keep.
Preserving the future

Database preservation

DPC training course

Advanced features
Day 2, morning

Trainers: Luis Faria, Miguel Guimarães

keep.
Preserving the future

Agenda for Day 2, morning

10:00 Welcome *by Jenny*

10:05 DBPTK advanced features *by Luís and Miguel*

11:05 Break

11:25 Real-world use-cases *by Luís*

11:45 Case study: Implementing database archiving at the National Archives of Estonia
by Kuldar Aas, National Archives of Estonia

12:15 Questions and discussion

12:45 Lunch

DBPTK

Advanced features

DBPTK Desktop

Advanced features

keep.
Preserving the future

SSH Tunnel

Database server may not have a display or enough resources

Database server may not allow/support remote JDBC connection

keep.
Preserving the future

Selection of tables and columns

Select which tables and columns to export

Optimize export time

Reduce storage usage

Anonymize archived content

keep.
Preserving the future

Selection and materialization of views

Select the views to document

Document view query and column data types

Select the views to materialize (into tables)

Select the columns of the view to document and materialize

keep.
Preserving the future

Custom views

Create views at the moment of export

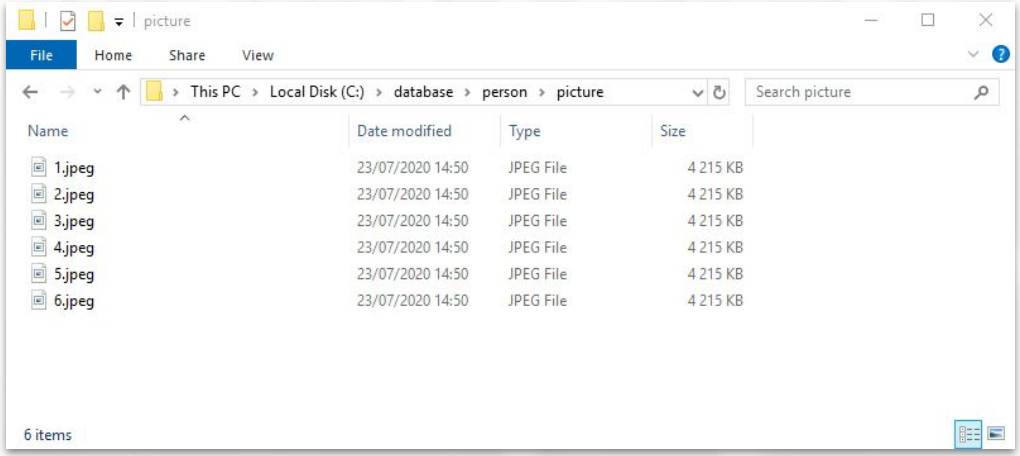
Select content and merge tables with join queries

Test the custom view before saving

keep.
Preserving the future

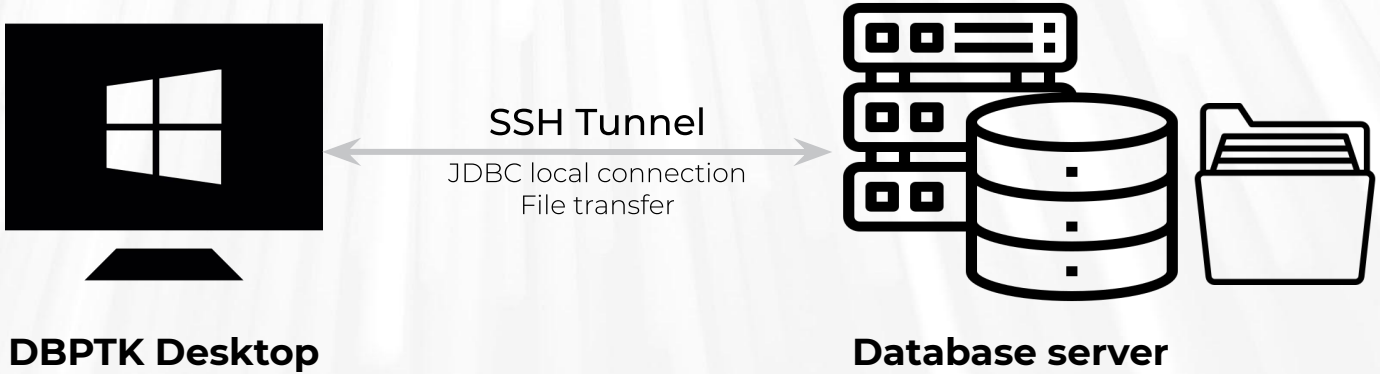
External files (files stored outside the database)

person		
Name	**Birth**	**Picture**
Mary	1986-03-28	1.jpeg
Phillip		2.jpeg



keep.
Preserving the future

External files (files stored outside the database) via SSH tunnel



Connect to both remote database and file server

keep.
Preserving the future

Automated quality assurance

Independently assess if information is complete and correct

Cope with silent errors due to network or database driver misbehaving

Verify that all columns are exported, the content of columns and files is complete and correct

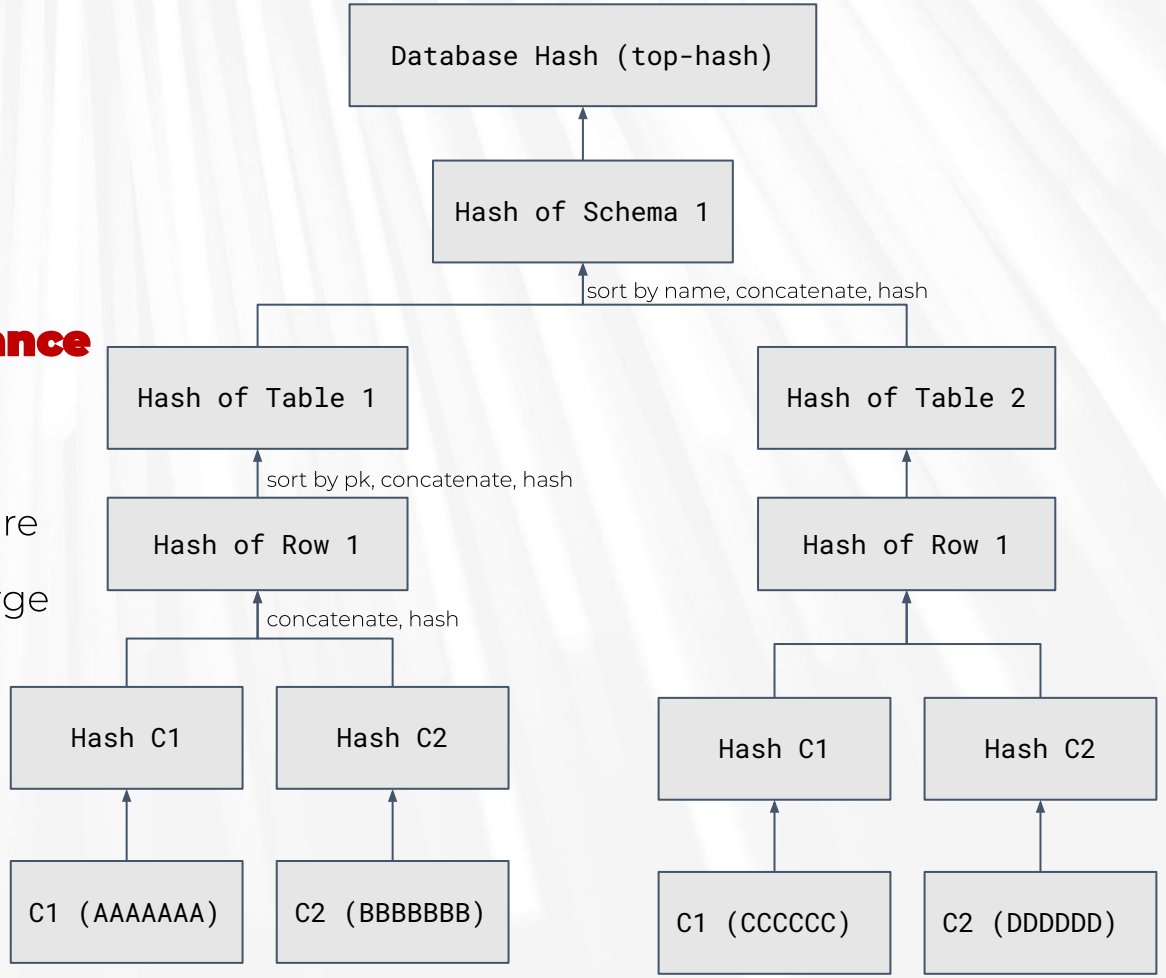
Ensure information in the database was not changed between export and archival confirmation

keep.
Preserving the future

Automated quality assurance

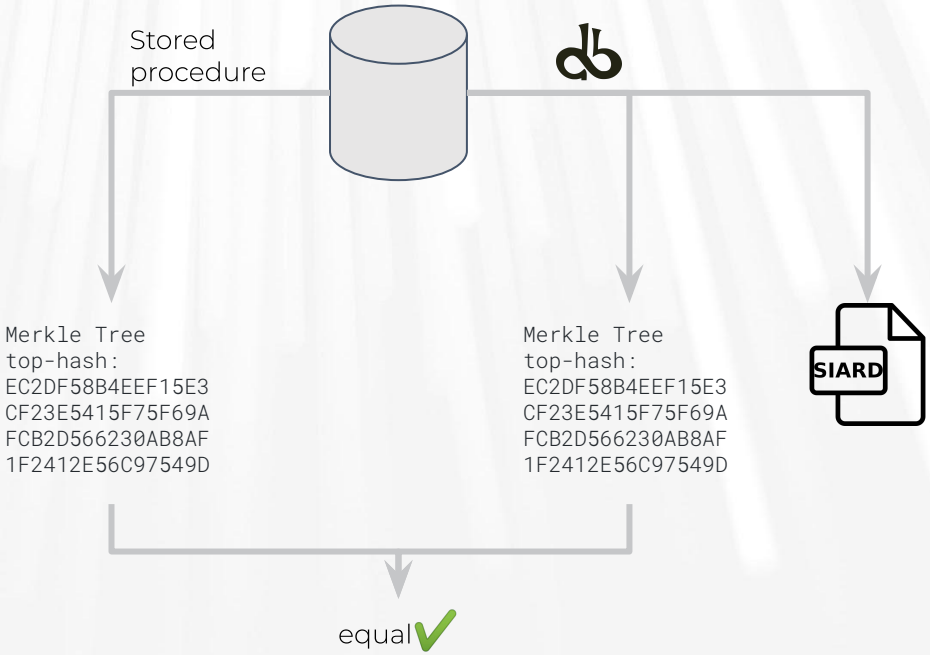
Merkle tree top-hash

A hash for the efficient and secure verification of the contents of large data structures.



keep.
Preserving the future

Automated quality assurance



keep.
Preserving the future

Save LOBs outside SIARD file

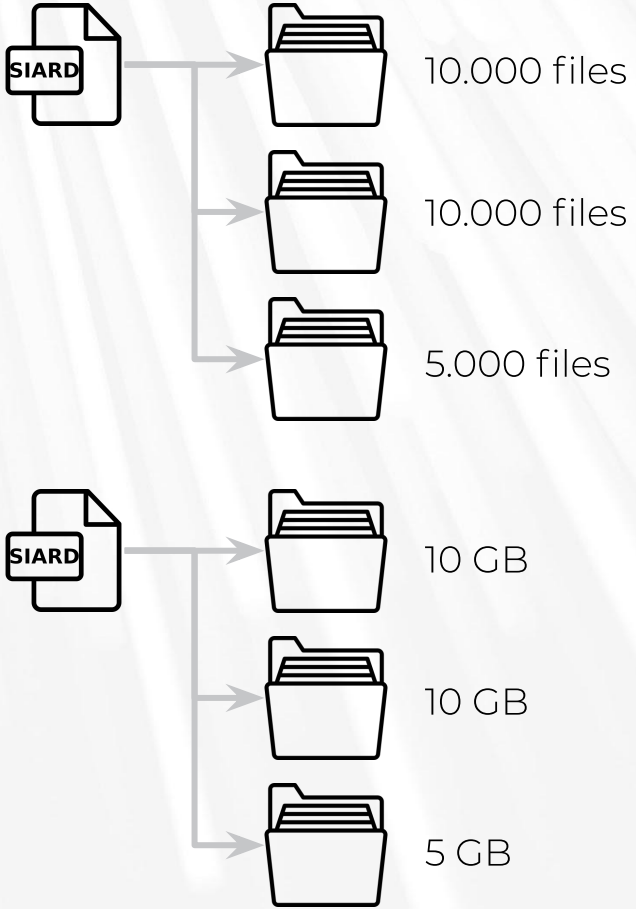
Large Objects (LOBs) are a set of data types that are designed to hold large amounts of data.

They can be large binaries or text files

Split very big SIARD files into manageable parts

Split by maximum number of files

Split by maximum storage size



keep.
Preserving the future

Migrate from SIARD to SIARD

Select tables and columns

Add local files

Upgrade SIARD version

Change SIARD options (compress, format XML, save LOBs outside, etc.)

keep.
Preserving the future

Migrate from SIARD to live DBMS

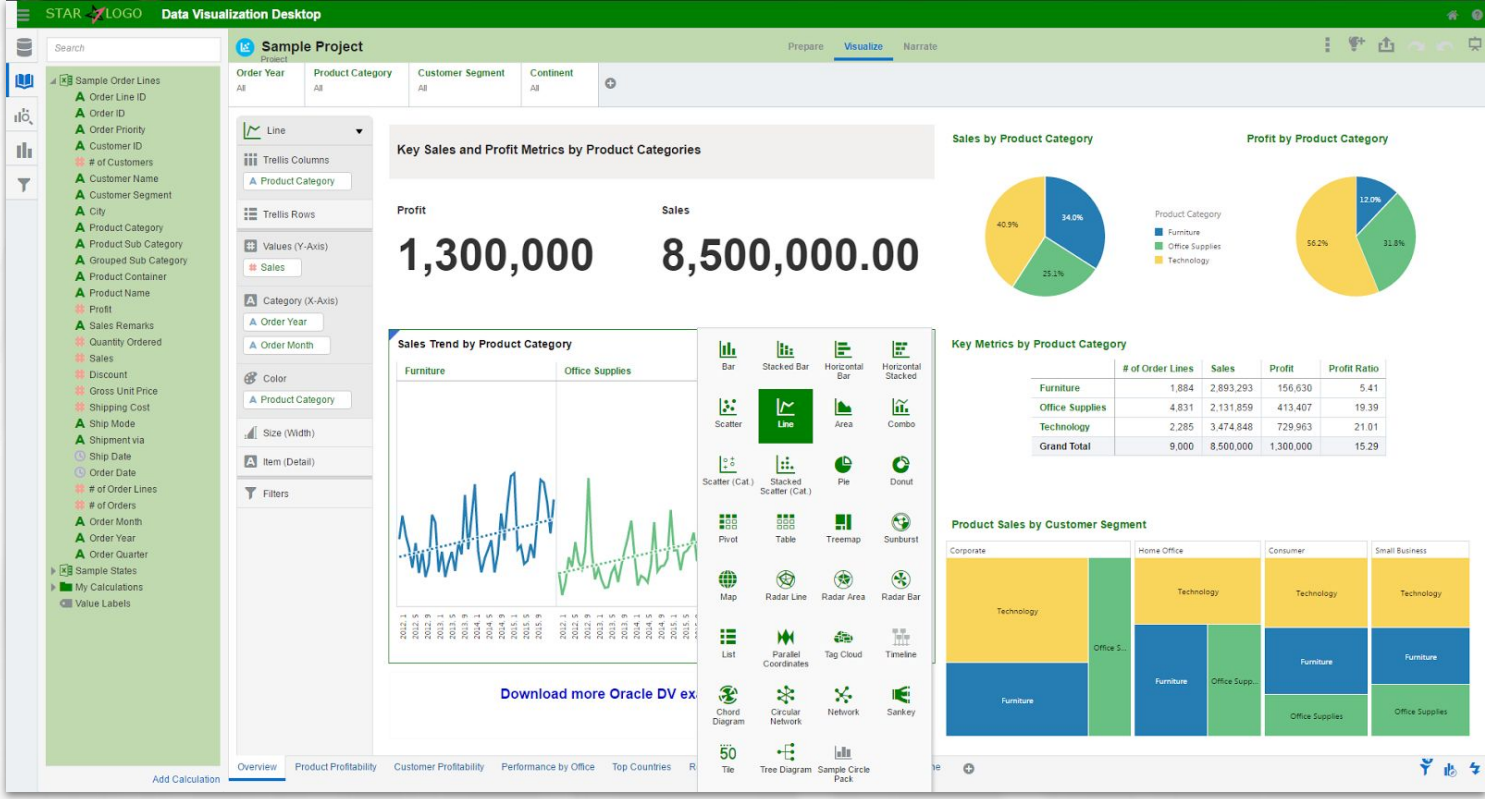
Import SIARD information into a database for restore or analysis

Supported: Microsoft SQL Server, MySQL, Oracle, PostgreSQL

NOTE 1: will not add behavior

NOTE 2: may not be able to add constraints, but will try

keep.
Preserving the future



Data load

Import archived data into modern database system

Use the full query power of a modern database engine and enable advanced analytics like data mining

www.keep.pt

DBPTK Desktop

Advanced features demonstration

DBPTK Enterprise

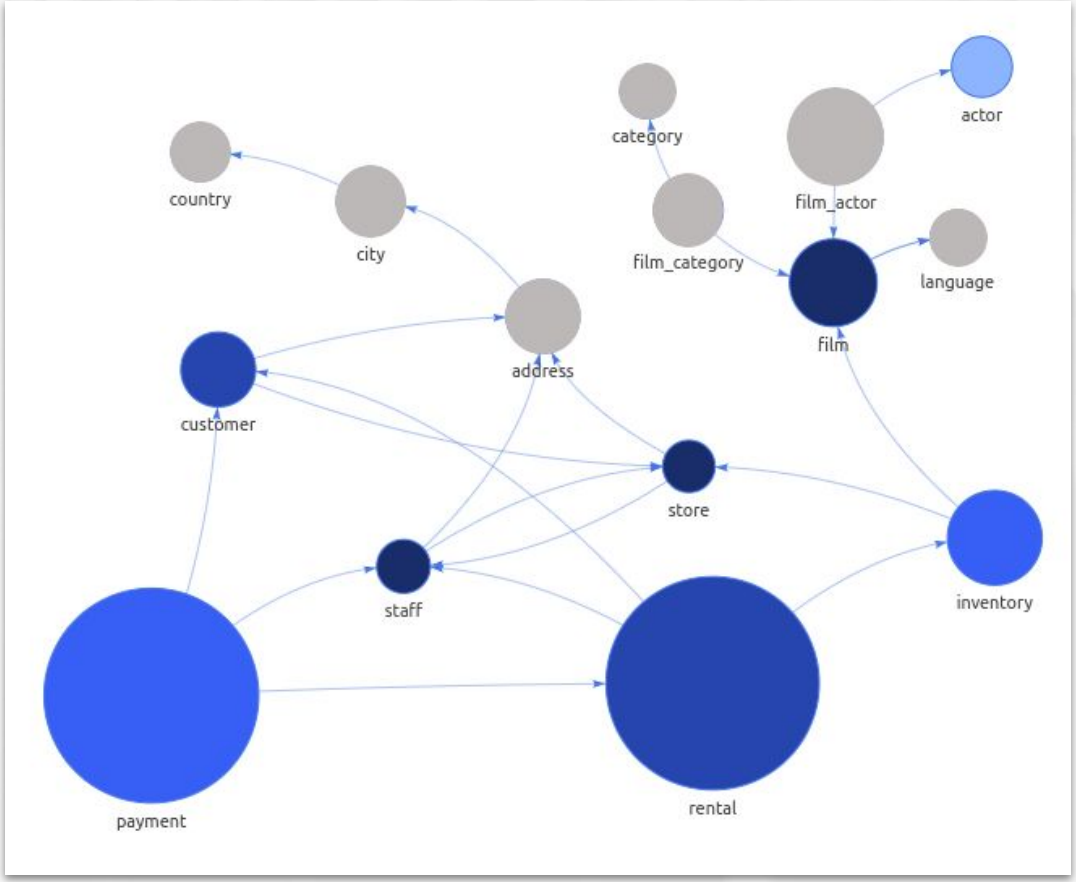
Advanced features

keep.
Preserving the future

Table management

Show/hide tables

Change table names and descriptions



keep.
Preserving the future

Column management

Show/hide columns

Change column name and description

Change column order

Show/hide column in search results, advanced search and row details

keep.
Preserving the future

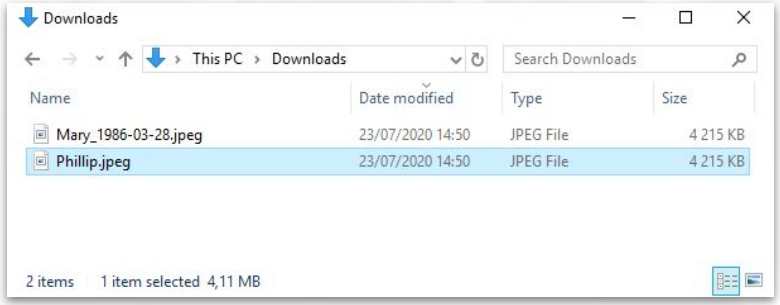
Binary (BLOB) column options

HTML template for the download link

Text template for the file name (can use the value of other columns in the same row)

MIME type

person		
Name	**Birth**	**Picture**
Mary	1986-03-28	download
Phillip		download

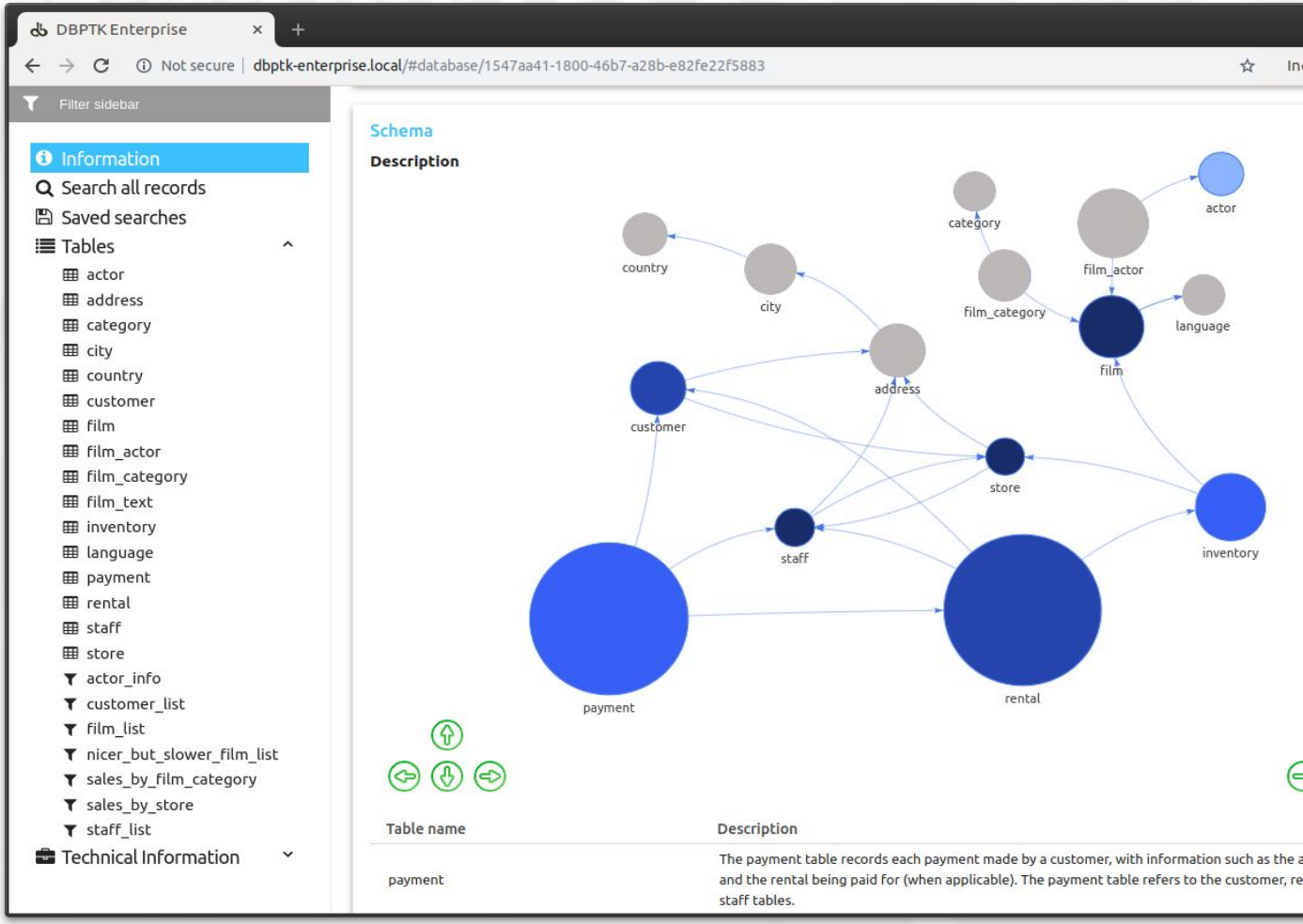


keep.
Preserving the future

Data transformation

Transform content to answer useful questions

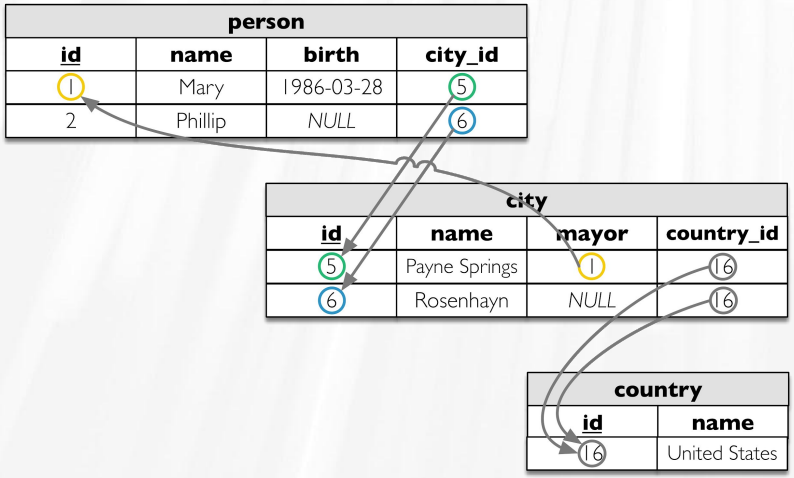
De-normalization and table and **column hiding**, to simplify browsing/search and allow **anonymization** of content



www.keep.pt

keep.
Preserving the future

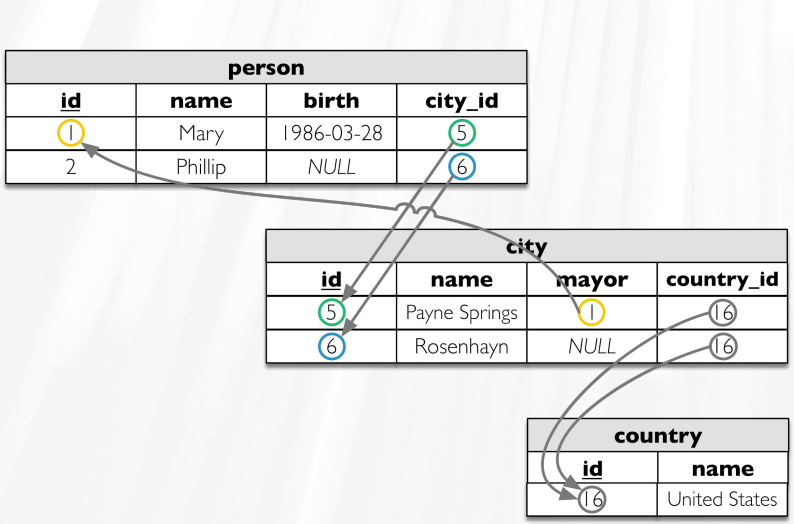
Data transformation (aka denormalization)



person				
Name	**Birth**	**City name**	**Mayor**	**Country name**
Mary	1986-03-28	Payne Springs	Mary	United States
Phillip		Rosenhayn		United States

keep.
Preserving the future

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

Country

Name	Cities
United States	Payne Springs, Rosenhayn

City

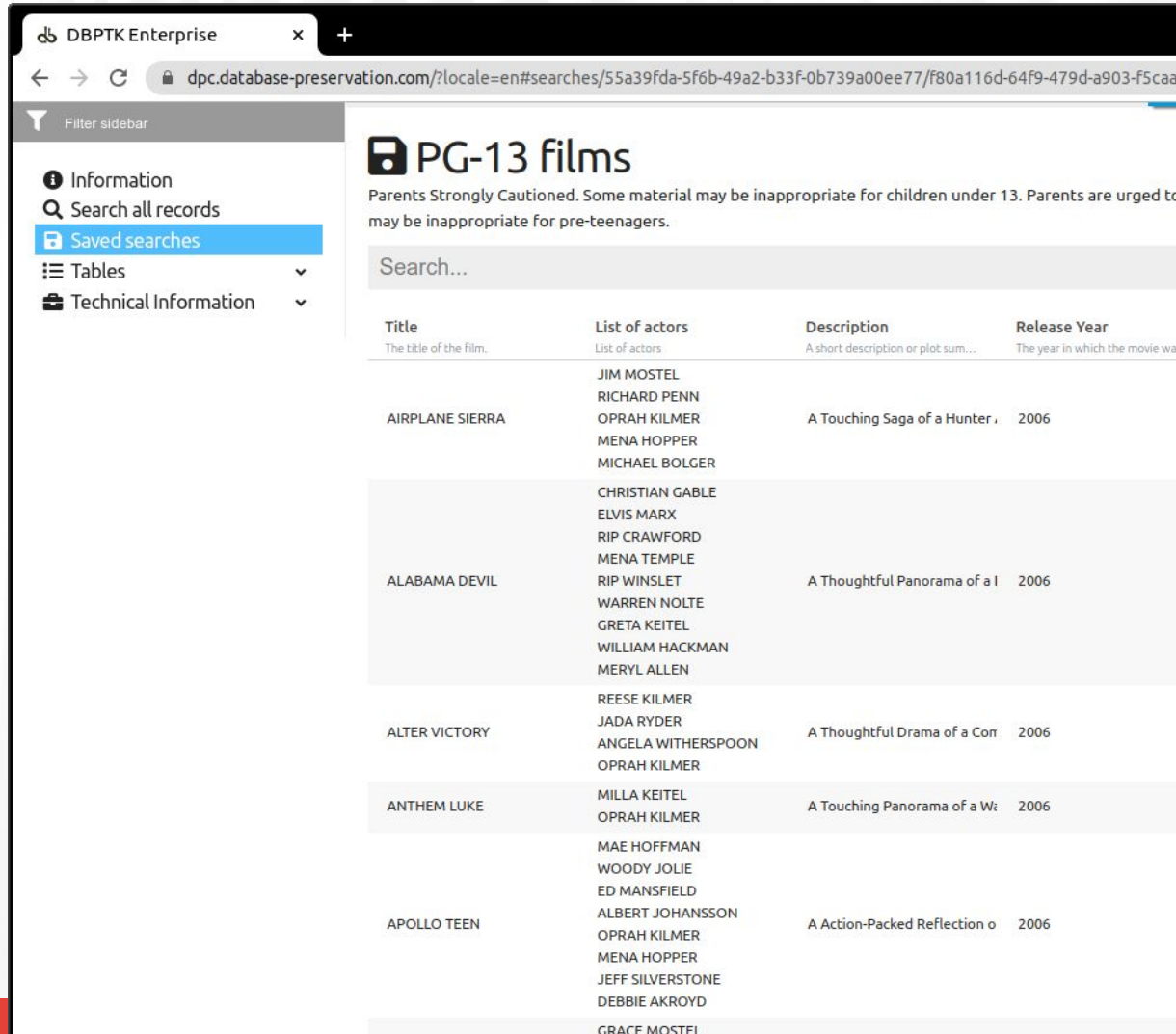
Name	Residents
Payne Springs	Mary
Rosenhayn	Phillip

keep.
Preserving the future

Saved searches

Use advanced search to filter the results

Save the search for reuse by all users

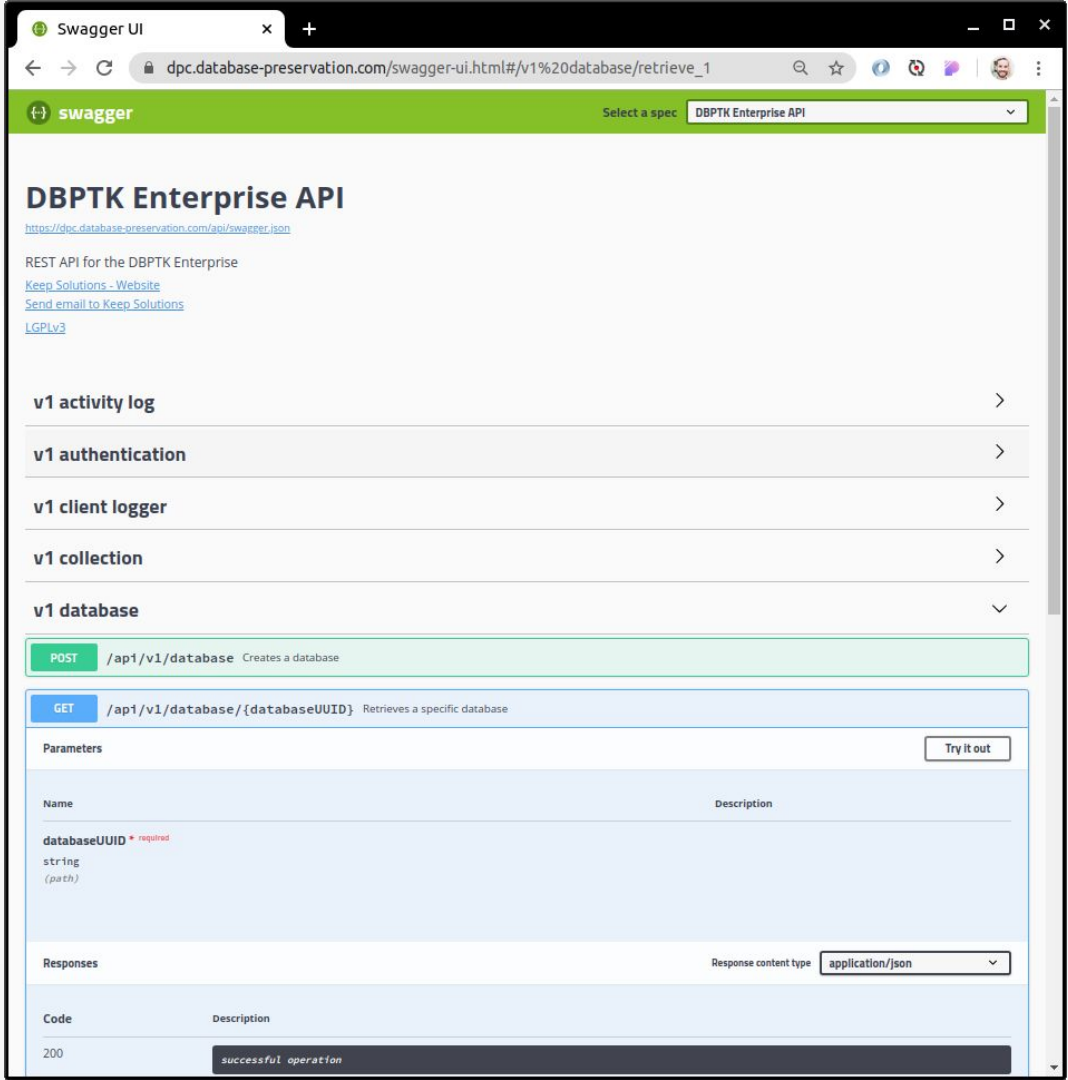


keep.
Preserving the future

REST API

Systems integration

Custom access portals



keep.
Preserving the future

Load on access and auto-unload

Option to **automatically load** databases when user accesses

Automatically unload after a configurable amount of time

Optimize used resources on systems with many databases

DBPTK Enterprise

Advanced features demonstration

DBPTK Developer

Advanced features

keep.
Preserving the future

Import config

Generate import config file

Edit import config file

Define parameters

Use and re-use the import config file

keep.
Preserving the future

Import config

YAML syntax

Select tables and columns

Select **where** and **sortBy**

Use **{{variables}}**

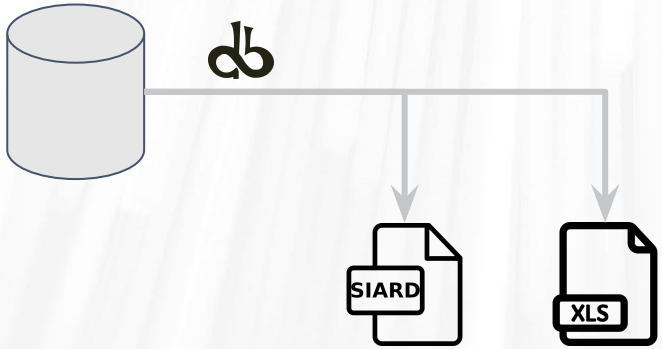
Set every **export parameter**

Set **filter parameters**

```yaml
import:
 module: "mysql"
 parameters:
   hostname: "localhost"
   port-number: "3306"
   username: "root"
   password: "123456"
schemas:
 sakila:
   tables:
     - name: "actor"
       columns:
       - name: "first_name"
       - name: "last_name"
       - name: "picture"
         externalLOB:
           basePath: "C:\\>database\\actor\\picture\\"
           accessMethod: "file-system"
       where: "last_update between '{{START_DATE}}' and '{{END_DATE}}'"
       sortBy: "actor_id ASC"
   views:
     - name: "actor_info"
       materialize: false
       columns:
       - name: "actor_id"
       - name: "first_name"
       - name: "last_name"
       - name: "film_info"
       where: ""
       sortBy: ""
```

keep.
Preserving the future

Inventory filter module



person	
Name	**Birth**
Mary	1986-03-28
Phillip	

Get a **list of what was exported** to SIARD, usually identifiers

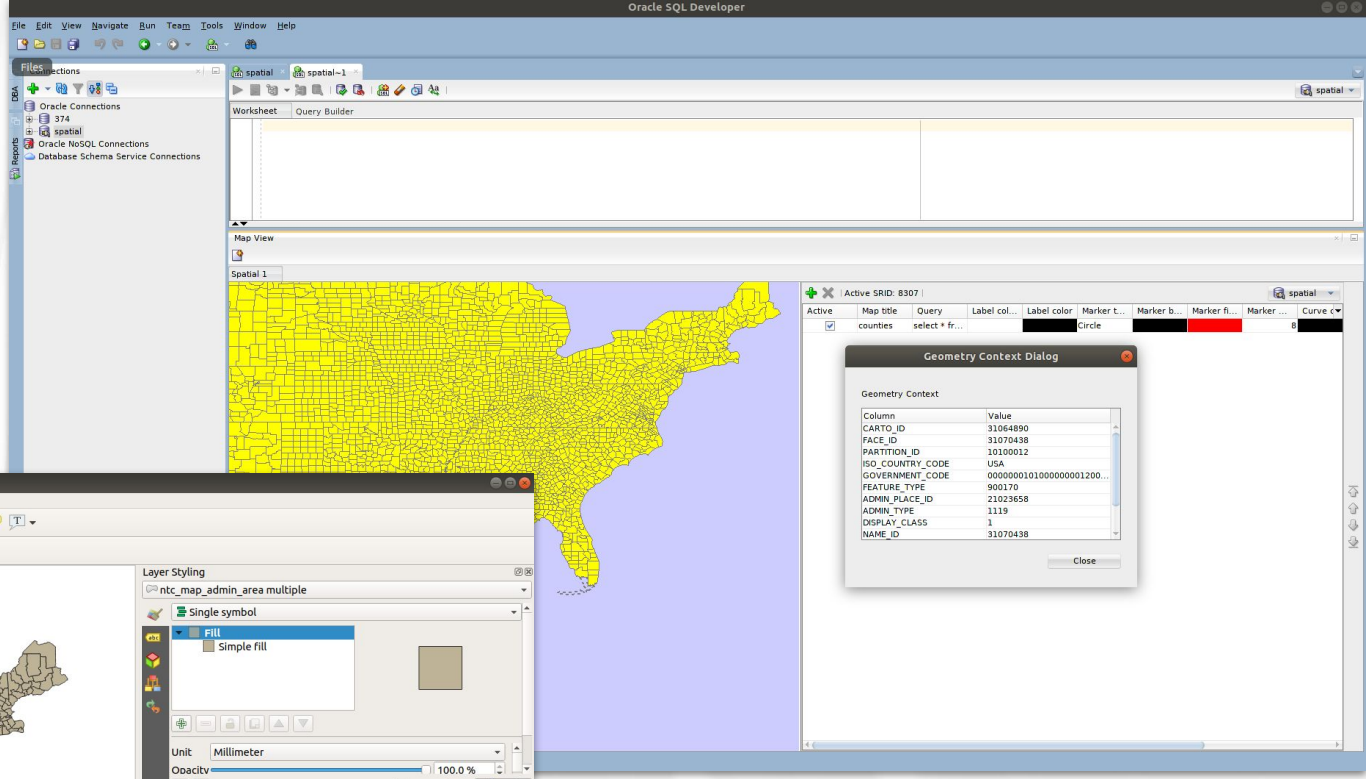
May later be used to mark records as archived

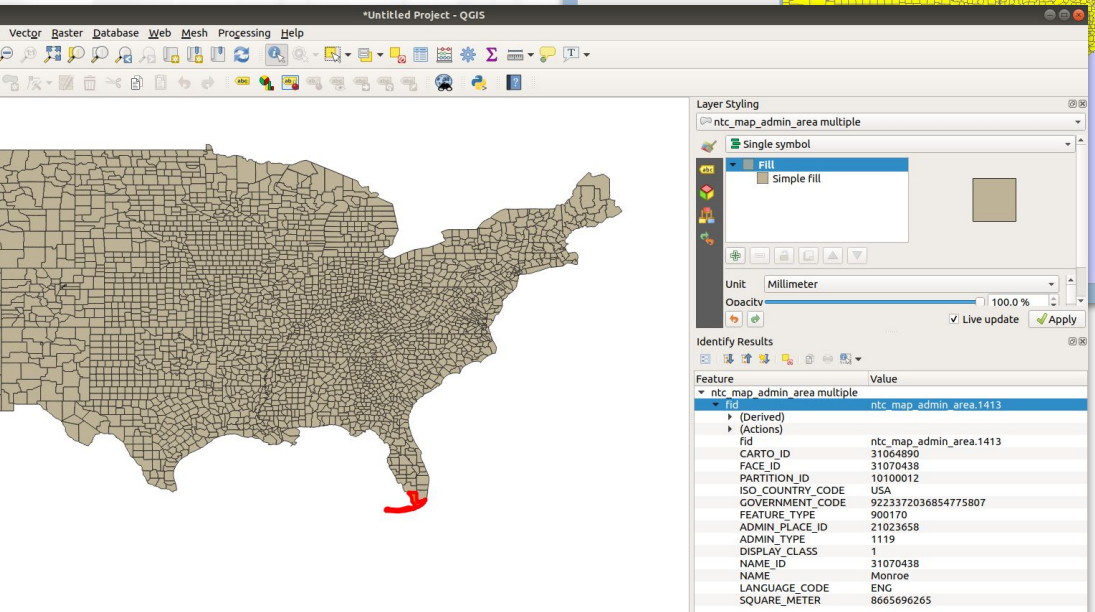
https://github.com/keeps/dbptk-developer/wiki/Inventory-Filter-Module

Convert Oracle geodata

Export Oracle geographical information

Closed specification only supported in specific Oracle license types





Into GML

An open international standard

Ready to be used with open-source desktop or Web tools.

keep.
Preserving the future

Custom import, export and filter modules

```
public interface DatabaseFilterModule extends DatabaseImportModule, DatabaseExportModule {
/**
  * The reporter is set specifically for each module/filter, so this call does
  * not need to be chained to the next DatabaseFilterModule
  *
  * @param reporter
  *          The reporter that should be used by this DatabaseFilterModule
  */
 @Override
 void setOnceReporter(Reporter reporter);


 /**
  * Import the database model.
  *
  * @param databaseExportModule
  *          The database model handler to be called when importing the database.
  * @return Return itself, to allow chaining multiple getDatabase methods
  * @throws ModuleException
  *           generic module exception
  */
 @Override
 DatabaseFilterModule migrateDatabaseTo(DatabaseFilterModule databaseExportModule) throws ModuleException;
}
```

keep.
Preserving the future

Fine-tuning

Fetch size

Controls the amount of rows that are retrieved from the database and stored in memory at once.

Oracle LOB **prefetch size**

Controls the amount of LOB that is prefetch for each row retrieved from the database and stored in memory at once.

Temporary files location and MapDB options

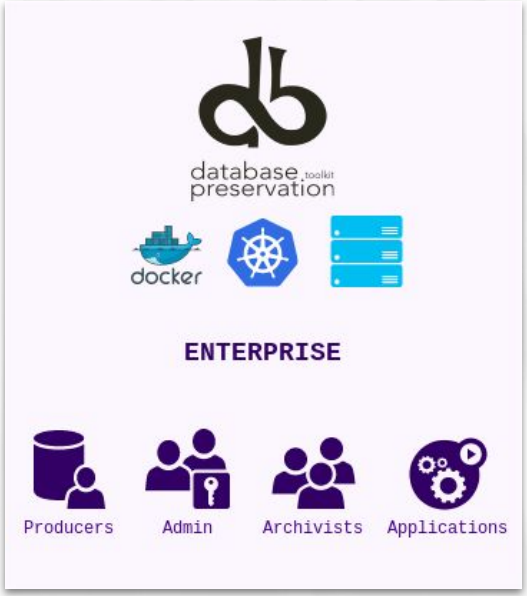
May need substantial space if SIARD is very large. MapDB is used for referential and entity integrity check.

Timezone options

Controls the timestamp field handling

How to choose?

keep.
Preserving the future



keep.
Preserving the future

	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

Issues and help desk

keep.
Preserving the future

Issues

Get information about the issue

DBPTK Desktop, menu Help > Logs

DBPTK Enterprise, docker logs

How to **reproduce** the issue

keep.
Preserving the future

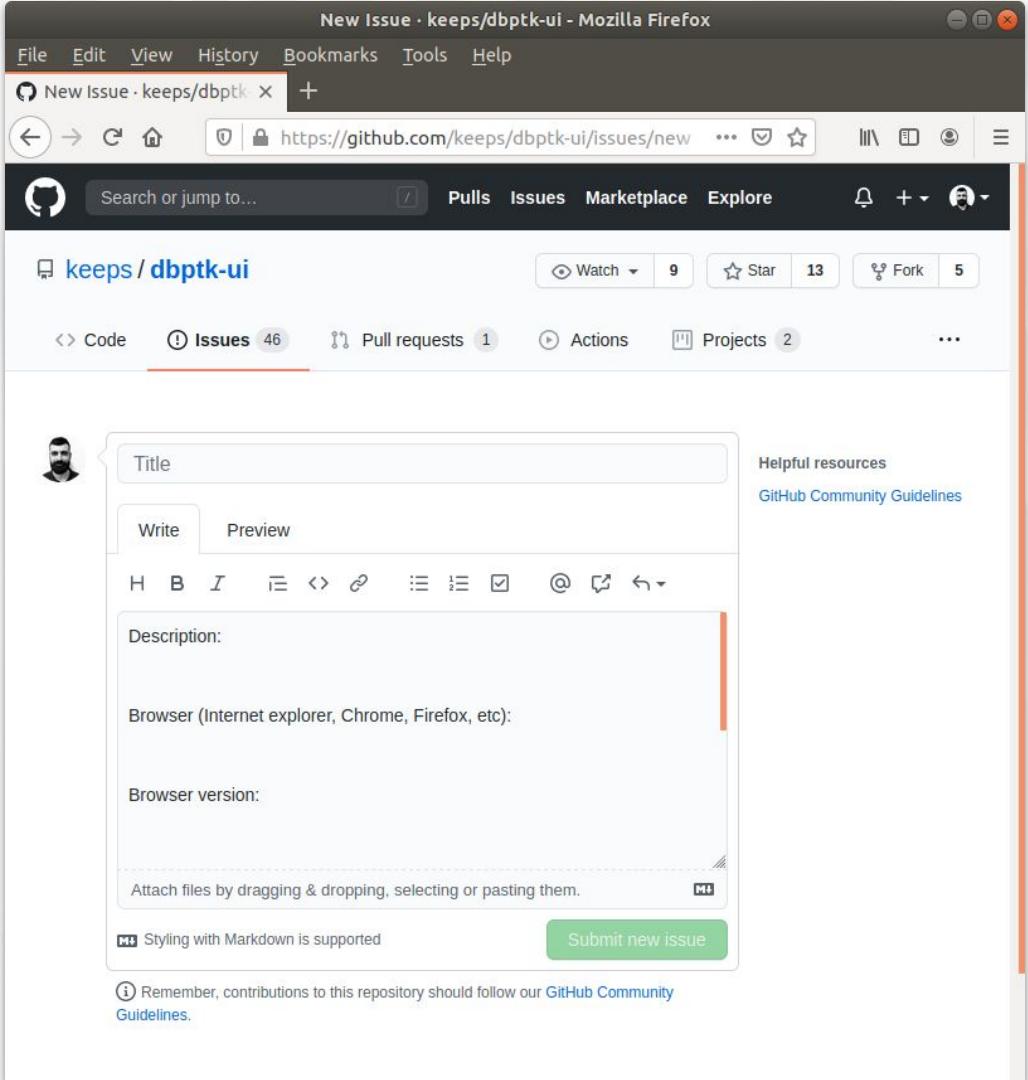
Issues

Submit the issue

Open-source support

Register free GitHub account

https://github.com/keeps/dbptk-ui/issues/new



keep.
Preserving the future

Issues

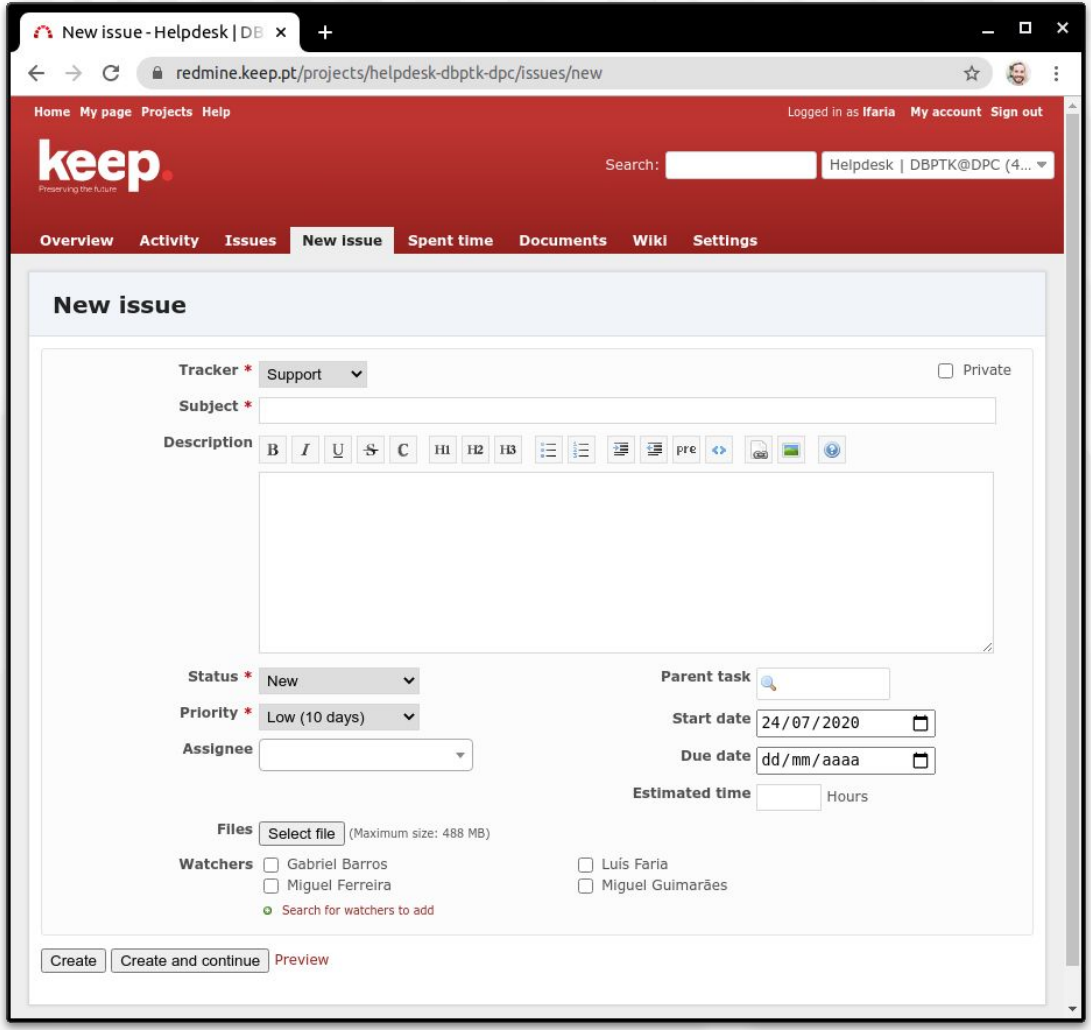
Submit the issue

Commercial support

Contact sales@keep.pt

Check DPC database preservation

commercial helpdesk service



20-min break

Back at 11h25

GMT+1

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**

keep.
Preserving the future

Hospital legacy databases

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

keep.
Preserving the future

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**.

keep.
Preserving the future

Hospital legacy databases

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

keep.
Preserving the future

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**.

keep.
Preserving the future

European Taxation and Customs Union: trader messages archive

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

keep.
Preserving the future

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

Third-party validation using Merkle Tree top-hash feature

Using inventory feature to mark messages as archived and ready to be purged/pruned

Using inventory to verify that no message missed being archived

keep.
Preserving the future

European Taxation and Customs Union: trader messages archive

Main software used

DBPTK Developer for continuous partial export to SIARD

RODA for archival, search and load into DBTPK Enterprise

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

Main features used

Import config with custom view with where filter and variables

Automated quality assurance with Merkle Tree top-hash

Inventory report with identifiers of archived messages

Load on access and auto-unload for browsable databases in DBPTK Enterprise

Case study: Implementing database archiving at the National Archives of Estonia

by Kuldar Aas, National Archives of Estonia

Questions and discussion

Lunch

Practical session starts at 13:45

GMT+1