Preserving databases using SIARD

Experiences with workflows and documentation practices

RDB SIARD

v1.0

CEF eArchiving Building Block, E-ARK3

CEF-TC-2019-3 eArchiving

Cover Sheet

Document Status:

Status
Preliminary Review to be discussed with A2 SIARD-group

Document Approver(s)

Name	Role

Document Reviewer(s)

Name	Role

Summary of Changes:

Version	Date	Created by	Short Description of Changes
0.1	03.06.2020	Ane Kamilla Hovdan, Arne-Kristian Groven, Lars Jynge Alvik, Lauri Rätsep,	Preliminary version

		Markus Merenmies, Vesa-Matti Ovaska	
0.2	18.8.2020	Ane Kamilla Hovdan, Arne-Kristian Groven, Lars Jynge Alvik, Lauri Rätsep, Markus Merenmies, Vesa-Matti Ovaska	National case studies
0.3	27.8.2020	Arne-Kristian Groven, Markus Merenmies	Introduction
0.4	28.8.2020	Arne-Kristian Groven	SIARD background
0.5	31.8.2020	Arne-Kristian Groven, Markus Merenmies	Final
1.0	30.9.2020	Arne-Kristian Groven, Markus Merenmies, Vesa-Matti Ovaska	Finalising the report, the introduction, the background, and the final sections, to make it ready for review

Experiences with workflows and documentation practices

Introduction

Preserving database information

Preserving databases comprises several elements. These are how to:

- preserve digital information which is in a database-specific format,
- preserve the structure of the database and the logical structure of information,
- preserve complex or large objects which are in the database, and
- wrap data, structure and related documentation into archival packets which can be managed in long-term preservation systems.

Each of these themes has its requirements, challenges and methods.

The method examined covers the use of the SIARD[1] (Software Independent Archiving of Relational Databases) file format and CITS.[2] The SIARD standard offers methods for archiving into a text-based format both database data, table structures, relations, triggers, procedures and table views. CITS is an acronym for Content Information Type Specification and is defining how to package-specific classes of content in the context of long-term specification. A CITS for relational databases based on SIARD is under development based on more generic package specification developed within the E-ARK portfolio of EU funded projects aimed at European (international) standardisation of long-term digital preservation.

One major preservation challenge related to SQL-based relational databases is that SQL, which has existed for 40 years, accepts the introduction of new vendor-specific constructs. Consequently, relational databases from one vendor cannot (easily) be opened by another. This problem is solved by using SIARD for long-term preservation. In addition, tools supporting SIARD provide easy-to-use interfaces for automatic transformation from vendor-specific databases into the SIARD format in a transparent and well-documented way.

For long-term preservation and access purposes, additional documentation is also needed to describe the context of the database and provenance of the database content. It is also essential to look for traces that can give valuable information about the original context of creation and use, and the rationale behind the data captured in the database using SIARD.

Workflows and practices in the case study

This case study-based report examines preservation workflows and documentation practices from selected archive organisations which have undertaken database preservation. The way the workgroup has solved this is by describing all relevant steps in the preservation process for five different National archives, namely the National Archives of Norway (NAN), the National Archives of Estonia (NAE), the National Archives of Finland (NAF), the Danish National Archives (DNA), and the

[1] https://dilcis.eu/content-types/siard or https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html
[2] https://earkcsip.dilcis.eu/

Swiss Federal Archives (SFA). Practical examples from each country are also presented, alongside a short analysis of the possible benefits and threats.

This collection of experiences and best practices from five European National Archives represents a novel contribution to the work on long-term preservation of relational databases based on SIARD. The different National Archives also represent different levels of maturity related to SIARD preservation. Two of them, the Danish National Archives and the Swiss Federal Archives have been using SIARD for 7 and 12 years respectively. The National Archives of Estonia has been using SIARD for four years, The National Archives of Norway has been using SIARD for two years, and the National Archives of Finland is in the early stages of using SIARD.

The purpose of the case study is to reveal workflows and establish a best practice for obtaining this kind of information and ensuring the semantics of the SIARD files (i.e. the meaning of the content and structure in the database represented in SIARD such as database descriptions, ER diagrams, coded value explanations, user guides from the original production system the database was a part of). Before presenting each of the national cases, SIARD preservation will be presented in some detail, focusing on what types of documentation are needed to understand and utilise a preserved database in the future.

The examined case studies are divided into following principal steps in the workflow: appraisal, pre-delivery, ingest and access.

- The appraisal step includes administrative decisions and processes to decide what information should be preserved;
- The pre-delivery step includes activities which the database owner (content provider) should handle and how to prepare data from the database to delivery to the archival institution;
- The ingest step is done by the archiving institution, and it includes both quality checks and activities which are necessary when transferring digital content to be preserved permanently;
- Access and usability of preserved database information.

A SIARD file needs additional information

What makes SIARD database preservation valuable?

Database preservation with SIARD is a tool-based approach for archiving both database data, table structures, relations, table views, keys, (traces of) procedures, and triggers into a text-based format as described in the SIARD-standard. The SIARD standard is an open standard, providing transparency. A SIARD file consists of different types of XML and XSD file pairs inside a ZIP64 container file, one XML/XSD file for each table, one XML/XSD file containing different kinds of metadata.

In addition to metadata related to database data and structure, authenticity and integrity enhancing metadata are stored inside a SIARD file. These are automatically generated by the SIARD tools when a SIARD file is generated.

SIARD is supported by two open-source software tools for generation and visualisation of SIARD files. The tools are the SIARD Suite[3], owned by the Swiss Federal Archives, and the Database Preservation ToolKit (DBPTK)[4] owned by KEEP SOLUTIONS from Portugal. Both partners in the E-ARK project consortium. Automatic generation of a SIARD file from a running relational database is a simple operation using these tools. All you need to do is to give the SIARD tool access to a database user account and then push a button in a GUI window, to generate a SIARD file of the database. Alternatively, performing the same type of commands in a text-based user interface. The access rights of the database user define which part of the database will be saved as a SIARD file. In addition to the two open-source tools, a proprietary tool called Spectral Core Full Convert has also been developed in the last couple of years to handle SIARD files.

In theory, SIARD file generation is actually a simple mapping of a database, from the original SQL representation of a source database system (e.g. Oracle, DB/2, Microsoft Access, SQL Server, MySQL) into one SIARD file in a format similar to other modern XML-based formats, like DOCX, etc. The main purpose of SIARD preservation is to save the original database, as a whole or partially so that both data and structure reflects that of the source database. Another purpose is to avoid vendor lock-in, which is a problem when it comes to database preservation. Vendor lock-in means that for example, an SQL database generated in Oracle cannot (easily) be exported to a MySQL database management system or vice versa. Because relational database vendors are allowed to invent new concepts and still call themselves SQL compliant, a lot of SQL dialects have developed over the 40+ years SQL has existed.

In the context of long-term digital preservation, vendor lock-in is counterproductive. Should every archival institution keep one DBMS active for each vendor long into the future? The answer is no. SIARD offers transformations from different vendors' SQL dialects into an international SQL standard, ANSI/ISO SQL:2008. These transformations are well-documented (serving as authenticity-enhancing information) and implemented in the SIARD tools. The resulting SIARD file is, therefore, an XML-representation of an SQL:2008 database, reflecting the original vendor-specific SQL dialect.

Visualisation and access of SIARD files are possible in three ways:

1. By using the SIARD tool support (e.g. SIARD Suite or DBPTK). In the same way that Microsoft Word makes it possible to read/write and move around in a DOCX file, the SIARD Suite makes it possible to read and move around in a SIARD file. It is also possible to write descriptive information into a SIARD file in almost every level (e.g. column-level, table-level, etc).
2. The SIARD tools can export an SQL:2008 database from the SIARD file format into an RDBMS of your choice. This enables the additional flexibility when accessing the data, where users can define their own queries.
3. It is possible to read the SIARD file in a standard internet browser. This may be necessary in cases when other options are not available, and urgent access to some data is required.

To summarise, (basic) SIARD preservation provides:

- Vendor-independent SQL:2008 preservation of the original relational database;
- Evidence supporting the authenticity and integrity of the preserved database;

[3] https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html
[4] https://database-preservation.com/

- Flexible access options for the preserved database;
- The possibility to add descriptive information (metadata) into the database (SIARD file).

Additional information needed to understand the preserved database

As is evident in the five cases presented in this report, covering SIARD practice from National Archives in five different European countries, additional information is both desirable and required from the database users/owners before the long-term preservation of the database. This is needed to understand the preserved database, the SIARD file, in the future.

Every table, every table column, and every data value set has to be properly understood. At least for the main tables. Otherwise, lots of X's of type Y's in table Z's can be preserved in the best possible way, using the SIARD method, without being able to understand what they all mean. Some column names of familiar types in some of the tables will probably be understood to some extent (e.g. ”Name”, ”Family name”, ”Mobile phone”, ”Address”, etc). But most of the database, represented as a SIARD file, will not be understood without additional (meta-) information, describing the database structure, semantics, and rationale. Therefore, to understand the preserved database, some kind of data dictionary or database catalogue is necessary.

A database catalogue can be found inside the database. If available, it consists of metadata containing definitions of database objects/constructs such as base tables, views (virtual tables), synonyms, indexes, value ranges, users, and user groups. The SQL standard specifies a uniform means to access the catalogue, called the information schema, a set of read-only views. But not all relational database products follow this standard.

A data dictionary can be something similar to a database catalogue, often with a wider scope. It can be realised in different ways:

- As a document describing the database;
- As an integral component of the DBMS;
- As a separate piece of software, either communicating with the DBMS, or not.

In the worst-case scenario, neither a (complete) database catalogue nor a data dictionary exists before the preservation of the database. Then it needs to be planned for and constructed as part of the preservation process. In addition to the data dictionary or the database catalogue, other types of database or system documentation are also presumably valuable to be able to understand the database and its context.

Additional information needed to understand the original context of creation and use of the data

Different types of documentation should be provided to capture the original context of data creation and use. This documentation comes from different domains:

- The operational system domain, including human-system interaction;
- The business domain, explaining more high-level functions and work processes;
- The legal domain, defining the legal rationale for performing certain functions.

Starting with the legal domain, laws, and regulations are constantly changing. Therefore the legal framework surrounding a system and its associated relational database during its life-span should be captured as part of the database preservation.

Business (process) documentation can also prove valuable for future understanding, including the use of ER diagrams (ER models). The name ER model is short for Entity-Relationship model. An ER model can be used as a tool when creating a database; it can also serve as a tool when preserving one. Without going into too much detail, three different types of objects exist in an ER model: Entities that exist in the real world, the relationship between entities, and attributes associated with both entities and relationships. Some National Archives require ER diagrams in addition to the SIARD file.

Finally, in the more operational domain, (high-quality) user manuals are examples of valuable additional information. If a high-quality user manual does not exist, video capture of system-user interactions with the system associated with the relational database can be valuable supplements to the SIARD file. Likewise, screenshots of the system GUI. But, to extract deep knowledge from the screenshots of the system GUI, the screenshots should be annotated. Each field in the GUI screenshot should include information about what place inside the database it corresponds to.

This deep knowledge may be documented before preservation, or, if not, it has to be documented as part of the database preservation process.

- If high-quality documentation is available, then the link between the GUI and the database is already documented there;
- Otherwise, it has to be produced during preservation, by taking screenshots of the GUI (the main system/database stakeholders should make the annotations on the screenshots).

In some databases, there are views in the database, representing original system-user interaction. But, in other databases views are more or less non-existent. This is because the system's application layer handles what could have instead be represented as database views. It is assumed that views will be valuable for future use of the material. The annotation of screenshots makes it possible to create views in the future, even for databases that do not originally contain views.

Case Study Estonia

The Estonian public sector's digital infrastructure consists of 2000+ individual databases, offering 3000+ e-services to citizens and businesses. In total, 95% of public sector information is created and managed in digital form, mostly as structured data. The principal issue for the National Archives of Estonia (NAE) is scalability and efficiency: how to appraise 2000+ databases, how to transfer data from a living information system, how to describe the data sufficiently and preserve for current and future users without spending too many resources on it.

More technically, there are some issues which make hard to apply the common "full snapshot archiving" scenario (i.e. taking SIARD snapshots of the full database every two or three years):

- the complexity of large database models significantly limits the number of researchers who can use a full database snapshot;
- the level of data duplication across consecutive snapshots further complicates effective reuse (e.g. when trying to produce time-series over data);

– the size of database snapshots (possible 10s of TBs) leads to performance issues in archiving, preservation and reuse.

To address these issues, NAE has been, in collaboration with KEEP Solutions and the Danish National Archives, looking into the possibilities of establishing a more selective database archiving approach. This includes selective archiving of tables and fields and archiving materialised database views, all supported by a GUI. By writing this case study, these usability and functionality updates now exist within the Database Preservation Toolkit, which is the primary database archiving software used at NAE.

It is worth noting that NAE is trying to do as much as possible of the work itself – we create the initial archival description, teach how to use the tools, and if necessary (and possible) also execute the SIARD creation and transfer. In short, we try to make database archiving as simple as possible for the transferring agency.

Appraisal

In 2016–2017 NAE started to implement a two-step appraisal process for national-level datasets. The first step was a macro appraisal decision to select "the databases which might include some information of archival value". The appraisal process gathered a list of all Estonian public sector information systems (based on the national information system registry RIHA[5]), and conducted an evaluation of the high-level functions and e-services provided by these systems. This high-level appraisal decision was formalised in October 2017 and identified a total of 26 potentially valuable databases. While the number might seem low, we must take into account that:

– such a huge appraisal is certainly not done without errors, and NAE has committed to reviewing the list annually. It is expected that new valuable databases will be discovered through standard agency consultation, growing the list to 50+ by around 2025;
– the appraisal did not include information systems used for electronic recordkeeping (i.e. systems which keep binary files and standard records and aggregation metadata), as records management undergoes classification scheme-based appraisal and archiving in Estonia;
– the appraisal process revealed that the majority of information systems in Estonia are used for operational purposes only (i.e. for the collection, short-term management and access to data). Often, the data is also sent to other systems, for example, nationwide central portals, which are more relevant as the target for archiving.

The second step of the appraisal process is done during pre-appraisal and is the so-called detailed data appraisal. This step is about analysing the components (i.e. data model, services and similar) of the database selected in the first step of defining the specific data components which are most valuable for future generations.

Pre-delivery

The database archiving process at NAE starts with arranging several meetings where all relevant parties (data owner, developers, database administrator, agency archivist, NAE technical staff and NAE archivist) are present. It is worth noting that the technical tasks are often outsourced, meaning

[5] https://www.riha.ee/Avaleht (in Estonian)

there are effectively three organisations present – NAE, the agency owning the data, and the public or private sector "IT house" which takes care of the management and development of the database. All these initial meetings are documented; the documentation is archived with the data.

The primary aim of these meetings is to brief the data owner and IT staff about the complete procedure: provide contacts, training on formats and archiving software, agree on a preliminary schedule, etc. The standard approach is to recommend the use of the Database Preservation Toolkit (DBPTK) software for creating the SIARD snapshot. NAE also encourages the database administrator to start experimenting with DBPTK as soon as possible, as agencies are often most concerned about the amount of time it takes to create a SIARD snapshot.[6]

The second purpose of these meetings is to understand the data and the business context of the database; ultimately to carry out the second step of appraisal (i.e. deciding what exactly to archive) and estimate resources needed for archiving. The last is especially relevant if the database administration and hosting are outsourced, meaning that there is an IT company that needs to be paid for their intervention. The aspects which we try to understand are:

- The level of documentation: how much documentation is available, what does it cover, how usable and understandable is it;
- The data layer: looking at the logical and technical data models, schemas, diagrams, tables, column types, queries, data description, etc.;
- The functionality: looking at software, GUI, applications, the services being offered;
- implementation of queries and views: are queries and views implemented in the data layer as SQL (therefore possible to be materialised and archived within the SIARD snapshot by default with DBPTK) or in the application layer as pieces of code (and thus not possible to be materialised with DBPTK);
- The amount and complexity of data: how large is the native dump (Oracle, PostgreSQL, MySQL, etc.), how many internal and external LOB's exist.

In addition, NAE maintains a list of "known issues in databases" (i.e. aspects of databases which can not be easily migrated into SIARD). Examples of such issues are geodata formats (i.e. coordinates), special data types for links to external LOBs etc. These get special attention at the meetings to solve potential problems as early as possible.

Following these initial discussions, archivists and the data owner start looking at the data and documentation to decide which components are reasonable to be archived and which not. Special attention is given to database views. The archivists evaluate which views are expected to be most useful for archival researchers, and create a list of "archival value views". One aspect here is access restrictions – archivists try to find views which do not include restricted data and can therefore be released to researchers immediately after archiving.

Also, archivists and the data owner decide on the archival date and time of data (i.e. the moment in time when the database is frozen for archiving). Here the main consideration is to find the time when the data is "as final as possible" (For example, try to ensure that there are not too many content workflows running (for example, citizen applications "being processed"), that automated

[6] The most extreme example for NAE is an agency who discussed different concerns in great detail over many meetings / hours, and once the database administrator got brave enough to try DBPTK it took him half an hour to actually create the first test SIARD snapshot and effectively learn to use and trust the archiving software.

data quality scripts are not running, if data gathering is seasonal try to select a date which is off-season, etc).

The technical staff at both NAE and the producer start in parallel setting up the IT environment for SIARD snapshot creation, experiment with different settings, evaluate the test snapshots for errors and try to solve these if present. Once the list of "archival value views" is received from the archivists, the process of archiving these views as materialised tables in a separate SIARD file is tested. It might also be possible that archivists demand for the definition of a new SQL view purely for archival purposes (for example, they want to archive a view which is only implemented on the application side and does not exist as SQL). In this case, the IT staff have to create the SQL view manually or otherwise define the required view in a way that it can be archived with DBPTK. At the date and time selected for archiving either:

a) the actual SIARD creation is done on the live database, or
b) an exact copy of the data is done for later SIARD creation.

In most cases the second option is used – for larger databases SIARD snapshot creation can take multiple hours or even days, and it is often impossible to guarantee that data in the live database is not manipulated (which in turn would possibly result in data inconsistencies in the SIARD snapshot). Another benefit of working with a copy of the data is that if errors still occur during SIARD creation, we can easily just repeat the process and not wait a few months until a next suitable date and time emerges. The main downside is the potential cost – setting up a duplicate database takes both effort and requires appropriate hardware. For example, 20 TB of disk space for the duplicate database is not always readily available, the cost going to the IT store and setting it up can be directly measured in (thousands of) Euros.

Once all problems are solved, and SIARD snapshots have been created, the agency has three representations of the database ready for delivery:

- Full native dump (e.g. Oracle dump). This representation is archived as the last resort backup (i.e. the data we can get back to if users discover problems with the SIARD snapshot);
- Full SIARD dump. This representation is archived as the main preservation representation and contains all important database elements. Views are not materialised in this representation. If there exist paths to external files then these files will be stored into this representation;
- SIARD file with materialised views. This representation includes tables with materialised views as selected within the second step of the appraisal. It is the main dissemination representation (i.e. by default, users can access and/or order the views, not the full dump).

The final steps in the pre-delivery phase are about documentation. NAE evaluates all the (written) documentation available at the agency and prepares a list of documents to be archived. Selected documentation usually includes a data model, training material and user guides, technical architecture documents, service offering descriptions and similar. We also try to find and evaluate older versions of the documentation to allow researchers to get a better idea of how the database was historically built and used.

In addition, NAE asks the content provider to create videos where the main pieces of functionality, queries and use cases are executed in the live system. This is done because the application layer is hard to archive as these pieces of code are related to specific software, versions, hardware, online

third parties, authentication methods and currently existing servers. So because we are not archiving the application layer, the videos are an important way to show to future historians how the database was really used and what it looks like. It also helps the average visitor of the archive to understand the main entities of the database without reading the documentation.

Once all of the above is done, the documentation and database representations are transferred to NAE using agreed-upon means (e.g. external hard drives, sftp, etc.). It is worth noting that currently there are no specific transfer structure or packaging requirements in place because the size, complexity and composition of a database delivery can vary significantly. Instead, the actual information package to be preserved is created after delivery and validation at NAE.

Ingest

Validation

After the delivery, all information is validated. The validation workflow has not yet been fully implemented at NAE at the moment of writing this case study. However, the following list describes all the steps intended to be implemented in 2020:

- delivery integrity check (not yet implemented)
  - preparing a script for agencies which allows them to easily create a manifest with all delivered filenames, paths and MD5 checksums;
  - preparing a script to compare the manifest with content arrived at NAE, and to highlight any inconsistencies;
- automated SIARD validation (as implemented within DBPTK)
  - DBPTK implements full validation against the SIARD 2.1 specification, as well as some additional checks (list of additional checks is available here: https://github.com/keeps/db-preservation-toolkit/wiki/Validation);
  - the errors and warnings in the validation log are evaluated by database preservation experts to see if the SIARD generation must be repeated (though this has not yet happened at NAE);
  - note that to avoid the unnecessary transfer of files we also recommend applying automated SIARD by the agency/data provider;
- manual validation by NAE staff (already implemented)
  - The log files from the SIARD creation (as created by DBPTK) are checked to find any additional errors which are not caught by the automated validation;
  - Some LOB paths within the SIARD file are checked (i.e. is the path correct, does the file/LOB exist in the correct folder);
  - Checking additional documentation – does it open correctly, is the content of the document the same as claimed within the title or file name;

Finally, it is worth noting that all validations are documented, logs and validation notes are added to the information package and archived with the database snapshots.

Archival description and official conclusion of the delivery

The creation of the archival description is usually started in parallel with the technical actions in pre-ingest. After initial kick-off meetings NAE archivists proceed to analyse available related appraisal decisions, legal acts which define the purpose and scope of the information system, and possibly other documents defining the content and nature of the system. Based on these sources, an initial archival description is created by NAE; this initial description is being discussed and extended by NAE and agency archivists both before and after delivery. However, the archival description must also include the exact numbers for tables, materialised views, documentation files being archived, which means that the description can only be finalised and approved by both the transferring agency and NAE once the validation process described above has been concluded.

In most cases the archival description of databases consists of two levels: a fonds level description for the database as such (consisting of rather lengthy descriptions of the activities which produce the data, history of the database, relations to other datasets, etc.) and three archival file-level descriptions, respectively:

a) the full SIARD snapshot;
b) materialised views;
c) documentation.

Effectively each transfer results in the addition of three archival files into the fonds, and future users can search and find “buildings registry full snapshot 2019”, “buildings registry documentation 2019”, “buildings registry full snapshot 2024”, etc. within the archival catalogue.

Once all validations have concluded successfully, and the archival description has been finalised, the responsible archivist prepares the official “acknowledgement of transfer” document. This document is the official basis for transferring the ownership of the data and reuse obligations to the National Archives of Estonia. This document is also the official conclusion of the delivery, meaning that the agency can start destroying transferred data and removing the software and hardware which has been potentially set up for the purposes of creating the SIARD snapshots.

As mentioned above, the infrastructure set up by the agency to support archiving can be quite complex and costly. Therefore the agencies are somewhat interested in signing the “acknowledgement of transfer” document as soon as possible. However, the standard practice is that first all validations must be concluded successfully (which can take many weeks for larger databases).

Storing to archive, actual preservation

NAE has not yet set up the workflow to ingest archived databases into its digital repository. The expectation is to start implementing the SIARD CITS as the basis for a database Information Package as soon as it becomes available, as such the current expectation is to retrospectively ingest all already archived databases by summer 2021. For now, the overall size of transferred databases at NAE is quite small, less than 50 TB. Therefore all database transfers are currently preserved in a simple folder structure on secure disks with regular backups.

Access

In NAE, the access part is still under development. The main issue is that most full database snapshots include personal data and cannot, therefore, be delivered to the users directly and online. Yet, the Estonian constitution demands proactive publishing of public information as widely as possible. The intended solution is to create a separate SIARD file with materialised views which do not include restricted information and to publish these online for anyone interested. The intended tool for this is DBPTK Enterprise (formerly Database Visualization Toolkit) which NAE wants to set up as a separate webpage within its Virtual Reading Room and where users can:

- see a list of database snapshots and unrestricted materialised views;
- browse or carry out a full-text search within these views.

For restricted full snapshots and views, the same technical setup is intended, but this is only going to be available in house for selected archivists who are in charge of answering user requests. We also foresee a need for creating new materialised views and anonymised representations for unrestricted access based on the full snapshot (in case user requests show patterns of interest on specific topics or databases).

Example: Estonian Buildings Registry

The principles and workflows described above were first piloted in 2019 at the Estonian Buildings Registry. The database is over 10 TB in size, consists of hundreds of tables and thousands of relations between those tables.

Let's build an archival reuse case on this database. For example, 50 years from now, a visitor wants to get information about a building located at a specific address. Looking at the full database snapshot reveals that it consists of 198 tables, which means that the full-text documentation of the data model is many hundreds of pages, and a visualisation of the data model looks like this:

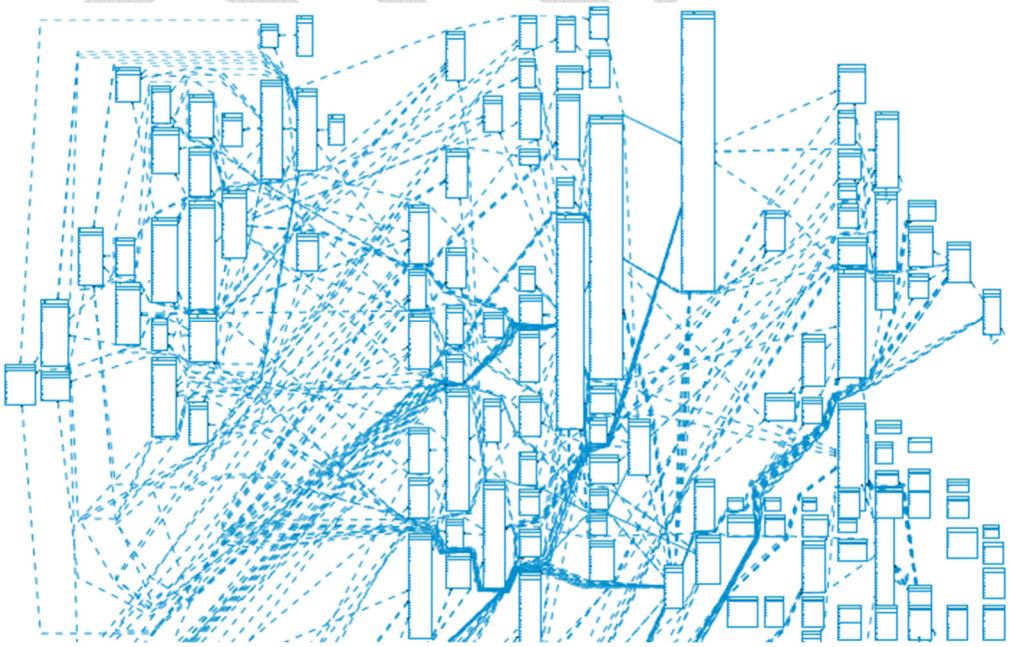


Figure 1:

After spending a few days trying to understand the data model, the user finally understands that there are three tables with the prefix AADR (called tables A1, A2 and A3), seemingly including some address related information. There is also a central table for detailed data on buildings (called table B), but the connection to AADR tables is not easy to understand because foreign keys are missing. After spending one more day, the user finally comes up with a query which (seemingly) puts all the necessary information about buildings and their addresses together, and therefore allows finding the data needed.

On the other hand, if we assume a view exists (named VIEW_B_A) with a description that it contains key data on all buildings and their addresses. The simplified SQL that generates that view is:

```
CREATE OR REPLACE FORCE EDITIONABLE VIEW VIEW_B_A ("B_ID", ... "FULL_ADDRESS",
"SHORT_ADDRESS", ..., "NIMETUS", ... ) AS
  SELECT B.ID, B_ADDRESS AS ..., B_ADDRESS_TEXT AS ...
  FROM B, A1, A2, A3
  WHERE A1.ID = A2.A1_ID AND A2.B_ID = B.ID AND A3.B_ID = B.ID;
```

Because this view exists in materialised form, the user does not have to consult the whole complex database, but only the view. Further, looking for a specific address is a simple search action and does not require specific knowledge on data models, relations and SQL. Given that the materialised view has been uploaded into the DBPTK Enterprise access GUI, the user searches for the address “Nooruse 3, Tartu”, and is presented with a record containing key data about the main office of the National Archives of Estonia.

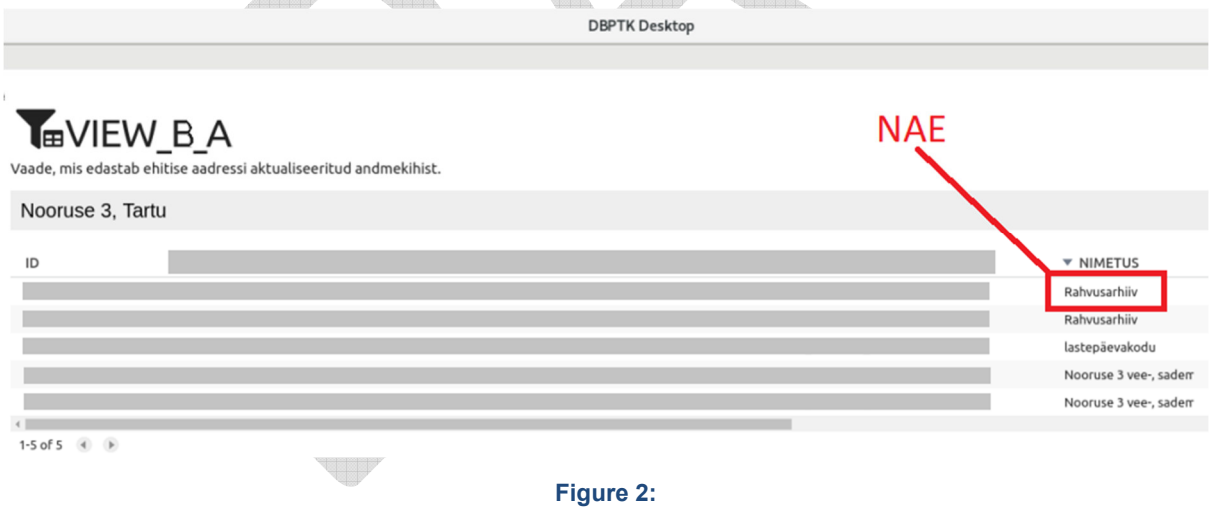


Figure 2:

This simple example shows that materialising views is useful even when aggregating only four tables. The Estonian Buildings Registry database had about 100 views implemented at the time of archiving. A list of views was extracted by the database administrator and submitted to the NAE archivists for evaluation. The views were analysed by archivists as part of the detailed data appraisal process in regard to the following questions:

- is the view technical and needed only for internal purposes within the database (e.g. ensuring integrity or quality of data), or is the view meaningful and includes data relevant to users;

- is the view used in specific user-oriented e-services;
- is the view complete, meaning that it includes a full record of an entity or action;
- is the view relevant to users (i.e. it answers a specific research question);
- how difficult would it be to reproduce the view in 5, 10 or 50 years;
- does the view include restricted data which cannot be released publicly;
- does the view overlap with other views (i.e. is there another view which covers this view entirely and includes additional data)?

During the review, archivists had access to the human-readable description of the views, SQL sentences and also examples (e.g. 10 rows of the materialised view). Finally, 13 views were decided to be relevant and were materialised and archived with DBPTK. In addition, all the materialised views were individually described.

Most of the selected views rely technically on complex joined queries, meaning that the data provided by the view cannot be easily found using SIARD browsers (like DBPTK). The only way to get the same result is to export the full SIARD to some live relational database and run the queries, which requires appropriate hardware, software and knowledge from the user of the archives. As such, we do feel that the extra effort of analysing, selecting, materialising and describing views will hugely benefit future users of the archives, because it allows a much wider audience to access the data.

Summary

The Estonian public sector is extensively digitised; most public records reside as data in more than 2000 information systems. The key method for handling such an amount of data is a two-step appraisal process which has been used to select a handful of most important databases for in-depth analysis and archiving.

The National Archives of Estonia uses SIARD as the format for database snapshots, and DBPTK as the tool to create SIARD snapshots. NAE has formed and aims to continue a strategic partnership with KEEP Solutions to continuously improve the quality of the software in terms of SIARD creation, validation and reuse.

In practice, NAE has learned that the complexity of current databases makes reuse of a full database snapshot extremely difficult for future users. Therefore NAE is analysing and selecting specific views to be materialised and archived with the full snapshot, to provide an easier entry point which does not require the setup of complex hardware and software or the consulting of hundreds of pages of documentation. Further, the selection of individual views allows to easily separate the content which is not restricted and publish it immediately online for all archival users.

In addition, NAE tries to find simple steps to allow for the preservation of the initial look-and-feel of the archived database for researchers interested in the evolvement of IT and services offered to users. The main practical method is the recording of videos or screencasts, demonstrating the main use cases of the applications and the native GUI.

The database archiving approach described in this case study has only been implemented since 2019. As such, while NAE has archived some databases and gathered valuable experiences, there are many tasks which have to be improved, automated or implemented. The main areas being tackled in 2020 and 2021 include more detailed validation, ingest of transferred databases into the NAE digital repository and setting up public access possibilities.

Strengths	**Weaknesses**
Reuse: NAE focuses on reuse, trying to make the database as easy to use as possible. Simple for agencies: NAE provides extensive support to agencies throughout the whole archiving process, therefore lowering the need for knowledge and resources, and making archiving manageable for the agency.	NAE resources: the amount of resources dedicated to one individual database is relatively large, up to 2–3 months (FTE). Current personnel are not sufficient to archive all valuable databases with a reasonable interval at this pace. Therefore more automation has to be included and additional staff hired. Missing solutions, tools or workflows: currently production-level solutions are missing for parts of validation, archival storage and access.
Opportunities	**Threats**
By default view-archiving: the materialised view approach can in future fully replace the archiving of a full snapshot, especially if the definition of “archival views” is integrated into the system design and/or data governance processes of an agency. Future-proof: the materialised view approach is compliant with the current government push towards microservices-based cloud infrastructures. As microservices are defined through domain-driven design, each microservice includes a view of the data of a specific process within a domain.	Missing crucial content in high-level appraisal: macro appraisal inherits the risk of evaluating poorly named or insufficiently described databases as not valuable, when, in fact, the content itself is crucial and valuable. NAE aims to double-check the value of agency databases within regular consultation, but it might happen that agencies destroy (some of) their data based on the macro appraisal decision. Undetected errors in validation: current validation regime is too basic, meaning that there is a possibility for not detecting errors which prohibit the use of the archived database in future. Insufficient selection criteria for views: the criteria for selecting views to be materialised and archived has not yet had time to mature, and has not been quantitatively verified in access.

Case study Norway

The production line method used by the National Archives of Norway (NAN) was originally developed by the municipal community in Norway. The goal was to develop a user-friendly process that produces good quality archive packages. In Norway, official reports from 2010 and 2017 have shown that we have a large amount of born-digital material which is at risk of being lost. These archives should be preserved for a long or short period and are defined as the backlog. A report from NAN estimates that the backlog consists of 2200 systems in the municipal archival community, which would take 500 work years to preserve with the traditional methods. Other reports claim that the actual number of systems is even higher. Bearing this in mind, the municipal community came up with this new way to preserve archives.

NAN has received born-digital systems since the mid-1980s. Most of the first archives received were character space-delimited flat files with a paper-based description. Eventually, NAN started developing their own standards for describing databases in the early 1990s, NOARK for journaling databases and ADDML for registry databases.

NAN started using SIARD as part of a pilot involving the production line concept in mid-2019. Currently, SIARD has only been tested for systems built for NOARK version 3 and 4. There is a current pilot looking into using SIARD also for NOARK version 5 systems and systems that previously have been described using ADDML (e.g. registry data, etc.).

At NAN, we are in the process of expanding our digital archive. We are conducting, or are planning to conduct pilots and POCs for sub-processes ranging from delivery to access. In all these sub-processes, we ensure useability for this method and the SIARD format.

Appraisal

In Norway, electronic records management systems are built upon the NOARK standard. For these systems, an appraisal is predetermined. For other systems, an appraisal process needs to be carried out.

Pre-delivery

It is the archive creator's responsibility to make the SIARD-file. NAN is currently advising the use of Spectral Core Full Convert (SCFC) to generate SIARD-files. This software is licenced, and the licence is currently at €639. The application is user-friendly, and so far, every archive creator has been able to produce SIARD-files without involving any third-party companies.

Once the SIARD-file has been produced, it is made into a SIP using Arkade 5. This is a packing and validation tool developed by NAN that can create SIPs according to the DIAS standard. It does not currently validate SIARD files, so in this process, it is only used for package creation. The following is

included in the SIP:

- SIARD-file
- Archive documents
- Any relevant extra documentation the archive creator possesses
- Archival description
- Package metadata (METS, PREMIS, etc.)

Ingest

The SIP is transferred to NAN either by physical medium (external hard drives) or through a file uploading service. Once the SIP is received at NAN, the SIARD-files are validated and further treated to ensure useability.

Adding semantics to metadata.xml

At NAN we use Documaster Decom for adding descriptions to metadata.xml. This tool is licenced. The tool is very efficient for popular systems with many instances because it allows for building and sharing of templates. The templates are merged with the SIARD-file and in such add descriptions to all tables, columns and fields and describe important relations.

The screenshot below shows the main window in Decom. To the left, can be seen all tables in the database. The colour of the flag indicates how important the table is. Tables with a red flag are most important. Some tables also have a notation D in front of the table name, indicating that this is important for access (DIP). To the right, we see all columns in the package table compared to the template table and how they match. Below this, there are descriptions of each column.

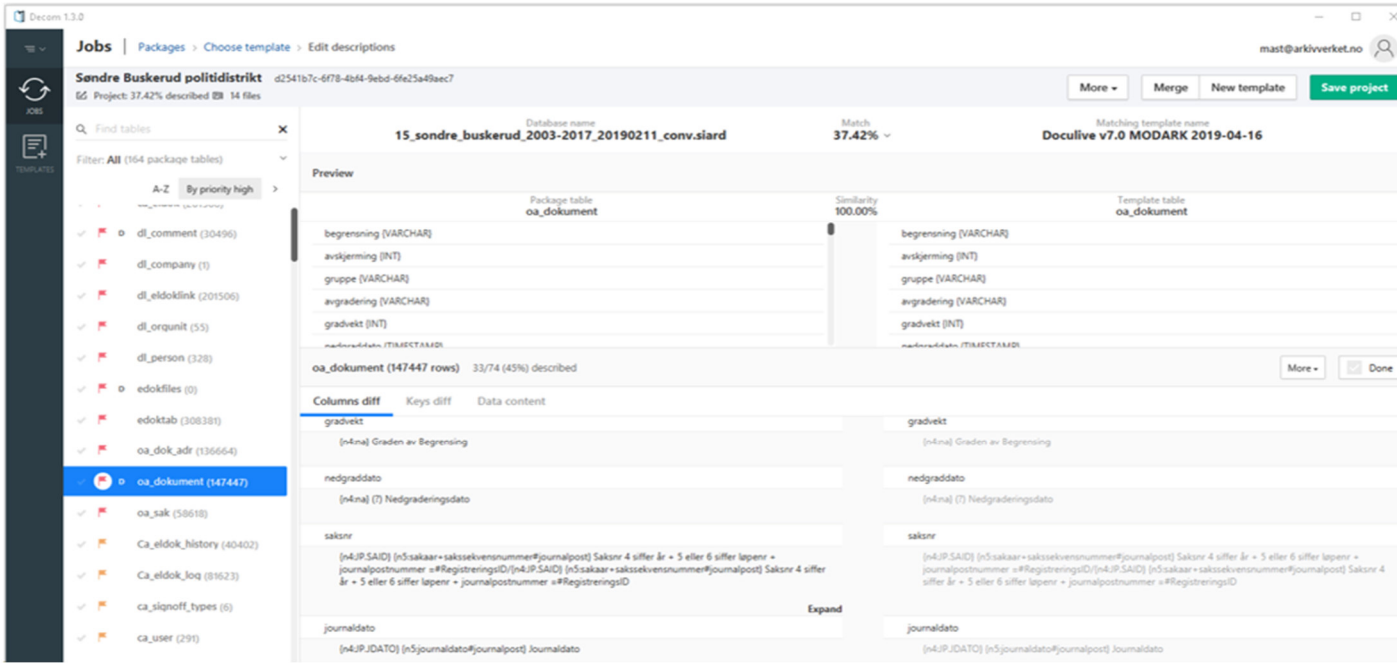


Figure 3:

In the Norwegian archival community, we currently have approximately 30 such templates for the

most common systems and more are being made. These have been developed with funding from NAN, and hence, they must be open and free to use by the entire Norwegian archival community. So even if the tool is vendor locked, the templates are accessible.

There are no written guidelines for making the templates. To ensure that we produce templates with good descriptions, and hence ensure usability for the SIARD-files, we use people who are highly familiar with the original systems to develop of these. We rely on their knowledge to produce good quality templates. We have looked into the use of entities and various tags, but this has not been standardised. As an example, we might use the entity DIP for tables that are useful for access or NOTE if there are additional written descriptions in external documentation.

Validation and document conversion

NAN uses two levels of validation, validation of the SIARD file itself and validation of the document structure of the archive.

All SIARD archives received by NAN so far have been document databases with files stored outside of the database with file references from the database. We have created a script to validate the file references and the general structure of the database. For databases with the files embedded inside the SIARD-file (either as LOBs inside the XML-files or LOBs in folders), we could use Documaster Decom for this purpose. In both cases, we do document conversion to PDF/A documents, either by a separate script using LibreOffice or Decom using LibreOffice. After validation NAN creates a new AIP SIARD-file with updated file references and semantics. The resulting SIARD-file is validated with DBPTK.

Additional information and archival description

In addition to the record descriptions added to the description field in metadata.xml, we gather additional relevant information about the system and the information within it.
As part of the SIP, the archive creator needs to fill out an archival description. As of today, this is an Excel spreadsheet with three different worksheets where they fill in the information. It would be more beneficial if this was more machine-readable.

The worksheets contain information about:

- The content creator and their history.
- The information contents. What is the system used for, how is the information related to laws, appraisal, exemptions from public, etc.?
- System information. Version number, conversions, time span, number of records, additional documents, etc.

If the content creators have other relevant documentation, we ask for this as well. This can be system documentation, guidelines for how the system has been used, screenshots of the user interface, the data model, etc. It varies a lot how much and what kind of information they can provide. All this additional information is stored within the AIP for long-term storage. We use the SIARD with the additional template descriptions as the format for long-term storage. The final AIP will include:

- SIARD with updated file paths and descriptions;
- Archive documents converted to a suitable format;
- Test report that also summarises changes compared to the SIP;
- Any relevant extra documentation the archive creator process;
- Archival description.

Code lists

We have not received a SIARD with coded lists, and we do not have a solution for this yet. For previous systems we have received containing code lists, the code values have been either incorporated into the ADDML-file or stored as a separate document. We assume that we would store such values in a separate document until we have a better solution.

Tagging of sensitive material

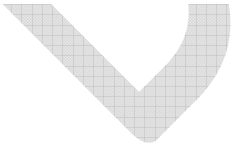
In NOARK systems, information that is sensitive or otherwise exempted from the public can be tagged. This means that such information can be tagged as part of the templates. For non-NOARK systems, we do not have a good way of tagging this. If this is a part of the database, it can be tagged in the template in a similar manner as for NOARK systems. Otherwise, we can tell from the archival description if the dataset contains sensitive material, but usually not for an individual table or record.

Access

We do not have a method to access and view the SIARD files. However, we are considering doing a pilot or a proof of concept with various database viewers as a part of the new Digital Archive we are currently developing. We are expecting to look into this more towards the end of this year. We have not yet decided on a structure for a DIP. The municipal archival community is currently working on a SIARD visualisation, based on DBVTK. They are looking into reusing the descriptions from Decom templates as descriptions for the visualisation.

Example

N/A



Summary

NAN has been using SIARD as part of a pilot for testing the production line method for about a year. The main motivation for using SIARD, and this method has been its user-friendly approach and that it is so beneficial for the public sector. The institutions need to create the SIARD-file from their databases and deliver the SIARD-file and the corresponding archive documents, any relevant extra documentation the archive creator possesses, archival description and package metadata (METS, PREMIS, etc.) to the archives. At NAN, the files are validated and treated to ensure useability. A critical part of this is using templates to add descriptions in metadata.xml, using Documaster Decom. The templates are merged with the SIARD-file and in such add descriptions to all tables, columns and fields and describe important relations. Also, NAN uses two levels of validation, validation of the

SIARD file itself and validation of the document structure of the archive. In addition to the record descriptions added to the description field in metadata.xml, we gather additional relevant information about the system and the information within it. As part of the SIP, the archive creator needs to fill out an archival description, which contains information about:

- The content creator and their history.
- The information content. What is the system used for, how is the information related to laws, appraisal, exemptions from public, etc?
- System information such as version number, conversions, time span, number of records, additional documents, etc.

If the content creators have other relevant documentation, we ask for this as well. This can be system documentation, guidelines for how the system has been used, screenshots of the user interface, the data model, etc.

Strengths	**Weaknesses**
User friendly: the method has proven to be both very user friendly and suitable for long-term preservation and access. Templates: reusability of templates across different archives increases efficiency.	Templates: because there are no guidelines for making the templates, they vary in form and level of description. Also, the descriptions are in plain text and not in a standardised machine-readable format. Supporting documentation: we have no guidelines for what kind of supporting documentation we ask for. This would also be very difficult to implement as there are huge variations in what kind of documentation the various content creators actually have.
Opportunities	**Threats**
Using SIARD for newer production databases (still in use): since we receive the entire system, this would sometimes include the current production database. This is documentation that is bound to change in the future, and we will have to receive it again when we receive the next period of the system. This means that we will get several versions of the same material. Code values as separate documents: code values are not yet machine-readable, connected to the corresponding fields in the SIARD-file.	Decom: using a commercial tool outside NANs control.

Case Study Denmark

Introduction

The vast majority of data in the Danish public sector are organised as databases with or without files in various formats. For this reason, the focus of digital preservation at the Danish National Archives (the DNA) has for decades been on archiving these data in a standardised, system-independent and cost-efficient manner.

The DNA began collecting digital-born archives as early as the mid-1970s. However, the large and systematic ingest of relational databases had its roots in the late 1990s. At this time, the first concepts of a national system for archiving relational databases were envisioned and implemented by the DNA. We have since sought to cooperate internationally on the challenges of digital archiving, and we are currently spearheading the development and implementation of common standards for preservation formats in the European Union's eArchiving Building Block.

Currently, anything digitally archived from data producers in the Danish public sector is archived as standardised information packages, and for databases, we use a Danish variant of SIARD, the format for System Independent Archiving of Relational Databases. SIARD-DK has proven to be a very resilient and encompassing format for the needs of large-scale archiving of public databases in Denmark.

Regulatory Preconditions

The DNA continue to support large-scale database archiving, and a very important precondition is our ability to create the regulatory frameworks for the archival creators and other partners to abide by. The Danish Archival Law defines the general rules and requirements for how public archiving operates in Denmark and mandates the DNA or other public archives to collect, preserve and disseminate data of historical value from the public administration. Subsequently, multiple Executive Orders issued by the DNA have throughout the years been instrumental in fulfilling the mandate of the Danish Archival Law. Correspondingly the Danish Archival Law was changed in 1992 and more significantly in 2000 and 2007 to facilitate the appropriate preservation of electronic records.

As with many national archives, the DNA has adopted – however, yet only informally – the OAIS model for how to structure and describe the digital archiving systems and processes. Among other OAIS differentiates digital archival items by their enrichment in the archiving process making sure that a collective and organisational understanding of the life cycle adaptations to data (specifically changes in content and metadata of information packages) applies for the three major obligations of collecting, preserving and disseminating data. Therefore, this contribution is structured by and tries to highlight whenever the *Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Package (DIP)* comes into play, but first, we want to focus on the presubmission steps of approval and appraisal.

Appraisal

Approval of National IT Systems

The traditional archival virtues of assessing the value of administrative records for posterity applies to digitally-born archives in completely normal ways, and we seek to make sure provenance is preserved and documented. However, an appraisal is complicated severely by the disparities and intricate database models of modern IT systems that are in use by Danish government authorities. The DNA has taken steps to counter and reduce the complexities involved with an appraisal which involves two steps that require a dialogue with and documentation from the governmental archival creators.

The first step usually begins even before any archival creator has switched the power on to the IT system in question. The step is mandated through an Executive Order from 2013 citing state and court authorities (municipal and regional authorities are exempt) to notify the DNA if a new or significantly overhauled IT system is to be taken into use, and notification is to be provided three months before this. The archival creator uses a contact form on the website of DNA for notification, and this includes describing and attaching documentation on what kind of data, data structures and data models are to be generated and/or hosted within the IT system. The documentation, if available, includes table and column descriptions. From this *documented notification,* a preliminary assessment of historical value is performed by an archivist at the DNA, and the archival creator is notified whether data in the IT system are to be preserved for posterity. For IT systems that contain information worth preserving further documentation, specifically ER diagrams and a description on how a SIP utilising a relational database model (in the form of SIARD-DK) can be produced from the data, is required by the archival creator to be sent to the DNA. Then approval of the IT system can be issued along with the requirement to create a submission of the data, usually every five years.

The Binding Appraisal of Data

Fast forward approximately five years later, the second step begins with an archivist contacting the records creator to make formal arrangements on the submission of data. The DNA determines which database tables are to be submitted, the period of data creation and extensive context documentation such as technical documents, user guides and system purpose notes are to be provided by the records creator. For cost-efficiency DNA appraises data at the database table level rather than column or row levels which inadvertently means the DNA also ingests data that are not necessarily of historical value.

Approval and appraisal of IT systems are anchored by archivists at the DNA in a dedicated organisational unit. They handle all initial dialogue with the data producers and make submission agreements. However, when the SIP is delivered to the DNA, another dedicated unit continues the process of test, validation and creating the AIP for long-term preservation.

Pre-delivery

Creation of the SIP

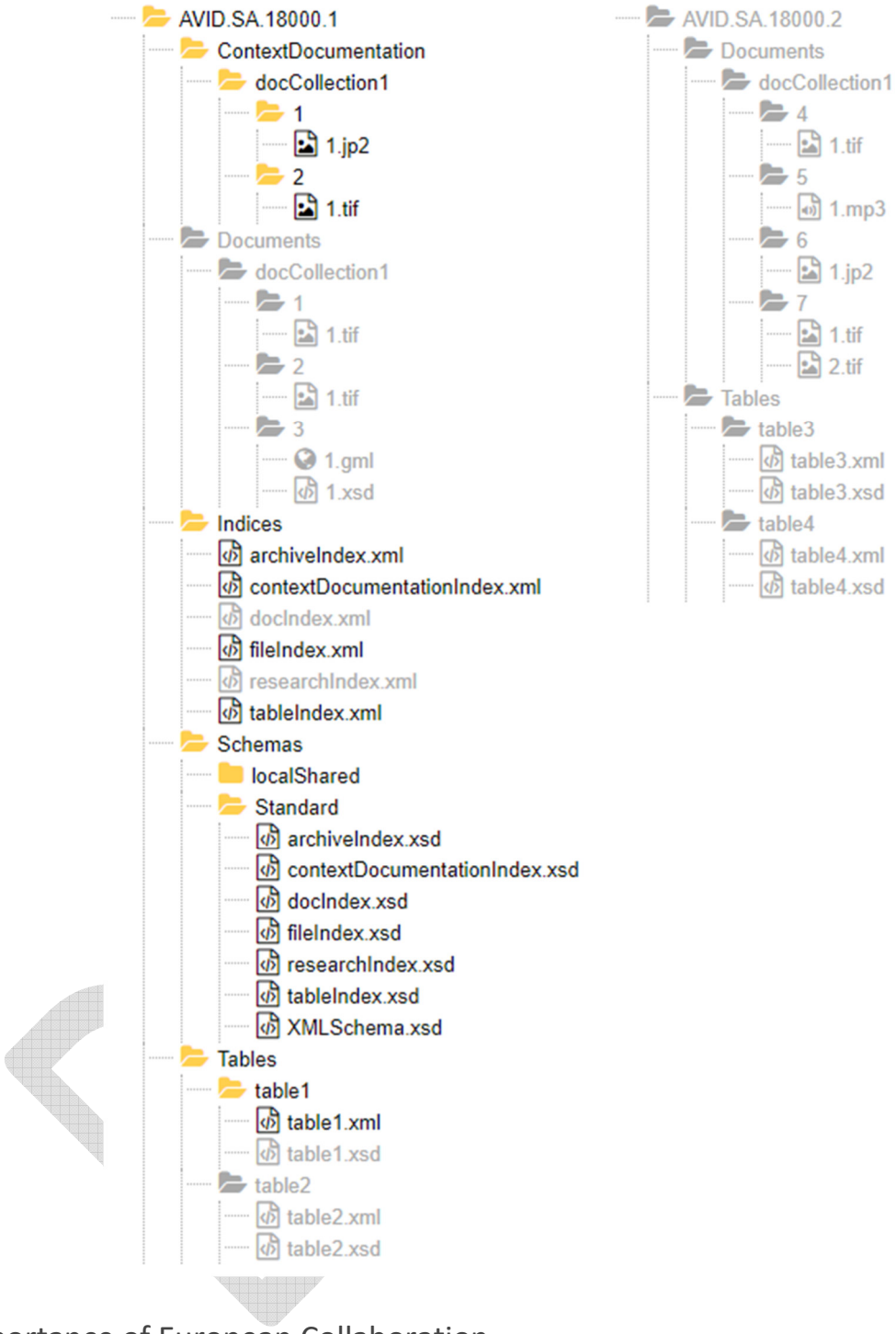
Either the archival creator or most commonly a supplier of their choosing creates the SIP according to an Executive Order issued on how to create SIPs which specifies in detail how to structure the information package, what metadata to provide and in which formats to provide content. It is within this Executive Order that the SIARD-DK format is specified. The archival creators or their suppliers are encouraged to use the Database Preservation Toolkit for creating the SIARD-DK database export, but many suppliers also use proprietary tools.

The DNA decided in 2009 to switch to SIARD because of its expected wider support compared to an internal format for relational database archiving (2000–2010). The previous year the EU Planets Project chose SIARD as the format for system-independent archiving of relational databases. SIARD-DK is a variant of SIARD 1.0 made in 2010 due to scalability for the specific purpose of supporting large objects (LOBs) stored outside the SIARD file. This feature is now supported in SIARD 2.0, and the DNA expects to switch to the latest version of SIARD in 2023 (the current version is 2.1.1).

The Danish SIP is composed of the following top-level folders:

1. **ContextDocumentation**: This folder contains all "meta" documents to the administrative use of the IT system. This is, for example, technical documents, manuals, description of purpose and use, and documents created from the process of creating the IP (such as archival notes, production notes, error lists, appraisal decision and contracts).
2. **Documents**: This folder contains all documents (i.e. text, video, audio) of archival value from the original IT system. The documents are stored in select preservation formats. The folder should only exist if the original IT system had documents.
3. **Indices**: This folder contains all the index files referring to files within the IP, both content and metadata. Indices make it possible to recreate the IP as a database that can be searched across folders and content and contains fixity numbers.
4. **Schemas**: This folder contains validation schemas for the IP. They make it possible for the recipient archive to test and validate through automated processes whether the IP has been created according to official regulations.
5. **Tables**: This folder contains all database tables of archival value exported from the original IT system.

Illustration of the folder structure as regulated through the Danish SIP. Greyed folders and files are optional:



Importance of European Collaboration

The SIARD-DK format solves the above goal of delivering cost-efficient, standardised and system-independent database archiving as outlined in the digital migration strategy of the DNA but it does not solve the ambition of creating a common European market for transactions of archiving services. It is an objective for the DNA to create a competitive market for the production of SIPs, thus providing data producers with a pool of suppliers to procure archiving services from. The objective is a driver for the involvement of The DNA in the creation of a sustainable eArchiving Building Block within the European Union.

When we look at the figures for ingest since 2017 and how they distribute by producer, we see that out of a total of 920 SIPs, they distribute between producers:

- 470 SIPs, amounting to 51%, were produced by a supplier.
- 379 SIPs, amounting to 41%, were produced by the record creators directly of which close to half (18%) were created by one archival creator, Statistics Denmark.
- 71 SIPs, amounting to 8%, were produced by the DNA.

Another important driver in the involvement of the DNA is the ambition to create a common market for tools and software across Europe. Here, validation of the SIP is of special importance, and with the prospects of having a *Common Specification for Information Packages* and a sub *Content Information Type for Relational Databases* (in other words SIARD), the possibilities of having shared validators as well as shared creation and dissemination software for information packages are within reach. For the last many years DNA has maintained its own free validation software, ADA, which has gone through several iterations most recently with a new workflow for the SIP testing staff to guide them through the test and validation process improving efficiency and decreasing error rates. We hope to be able to use a common validator for SIARD in 2023 created in open source and financed by all major national archives which use SIARD.

Ingest

From SIP to AIP

If the SIP does not validate it will most commonly be returned to its producer along with a testing note detailing specifics on error correction and then resubmitted to DNA, however, if errors are minor they can in some cases be rectified by the DNA test staff if accepted by the records creator. The following table shows the number of submissions of information packages for validation since 2017 and the number of validation attempts before final approval of the SIP was given by testing staff. The numbers show that 80% of all SIPs are approved in either the first or second validation attempt.

Attempts before approval	**AIPs**
1	355
2	302
3	71
4	18
5	4
6	3
7	0
8	1

9	1
10	0
Total	**755**

When the SIP has been validated, it can be officially accepted and ingested into the digital collections of The DNA and become an AIP. The task to migrate the SIP to AIP is quite straightforward as most of the regulatory work on how to create the information package is complied with during the creation of the SIP. The SIP is for this reason close to identical with the final AIP that is preserved.

We store AIPs in an in-house developed storage system which creates two copies for cold storage in optical and magnetic storage mediums and one remote copy for storage at another location and another organisation in Denmark. We continuously supervise the longevity of ingested formats and make sure to migrate data if they are in danger of becoming technologically obsolescent or new information package standards are adopted.

The yearly 2019-figures of submissions (submissions are equal to databases because in fact all of our information packages are relational databases with or without LOBs) are as follows:

- 366 submissions totalling 172 TB.
- 417 submissions were tested totalling 171 TB.
    - 223 validated and approved submissions totalling 62 TB.
    - 194 invalidated and rejected submissions totalling 109 TB.

We received a lot of what we consider to be small databases and some we consider to be very big. Of the 366 databases we received in 2019, 259 of those were less than 100 GB, 60 were between 100-999 GB, and 47 were above 1 TB. The largest databases were 9.1 TB and 10.6 TB. Soon, we are expecting to see databases with documents to be 20+ TB of which documents absorb most of the data. We have previously ingested databases with tables ranging from 3,000-5,000 tables, but the normal range is between 20–250 tables. The largest amount of data absorbed by tables in one database has so far been 1 TB, of which one table absorbed 700 GB of data.

The most recent figures since 2017 are divided between state, municipal/regional and private data producers in the table below:

	State government	**Municipal/Regional**	**Private**	**Total**
SIPs	577	279	76	932

As of this date, the DNA has more than 5,500 AIPs shelved for long-term preservation.

Access

The access process begins with The DNA receiving an access request from a user whether that being a citizen, researcher, the original records creator or a member of the DNA staff. If DNA agrees with

the request, it is in some cases necessary to ask the Danish Data Protection Agency for permission to grant access. The AIP is then fetched from storage, if it has not previously been requested, and migrated to a DIP by converting the SIARD-DK format to the proprietary format in the RDBMS, Microsoft SQL Server, used by our dissemination software.

Also, with access to archived data, the DNA utilises in-house developed software, named SOFIA. The software recreates the database tables and provides a document viewer. It can be accessed remotely, but typically data are extracted from a recreated copy of the AIP (which *de facto* makes it our DIP), but only the data relevant to the archival user's questions are disseminated in a CSV export file along with documents (effectively making it a minor DIP). The hope is that also with dissemination we can switch to a European developed software, and we are considering Database Visualization Toolkit. Access has previously been a minor focus area for the DNA, but this field is currently developing rapidly in line with our director's newly adopted strategy of getting "data into play".

Closing Observations

Database archiving is a complex and multifaceted challenge for a number of reasons outlined in these paragraphs, and it has taken the DNA decades of continuous development and accumulation of competencies to get to where we are today. However, it should also be clear that database archiving is not a constant which you invent once, but rather it has taken DNA many iterations, and it will still require a plethora of iterations over time to effectively combat the challenges of disparate, dependent and potentially technologically obsolescent IT systems in use by the Danish public sector.

We would dare to say that it has become part of our DNA to archive databases considering the decades-long span of our experiences and continuous competencies within the field. Our current, and most likely persistent, efforts involve reaching out and collaborating with a European field of public archives sharing the ambition of cultivating a sustainable and empowered digital archiving community.

Case study Switzerland

Background

SIARD originates in Switzerland and the Swiss Federal Archives, where the SIARD standard is Swiss E-Government Standard. SIARD preservation has been practised for more than ten years now. The first available SIARD tool, SIARD Suite[7] developed by Enter AG, is owned by the Swiss Federal Archives. The SIARD Suite tool is now available as free and open-source software.

The Swiss Federal Archives are serving Swiss Government bodies, while regions, cantons, and municipalities are serving public bodies at their level. They are also, to a large extent using SIARD for database preservation.

[7] Source: https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html

Appraisal and pre-delivery

The number of SIARD files preserved at the Swiss Federal Archives from Government bodies per year is not very large, about 10–30. The Government bodies are themselves responsible for producing the SIARD files and collecting additional information. But in many cases, the work is outsourced to private IT-companies. The Swiss Federal Archives can provide advice in the process, but they are not actively supporting the SIARD-production.

It is required that additional documentation is provided together with the SIARD-files. This documentation has to be as complete and understandable as possible, making it possible to understand the SIARD-file without prior knowledge about the application used on top of the database. What this means exactly may differ from case to case, and is often the result of an iterative process.

A wide range of documentation is required to capture the original context of the database and its associated system. Contextual documents include user manuals, technical system documentation, ER-diagrams of the system, plus journal instructions, and relevant legal context.

Another requirement is that all coded values in the database (represented by the SIARD file) have to be explained. It is now required that all tables in the SIARD file are described/explained.

But, contrary to what, e.g. the Danish National Archives requires, the SQL queries are not among the requirements. (If views exist (named queries) in the database, they will be reflected in the SIARD file.) As a consequence of this, various guidelines associated with queries (e.g. a guideline explaining how views and SQL queries, are not among the requirements). The tool used for producing SIARD files in Switzerland is the SIARD Suite.

Ingest

At the time of Ingest, it is expected that the SIARD file is supplied with an explanation of all the tables, fields, and coded values. Also, it is expected that the rationale behind excluding some of the original tables is documented. In the case when tables in the original database are left out.

Descriptions of the original application interacting with the relational database to be preserved in SIARD is also expected to follow the SIARD file, together with user manuals.

Screenshots of system-user interactions are not obligatory but are also welcome.

Validation consists of manual and visual control of the additional documentation, and of the relationship between the additional documentation and the SIARD file.

For validation of the SIARD-file, a tool called Kost-Val[8] is used:

[8] Source: https://kost-ceco.ch/cms/kost-val.html?highlight=kost-val

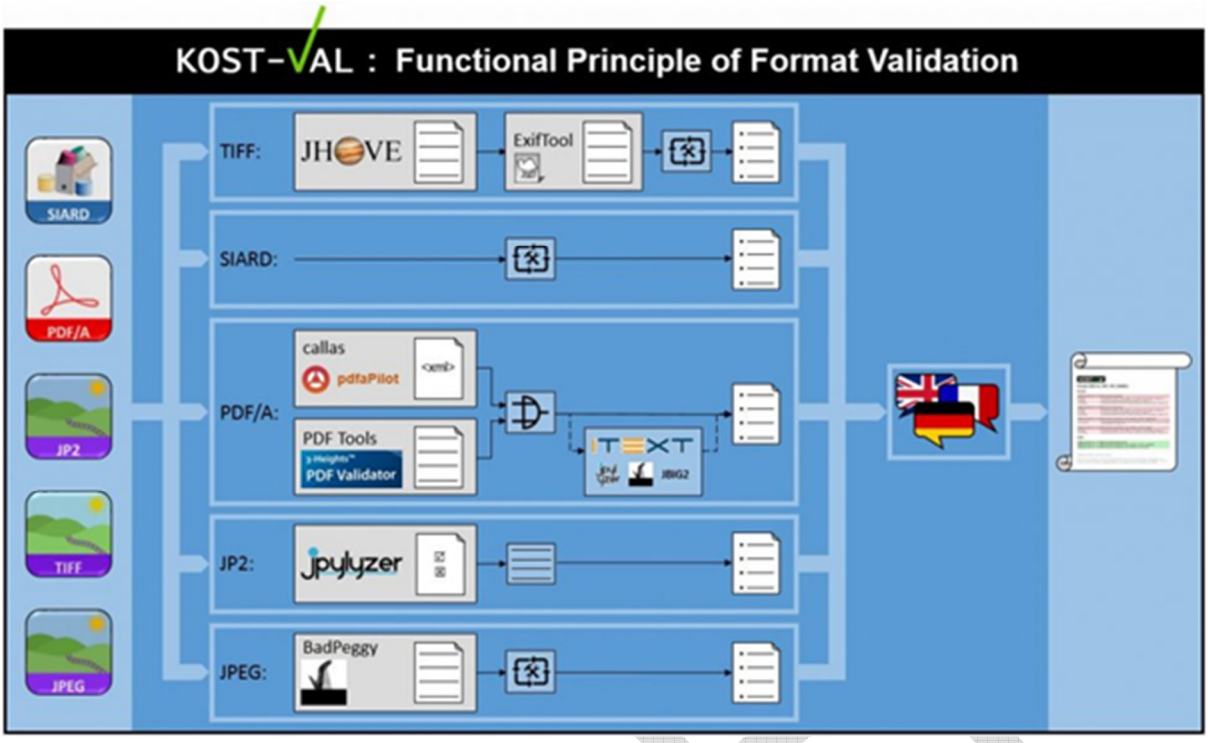


Figure 4:

Case Study Finland

The National Archives of Finland (NAF) started developing its digital archiving services with requirements for records management systems. The principle is to provide guidelines and requirements for content providers to produce born-digital records using accepted file formats and a comprehensive set of administrative and structural metadata. The strategic goal of the National Archives of Finland is to support the usability of digital records for research purposes and not to archive original usability of information as it was in records management systems. Aligned with this long-term preservation strategy, NAF developed its practices for archiving structured information from databases.

Appraisal

Appraisal of information in the databases of public administration is carried out according to the guidelines published in the appraisal policy of the National Archives of Finland. The current appraisal policy dates from 2012, but it will be updated by the end of 2020. NAF gives annually approximately 60 appraisal decisions to public agencies. Appraising registries and databases of public administration has become a central issue in recent years.

The appraisal process in Finland is based on the appraisal proposal prepared by an agency. This is to highlight that appraisal is by nature, a co-operative process between NAF and an agency.

One of the primary issues in the appraisal of databases is to analyse how a database is related to business functions, activities and transactions of an agency. The key point is understanding the information value of databases and the role of databases in the functions and business process of an agency. Besides these aspects, there is analysed the role of different stakeholders participating in data production. Is the information in a database produced in a unique business process of an agency or is data aggregated from several information sources?

In current appraisal practice databases and registries are approached mainly as logical data sets in the context of data production. This is especially the case with complex database information. When appraising relatively small and closed databases, it can be possible to approach detailed database elements also as a part of the appraisal process. However, in many cases, it is not possible to extend the appraisal analysis to database level (i.e. to analyse tables or specific database elements). The in-depth analysis, which connects database content, database structures and the appraisal decision have to be conducted as a part of pre-ingest negotiations (transfer planning process).

In updating our appraisal policy, we have identified that the current trends of data protection regulation have a substantial impact on the appraisal process. The archival value of personal information and especially of any sensitive personal information needs to be defined more accurately than before. That also affects the way how appraisal decisions are applied when selecting the methods used in archiving database information.

In current appraisal, the goal is to avoid archiving duplicate information in databases. This goal, combined with the more significant role of data protection issues emphasises the need to develop and adopt selective approaches to database archiving. There will be more focus on questions, how restrictions regarding the use of personal information are taken into account in the context of database archiving?

As a part of the in-depth analysis of the database, the following questions can be analysed. A part of the questions can be emphasised and or given less consideration depending on which kind of case we are dealing with:

- What database elements are linked to the key functions of an agency?
  - Classifications
  - Code lists and values
  - External linked databases
    - Code services: who has maintenance responsibilities for coded lists
    - Semantic interoperability
    - Open data principles
  - How are different separate databases connected?

- Is data in databases collected and processed by several stakeholders?
  What kind of law-based (or other) deliveries of information to other registries/databases/electronic services (e.g. for statistical purposes) are possible to identify?

Usually, the in-depth analysis of database information is carried out after the first appraisal of databases. In many cases, the time span from the appraisal process to the beginning of pre-ingest negotiations may be from several years or even ten years.

During the transfer planning negotiations with agencies, the following aspects are central:

→ interpreting how to apply the appraisal decision in the context of database elements and metadata

Currently, the methods and guidelines for database archiving (XML/CSV extracts, or SIARD and what kind of additional documentation is needed) are selected as a part of transfer planning (pre ingest negotiations).

Pre-delivery

The current archival legislation in Finland does not determine exactly when public agencies should contact the National Archives about the transfer of databases. Pre-ingest negotiations can be launched by an application retirement in an agency. These kinds of cases are often related to migration projects, where the agencies need to analyse what information is migrated to new information systems, which can be deleted and what needs to be archived. In migration projects, the agencies need to analyse different data layers and identify the relevant information to be retained and archived. That is why migration projects serve as a good starting point for pre-ingest negotiations.

In appraisal and pre-ingest negotiations with agencies we are facing with the reality of complex and several interconnected databases in public administration. In this context, our toolbox and methods of archiving database information need to be updated.

Practical methods for archiving structured information and databases in the National Archives of Finland (NAF) are based on normalisation of data to be used without the original information system. Since 2012 NAF has received different kinds of database extracts in XML- or CSV-format from state agencies. The information packages include (e.g. table files in XML or in CSV-format and additional context and content documentation). In a part of the deliveries, the database structures of the source system are described with ADDML. In many cases, archival packages also include free text descriptions, field lists, code lists or a variety of context documentation.

A special case in pre-ingest negotiations: code lists

In many cases, we have identified that the content (tables) of the database are very much dependent on different code lists used in the original application environment. That is why we need to analyse what kind of code lists are used as a part of database information. What versions of used code lists are relevant to ensure the usability of database information? It is important to take into account different versions of code lists, which are linked to archived information. This analysis needs to to be carried out in co-operative dialogue with an agency and its substance specialists.

Are agencies using their own organisation specific code lists or code lists, which are available in common code list services/reference services? How are the used code lists maintained?

How to get the content of code lists into archival packages and support re-use of database information? Do we offer the code list as separate documents in connection with database information, or are there some other options?

Semantic interoperability

Promoting semantic interoperability has been a pivotal issue in public administration in recent years. This has resulted in building common reference services in Finland. We have noticed that the discussion about semantic interoperability has raised awareness about the importance of documenting the code lists used in public administration. In a larger sense, this awareness helps to support the goals of long-term preservation.

Digital and Population Data Services Agency of Finland maintains an open reference service for some of the most used code lists. https://koodistot.suomi.fi/

Code lists maintained by Statistics Finland are widely used in public databases, from which there are information deliveries to law-based statistical purposes. Besides Statistics Finland there are also other statistical agencies by law. The use of code lists of Statistics Finland has been central in promoting semantic interoperability. For example, the code list of occupations (2010) of Statistics Finland is one of the most used common code lists in several databases. Code lists are available as data sets to download, or they can be used through APIs from reference services.

In the future, how code lists are archived and can be accessed as part of the use of SIARD or archiving database extracts with other methods, needs to be analysed. Currently, there is no solution to this issue.

Example – capturing “not active” research database

This example focuses on the practical findings of how to preserve and provide access to research databases which are transferred to archive only as data without user interface. The database contains historical information about persons who lived in the old Karelia area in Eastern Finland. Information was stored and updated in research projects during several phases. The project stopped in 2019, and it was decided to transfer older versions of the database which are not updated to NAF. The structure of the database is simple; collecting information from each area or location to separate tables. The table structure is a result of how information was collected from different sources. The result of this database design is that it contains 1,700 tables; most of them are using the same table structure. The logical structure between tables and related code tables is handled only in the application layer. There was no database-level structural information in the database.

The starting point for the preservation actions was the SQL file, which was produced from the MySQL database when it was still active. We could not get access to the original database. We

established a database instance to development infra and created a database using SQL script, which was about 11GB. The process succeeded without errors. SIARD was created using DBPTK. The process took quite a long time because of limitations in the development server. The total number of rows in all tables was about 45,000,000, and it took about 12 hours to finish (generate and validate) the process. The size of the SIARD file is 65GB. We had no external documentation available related to database structure and definitions of tables and columns. When examining content and column names, we could use the DBPTK tool to add some descriptions to support further use. The biggest problem was the huge number of similar tables (as described earlier). If we could have some logical level functions to connect data-tables to code tables and create logical groups of selected tables that would have help us and research purposes. NAF is going to open this database for researchers who can download the content from selected tables in CSV format.

As a result, we noticed that DBPTK and SIARD could be used for this kind of database rescue/capture projects. We noticed during the project that application logic which is not documented in database definitions should be available as external documentation and it would help further use of database if that structure would be possible at some level to created to SIARD-file and connect/group tables in a logical way.

Access

Currently, NAF does not provide access to born-digital content from our digital reading room system. We will start piloting this kind of research service in late 2020 with research databases. By request and with permission it is possible to have off-line copies of preserved database files in CSV- or XML-format.

Summary

NAF has received data to be preserved in a structured format, but so far NAF has not carried out any SIARD deliveries from agencies. Our appraisal principles point out that preserved information should be able to be used for modern digital research purposes, which promotes the use of preservation of information as structured digital format as possible. So far the preservation of databases has been based on content-specific XML- or CSV structures and external documentation. In some cases with CSV extracts, we have used ADDML as technical structure. Inspired by participation in E-ARK3, we are currently analysing the SIARD-based method as a next step when developing our methods with archiving of databases. First, we are internally piloting the use of SIARD with some SQL-based research databases and some legacy databases owned by (in the custody of) the NAF. After internal pilots, the next step would be piloting SIARD with some agency. We think that co-operation and shared good practices from countries with substantial experience with SIARD are valuable for our development work.

Strengths	**Weaknesses**
Well-established practices for appraisal principles of digital records.	No resources for own development.

NAF has a relatively strong mandate to dictate technical practices for preserving structured data. SIARD offers an efficient solution for capturing a database and its structure.	No ready guidelines for context and additional/supporting documentation in NAF (ER-diagram, relevant systems documentation etc.). currently no access to archived databases.
Opportunities	**Threats**
Implementing SIARD offers possibilities to be involved in international co-operation. SIARD user groups lead to more co-operation (common tools, toolbox development) in the European context. Implementing common solutions and guidelines will be efficient in the long run.	Archiving database extracts is not an efficient solution in every case. We should be aware of the limitations and not to use it automatically in every case. Agencies need help and guidance about issues of database archiving. Agencies need practical solutions to questions of database archiving.

Summary – findings and conclusions

About the study and the findings

This case study collected database preservation practices from the National Archives in five European countries using SIARD. Practices are examined and presented using appraisal, pre-delivery, ingest, and access steps in the long-term preservation workflow from both content providers and archiving organisations viewpoints.

The main focus in this study is on capturing information which has archival value related to database preservation in such a way that it can be used and understood in the future. How information should be preserved, and how it can be disseminated from database management systems, how actual content should be structured and documented and what extra documentation would be needed to preserve usability and semantics. SIARD-based preservation consists of three fundamental layers of documentation:

1) The SIARD file represents a standardised SQL version, ANSI/ISO SQL:2008, of the original relational database to be preserved, containing data, data types, table structure and relations between them, (sometimes) views (virtual tables), etc., represented as XML/XSD pairs inside a ZIP64 file.
2) Documentation enables understanding of the database as such, including how to understand the tables, relations, data (columns in tables), data types, views etc. This involves structural and semantic documentation, both external and internal to the database.
3) Documentation enabling understanding of the original context of creation and use. This includes everything from database views to ER models and laws and regulations defining the rationale behind the database and its associated system.

Without any context documentation, the content of a SIARD file will be impossible to interpret. Information, such as database descriptions, ER diagrams, coded value explanations, user guides from the original production system and the database was a part of is key to obtain usability. It is also recommended that all documents are converted to a format suitable for archiving and/or viewing to ensure that one is able to present the content.

Database system documentation should be required to support the appraisal process and long-term usability of the database content. The documentation should explain how information in the database is related to the business functions of an agency. Is information produced for certain key functions or is data only for support purposes. A SIARD file needs additional information and handling to obtain usability. ISO 15489 defines, (e.g. a usable record as one that can be located, retrieved, presented and interpreted within a time period deemed reasonable by stakeholders).

The examined case studies are divided into following main steps in the workflow: appraisal, pre-delivery, ingest and access.

- The appraisal step includes administrative decisions and processes to decide what information should be preserved.
- The pre-delivery step includes activities which database owner (content provider) should handle and how to prepare data from database to delivered to archival institution.
- The ingest step is done by the archiving institution, and it includes both quality checks and activities which are necessary when transferring digital content to be preserved permanently.
- Access and usability of preserved database information.

In the following, each step is discussed more in detail, based on experiences from the different case contributors.

Appraisal

First, in the preservation workflow, administrative decisions should be made in the appraisal process to decide both at macro and micro-level what should be preserved and what should not. At a macro-level, decisions are made based on evaluation of different parameters:

- Are specific functions of an agency or (public) institution so valuable that it should be documented for the future?
- Is the information valuable as part of the historical documentation of the society in a certain period of time?
- Is documentation of persons or organisations rights and duties needed in the future?

If database preservation is the answer to these questions, one has to move to the next level of decisions. Especially at this micro-level, organisations should be aware what techniques should be used, and documentation is needed for preserving the database content so that the data can be understood in the future, within its original context of data creation and use.

SIARD tools offer good possibilities for a full snapshot of the database but have to be supplemented with several types of additional information. The complexity and the size of the database might cause technical and performance problems, which should, ideally, be explored at these early stages.

Findings and recommendations:

- Information and schemas for performing appraisals should be available at the National Archives websites to make it easier for the database owner to prepare content.
- The existence or non-existence of various types of internal and external documentation should be clarified at this early stage. Documentation on what kind of data, data structures and data models are to be generated and/or hosted within the IT system includes table, column, and data type descriptions, ER diagrams, technical documents, user guides, training manuals, and documentation describing the purpose of the system/database
- The preservation time-line defining how often the data should be preserved from the database (is it every year, every five years, etc?).
- The appraisal process should clearly define what elements of the database are considered valuable and what elements (tables) are not. Two approaches can be taken from here:

    - Save the whole database, even though only parts of it is considered valuable. This approach does not impose any risk of breaking relations in the database and makes SIARD-file generation easy, push the generation button, and low cost.
    - Save only parts of the database (e.g. because of size issues); This approach increases the risk of breaking relations (that will be lost forever) and requires QA (costly) before the preservation of a database subset.
- Strategies for big databases/tables should define practice for managing big databases/tables needed.
- The result of an appraisal process is a submission agreement.

The national requirements for the appraisal process differ from country to country: In the most proactive cases, it is required by law to perform an appraisal of public IT-systems before the system can start running. But then you have old IT-systems, pre-existing such requirements, you have countries not having such strict requirements, and you have the cases when strict requirements have not been followed-up.

As a rule of thumb, the more clarifications in the Appraisal phase, the easier and less costly will the rest of the preservation be.

Pre-delivery

The organisation/agency owning a database (also called, for example, the content provider or the creator) has the responsibility for creating valid SIARD files and providing supplementary external documentation (in accordance with the submission agreement). But the preservation work is in many cases performed by consultants or IT-suppliers, and sometimes by the National Archives (often as part of some kind of crisis management).

Findings and recommendations:

- Information about the SIARD-based preservation process, the SIARD tools, where to get assistance, and requirements on additional documentation should be made available at the National Archives (website). Even though certain National Archives recommend one specific tool, information about all available tools should be present.
- The production of the SIARD file and additional documentation should follow the submission agreement.
- Where to place additional descriptions (of data, tables, views etc.) has to be decided upon
    - Either in the descriptive fields in the SIARD file, using, for example, SIARD Suite or DBPTK (Enterprise).
    - In external documents, or in case of pre-existing database catalogues inside the SIARD file.
    - One of the national archives is using a special-purpose (proprietary) tool for the production of templates describing classes of similar-looking databases, merging the resulting template into the SIARD file.
- In case of inadequate or missing database catalogue or data dictionary, a process for capturing these has to be initiated as soon as possible.

    - Involving different stakeholders at the creator organisation, (e.g. users of the system, database administrators, etc.), together with stakeholders from archives.
    - In case of inadequate or missing database views, a process for capturing these has to be initiated as soon as possible.
    - Involving different stakeholders at the creator organisation, (e.g. users of the system, database administrators, etc). together with stakeholders from archives.
- In case of missing descriptions/illustrations of user-system dialogue in user manuals, a process for capturing these has to be initiated as soon as possible.
    - Involving different stakeholders at the creator organisation, (e.g. users of the system, database administrators, etc), together with stakeholders from archives.

Ingest

Ingest is the process of transferring submitted data, metadata, and additional documentation packaged in a submission package (SIP) into an archive or repository for long-term preservation and dissemination. The workflow consists of several steps, including:

- Identifying and understanding the received data.
- Validating the received data (content), (e.g. against associated metadata).
- Analyse the associated metadata and descriptions, and include it into metadata management systems.
- Package the received data, metadata, and additional documentation for long-term storage (AIP).
- Package the received data, metadata, and additional documentation for dissemination (DIP).

A submission is in many cases rejected one or more times, because of the result of validations involved. The improvements have to be made by the producer/creator/database owner, and re-submission has to take place.

Findings and recommendations:

- Validate if a (complete) database catalog/data dictionary exists. The type of validation will depend on the format of descriptions.
- Validate the existence and the content of views. If no views, check if screenshots of system-user interaction exist, annotated with reference to the database.
- Validation is performed using different types of tools:
    - Visual validation of SIARD files.
    - A simple validation is to import SIARD files produced in one tool to be imported to another tool, to see if it validates.
    - Validation of document structure.
    - Validation of the SIARD file structure.
    - Validation against templates, if existing.
- Validation of SIARD files should be explored further.

Access

How to access preserved databases in the future is also an important theme, not covered in much depth here because several of the contributors to this report have so far limited experience with the access part of SIARD preservation. To summarise access options for SIARD preserved databases:

- A SIARD file, the XML representation of the original database converted to ANSI/ISO SQL:2008, can be exported into a relational database. From here for example, additional views might be added, and a new SIARD file can be generated, making it flexible for new and extended use compared to the original use;
- Both SIARD Suite and DBPTK Enterprise (Database Visualization Toolkit) can be used for accessing SIARD files, and perform editing on descriptive fields;
- A SIARD file can also, in case of urgency (be opened by a ZIP64 tool) and accessed using a standard web browser.

Topics for future work

To improve SIARD-based database preservation, further focus should be put on some of the findings in this study, for example:

- Evaluation of available solutions and tools to produce, handle and use SIARD-files according to usability requirements found in case studies. This would be discussed with SIARD user group.
- Need to develop guidelines, how to present content documentation (e.g. explanations of code values used in the database), in a coherent way.
- Compliance with semantic interoperability efforts and services in public administration.
- Develop more sophisticated methods for validation.