2

Cell

Row

Column

4

www.keep.pt

Information to preserve

Within the relational database:

- Information in tables

- Column data types

- Relations and constraints

- Projections (views)

- Behaviour (triggers and routines)

- Other (users, permissions, etc.)

Outside the relational database:

- External resources
  (e.g. files in filesystem)

- Submission forms

- Presentation interfaces

- Application logic and queries

Preservation strategies

- Hardware and software museums

- Emulation

- File format migration

- Encapsulation

Hardware and software museums

Preserve the **whole technology stack** needed to render the original content.

⊕ reproduction accuracy	⊖ great difficulty to maintain
	⊖ restrictions on the access to information
	⊖ need for users to understand how to operate long gone systems

www.keep.pt

Emulation

Use of a software system that **emulates the behaviour** of an older hardware and/or software platform within a newer one.

⊕ reproduction accuracy	⊖ difficult to maintain
⊕ no need to maintain hardware	⊖ difficult to set up
	⊖ need for users to understand how to operate long gone systems

www.keep.pt

File format migration

Transfer of digital information from one hardware and software configuration into another.

Convert information encoded in a file format, tied into an **obsolete technology stack**, into another **more current or better suited for long term preservation**.

⊕ easier to use and reuse information	⊖ possible data loss during conversion (can be mitigated by quality assurance)
⊕ no need to maintain hardware	⊖ might need to migrate again in the future
⊕ no need to maintain software	

www.keep.pt

Encapsulation

Keep files together with all necessary documentation needed for future development of emulators, file format migrators or software renderers.

⊕ postpone actions that can be costly	⊖ may hinder timely access to information
⊕ no need to maintain hardware or software	⊖ difficult to gather documentation of complex or closed file formats
	⊖ difficult to ensure quality and completeness without hindsight

www.keep.pt

Problem with preserving databases

Every vendor has his data types and export formats

It is rare that information exported from one vendor's system works on another

Sometimes doesn't work on different versions of the same product

Need for a vendor-agnostic format based on standards

Preservation format criteria

Ubiquity	Stability	Complexity
Support	Ease of identification and validation	Interoperability
Disclosure	Intellectual Property Rights	Viability
Documentation quality	Metadata support	Re-usability

https://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

SIARD: Software Independent Archiving of Relational Databases

Database preservation format

Based on international standards

For database data, structure and behaviour

Swiss national standard eCH-0165

Now managed by DILCIS board and the EU eArchiving building block
https://dilcis.eu/content-types/siard
https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eArchiving

More about this on the next session
Kuldar Aas (National Archives Estonia): E-ARK standardisation efforts for databases

Simple database archive flow

Producer

Database

Related files

Archive with DBPTK Enterprise

SIARD

Preservation format

SIARD format validation

Simple catalogue

Database viewer

User
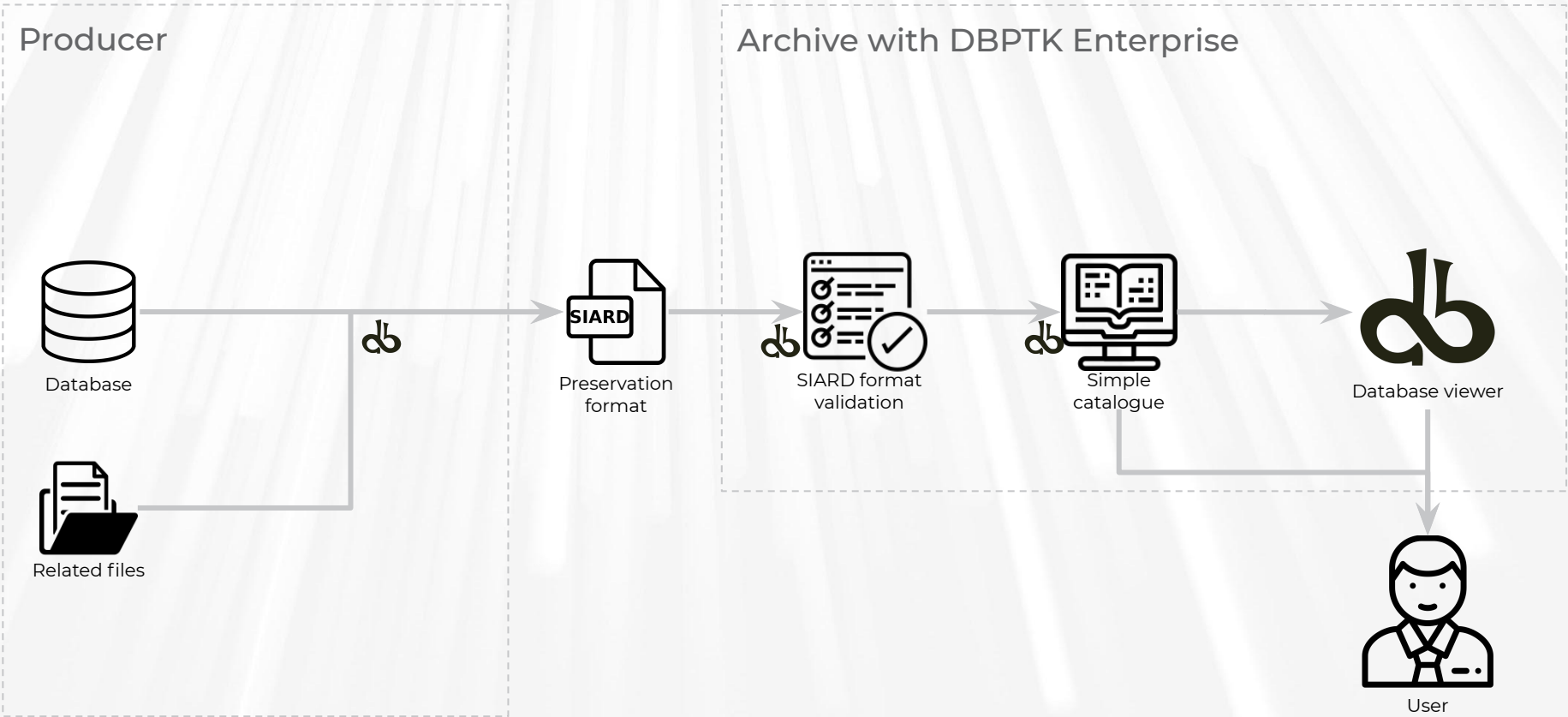
Icons from www.flaticon.com

www.keep.pt

Full database archival flow

Producer

Original application

Documentation

Database

Original backup format

Related files

Preservation format

SIARD

Finding aids

Archive with RODA + DBPTK Enterprise

EARK SIP

Ingest validations

Catalogue

Database viewer

User
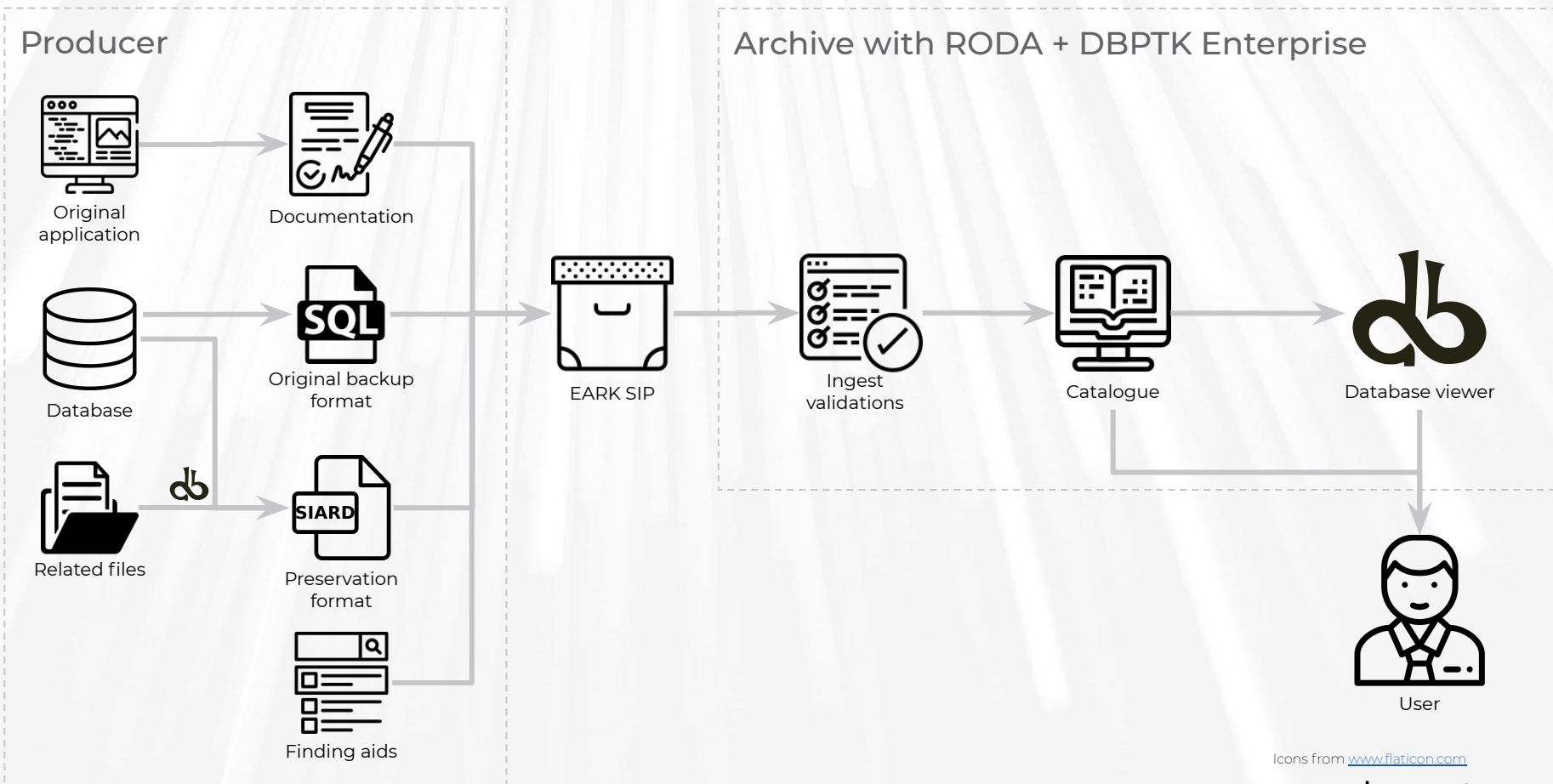
Icons from www.flaticon.com

www.keep.pt

DBPTK

Database Preservation Toolkit

Set of tools to store relational databases in a standard archival format.

https://database-preservation.com

DBPTK Desktop
Desktop application to save database to preservation format, validate it, and browse and search the content

DBPTK Enterprise
Web application to browse and search on the content of multiple large preserved databases
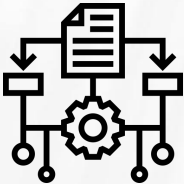
DBPTK Developer
A command-line tool and development library for automation and system integration

www.keep.pt

DBPTK Desktop

Top features

DBPTK Desktop features

Windows

macOS
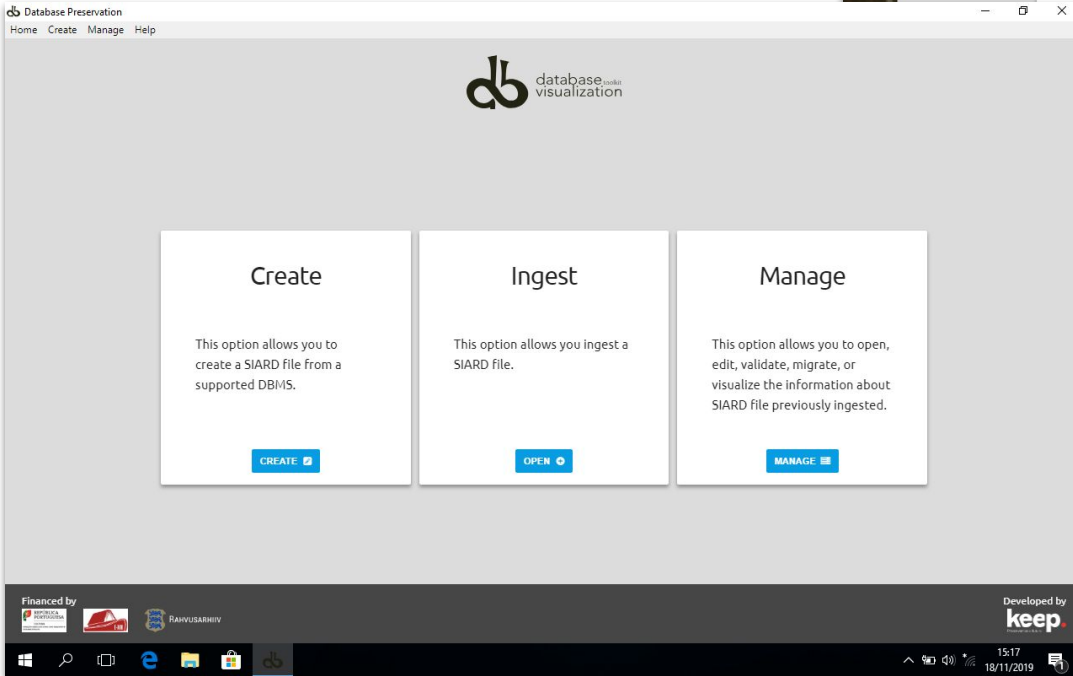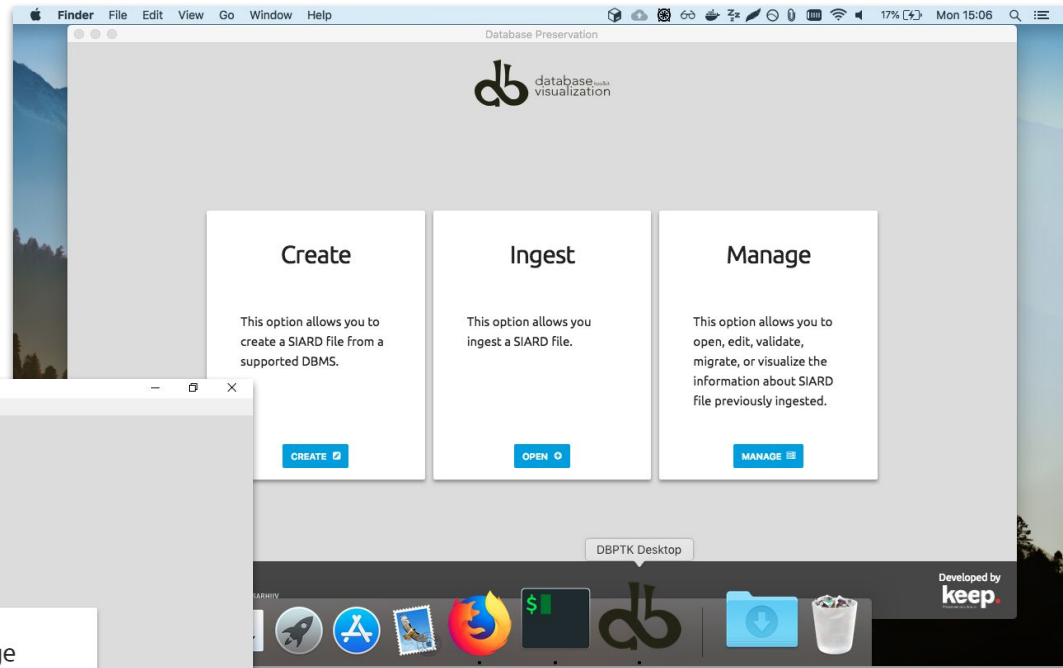
Also available on Linux

DBPTK Desktop features

SIARD creation

Export database to a preservation format

- Connect to a local or remote database and save all content into a preservation format like SIARD

- Test connection will diagnose most common problems and provide you with helpful hints to solve them

Supported DBMS:
- Microsoft Access
- Microsoft SQL Server
- MySQL / MariaDB
- Oracle
- PostgreSQL
- Progress Openedge
- Sybase

www.keep.pt
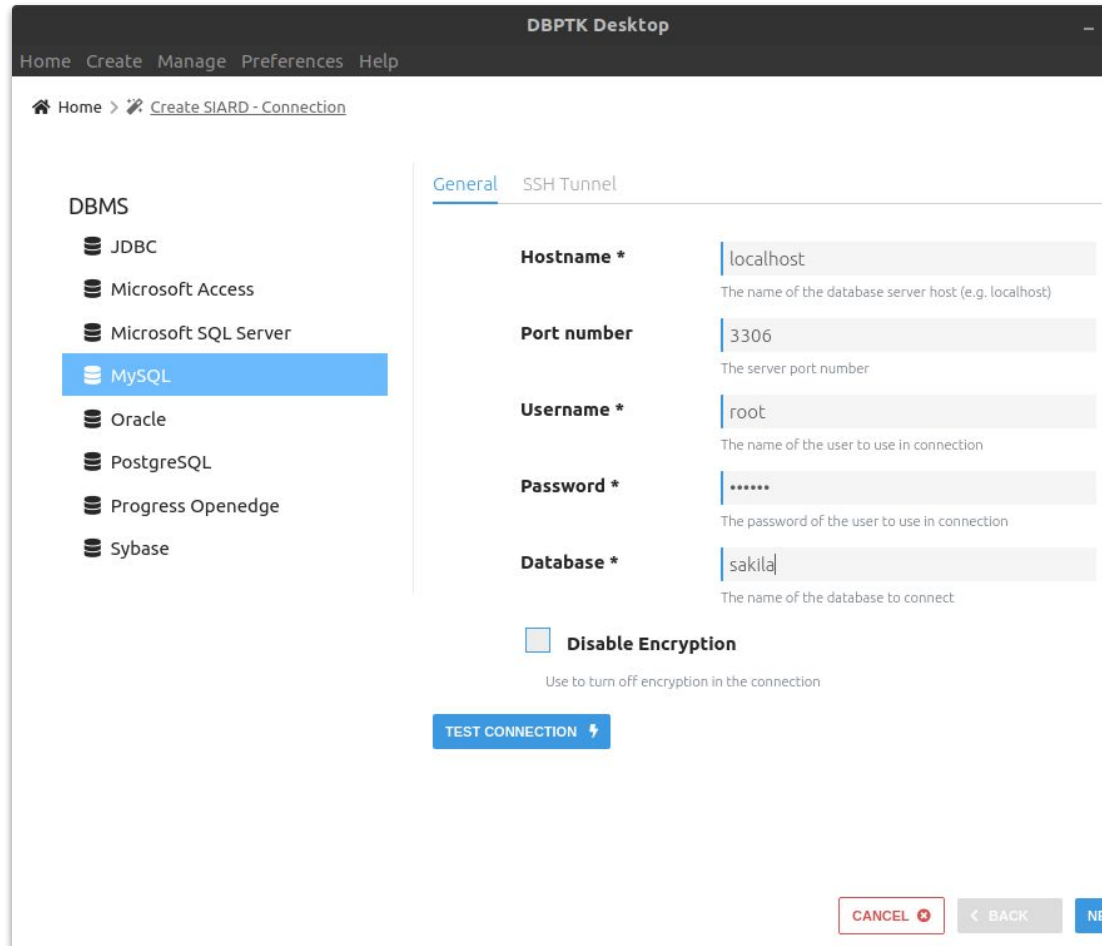
DBPTK Desktop features

Migration report
Detailed report of migration changes and losses

- All export and selection parameters are presented.

- All column data types mapping to standard types are recorded.

- All compromises are documented.

Database Preservation Toolkit (version 2.8.2) – Conversion Report

Parameters

Import module: mysql

- hostname = dpc.database-preservation.com
- database = sakila
- username = mguimaraes
- password =
- port-number = 3306
- disable-encryption = false

Export module: siard-2

- version = V2_1
- file = /home/mguimaraes/Desktop/sakila-dpc.siard
- compress = true
- pretty-xml = false
- external-lobs = false
- external-lobs-per-folder = 1000
- external-lobs-folder-size = 0
- digest = SHA-256
- font-case = lowercase

Date: 2020-07-22

Details

- Type conversion in import module: in `sakila.address.address` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(50)`
- Type conversion in import module: in `sakila.address.district` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(20)`
- Type conversion in import module: in `sakila.city.city` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(50)`
- Type conversion in import module: in `sakila.country.country` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(50)`
- Type conversion in import module: in `sakila.actor.actor_id` (format: schema.table.column) has original type `SMALLINT UNSIGNED` and was converted to the standard type `SMALLINT`
- Type conversion in import module: in `sakila.actor.first_name` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(45)`
- Type conversion in import module: in `sakila.actor.last_name` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(45)`
- Information: check constraints is not yet supported for MySQL. But support may be added in the future
- Type conversion in import module: in `sakila.address.address_id` (format: schema.table.column) has original type `SMALLINT UNSIGNED` and was converted to the standard type `SMALLINT`
- Type conversion in import module: in `sakila.address.address` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(50)`
- Type conversion in import module: in `sakila.address.address2` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(50)`
- Type conversion in import module: in `sakila.address.district` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(20)`
- Type conversion in import module: in `sakila.address.city_id` (format: schema.table.column) has original type `SMALLINT UNSIGNED` and was converted to the standard type `SMALLINT`
- Type conversion in import module: in `sakila.address.postal_code` (format: schema.table.column) has original type `VARCHAR` and was converted to the standard type `CHARACTER VARYING(10)`

DBPTK Desktop features

Edit SIARD metadata
Enrich archived database with descriptions

- Add descriptions to database, tables and columns to better understand its contents

www.keep.pt
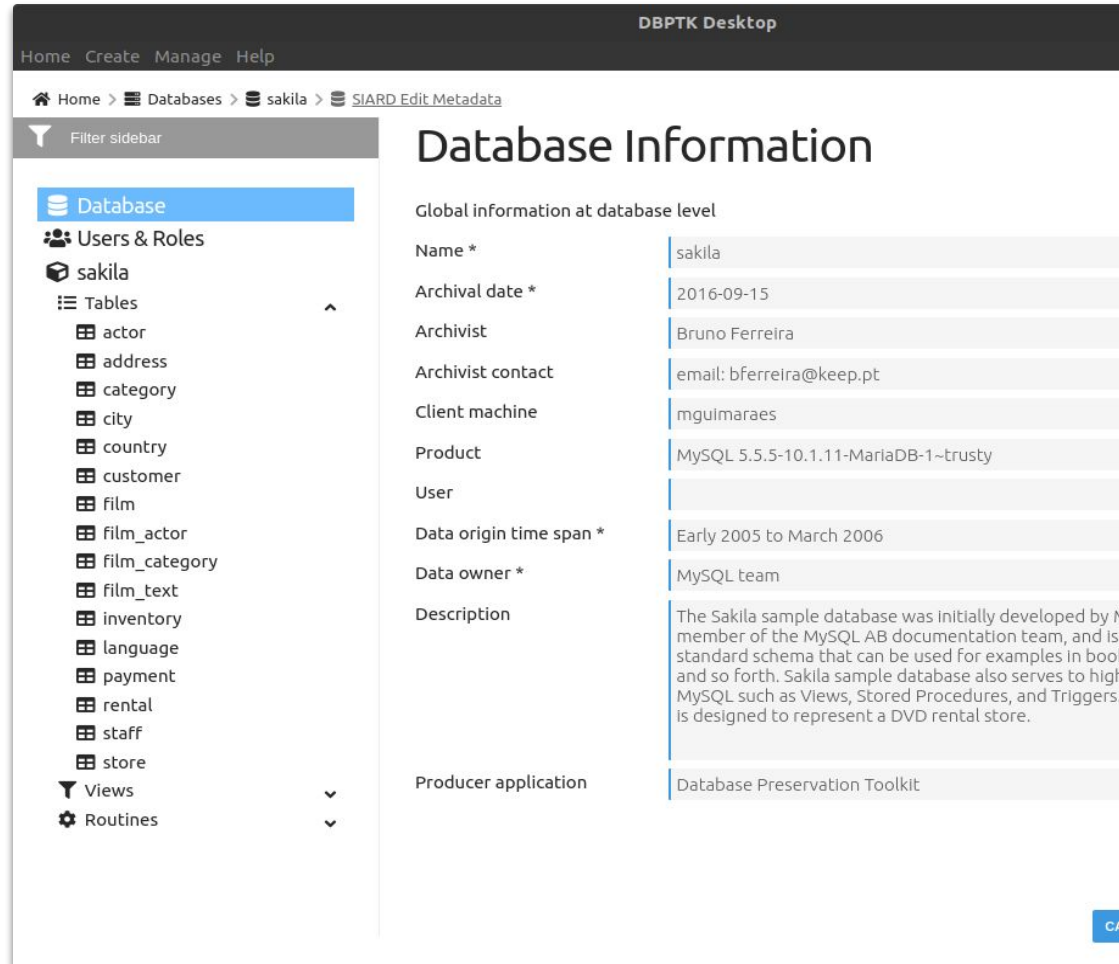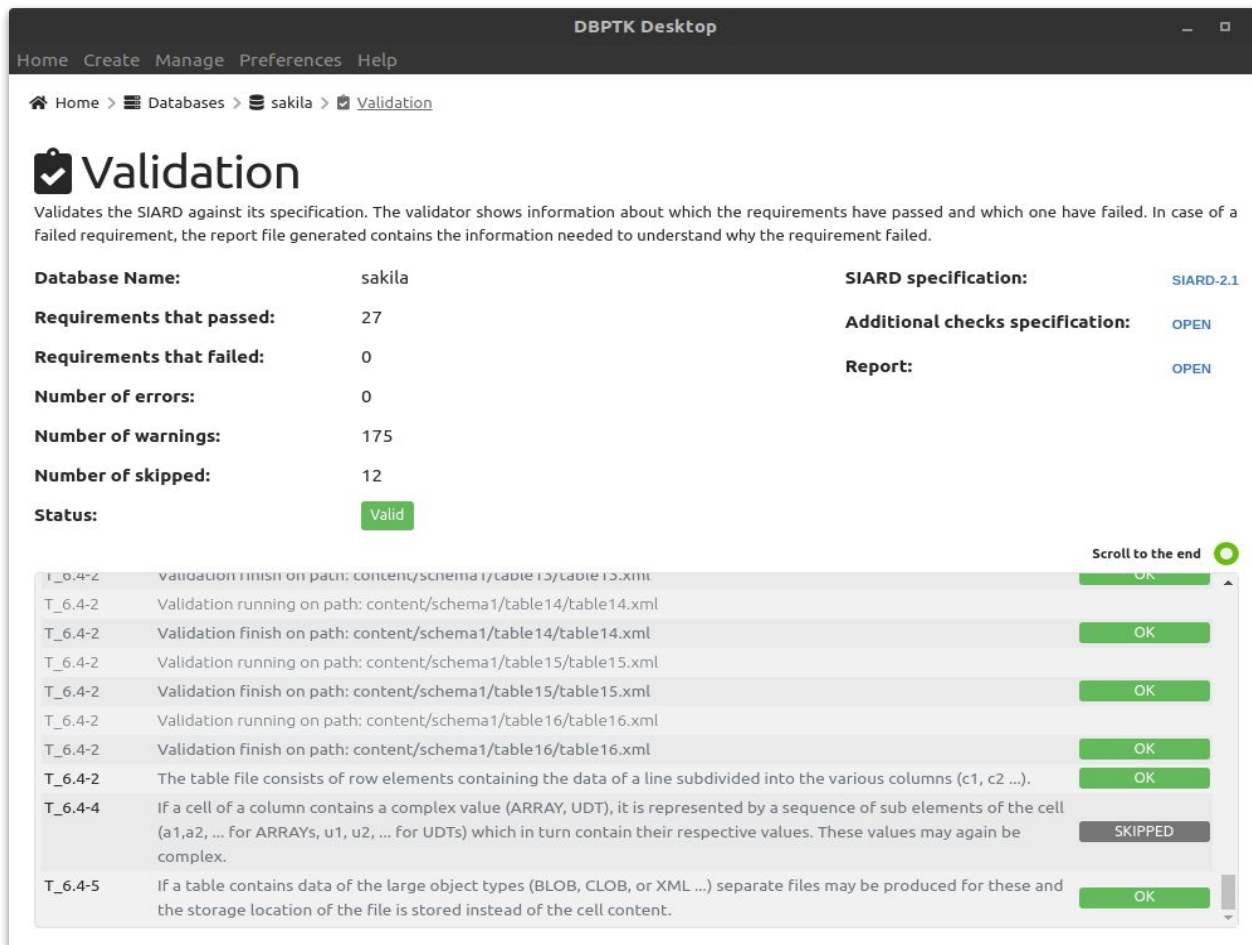
DBPTK Desktop features

SIARD validation
Validate archived database

- Validate SIARD against specification plus many additional checks for a thorough validation

www.keep.pt

DBPTK Desktop features

Search records
Browse and search database content

- Google-like search on the database content.

- Drill down on specific tables and do advanced search for specific fields to find exactly what you are looking for.

www.keep.pt

DBPTK Desktop features

Auto-update
Automatic check of updates

- Stay up-to-date with automatic update check on startup and installation of new versions.

Found Updates

Found updates, do you want update now?

Yes   No

Install Updates

Updates downloaded, application will be quit for update...

OK

www.keep.pt
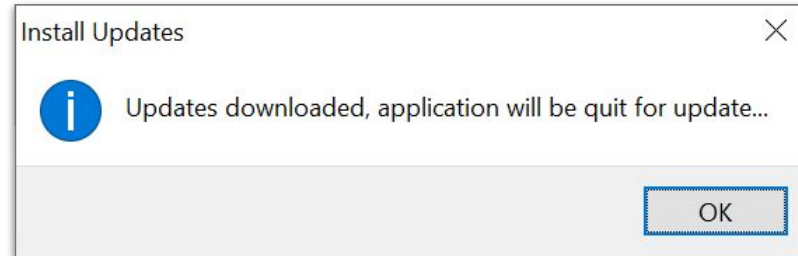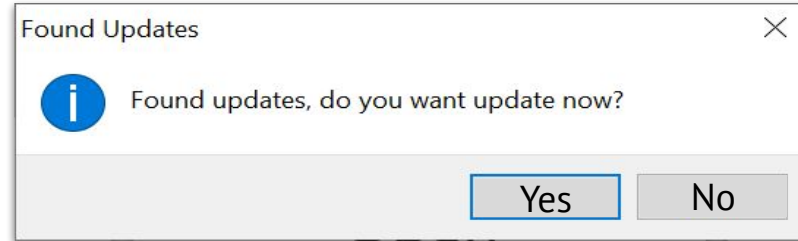
DBPTK Enterprise

Top features

DBTPK Enterprise features

Enterprise architecture
For large institutions with many databases and users

- A web application that can be horizontally scaled to support many very large databases being accessed by many users

www.keep.pt
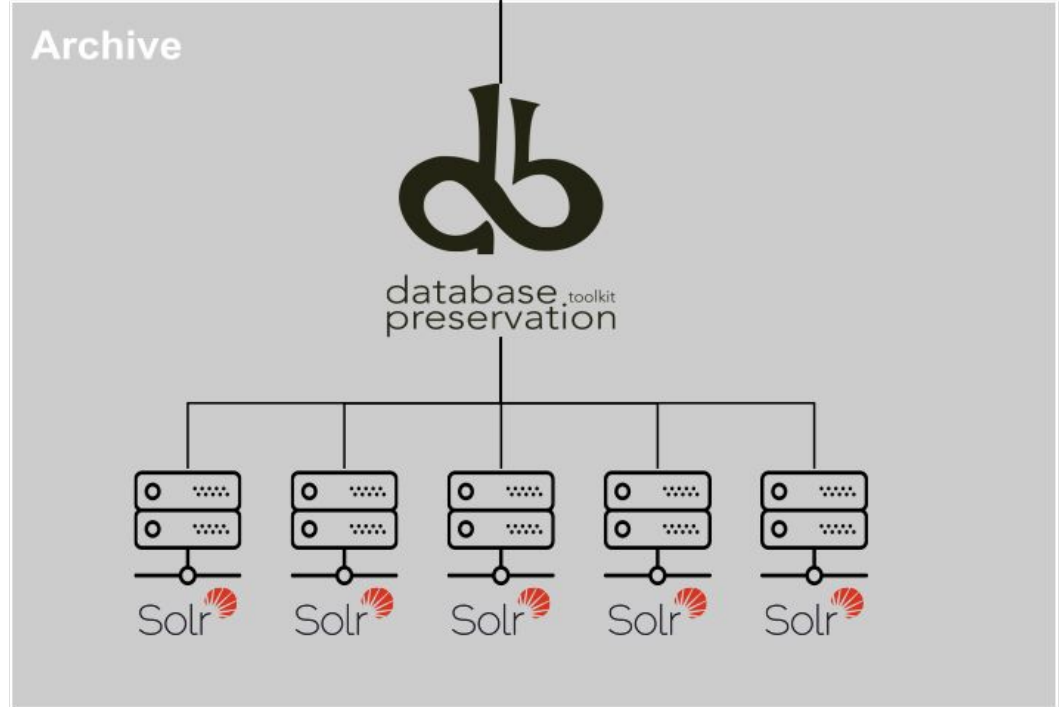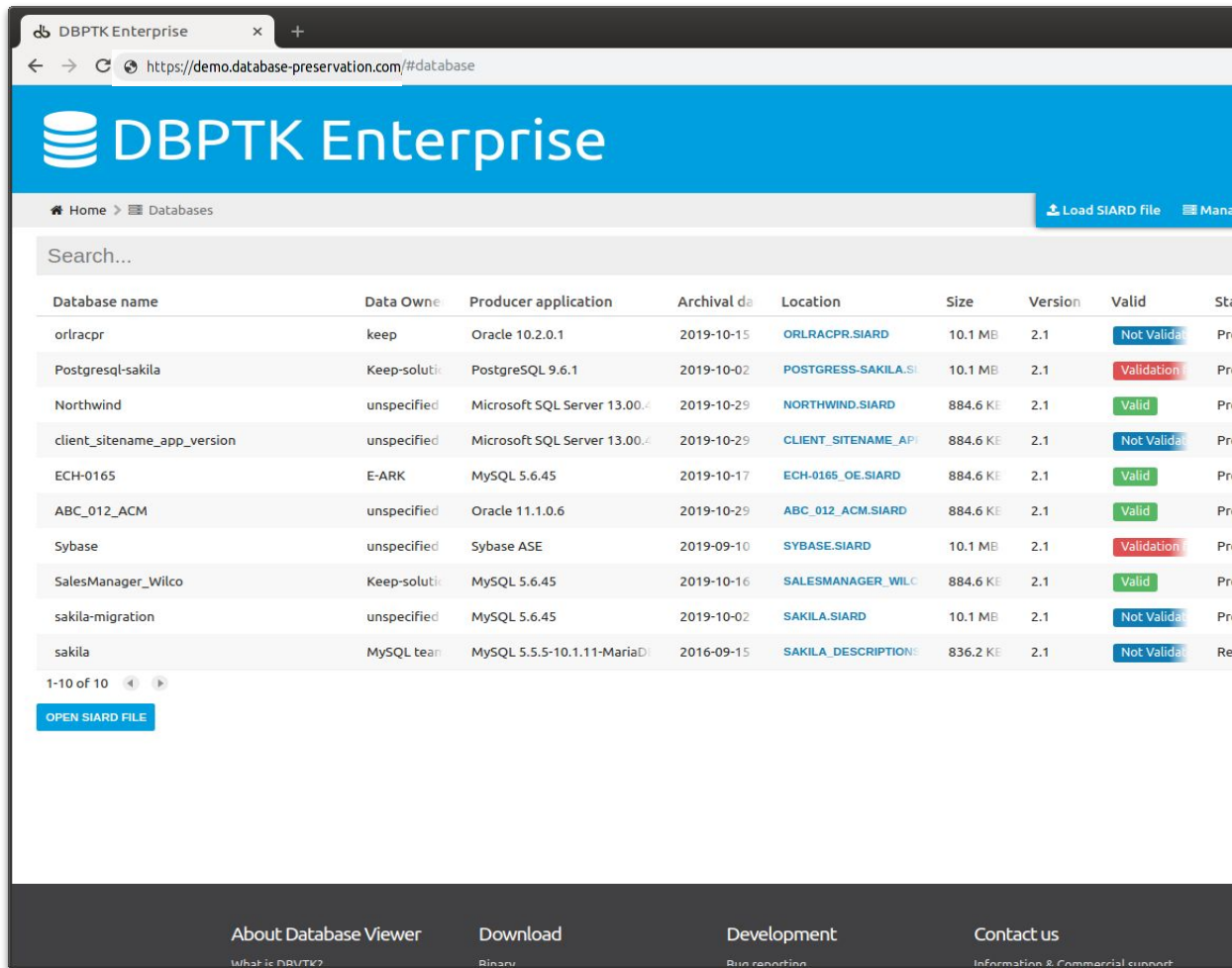
DBTPK Enterprise features

Manage multiple databases
Single system, multiple databases

- Search through the databases, manage their status, enrich their metadata, validate them, make them ready for users to search.

www.keep.pt

DBTPK Enterprise features

Data transformation
Transform content to answer useful questions

- **De-normalization** and table and **column hiding**, to simplify browsing and allow **anonymization** of content

www.keep.pt

Data transformation (aka denormalization)

person			
id	**name**	**birth**	**city_id**
1	Mary	1986-03-28	5
2	Phillip	*NULL*	6

city			
id	**name**	**mayor**	**country_id**
5	Payne Springs	1	16
6	Rosenhayn	*NULL*	16

country	
id	**name**
16	United States
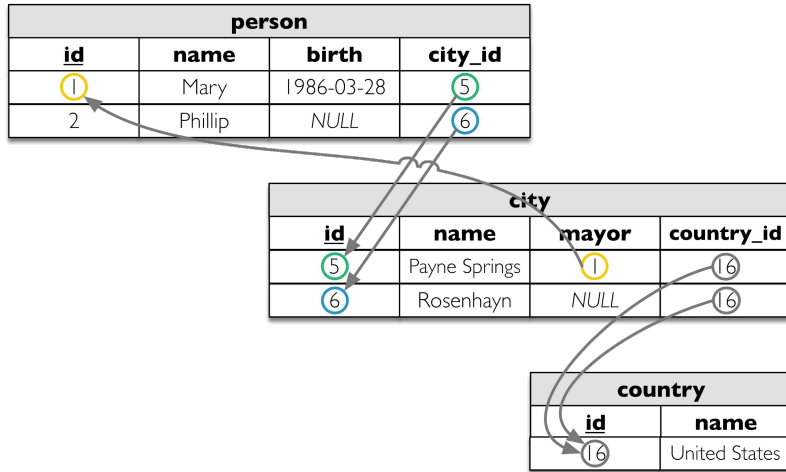
person				
Name	**Birth**	**City name**	**Mayor**	**Country name**
Mary	1986-03-28	Payne Springs	<u>Mary</u>	United States
Phillip		Rosenhayn		United States

31

www.keep.pt

DBPTK Enterprise features

Single sign-on
Support for multiple protocols

- LDAP, Active Directory, Database, SAML, ADFS, OAuth2, OpenID, Google, Facebook, Twitter, FIDO U2F, YubiKey, Google Authenticator, Authy, etc.

- Supports internal authorization definition or configurable external authorization
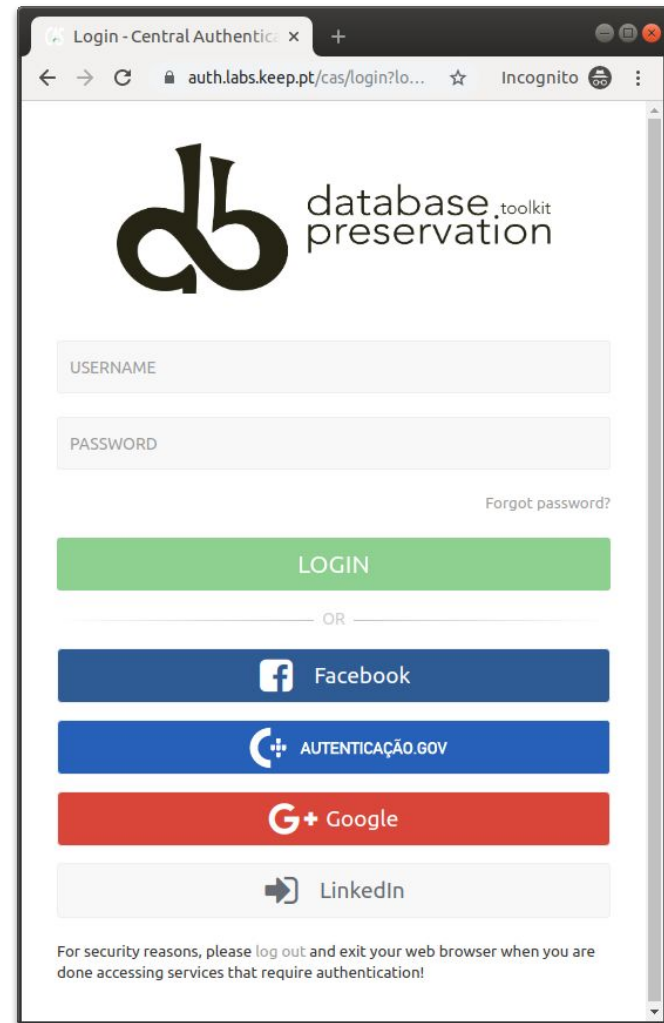
Browse and search
Allow users to access database content on the Web

- Allow them to search on a prepared, user-friendly and anonymized database content

33

www.keep.pt

DBPTK Enterprise features

Export features
Export data into tabular data

- Allow users to save search results in Microsoft Excel or other spreadsheet software format for easy analytics and diagrams

www.keep.pt

DBPTK Enterprise features

Activity log
Audit every access

- Who has done what, when and from where.

- Requirement for ISO 16363 certification.

www.keep.pt

DBPTK Enterprise & Desktop

Multiple languages supported

Interface translated into:

English, German, Estonian, Czech, Portuguese

Search stemming and stopwords support for:

English, Arabic, Bulgarian, Catalan, Czech, Danish, German, Greek, Spanish, Estonian, Basque, Persian, Finnish, French, Irish, Galician, Hindi, Hungarian, Armenian, Indonesian, Italian, Latvian, Dutch, Norwegian, Portuguese, Romanian, Russian, Swedish, Thai, Turkish, Japanese (using morphological analysis), CJK bigram (Chinese, Japanese, and Korean languages)

www.keep.pt
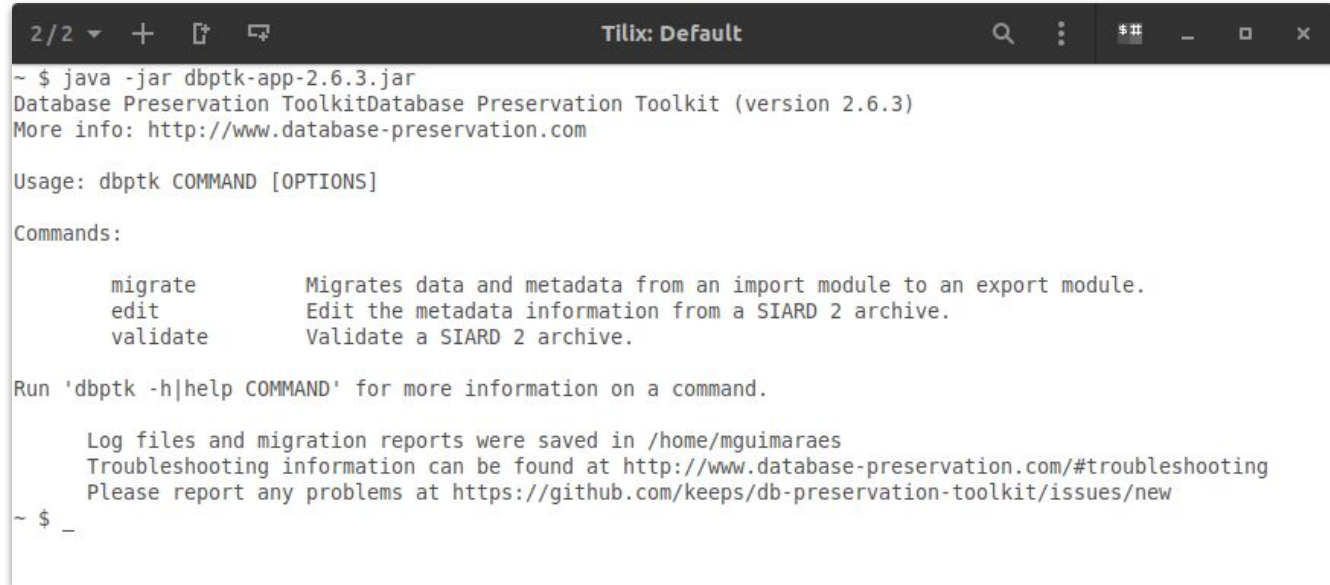
DBPTK Developer

Top features

DBPTK Developer features

Command line interface
Automation of periodic preservation tasks

- Command line interface allows easy automation of periodic tasks like saving database to preservation format, validating, and editing metadata.


```
2/2 ▼   +   ⥂   ⥡                    Tilix: Default              Q   ⋮   $#  _  □  ×

~ $ java -jar dbptk-app-2.6.3.jar
Database Preservation ToolkitDatabase Preservation Toolkit (version 2.6.3)
More info: http://www.database-preservation.com

Usage: dbptk COMMAND [OPTIONS]

Commands:

      migrate       Migrates data and metadata from an import module to an export module.
      edit          Edit the metadata information from a SIARD 2 archive.
      validate      Validate a SIARD 2 archive.

Run 'dbptk -h|help COMMAND' for more information on a command.

      Log files and migration reports were saved in /home/mguimaraes
      Troubleshooting information can be found at http://www.database-preservation.com/#troubleshooting
      Please report any problems at https://github.com/keeps/db-preservation-toolkit/issues/new
~ $ _
```
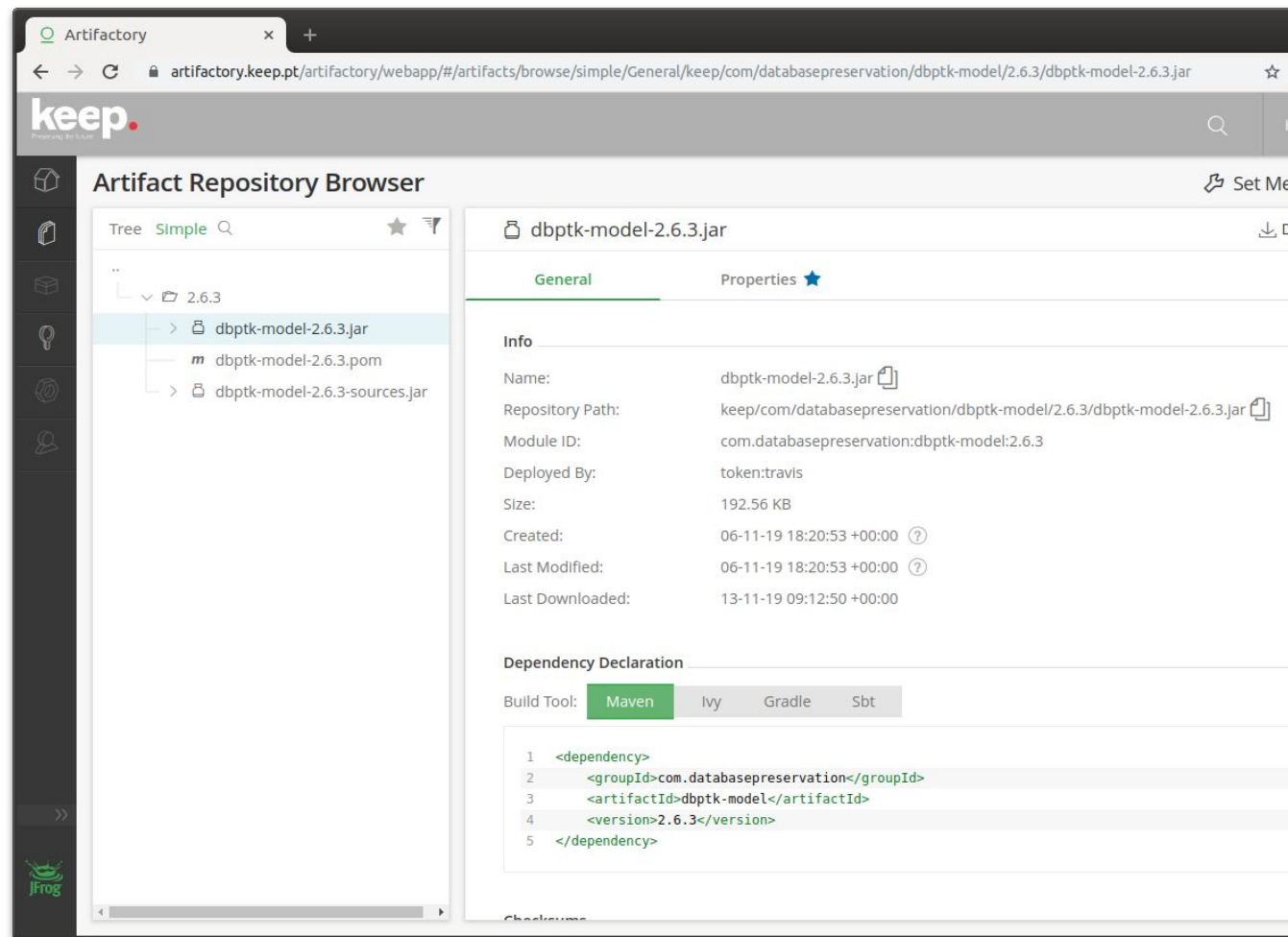
www.keep.pt

DBPTK Developer features

Systems integration
Java library

- Library to allow integration of production systems to directly use database preservation features.

www.keep.pt

DBPTK Developer features

Open source
For custom development

- Code base that allows custom development of new features or specialized support for new or legacy database systems.

www.keep.pt

And many more features

For archiving databases:

- SSH Tunnel
- Selection of tables and columns
- Selection and materialization of views
- Custom views
- External files (files stored outside the DB)
- External files via SSH tunnel
- Automated quality assurance
- Save LOBs outside SIARD file
- Migrate from SIARD to SIARD
- Migrate from SIARD to live DBMS
- Convert ORACLE geodata

For accessing archived databases:

- Configure visible tables
- Configure visible columns
- Set column name, description and order
- Binary columns advanced options
- REST API
- Load on access and auto-unload

www.keep.pt

DBPTK

	Desktop	Enterprise	Developer
Save to preservation format	✔	✔*	✔
Quality assurance (merkle tree)	✔	✔*	✔
Validation	✔	✔	✔
Enrich descriptions	✔	✔	✔
Browse and search	✔	✔	✘
Transform (de-normalization)	✘	✔	✘
Export to live databases	✔	✔*	✔
Activity Log	✘	✔	✘
Authentication	✘	✔	✘
Number of users	one	many	one
Number of loaded databases	few	many	N/A
Integration with repositories	✘	✔	N/A
Embeddable in Web portals	✘	✔	N/A

* Enterprise feature done via the upload/download of SIARD and usage of related tools

www.keep.pt

Database preservation

Real-world use cases

keep.
Preserving the future

Hospital legacy databases

Context

Set of **database systems** created to support **specific hospital services** (cardiothoracic, neonatology and neutropenia)

They contain **crucial information** about the **history of some patients** that may be needed for **urgent interventions**

Problem

- Databases were **replaced** by newer systems

- Information was **never migrated** to newer systems

- Original Database Management Systems are **obsolete**

- Original developers and submitters are **gone**

- **Not enough documentation** is available

Hospital legacy databases

Solution

- **Export** of all information into SIARD

- **Expert analysis** of original database and interfaces to create **documentation**

- Using RODA to keep documentation and DBPTK Enterprise to provide access

- Use table and column management and data transformation to make databases more **user-friendly** and **better documented**.

Main software used

- DBPTK Desktop for export into SIARD

- RODA for catalogue and archiving representation information (documentation)

- DBPTK Enterprise for access to database content

Main features used

- Custom views and materialized views

- SIARD metadata edition

- Table and column management

- Data transformation

European Taxation and Customs Union: trader messages archive

Context

New EU service that will provide a centralized interface with customs authorities for thousands of economic operators that bring the goods into the European Union.

All transaction messages will need to be archived for a decade.

Problem

- Estimated 10 million messages per day

- Production database needs to offload to archive daily and purge information

- Must ensure no message is lost or mangled in the archival process

- Archive process must keep up with production

European Taxation and Customs Union: trader messages archive

Solution

- Archive partial exports of database into SIARD (e.g. 1-hour timespans)

- Archive into RODA and load into DBPTK Enterprise when access is needed

- Continuous extraction, archive and validation workflow

- Quality assurance is key

Main software used

- DBPTK Developer for continuous partial export to SIARD

- RODA for archival, search and load into DBPTK Enterprise

- DBPTK Enterprise to access on request and retrieve original message(s)

Main features used

- DBPTK developer automation scripts

- Automated quality assurance

Questions?

Luis Faria
Research & Innovation Director

KEEP SOLUTIONS
lfaria@keep.pt

https://database-preservation.com

keep.
Preserving the future

www.keep.pt