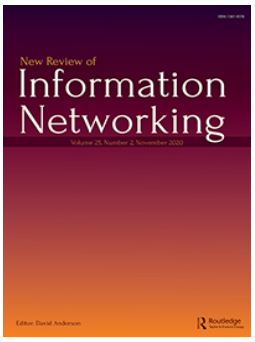

Routledge
Taylor & Francis Group

New Review of Information Networking

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rinn20

Beyond Relational Databases: Preserving the Data

José Carlos Ramalho, Bruno Ferreira, Luis Faria & Miguel Ferreira

To cite this article: José Carlos Ramalho, Bruno Ferreira, Luis Faria & Miguel Ferreira (2020) Beyond Relational Databases: Preserving the Data, New Review of Information Networking, 25:2, 107-118, DOI: 10.1080/13614576.2021.1919398

To link to this article: https://doi.org/10.1080/13614576.2021.1919398

Published online: 14 May 2021.

NEW REVIEW OF INFORMATION NETWORKING
2020, VOL. 25, NO. 2, 107–118
https://doi.org/10.1080/13614576.2021.1919398

Routledge
Taylor & Francis Group

Beyond Relational Databases: Preserving the Data

José Carlos Ramalho[a], Bruno Ferreira[b], Luis Faria[b], and Miguel Ferreira[b]

[a]Department of Informatics, CCTC Research Center, University of Minho, Braga, Portugal; [b]KEEP SOLUTIONS, LDA, Braga, Portugal

ABSTRACT

Relational databases are one of the main technologies supporting information assets in today's organizations. They are designed to store, organize and retrieve digital information, and are such a fundamental part of information systems that most would not be able to function without them. Very often, the information contained in databases is irreplaceable or prohibitively expensive to reacquire; therefore, steps must be taken to ensure that the information within databases is preserved. This paper describes a methodology for long-term preservation of relational databases based on information extraction and format migration to a preservation format. It also presents a tool that was developed to support this methodology: Database Preservation Toolkit (DBPTK), as well as the processes and formats needed to preserve databases. The DBPTK connects to live relational databases and extracts information into formats more adequate for long-term preservation. Supported preservation formats include the SIARD 2, created by a cooperation between the Swiss Federal Archives and the E-ARK project that is becoming a standard in the area. DBPTK has a flexible plugin-based architecture enabling its use for other purposes like database upgrade and database migration between different systems. Presented real case scenarios demonstrate the usefulness, correctness and performance of the tool.

KEYWORDS

Digital archiving; eArchiving; database preservation; database preservation toolkit; E-ARK; SIARD; databases; relational databases; RDMS

Introduction

It is common for governments to maintain databases with important and invaluable information, such as tax data, pension records and judicial information. This information is crucial to maintain both the rights and obligations of citizens. Other organizations, such as hospitals and laboratories, also have critical medical information for researchers attempting to understand and solve health-related problems. This valuable information must be kept authentic and accessible over time.

The digital preservation field tries to define the policies and processes necessary to keep information accessible and authentic through long periods of time. However, for databases, this task is especially challenging because they

CONTACT José Carlos Ramalho ✉ jcr@di.uminho.pt Department of Informatics, CCTC Research Center, University of Minho, Campus de Gualtar, Braga 4700-057, Portugal

are complex objects and present unique challenges to traditional digital preservation methods. The heterogeneity of information in databases is difficult to transport through time, and the variety and popularity of Database Management Systems that make use of proprietary storage formats or use nonstandard query languages further hinders the preservation process.

The next section describes the scope of this work and how it may be used to improve existing scenarios. Following this, databases and their significant properties are described, followed by the SIARD 2 format. The tool that is able to convert databases between different database systems and formats is described in the toolkit for archiving databases section and is followed by the testing and quality assurance of the same tool. The final section wraps up this work and includes some ideas of what can be expected from future Database Preservation Toolkit versions.

Use cases and scope

Today there are many kinds of databases: relational, object, graph, noSQL, and others. This article focuses on relational databases because they are currently the most widely used. A relational database is a digital object whose organization is based on the relational model of data, as proposed by E. F. Codd in 1970 (Codd 377–387). The various software systems used to maintain relational databases are known as a Relational Database Management System (RDBMS).

In producer institutions or agencies, relational databases are rarely alone. They are usually coupled with an application (web, desktop or other) that provides an interface to the user and allows to create and access the information kept in the database. In some cases, the RDBMS itself can develop and host such applications, providing an integrated solution.

These systems can become obsolete and be replaced, where information can be partly migrated, or simply cease to be of service to the institution and be decommissioned. The producer institution might then want or need to archive this information. Also, there are cases where long-lived systems might need to archive information periodically.

Whatever the case, the producer institution needs to package all information contained in the systems and transfer it to the archive (which may be in the same institution or another). The archived information needs to be preserved for a long time. Legal obligations often define this limit in the order of decades. Other cases, such as health records, define this limit in the order of generations.

After the archival process, consumers may request access to the preserved data. The way they will use the information will be inherently different from its original purpose. Consumers want to understand the data, how it was used, and be able to reuse it, possibly employing state-of-the-art analytical tools.

Significant properties

Digital preservation refers to the sum of activities (procedures, standards, best practices and technologies) necessary to ensure the long-term access to digital information. Information, in the digital preservation research field, is formally referred to as a digital object, which is an information object, of any type of information or any format that is expressed in digital form (Thibodeau), which in this case would be a relational database.

In an OAIS-compliant archive (ISO 14721:2012), the producer needs to package the information in a Submission Information Package (SIP) and send it to the archive, which creates an Archival Information Package (AIP) from the SIP and preserves the information for the following decades (possibly indefinitely). In the future, when someone requests access to the preserved database, the archive should provide the database (in a format usable by the designated community) packaged in a Dissemination Information Package (DIP), so that the person can access the information in the database.

Digital objects possess some properties that are essential to experience the digital object in the future; these are called significant properties. These characteristics may vary depending on the object type, organization policies and the object's future purpose. Looking at our case, when preserving a database the data is commonly a significant property; but other properties can be considered significant, such as the structure (tables, attributes, keys) or the metadata (title, creator, preservation begin date) (Dappert and Farquhar; Grace et al.; Hockx-Yu and Knight; Wilson).

Not all properties are significant, and while some may be significant for a purpose or organization, others may be optional for other applications. There is widespread agreement on the important role significant properties play in digital preservation processes, but no consensus on what is significant. In most cases, the ideal significant properties are the ones required by the designated community, but as there is no current way to determine those either. Organizations try to make educated guesses based on past experiences and opt for some property subset that is allowed by current technical and financial limitations (Dappert and Farquhar; Freitas and Ramalho; Grace et al.; Hockx-Yu and Knight).

Databases are complex digital objects that contain heterogeneous information, often accompanied by structural definitions and even documentation. Their complexity makes it difficult to preserve this kind of object whilst maintaining all its significant properties. Formally, a database is *a shared collection of logically related data, and a description of this data, designed to meet the information needs of an organization*. But they are usually enclosed in a DBMS, which is *a software that enables users to define, create, maintain and control access to the database*; and provides specific languages to define and manipulate the database, which is usually some variation of SQL (Connolly and Begg).

In the RDBMS data model, known as the *relational model*, all data is structured in **tables** (or relations), **table columns** (or attributes) and **table rows** (or tuples) containing one value per column (corresponding to a **cell**). Each column has a specific **type** (or domain) which defines the set of allowable values for cells in the column. The relational model also includes **relational keys**, in which **primary keys** are used to uniquely identify a row in a table, and **foreign keys** are used to relate a pair rows (typically from different tables). These keys are essential to relate pieces of information inside a relational database and may enforce rules to keep the database consistent. Also, there are **views**, which is a virtual/derived table that is dynamically created from the underlying base tables when required. Views may be used to hide confidential information or join, transform and display information from multiple tables in a way that is easier to understand.

In order to improve on current database preservation solutions, the E-ARK project attempted to define the significant properties for relational database objects, since there was no consensus on which properties were significant. After contributions from various institutions, public and private, as well as creators and users of existing database preservation formats, the project was able to decide on which properties were significant. These were:

- Content: It is essential to save the database content information (i.e. the cell values). This content might have to be transformed to comply with the format guidelines, and that transformation is acceptable as long as no information is lost (e.g. converting a date to a different date format).
- Binary content: Cell values can also contain binary data, like pictures or audio. The binary data should also be preserved.
- Relational structure: The relational structure is needed to keep database contents organized and relatable. Therefore, the tables, their columns and data types, the relations between tables, and other structural information are essential to correctly perceiving the database contents in the future.
- Behavior information: The significant behavioral properties are users, roles, permissions, views, triggers, routines and constraints; as they document how the database operated.
- Descriptive information: Descriptions for tables, columns, types, triggers, and other database elements are valuable assets to understand the database.
- External software interactions: Original application that interacted with the database, like a website or a set of forms used to change values in the database.

All of these properties can be obtained from the database, except for the external software interactions. However, from a long-term digital preservation standpoint, it could be argued that the designated community might be only interested in the database data and not even be able to use the outdated technologies involved in the original software. For example, when using a contacts application backed by a database, its present uses would include using the software to add and update contacts, while in the future the designated community might be interested in using the data for demographic studies (which would not be possible to do in the original software). Furthermore, attempting to preserve the original software and database system would require using an emulation strategy, which would probably be more expensive than format migration, and it might be impossible if the software (or part of it) is protected by intellectual rights laws (Ferreira).

Since the designated community will probably not have the need to use the original software that interacted with the database and preserving software is a complex subject with approaches of their own (and a whole research domain that studies it), this work focuses on the preservation of information within the DBMS. For cases in which the software interactions might be important to understand the database, information about the software can be added to the Submission Information Package as documentation (e.g. in the form of screenshots and textual descriptions).

Archival format

To transport all relational database significant properties (except for the external software interactions) through time, the SIARD 2[1] format was developed by a cooperation between the Swiss Federal Archives and the E-ARK project, using lessons learnt by creators and users of database preservation formats. The format relies on stable and widely accepted technologies, for instance, XML (Sperberg-McQueen et al.), Unicode (The Unicode Consortium) and URI (Masinter et al.). Compared to the original SIARD, the second version of the format uses an updated version of the SQL standard (SQL ISO Standard 9075:2008 (ISO 9075:2008)) and supports a wider range of database elements, such as arrays and user-defined types.

Figure 1 represents the actions happening on the producer's side during the pre-ingest and ingest phases. The producer gathers information to document the application that uses the database and migrates the database to SIARD 2 format using the Database Preservation Toolkit. The producer then packages the documentation and SIARD 2 file in a Submission Information Package and submits it to the archive. After validating the Submission Information

[1] Software Independent Archiving of Relational Databases 2, standard eCH-0165 https://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165&documentVersion=2.0.

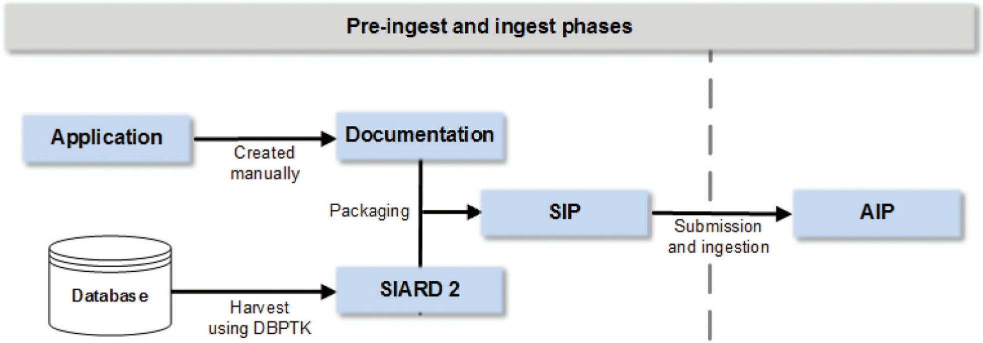

Figure 1. Packaging and submission of a SIARD 2 database and application documentation.

Package, the archive ingests it, producing the Archival Information Package which is maintained according to the archive's policies.

The access phase is depicted in Figure 2. A consumer can request access to a database, and the archive will handle the creation and delivery of a DIP containing the documentation about the original application and the database in SIARD 2 format. The consumer can then use the Database Preservation Toolkit to load the SIARD 2 database into a Database Management System. If the system supports Online Analytical Processing (OLAP) or some type of Knowledge Discovery technology, these can be used on the preserved database.

A toolkit for archiving databases

The DBPTK was developed to support information extraction, and relational database format migration,[2] since these processes, if done manually, are impractical, time-consuming and error-prone; therefore the need arose for a tool that could provide this functionality in an automated way, requiring as little human input as possible to avoid human-related errors. The tool is open source and free software which uses a modular architecture, allowing the combination of an import module and an export module to enable the conversion between database formats. The import module is responsible for retrieving the database information (metadata and data), whilst the export module transcribes the database information to a target database format. Each module supports the reading or writing of a particular database format or DBMS and functions independently, making it easy to plug in new modules to add support for more DBMS and database formats. The conversion functionality is provided by the composition of data import with data export. As all the import and export modules can be used interchangeably, any import module can provide information to any export module, and thus the software can be used to convert databases from DBMSs to preservation formats and from preservation formats back into DBMSs (B. Ferreira). The most popular relational DBMSs are[3] Oracle, MySQL, Microsoft SQL Server, PostgreSQL, DB2 and Microsoft Access. From this list, Oracle, MySQL, Microsoft SQL Server and PostgreSQL are supported as input and output systems, and Microsoft Access is supported as an input system.

An essential pair of modules for the preservation of relational databases are the SIARD2 import and export modules, which are used to convert databases to and from this database preservation format.[4] The main framework that supports the import and export modules includes an internal data structure

[2] The Database Preservation Toolkit is available for download at http://www.database-preservation.com/.
[3] According to the ranking at http://db-engines.com/en/ranking.
[4] Import and export modules for the SIARD1 format also exist, which allow the upgrading of databases in SIARD1 format to its most recent version.

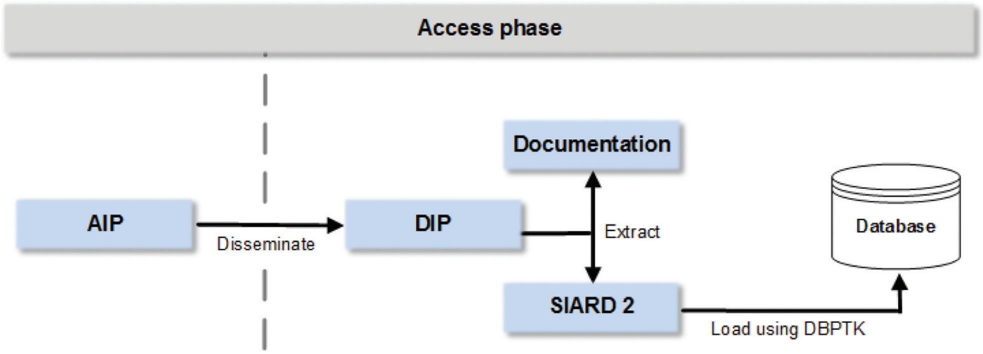

Figure 2. Extracting and loading of a SIARD 2 database.

based on the SQL ISO Standard (ISO 9075:2008) that is able to maintain all the information from the database. During the conversion, the import module extracts the database to this internal data structure, and the export module converts it to the output system or format. The conversion was engineered to process small chunks of data at a time, ensuring minimal processing resources are used during the conversion.

During the conversion, the software generates a report of all potential changes and other information about the conversion that may be relevant to the information producer or archivist. The report is needed because specificities present in some DBMS are not fully compatible with the SQL ISO Standard and need to be transformed before being added to SIARD. Registering these transformations in the report improves the chances to successfully preserve the database as well as improving its authenticity by documenting the conversion process.

The DBPTK also includes support for the extraction of binary contents present in databases. These databases are often very large, and when converted, they can become large SIARD files that may be hard to manage by the archive (e.g. they might not fit on a single media). To mitigate this problem, both the SIARD2 format and the application support distributing the database binary contents across multiple folders. The software can limit the number of binary files per folder and the total size of binaries in a folder and starts a new folder whenever a limit is reached. In practice, with the right settings, the application can produce a SIARD file (without binary contents) and several more easily manageable folders with the database's binary contents.

Some databases contain sensitive data that may not be freely distributed or might have some tables containing private corporate or personal information that may not be sent to the archive. To archive these databases and still cope with the data owner's data protection requirements, the tool supports filtering out one or more tables in a database.

Testing and quality assurance

To ascertain the quality of the conversions made using the Database Preservation Toolkit, a testing system was developed. The system was named *roundtrip testing* and is used to check if converting a database to a different database management system or format and then converting it back to the original database management system or format results in any data changes or losses. For most purposes, this test is equivalent to an identity test, validating the completeness and correctness of the conversions.

The Database Preservation Toolkit test suite contains the roundtrip tests, along with unit and integration tests. These tests are executed when the application is compiled, ensuring that any change that breaks existing functionality is noticed. The code is also submitted to continuous integration and continuous analysis systems, to ensure the tests also pass in different environments and to inform the developers of variations regarding the code quality.

The Database Preservation Toolkit was developed in the context of the E-ARK project, and as such, was subject to validation through piloting. The Danish National Archives piloted the Database Preservation Toolkit to make four successful data extractions from live databases into the SIARD 2 format. These were databases from:

- The Health System of the Danish Serum Institute, containing information from reported infectious diseases at the national level, and infectious diseases for all Danish citizens. This is a Microsoft SQL Server database containing 90 tables and around 500,000 records.
- A private sector business, Kultunaut Aps., containing information about Danish cultural events, from the smallest local gatherings to international exhibitions and events. This is a MySQL database containing five tables and 33 million records.
- The administrative system of the Danish National Archives, containing information about all the incoming scientific research data, and public deliveries of research data. This is a Microsoft SQL Server database containing 289 tables and 1.3 million LOBs.
- The administrative and health records system, from the Ministry of Higher Education and Science, containing information about social, psychological, and psychiatric counseling to students. This is a Microsoft SQL Server database containing 180 tables and 100,000 LOBs.

The National Archives of Hungary also piloted the Database Preservation Toolkit to convert several databases from Oracle DBMS to SIARD 2, including a database that is not normalized and contains more than 300,000 cases of the Hungarian Prosecution Office, and at least two more databases with different characteristics and containing both restricted and open content. The Database Preservation Toolkit was then used to convert the databases from SIARD 2 back to the Oracle DBMS. The National Archives of Estonia pilot used the Database Preservation Toolkit to migrate a database with a complex structure and around 200,000 records to the SIARD 2 format. The National Archives of the Republic of Slovenia migrated Microsoft Access databases to SIARD 2 and provided valuable feedback to improve the tool.

Conclusions

Databases are commonly used to store critical information that would be very hard to recover if it was lost or destroyed. Some of that information must remain accessible for many decades (e.g. for legal purposes). And since the Database Management System and external applications used to interact with the database are subject to software deprecation, an effort to create the processes, tools and formats needed to preserve databases was made by the E-ARK project.

The project was able to establish the processes for database archiving and access, as well as the software needed to support those processes. In cooperation with the Swiss Federal Archives, the E-ARK project developed the SIARD 2 format: the format used in these processes to save the database in a way that can be understood in the future.

The Database Preservation Toolkit, the software created to extract data from live databases, was made flexible to support different DBMS and formats, including their specific features and optimizations; as well as to perform on low computing hardware requirements, even when converting databases containing millions of records. This software is also capable of retrieving the contents of a preserved database and load them into a live database system.

The quality assurance systems in place during the tool's development and the validation provided by the E-ARK Pilots determined that the processes and tools can be used in archival institutions to support the long-term preservation of relational databases.

Since the Database Preservation Toolkit has a flexible plugin-based architecture, more modules can be added that do not necessarily convert from a database to another database. For instance, if a Linked Data export module were to be developed, this would enable extracting database information from any of the supported input systems and formats into Linked Data. By continuing the development of appropriate modules, the core of this tool may be used to convert any relational database from any system or format into any other system or format, and become a major asset in preserving relational databases.

Acknowledgments

The initial E-ARK project was in part supported by the European Commission within the Competitiveness and Innovation Programme 2007–2013, Grant Agreement no. 620998 under the Policy Support Programme.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Works cited

Codd, E. F. "A Relational Model of Data for Large Shared Data Banks." *Communications of the ACM*, vol. 13, no. 6, June 1970, pp. 377–87. doi:10.1145/362384.362685.

Connolly, Thomas M., and C. E. Begg. *Database Systems: A Practical Approach to Design, Implementation and Management (4th Edition)*. Pearson Addison Wesley, 2004.

Dappert, A., and A. Farquhar. "Significance Is in the Eye of the Stakeholder." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5714 LNCS:297–308*, 2009.

Ferreira, B. *Database Preservation Toolkit – A Relational Database Conversion and Normalization Tool*. 2016. U of Minho, Master's thesis.

Ferreira, M. *Introdução À Preservação Digital: Conceitos, Estratégias E Actuais Consensos*. 2006.

Freitas, R., and J. C. Ramalho. "Significant Properties in the Preservation of Relational Databases". *ECDL*, Springer, 2010. http://hdl.handle.net/1822/13702 .

Grace, S., et al. "Investigating the Significant Properties of Electronic Content over Time: Final Report." Technical report, King's College London, 2009.

Hockx-Yu, H., and G. Knight. "What to Preserve?: Significant Properties of Digital Objects." *International Journal of Digital Curation*, vol. 3, no. 1, Aug. 2008, pp. 141–53. doi:10.2218/ijdc.v3i1.49.

ISO 14721:2012. *Space Data and Information Transfer Systems – "Open Archival Information System (OAIS) – Reference Model"*. Standard, International Organization for Standardization, 2012.

ISO 9075:2008. *Information Technology – "Database Languages – SQL"*. Standard, International Organization for Standardization, 2008.

Masinter, L., et al. *Uniform resource identifier (uri): Generic syntax*. 2005.

Sperberg-mcqueen, M., et al. "Extensible Markup Language (XML) 1.0 (Third Edition)." *W3C recommendation, W3C*, Feb. 2004. http://www.w3.org/TR/2004/REC-xml20040204 .

Thibodeau, K. "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years." *The State of Digital Preservation: An International Perspective*. Council on Library and Information Resources, 2002.

The Unicode Consortium. *The Unicode Standard, Version 2.0*. Addison-Wesley Longman Publishing Co., Inc., 1996.

Wilson, A. *Significant Properties*. 2007.