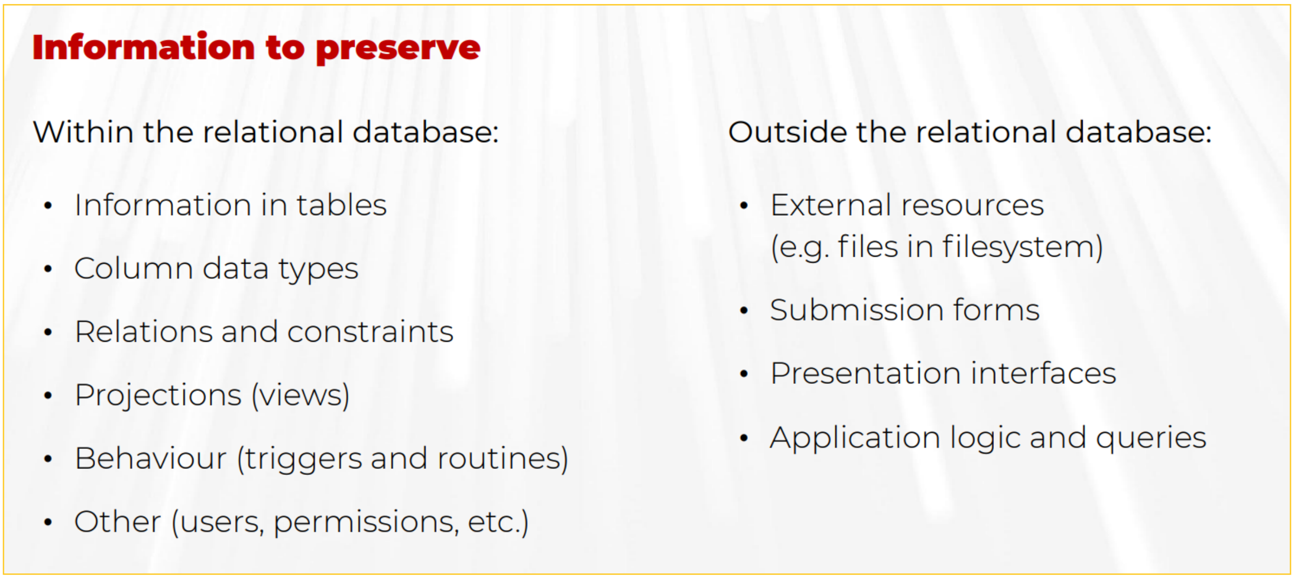
Preservation through standardisation with DBPTK

Luis Faria (KEEP Solutions LDA Company, Porto, Portugal)

Luis Faria is Research and Innovation Director at KEEP SOLUTIONS, working in research and development of solutions for digital preservation and information management since 2005. He holds a PhD in Computer Science with specialization in Digital Preservation by the University of Minho, has done his degree in Computer Science at the same University in 2005. Has participated in several research and development projects in the area of digital preservation, such as SCAPE, E-ARK, 4C and VeraPDF. He is co-author of preservation formats specifications SIARD 2 and E-ARK IP, and is manager of the open-source project RODA and Database Preservation Toolkit (DBPTK). Faria has been working on the challenge of database preservation for the last eight years, particularly looking at preservation through standardisation.

There is a lot of different information that may need to be preserved (see figure 16). Every database preservation strategy has advantages and disadvantages.

- Hardware and software museums have reproduction accuracy but are difficult to maintain
- Emulation also has good accuracy and an advantage of not needing to maintain hardware but is also difficult to maintain and set up and needs users to understand old systems.



Information to preserve

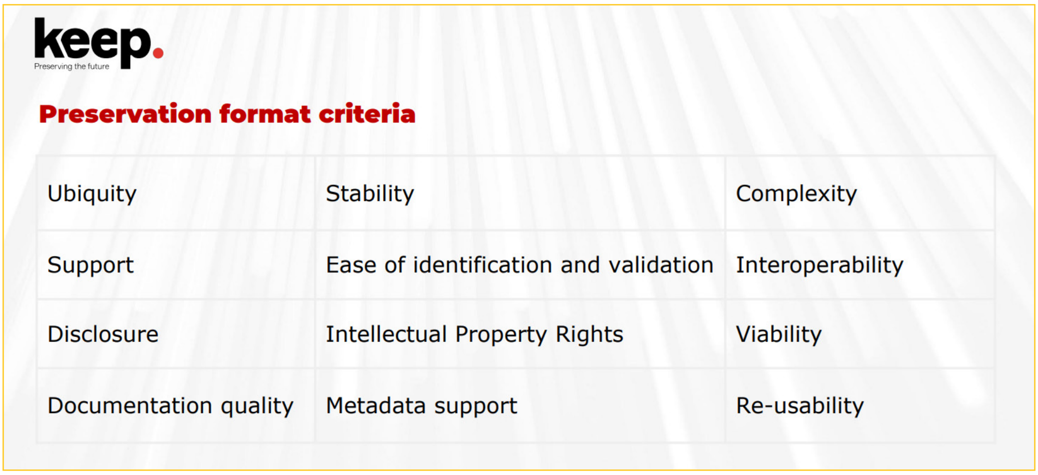
Within the relational database:

- 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

Figure 16: Information to preserve within and outside a database management system (DBMS)



keep.
Preserving the future

Preservation format criteria

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

Figure 17: Preservation format criteria according to Brown (2008) as presented by L. Faria

- File format migration makes it easier to use and reuse information, and there is no need to maintain hardware or software. The risk is that information may be lost in migration.
- Encapsulation keeps files together with all necessary documentation. This can postpone costly actions and means no need to keep hardware and software. Disadvantages are a huge cost for timely access to information, difficulties to gather documentation on all formats and to ensure quality.

When looking for a preservation format, Faria followed Adrian Brown (2008, see figure 17). SIARD scores highly on most of these criteria. It is based on international standards and is for database data, structure and behavior. A simple database archive flow would be that a producer submits a database file to the archive in SIARD formats. This is validated, put into a simple catalogue and may be viewed by the end user with the database viewer. A fuller workflow would include Keep's repository software RODA and more robust processes by the producer to document and capture the database.

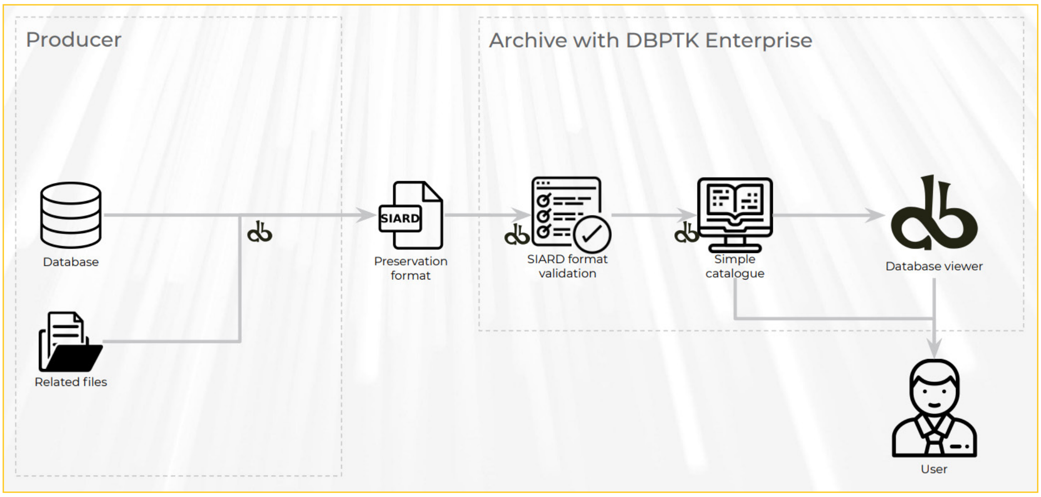


Figure 18: DBPTK simple database archive flow

The Database Preservation Toolkit (DBPTK) can be downloaded and used by anyone. There are three separate tools that can be used – DBPTK Desktop, Enterprise and Developer. The Desktop version features a connection to a database, stores the content in SIARD, and provides some problem-solving help. It can work with several formats. Users can create a migration report which notes migration changes and losses in the export. The user can then edit the SIARD metadata and enrich it. This will also allow for a full validation of the SIARD data.

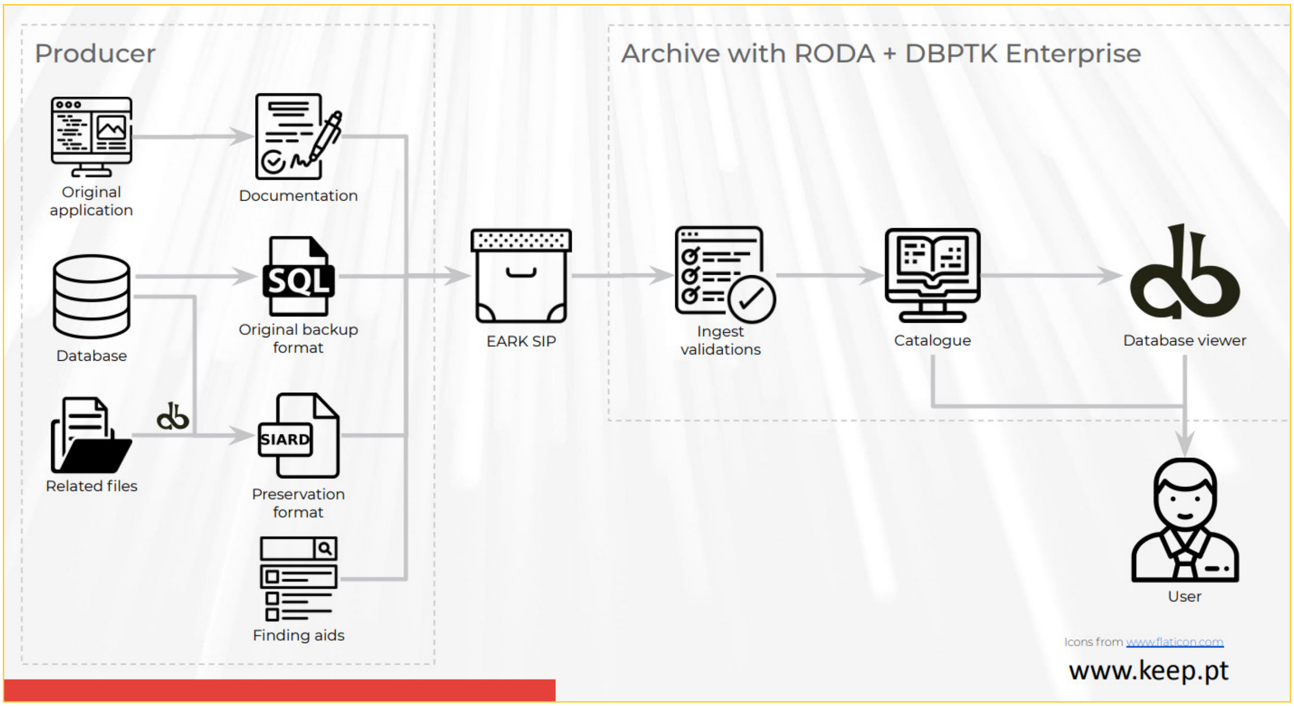
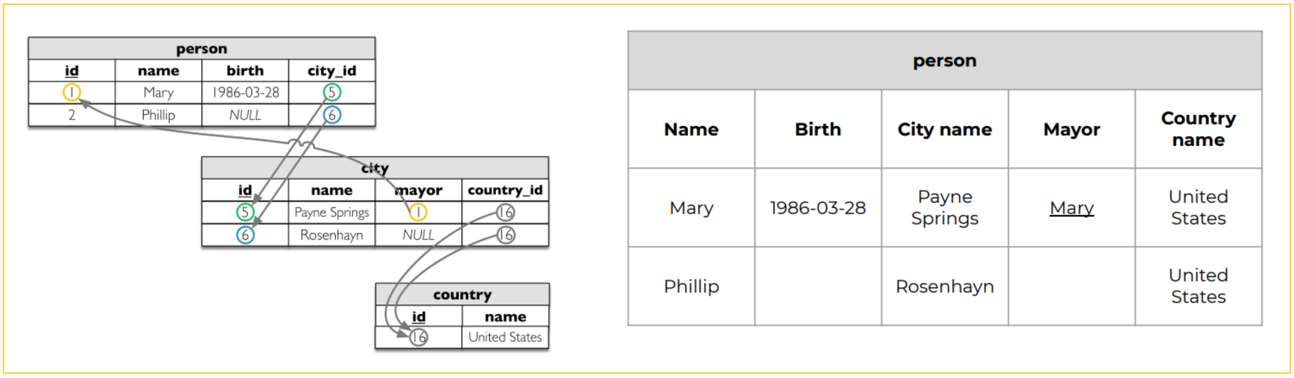


Figure 19: DBPTK full database archival flow

The Enterprise version aims at larger institutions with multiple users. These institutions can do data transformation (de-normalisation) and can provide access straight to their users through the web, allowing them to browse and search the database content. It also allows them to put SIARD back into a database including an activity log and supports multiple languages.

The Developer version features a command line and a java library. It is open source for custom development and allows specialised support for new or legacy database systems. Includes many other features for archiving databases and accessing archived databases.



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	Mary	United States
Phillip		Rosenhayn		United States

Figure 20: Data transformation (aka denormalisation) feature of DBPTK Enterprise

Faria brought two case studies: First, hospital legacy databases to support specific hospital use cases. Lacking sufficient documentation, the information is exported to SIARD where expert analysis creates the missing documentation. Second, a European taxation and customs union: trader messages archive – this is a new European Union (EU) service that will provide a centralised interface with customs authorities. All transactional messages must be archived for a decade. The productive system exports parts of the database to SIARD every hour, through RODA services, in a continuous extract/archive/validation workflow.

Questions and discussion

- Kai Naumann asked where DBPTK and RODA are installed – they mostly sit at the computing centres chosen by the agencies they support but can be deployed elsewhere if networks allow.
- Torbjørn Aasen asked how customers could create a migration log and preserve it alongside the data. – Currently, DBPTK has no way to include the report within the SIARD file as it is produced after the SIARD file has been created. The report can be taken separately but could use an information package such as the E-ARK SIP to package it up with the database content.
- Boris Domajnko asked which log would be best to keep with the SIARD file. Who will look at these after 10 years? What information is relevant to future users? Faria pointed out that the reports should also be subject to some standardisation. There is an execution log that is not intended as a report but will be useful if something goes wrong. He said you should keep almost everything (just in case) but some things are easier to keep than others.
- Faria mentioned Merkle Hash Trees (see references) which allow you to create checksums for every row and column that can be joined hierarchically allowing easy validation against manipulations.