http://doi.org/10.7207/twgn21-06

Preserving Databases

Data Types Series

Artefactual Systems and the Digital Preservation Coalition

DPC Technology Watch Guidance Note

July 2021

DigitalPreservationCoalition

ISSN: 2048-7916

First published in Great Britain in 2021 by the Digital Preservation Coalition.

The Data Type Guidance Note Series

Each Guidance Note in the Data Types series is designed to provide a primer on the current state of community knowledge about data types commonly encountered by those seeking to preserve digital holdings. Digital preservation is about keeping information findable, usable, and trustworthy over the long-term. The best approach for any repository will vary according to the scope and content of its holdings, available resources, and the expectations of its funders and users. There are however, broadly applicable good practices that have been established as a result of many years of research, practical implementation, and consensus building. These are presented here as a starting point, along with additional resources for further exploration.

This series of Data Type Guidance Notes has been authored by staff at Artefactual Systems in collaboration with the Digital Preservation Coalition. These notes have been developed in conjunction with the UK Nuclear Decommissioning Authority.

Digital preservation is an evolving field and continues to change and develop in response to external drivers and fresh challenges. New formats, standards, and examples of good practice will emerge over time and the information contained within this report will need to be updated. We welcome comments and feedback to: info@dpconline.org.

1 Overview

A database is a 'collection of data items and links between them, structured in a way that allows it to be accessed by a number of different applications programs.' (BCS, 2013).

A database management system (DBMS) is a software tool or set of software tools that manages data in a database (BCS, 2013). Microsoft Access and Claris FileMaker are examples of these DBMS (Wikipedia, 2021a). Database structures can range from simple tables to more complex sets of schemas, queries, views, tables, and other elements that work together to allow data to be added, deleted, changed, stored, and interpreted by users. These databases are referred to as 'relational databases' and are queried using SQL (Structured Query Language). Relational databases are the most widely available and most used databases (Freitas et al, 2009).

There are other implementations of databases that do not rely on related tables, such as non-tabular databases, flat databases, or collections of multiple complex data structures. These all have specific properties and fulfil a set of requirements known as ACID, an acronym for Atomicity, Consistency, Isolation, and Durability (Haerder et al, 2020).

Databases come in many different formats, structures, sizes, and levels of complexity. In production environments, databases can be part of larger software systems that provide functionality and access to the data.

This table lists some of the formats and database management systems that have been widely adopted.

System	**Brief summary**
Microsoft Access	Microsoft Access is a database management system that includes a front-end user interface for creating, viewing, querying, and modifying data. This system is targeted at users with minimal technical skills (Microsoft, 2021a). Microsoft Access data is stored in a proprietary format, although the software is capable of importing and exporting data in a variety of formats (Microsoft, 2021b).
Oracle Database	Oracle Database is a commonly used enterprise database management system used for transactional data, available in on-site, cloud-based, or hybrid configurations (Oracle, 2020; Wikipedia, 2021b).
MySQL and PostgreSQL	MySQL and PostgreSQL are free and open-source relational database management systems based on Structured Query Language (SQL). Standardized through ISO as ISO/IEC 9075-1:2016 (ISO, 2016), SQL is a framework for managing data stored in relational database systems, and is the language used to communicate with and store data in the database. Note that MS Access also uses SQL for querying, although its database is stored in a proprietary format.

2 Preservation Challenges

Though the preservation of a small and simple database may be a relatively straightforward task, the challenges increase when preserving large, complex and constantly changing databases. Some of the key issues are described below.

2.1 Complexity

Database preservation involves many moving parts: preserving digital information in a database-specific format; preserving the structure of the database and the logical structure of information; preserving complex or large objects in the database, and wrapping data, structure, and related documentation into archival packets for long-term management (RDB SIARD, 2019). Preservation of related documentation is important for describing database context and data provenance.

2.2 Volatility

Information held within databases can be volatile, meaning it is changed or updated frequently (Thomson, 2016).

- Making many backups of living data in a short period of time may increase disk space requirements.
- The number of columns and rows, complexity of the relationships, and quantity of data can create challenges when storing or attempting retrieval of the data.
- Volatility of the data also poses challenges for appraisal, particularly if data are being modified or deleted rather than simply added to over the life of the database.

2.3 Legal issues

Reuse and retention of personal data may be affected by varying legal frameworks in different jurisdictions. For example, reuse may be governed by the General Data Protection Regulation (GDPR), the German Bundesdatenschutzgesetz (BDSG), the UK Data Protection Act (DPA), or the US Federal Trade Commission Act (FTCA). These legal frameworks will be reflected in organizational privacy and security policies, in addition to retention schedules and other policies that govern whether and how archives can preserve databases and make them available to future researchers.

3 File formats

There is no single perfect file format for the preservation and future use of databases. Decisions made on file formats should be dependent on the features and functionality to be preserved and the future use cases to be supported. The Library of Congress (2020-2021) does not indicate recommended database file formats for preservation or access, but instead recommends a 'complete set of the content contained within the database'. Practitioners of database preservation may use simple text-based formats based on open standards, keeping the data vendor-neutral, transparent, and therefore more accessible whenever possible. (Thomson, 2016). The following table describes file formats that may be suitable for preservation and access in particular circumstances.

File format	Extensions	Brief summary

Delimiter Separated Values	.csv .tsv .txt	Delimiter Separated Values store data in rows and columns of data using characters such as commas, tabs or even pipes. Comma Separated Value (CSV) and Tabular Separated Value (TSV) are two common delimited file formats which can be used to export data from databases. For example, a relational database can be exported as one or more CSV files. This may be appropriate only for small databases; CSV does not have a file size limit but the format does not natively allow users to query the data or generate reports. The ability to open the data in commonly used software such as spreadsheet programs or data manipulation tools such as OpenRefine can be constrained by CPU processing limits. Delimiter Separated Value format preserves the data in the tables, but not other features of a database such as formulae, user interface elements, reporting features, and complex relationships. When selecting one of these formats, practitioners should consider which features of the database should be preserved or documented.
SIARD 1.0	.siard	Software Independent Archiving of Relational Databases (SIARD) is a vendor-neutral preservation format developed by the Swiss Federal Archives (PRONOM, 2010; Library of Congress, 2015). SIARD is the most established open database preservation format. Developed in the 2000s, it was later adopted and republished as a Swiss e-Government standard (CH-0165) (Swiss Federal Archives, 2013). SIARD encompasses four internationally recognised standards: XML, SQL:2008, UNICODE and ZIP64 (Swiss Federal Archives, 2020). Descriptions of sample case studies and workflows are provided by *Preserving databases using SIARD: Experiences with workflows and documentation practices* (RDB SIARD, 2019). The SIARD Suite can be used for Oracle, Microsoft SQL Server, MySQL, DB/2, and Microsoft Access databases (Swiss Federal Archives, 2020).
SIARD 2.0	.siard	SIARD 1.0 was extended in collaboration with the Swiss Federal Archives as part of the E-ARK Project (E-ARK, 2017a) to become SIARD 2.0 (E-ARK, 2017b). Version 2.0 introduced enhanced SQL support, more explicit data validation rules, support for storing large objects outside of the SIARD file itself, and changes to the compression mechanism.
SQLite	.sqlite, .sqlite3, .db	SQLite is a lightweight relational database file format contained in a single file (SQLite, 2020). SQLite's simplicity and the fact that it is natively a single file makes it a reliable and stable resting or storage format. The Library of Congress (2017) recommends SQLite as a preservation format for datasets.

4 Tips for creators

4.1 Data retention and legal issues

- Consideration of whether or not to transfer a database to an archive is dependent on a number of factors, including how and why the data were created and used by an organization, the value of the data, regulations and retention policies, and how the data will be used by others in the future.
- Familiarization with relevant regulatory environments will ensure that the collection and retention of data is permitted.
- If transfer to an archive seems likely, close coordination between database creators and users, records managers and archivists, IT personnel, and legal department personnel may be required to lay the early groundwork for successful transfer.

4.2 Documentation

- Expect the archive to require transfer of not only database contents but also related documentation that is necessary for understanding the data and the context in which it was created. This includes the user, architecture, and schema documentation, as well as any legal documentation that impacts retention policies, the protection of personally identifiable information, and any legal constraints on the use and dissemination of the data.
- Consider file-naming conventions when saving exported data. Filenames should reflect the name of the original database, the name of the worksheet or table that the data came from, and the date of the export or snapshot (Archaeology Data Service (ADS), 2009).
- If a database includes links to externally stored resources, consideration should be given to transferring them to the archives along with the database.

4.3 Security

- IT personnel may be required to provide documentation related to how the database has been secured against unauthorized access and use.
- IT personnel should be aware that archivists may need to have appropriate top-level ('root') administrative permissions to allow them to work freely with the database and extract data.
- IT personnel may be responsible for undertaking exports of the data when required by the archives.

5 Tips for archivists

5.1 General guidance

The following resources provide guidance on preserving and providing access to databases:

- Archaeology Data Service's *Guide to Good Practice for Databases and Spreadsheets* (2009).
- DPC's Technology Watch Report, *Preserving Transactional Data* (Thomson, 2016).
- The Software Sustainability Institute (2020) has a set of questions that an archivist can ask database owners about their digital materials to determine if software preservation is recommended.
- The Software Preservation Network (2020) provides guidelines and additional resources for software preservation.
- A number of software tools are available for working with Database data (COPTR, 2021).

5.2 Acquisition and appraisal

- Work to establish close communications with database creators and users as early in the lifecycle of the database as possible, in order to ensure that all parties understand the appraisal and acquisition practices that will be applied to the database. Consider creating some guidance documentation or training materials to make database creators and users aware of what they can do to mitigate some of the long-term preservation risks.
- Decide when and how to capture data. In some cases, "live" data may be captured multiple times during the life of the database, while in others inactive or "historical" data will be acquired. These decisions should be based on organizational retention and disposition schedules, IT practices, and archival acquisition policies.
- Be prepared to preserve and provide access to database schema(s), documents that describe all of the columns, headers, and value types (numbers, text character limits, formatted dates, etc.) in database tables (RDB SIARD, 2019).
- Determine what information content of a database, as well as which functionality provided by the utilised DBMS or user front end, is desirable to preserve to meet the needs of users. This will aid acquisition decisions and the selection of an appropriate preservation approach.

5.3 Preservation action

- An emulation approach (Morrissey, 2020) can be applied to preserve data and relevant system software.
- A migration approach can be applied to transform data to another database format, or a preservation format such as SIARD (SFA SIARD, 2020).
- Creation of a static snapshot (Microsoft, 2016), may provide some preservation value but will not remove the dependency on the source DBMS.

5.4 Characterization

- Identify file formats with a tool such as DROID (The National Archives, n.d.), FIDO (Open Preservation Foundation, 2020), or Siegfried (Lehane, 2020) that uses the PRONOM file format registry (The National Archives, 2020).
- If using the SIARD format for preservation, tools such as the SIARD Suite and Database Preservation Toolkit (DBPTK) (KEEP, 2020) can perform database validation automatically (RDB SIARD, 2019).

5.5 Quality assurance

It may be useful to perform quality assurance on incoming databases, checking that the data structures, tables, relationships, and value types conform to the schema and database architecture documentation.

- Perform checks on layout and formatting; tables and sheets; formulae, queries, macros; comments or notes; hidden or protected data; special characters or delimiters; and links (ADS, 2009). These file properties should also be checked when migrating databases to another format.
- Consult with records creators to understand and document any identified discrepancies between the preservation copy and the live database. If the discrepancies arose during a data export, the export may need to be reperformed.

6 References

Archaeology Data Service (2009) *Databases and Spreadsheets: A Guide to Good Practice.* Available at:https://web.archive.org/web/20201215095638/https://guides.archaeologydataservice.ac.uk/g2gp/DbSht_Toc

BCS Academy Glossary Working Party (2013) *BCS Glossary of Computing and ICT 13th edition*. Available at: https://learning.oreilly.com/library/view/bcs-glossary-of/9781780171500/11_GlossaryofICT_partA9.xhtml [accessed 24 March 2021]

Claris (2021) *Use Claris FileMaker to Build Business Applications — Claris.* Available at: https://web.archive.org/web/20210101004554/https://www.claris.com/filemaker/

COPTR (2021) *Database.* Available at: https://web.archive.org/web/20210706064251/https://coptr.digipres.org/index.php/Database

E-ARK (2017a) *Welcome to the E-ARK Project.* Available at: https://web.archive.org/web/20201218052540/http://eark-project.com/

E-ARK (2017b) *SIARD 2.0.* Available at: https://web.archive.org/web/20200925185826/https://eark-project.com/resources/specificationdocs/32-specification-for-siard-format-v20/STAN_e_FINAL_2015-07-04_eCH-0165_V2%200_SIARD-Format.pdf

Freitas R. and Ramalho, J. (2009) *Relational Databases Digital Preservation*. Available at: https://web.archive.org/web/20210205195333/https://www.researchgate.net/publication/239928984_Relational_Databases_Digital_Preservation

Haerder, T. and Reuter, A. (1983). *Principles of transaction-oriented database recovery.* Available at: https://doi.org/10.1145/289.291. DOI: 10.1145/289.291

ISO (2016) *ISO/IEC 9075-1:2016 Information technology — Database languages — SQL — Part 1: Framework (SQL/Framework).* Available at: https://web.archive.org/web/20210114164841/https://www.iso.org/standard/63555.html

KEEP Solutions (2020) *DBPTK (Database Preservation Toolkit).* Available at: https://web.archive.org/web/20210112095121/https://database-preservation.com/

Lehane, R (2020) *Siegfried.* Available at: https://web.archive.org/web/20201028192837/https://github.com/richardlehane/siegfried

Library of Congress [LC] (2020-2021) *Recommended formats statement: Datasets.* Available at: https://web.archive.org/web/20201116041423/http://www.loc.gov/preservation/resources/rfs/data.html

Library of Congress (2017). *SQLite, Version 3.* Available at: https://web.archive.org/web/20201117025418/https://www.loc.gov/preservation/digital/formats/fdd/fdd000461.shtml

Library of Congress (2015) *SIARD (Software Independent Archiving of Relational Databases) Version 1.0.* Available at: https://web.archive.org/web/20201101014106/https://www.loc.gov/preservation/digital/formats/fdd/fdd000426.shtml

Microsoft (2016) *Database Snapshots (SQL Server).* Available at: http://web.archive.org/web/20210506145800/https://docs.microsoft.com/en-us/sql/relational-databases/databases/database-snapshots-sql-server?view=sql-server-ver15

Microsoft (2021a) *Database Software and Applications | Microsoft Access.* Available at: https://web.archive.org/web/20201223120917/https://www.microsoft.com/en-us/microsoft-365/access

Microsoft (2021b) *Introduction to importing, linking, and exporting data in Access.* Available at: https://web.archive.org/web/20201109011756/https://support.microsoft.com/en-us/office/introduction-to-importing-linking-and-exporting-data-in-access-08422593-42dd-4e73-bdf1-4c21fc3aa1b0?ui=en-us&rs=en-us&ad=us

Open Preservation Foundation (2020) *Format Identification for Digital Objects (FIDO).* Available at: https://web.archive.org/web/20200916134739/https://github.com/openpreserve/fido

Oracle (2020) *Oracle Database.* Available at: https://web.archive.org/web/20210105064333/https://www.oracle.com/database/

RDB SIARD (2019) *Preserving databases using SIARD: Experiences with workflows and documentation practices*: *CEF eArchiving Building Block, E-ARK3 [CEF].* Available at: https://web.archive.org/web/20201230133036/https://dilcis.eu/images/2020review/9_Draft_SIARD_Case_Study_1.pdf

Software Preservation Network (2020) *Software Preservation Network.* Available at: https://web.archive.org/web/20210107085536/https://www.softwarepreservationnetwork.org/

Software Sustainability Institute (2020) *Digital preservation and curation - the danger of overlooking software.* Available at: https://web.archive.org/web/20191128143837/https://www.software.ac.uk/resources/guides/digital-preservation-and-curation-danger-overlooking-software

Swiss Federal Archives (2013-03-21) *eCH-0165 SIARD-Formatspezifikation*. Available at: https://web.archive.org/web/20201028110949/https://www.ech.ch/de/dokument/2760f452-6e56-48ef-bb0d-20f68638a825

Swiss Federal Archives (2020) *"SIARD Suite" -- Tools -- Archiving.* Available at: https://web.archive.org/web/20201030171732/https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html

The National Archives (n.d.) *Digital Object Record Identification (DROID).* Available at: https://web.archive.org/web/20201015033155/https://github.com/digital-preservation/droid

The National Archives (2020) *The Technical Registry: PRONOM.* Available at: https://web.archive.org/web/20201111032324/http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

The National Archives (2010) *PRONOM: SIARD (Software-Independent Archiving of Relational Databases) 1.0*. Available at: https://web.archive.org/web/20201101173235/https://www.nationalarchives.gov.uk/PRONOM/fmt/161

Thomson, Sara Day (2016) *Preserving Transactional Data.* Available at: http://dx.doi.org/10.7207/twr16-02. DOI: 10.7207/twr16-02.

Wikipedia (2021a) *Databases.* Available at: https://web.archive.org/web/20210108180820/https://en.wikipedia.org/wiki/Database

Wikipedia (2021b) *Oracle Database.* Available at: https://web.archive.org/web/20201218205237/https://en.wikipedia.org/wiki/Oracle_Database