Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Schweizerisches Bundesarchiv BAR

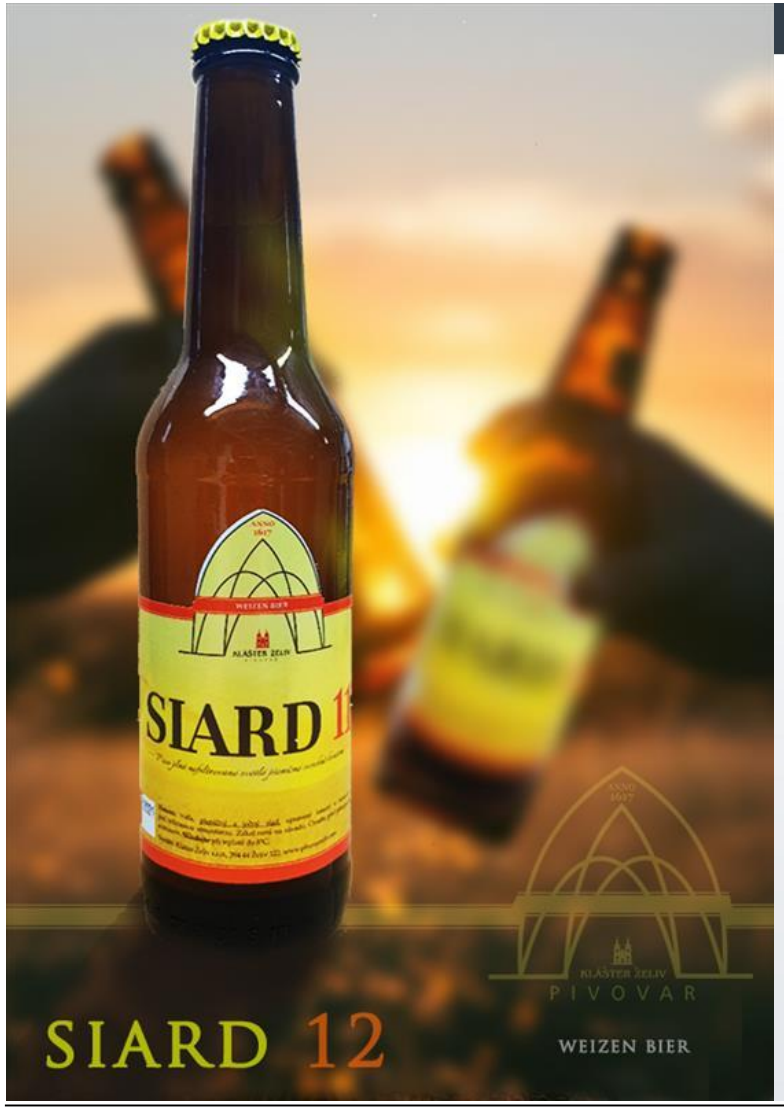
SIARD as a file format for long-term archiving of relational databases

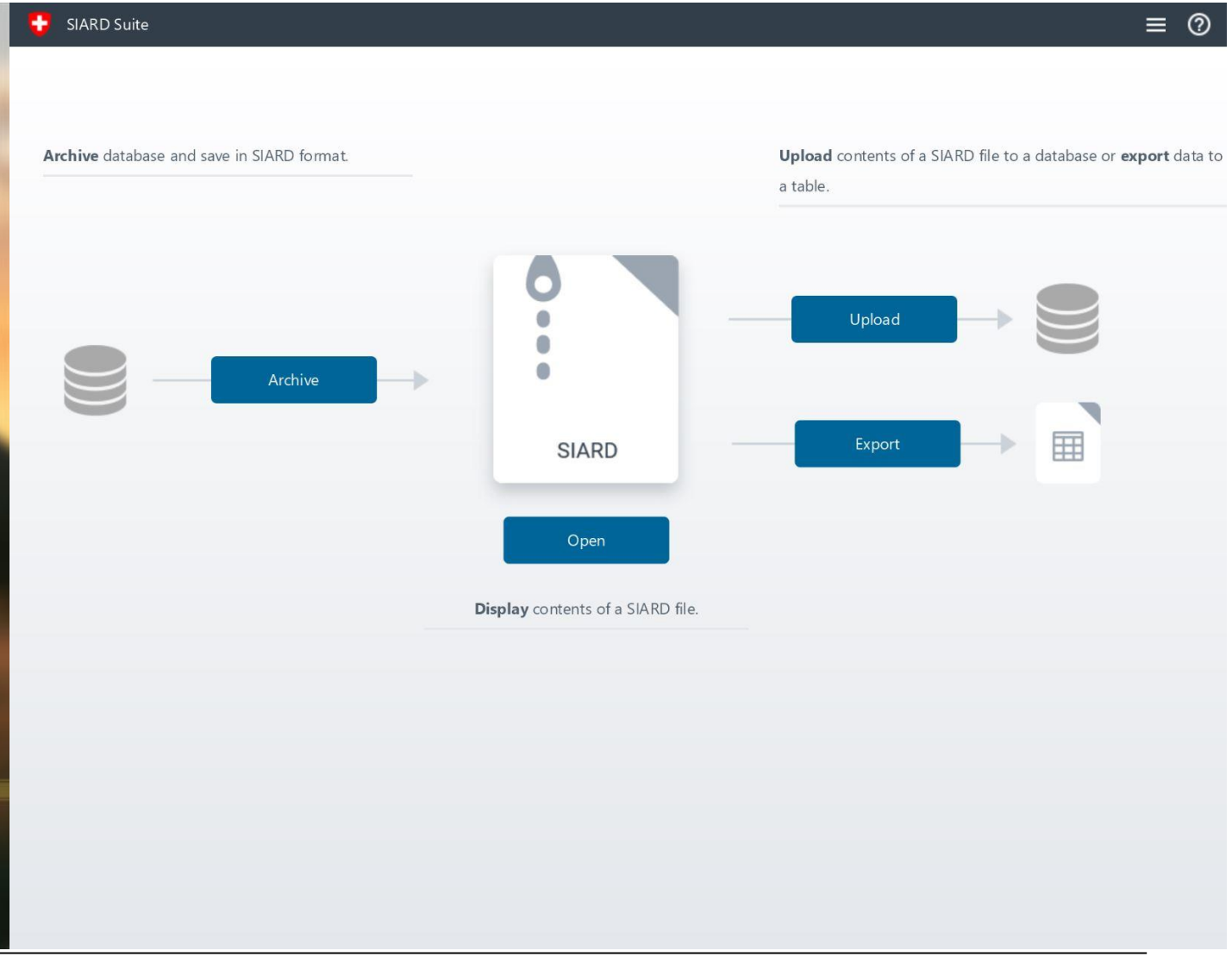
11.12.2023 / Nora Wyler, Silvan Auf der Maur & Audun Lund

Agenda

- History & Background
- SIARD as a file format for long-term archiving of relational databases
- Live Demonstration
- SIARD RDF
- Questions & Answers

What is SIARD?





How did it start?

Dr. Krystyna W. Ohnesorge

Dr. Hartwig Thomas

- 2004: Analysis of a system for long-term archiving of databases -> SIARD
- 2007: SFA develops version 1.0 of the SIARD format
- 2008: SIARD 1.0 is accepted as the official format for archiving relational databases by the European Open PLANETS project
- 2010: DNA creates a Danish variant named SIARDDK
- 2013: SIARD format version 1.0 as Swiss standard → eCH-0165
- 2016: Design and development of the SIARD preservation format 2.0 by SFA under the auspices of the E-ARK project
- 2019: Release of SIARD format version 2.1.1 correcting errors and ambiguities in Version 2.0 that might have led to practical problems
- 2021: DILCIS Board further develops this format during the E-ARK3 project and releases SIARD format 2.2

EMULATION – MIGRATION – NORMALIZATION

Currently the discussion about archiving digital material is dominated by the two strategies Emulation and Migration. We propose to add Normalization as a third option. […] Migration as an archival strategy advocates transforming the archived materials into a new format whenever their format is no longer supported by the archive's IT environment. Normalization as an archival strategy aims to keep archived material unchanged forever. This is possible, if the material is archived in a “normal” form for the type of objects archived.

Why is it so difficult to archive a DB?

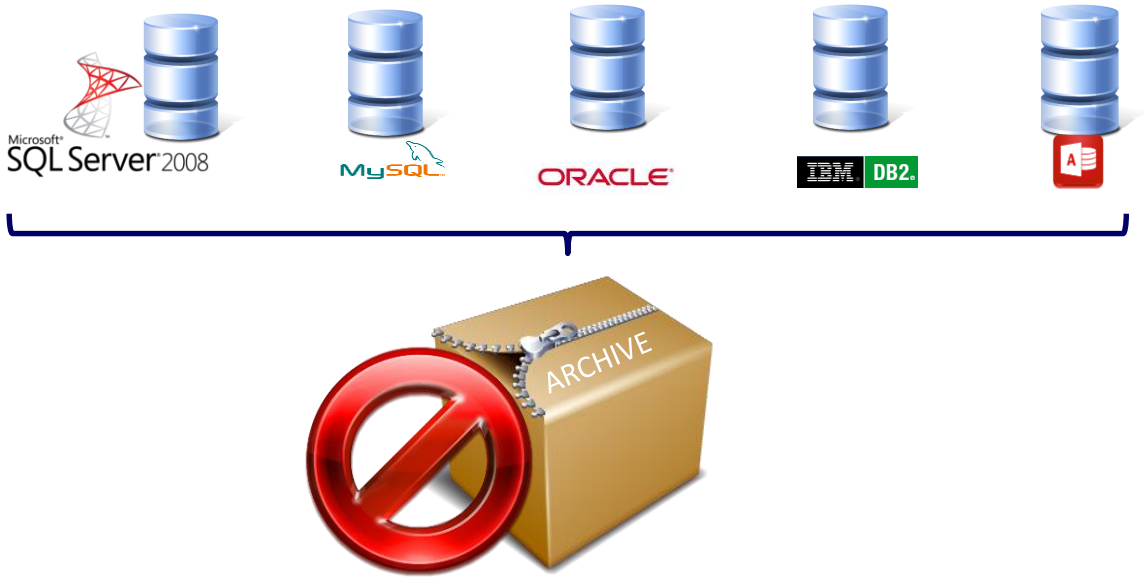
Several major database management systems exist and their storage formats are not compatible.
… And these systems and storage formats evolve over time!

Conclusions:

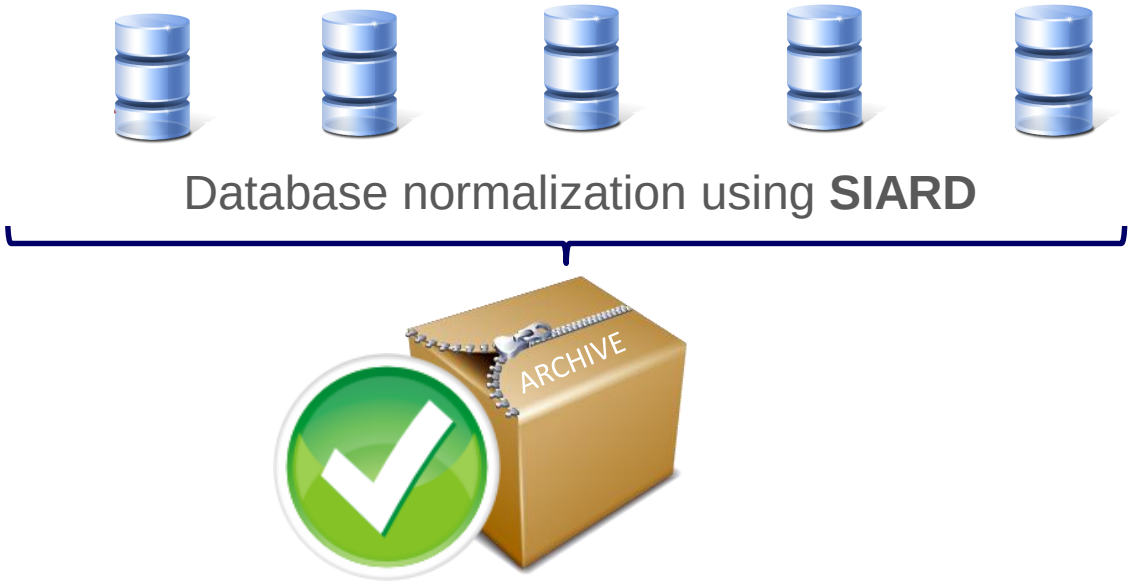
- we cannot archive proprietary storage formats
- today's storage files are almost certainly

  not readable with tomorrows applications

Proprietary database management systems cannot be preserved for the long term



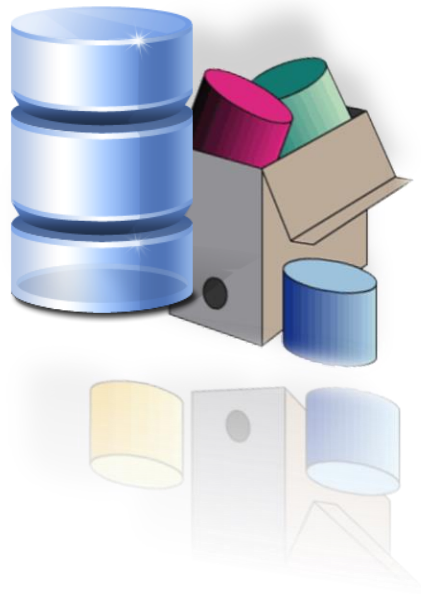
The SFA developed the format SIARD for archiving the contents of relational databases



SIARD Suite – supported RDMBS

- (Excel)
- **Microsoft Access**
- **Oracle**
- **Microsoft SQL Server**
- **MySQL / MariaDB**
- **IBM DB2**
- **PostgreSQL**
- Sybase
- SQLite

SIARD Suite



Archivable File Formats

Area of application	Archivable formats	Remarks
Text (unstructured)	plain text	UTF-8, UTF-16 ISO-8859-1, ISO-8859-15 US-ASCII
“Office” documents	PDF/A	Corresponds to PDF 1.4 (PDF/A-1) and PDF 1.7 (PDF/A-2) with limitations
Tables	CSV	Text file with delimiters, encoding as with text (unstructured)
Relational databases	**SIARD**	Version 2.2 PUID fmt/1777
Raster images *Geo-Rasterimages*	TIFF GeoTIFF	 *TIFF + XML*
Audio	WAVE	
Video	MPEG-4	

SIARD – what does it mean?

Software **I**ndependent **A**rchiving of **R**elational **D**atabases

- **Nature of data**: databases
- **Type**: relational databases
- **Convert content** into archivable format
- **Detachment** of data from executable applications

SIARD principles

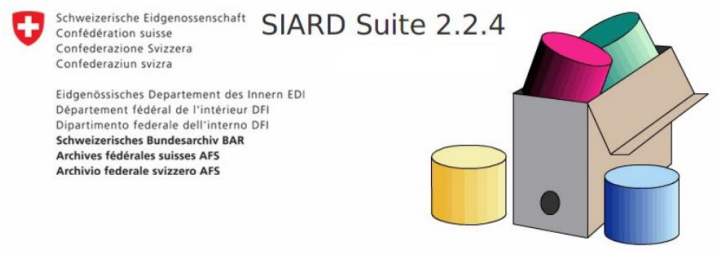
- **Preserve Information**, not layout or interaction
- **Preserve primary data**, not code
- Preserve tables with their **relations**

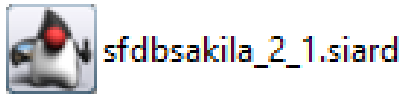
Functionality Preservation

Constraints
Archived databases are consistent when they are created from consistent databases.
Since, once archived, they will not be changed, preserving constraints is not mandatory.

Format vs. Software

- SIARD is an archivable format – SIARD Suite is a software – Please do not confuse the two!





«SIARD Suite extracts content from relational databases and stores it in the SIARD format suitable for archiving. This file format makes it possible to keep the data in the archive for a long time and independently of the original software. If necessary, the data can be loaded into a new database. This means that they can be stored independently of the original database and can also be reused in the future in modern database systems.»

SIARD Format – Technical details

- The SIARD format saves database-content in a **SIARD-file (SIARD-archive)** xxx.siard
- A SIARD-file is **ZIP-folder (ZIP64)** which contains several XML-files
- There is a single **XML-file** that documents all metadata for the database content, based on **SQL:2008**
- The remaining **XML-files** contain data from the tables (the actual database content)
- The format SIARD is based on **open standards**: SQL:2008, XML, XML Schema, UNICODE

SUSTAINABILITY

[digital-preservation/Digital_Preservation_Risk_Matrix at master · usnationalarchives/digital-preservation · GitHub]

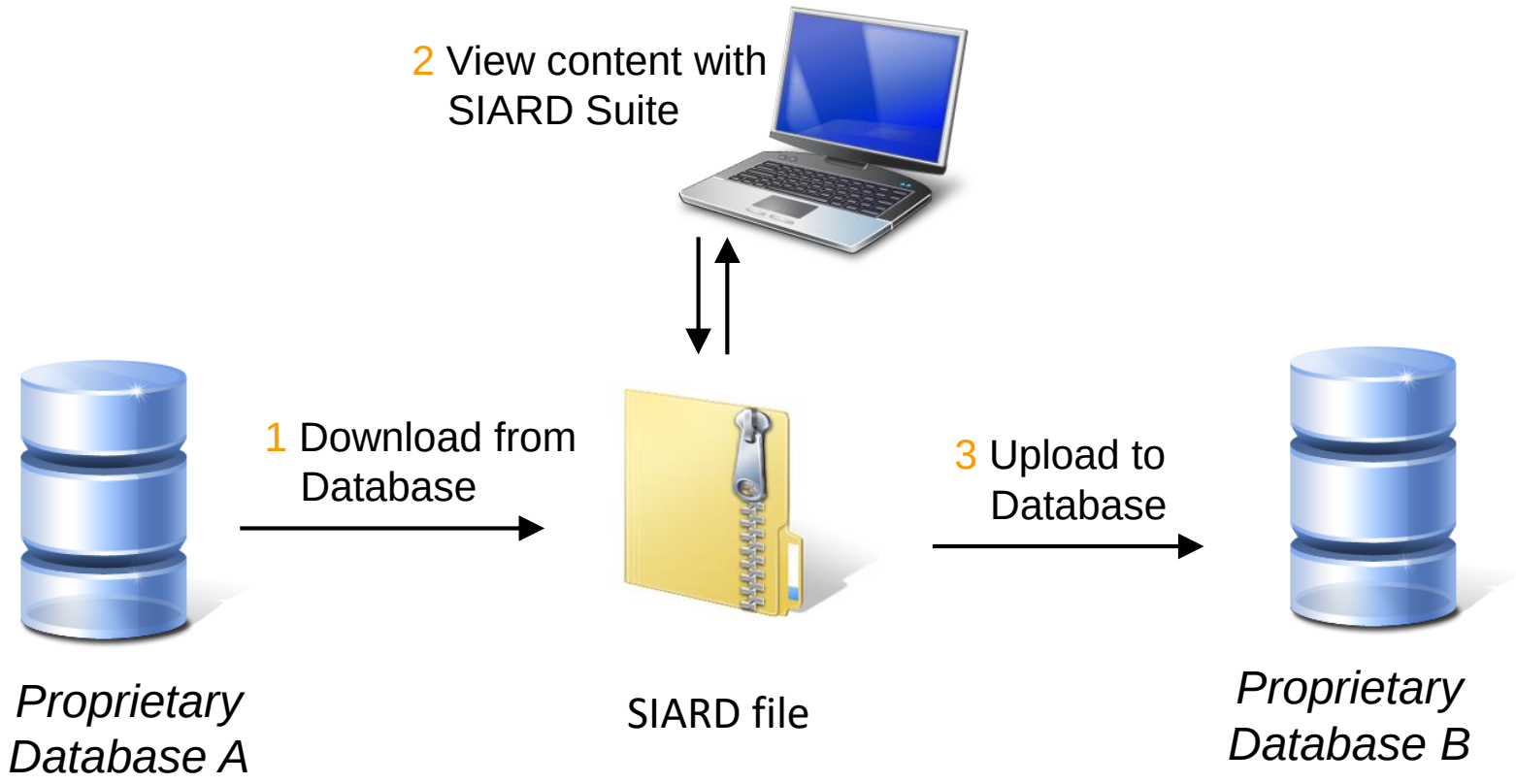
Format Name	File Extension(s)	Category/Plan(s)	Risk Level	NARA TOTAL
Microsoft Access 2019	accdb	Databases	Moderate Risk	12
SIARD 2.1	siard	Databases	Low Risk	51
SIARD 2.2	siard	Databases	Low Risk	51
JPEG 2000 File Format	jp2	Digital Still Image	Low Risk	24
JPEG File Interchange Format 1.02	jpg|jpeg	Digital Still Image	Low Risk	37
Tagged Image File Format for Internet Fax (TIFF-FX)	tif|tiff|tfx	Digital Still Image	Low Risk	25
Windows Bitmap 5.0	bmp	Digital Still Image	Moderate Risk	1
Microsoft PowerPoint Presentation OpenXML (Windows 2007-onwards)	pptx	Presentation and Publishing	Low Risk	25
Portable Document Format/Archiving (PDF/A-1a) accessible	pdf	Presentation and Publishing|Textual and Word Processing	Low Risk	38
Portable Document Format/Archiving (PDF/A-1b) basic	pdf	Presentation and Publishing|Textual and Word Processing	Low Risk	38
Microsoft Excel Office Open XML	xlsx	Spreadsheets	Low Risk	30
OpenDocument Spreadsheet 1.3	ods|fods|ots	Spreadsheets	Low Risk	45
Microsoft Word for Windows 2007-onwards (OOXML)	docx	Textual and Word Processing	Low Risk	30
OpenDocument Text 1.3	odt|fodt|ott	Textual and Word Processing	Low Risk	45
Plain Text	Plain_Text|txt|text|asc|rte	Textual and Word Processing	Low Risk	26
Rich Text Format 1.9	rtf	Textual and Word Processing	Moderate Risk	5
eXtensible Hypertext Markup Language 1.1	xhtm|xhtml	Web Records|Software and Code	Low Risk	33
eXtensible Markup Language 1.1	xml	Web Records|Software and Code|Structured Data|Textual and Word Processing	Low Risk	42

SIARD and its Tools

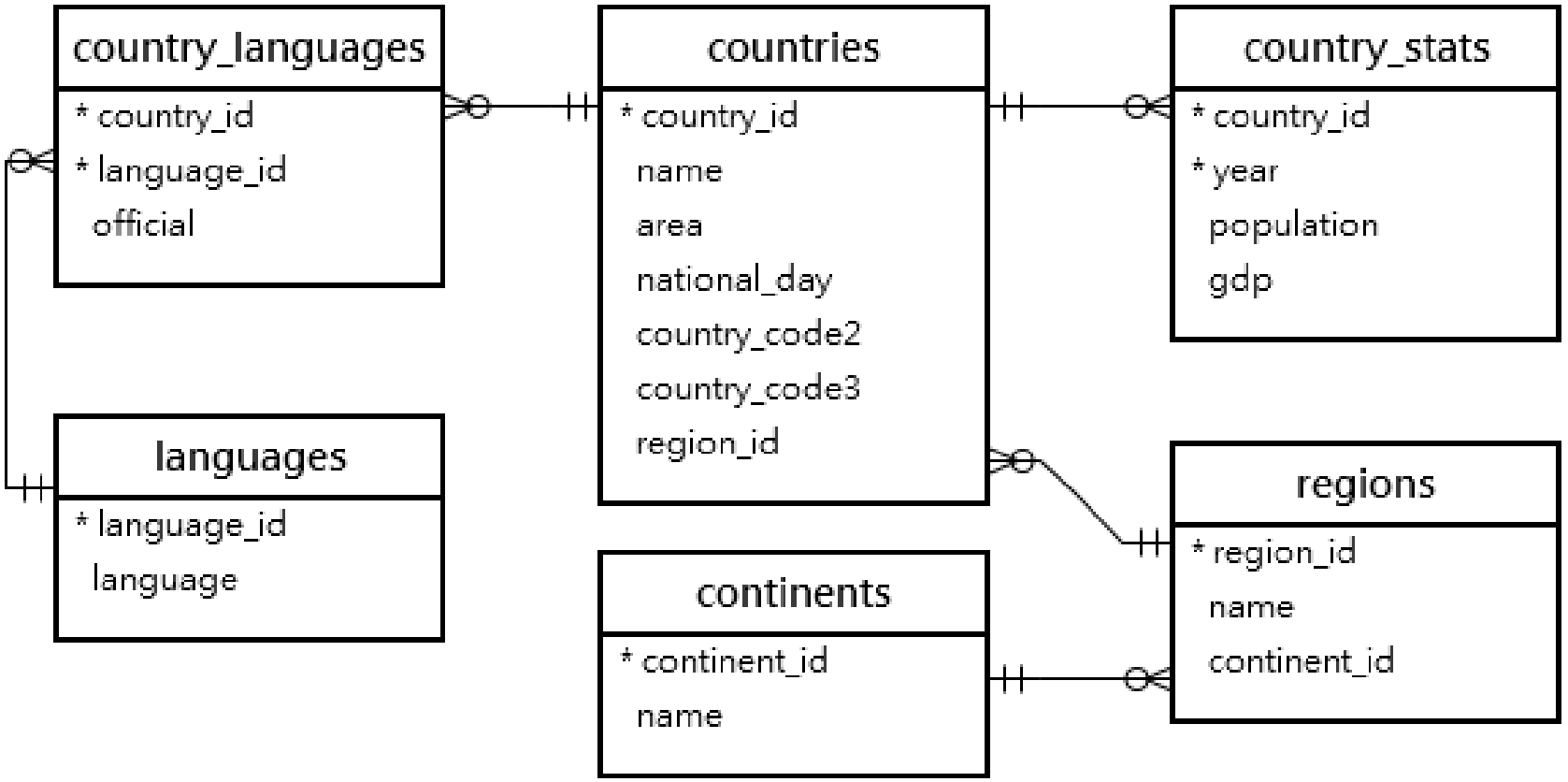
SIARD Format - DILCIS Board - OSS GitHub - GitHub – DILCISBoard

- SIARD Suite - download, explore, upload - SFA - OSS GitHub GitHub - sfa-siard
- DBPTK - download, explore, upload - Keep Solutions – OSS GitHub GitHub - keeps/dbptk-ui
- KOST-Val - validation - KOST - OSS GitHub - GitHub - KOSTKOST-Val
- SIARDexcerpt - access individual records - KOST - OSS GitHub GitHub - KOST/SIARDexcerpt
- dbDIPview - access - ARS - OSS GitHub - GitHub - dbdipview

SIARD Suite – Features



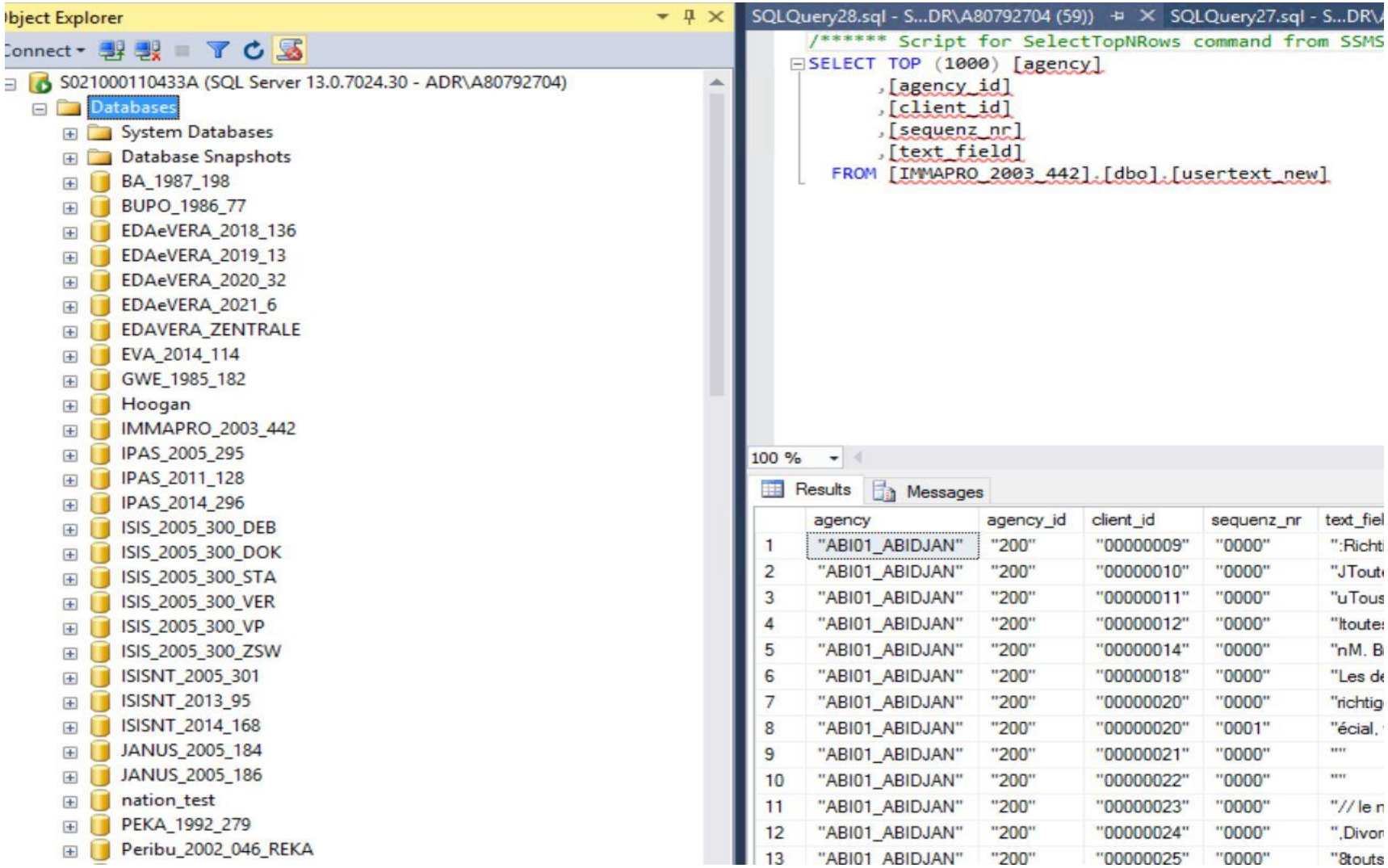
«Nation» MariaDB Database – live demonstration



SIARD Access

- SIARD Suite (SFA) / DBPTK (Keep Solutions)
- Mediation Server
- iDA (from Piql)
  - preserve the database data and query functionalities.
  - An API that executes queries in the SIARD database.
  - The data in SIARD format, the iDA query engine and the iDA API are integrated into an Archival Information Package (AIP).
- SIARD RDF

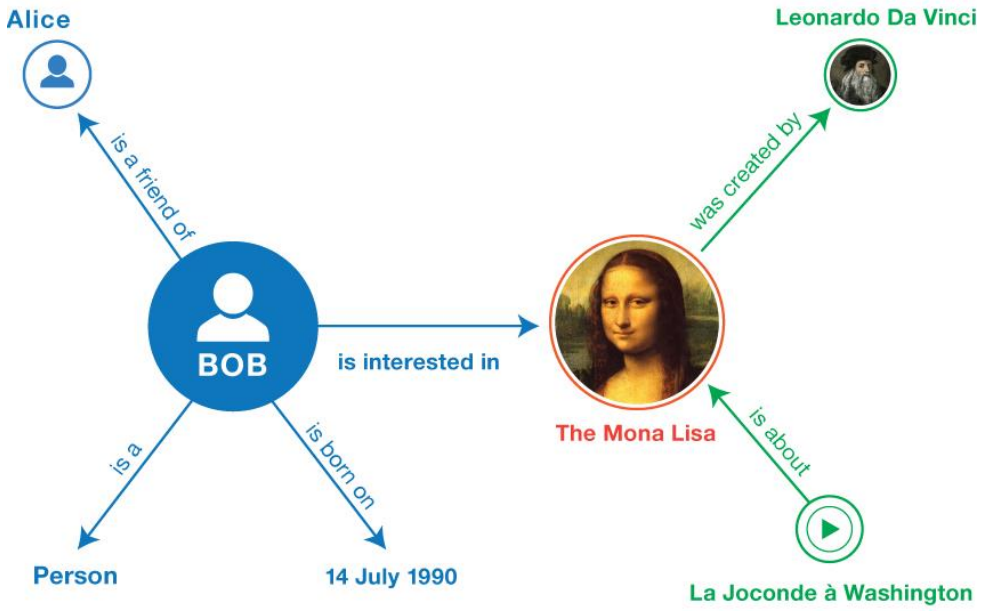
Mediation Server



SIARD to RDF

Linked Data 101

- Information is defined by relationships to other information points
- Each information point has a unique identification: Unique Resource Identifier
- Also called "Semantic Web" because the logic equals a phrase structure: subject URI -> predicate -> object URI
- Standardised naming schemes enable interoperability of data records

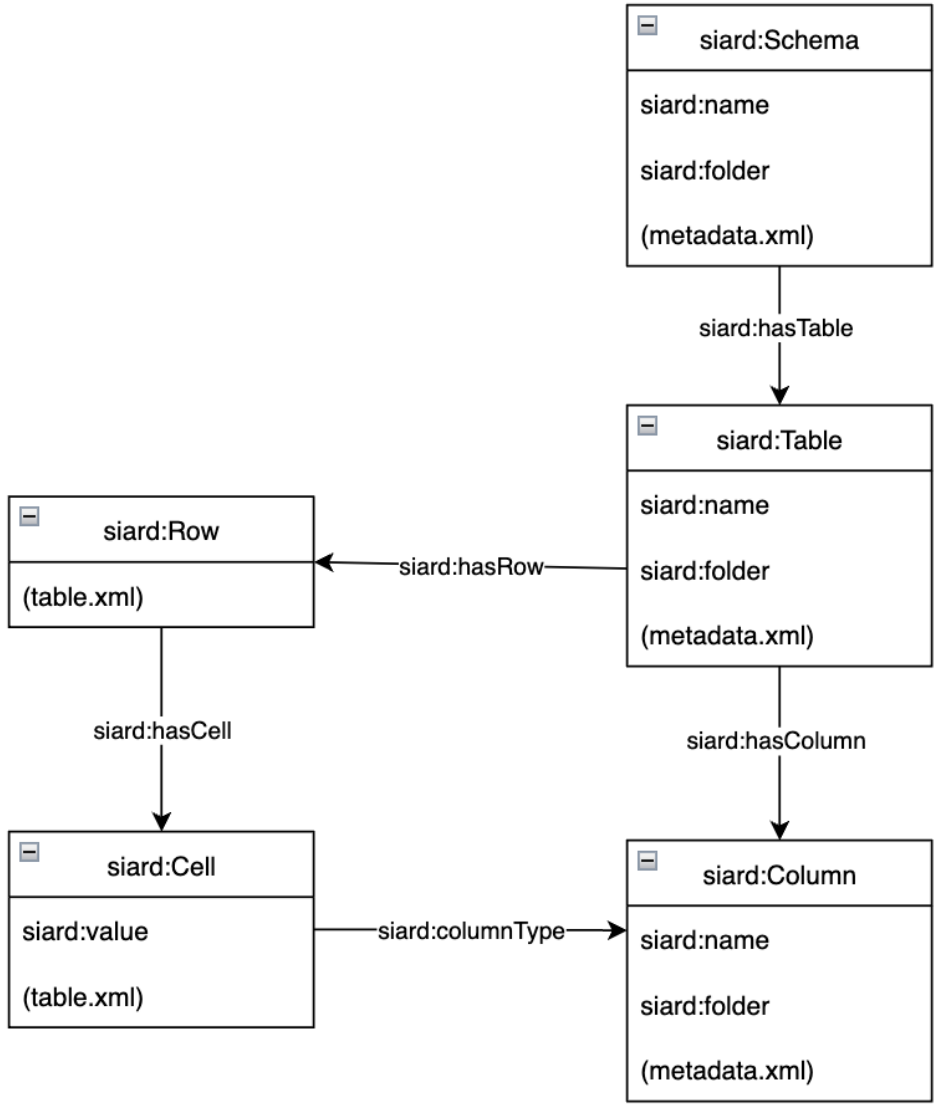


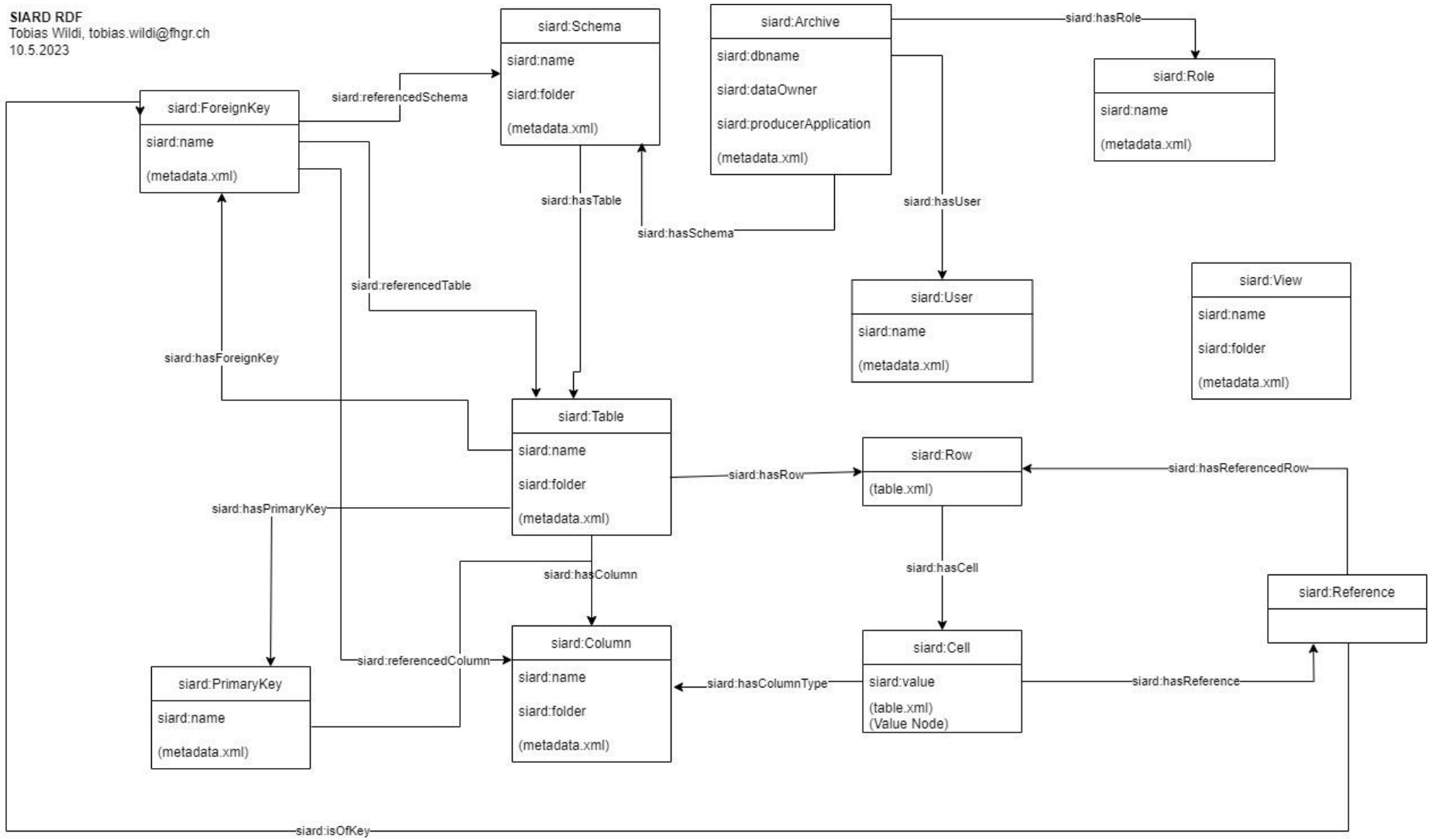
SIARD to RDF objectives

- Transforming data from relational databases into linked data
- Enabling interoperability with other datasets

Schema Mapping

- Components have properties with defined values (literals)
- Components are linked by predicates / properties
- The direction of relations must be maintained

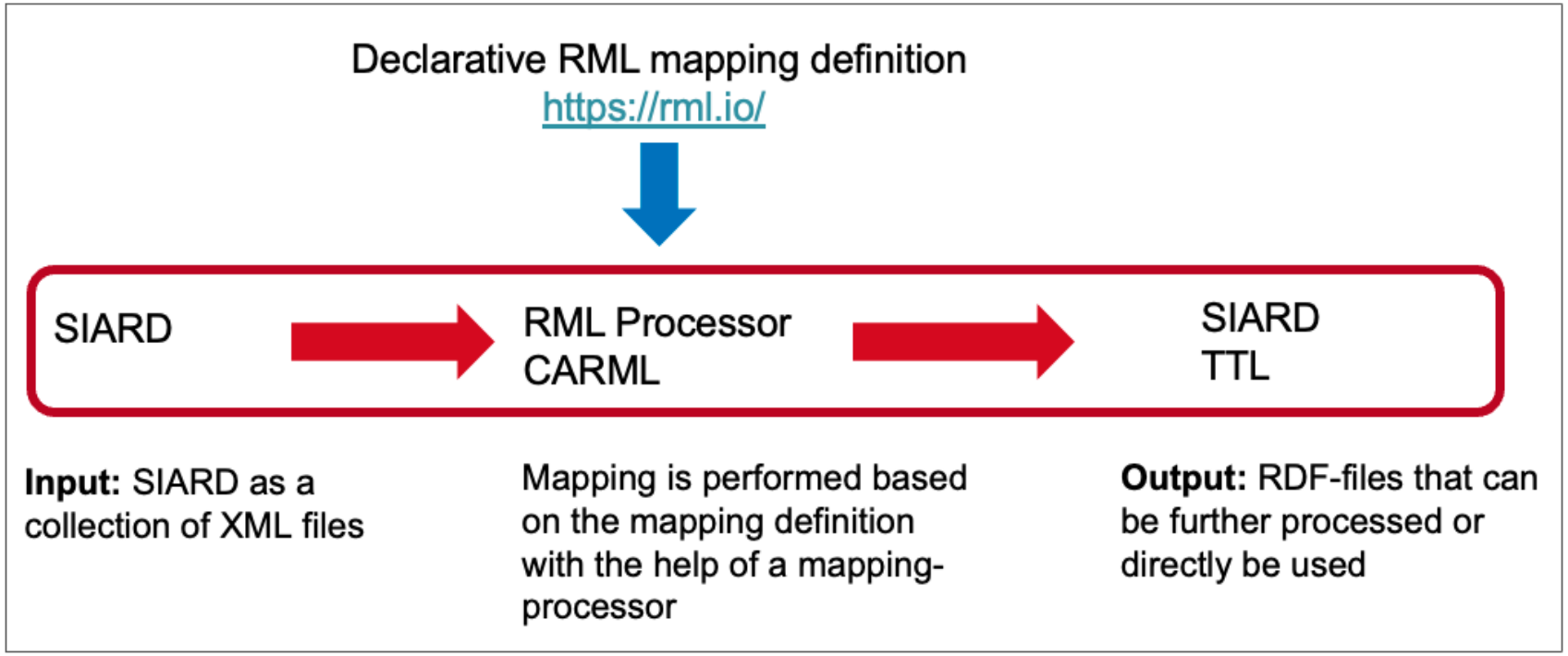



SIARD RDF
Tobias Wildi, tobias.wildi@fhgr.ch
10.5.2023
siard:Schema
siard:name
siard:folder
(metadata.xml)
siard:Archive
siard:dbname
siard:dataOwner
siard:producerApplication
(metadata.xml)
siard:hasRole
siard:Role
siard:name
(metadata.xml)
siard:ForeignKey
siard:name
(metadata.xml)
siard:referencedSchema
siard:hasTable
siard:hasSchema
siard:hasUser
siard:referencedTable
siard:User
siard:name
(metadata.xml)
siard:View
siard:name
siard:folder
(metadata.xml)
siard:hasForeignKey
siard:Table
siard:name
siard:folder
(metadata.xml)
siard:hasRow
siard:Row
(table.xml)
siard:hasReferencedRow
siard:hasPrimaryKey
siard:hasCell
siard:hasColumn
siard:Reference
siard:Column
siard:name
siard:folder
(metadata.xml)
siard:Cell
siard:value
(table.xml)
(Value Node)
siard:referencedColumn
siard:hasColumnType
siard:hasReference
siard:PrimaryKey
siard:name
(metadata.xml)
siard:isOfKey

Implementation of SIARD RDF mapping

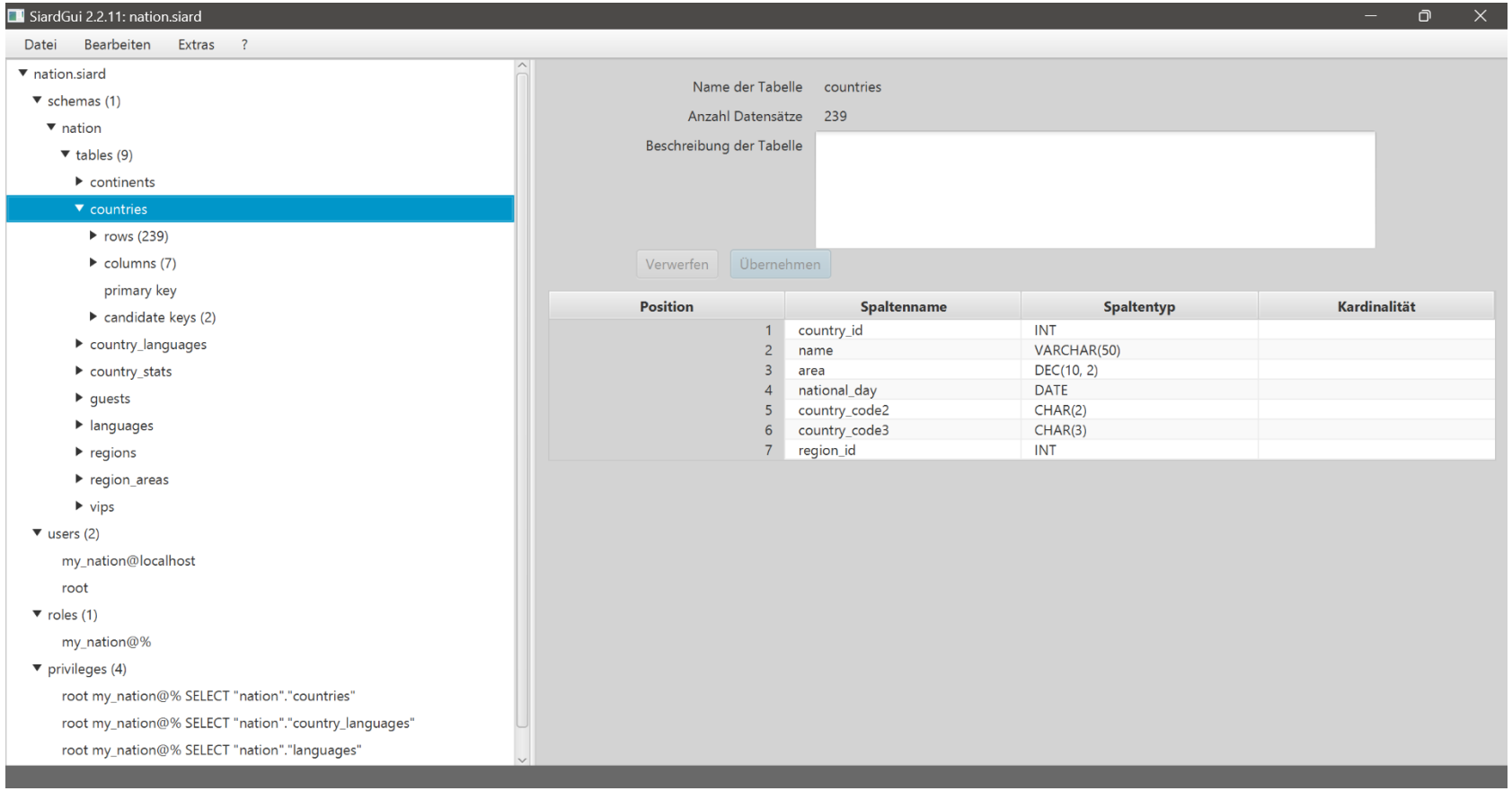
Declarative mapping with RML

To implement the mapping, RML (https://rml.io/) was used as a declarative mapping language. Based on the declaration in RML, the transformation from XML-based SIARD to RDF is carried out by CARML (https://github.com/carml/carml). CARML produces the triples in various formats (TTL, NT, JSON)



Demonstration

[YASGUI (admin.ch)]



Sample queries

Triples

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix dcat: <http://www.w3.org/ns/dcat#>
SELECT (count(?s) as ?TripleCount)   WHERE {
  graph <http://lindas.admin.ch/sfa/nationDB> {?s ?p ?o}
} LIMIT 10000
```

Properties

```
SELECT distinct ?p   WHERE {
  graph <http://lindas.admin.ch/sfa/nationDB> {?s ?p ?o}
}
LIMIT 10000
```

Languages

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?language ?official_status WHERE {
  #Get ID from countries table (replace country name in " ")
            ?cell <http://siard.link#value> 'Switzerland'.
            ?rows <http://siard.link#hasCell> ?cell;
            <http://siard.link#hasCell> ?cells.
            ?cells <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/countries/country_id>;
            <http://siard.link#value> ?swiss_id.
  #Use ID to get language IDs from country_languages table
    ?id_cell <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/country_languages/country_id>;
            <http://siard.link#value> ?swiss_id.
            ?row <http://siard.link#hasCell> ?id_cell;
            <http://siard.link#hasCell> ?language_id.
            ?language_id <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/country_languages/language_id>;
            <http://siard.link#value> ?values.
  #Use language IDs to get language from languages table
            ?lang_id <http://siard.link#value> ?values;
    <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/languages/language_id>.
            ?lang_row <http://siard.link#hasCell> ?lang_id.
            ?lang_cell <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/languages/language>.
    ?lang_row <http://siard.link#hasCell> ?lang_cell.
            ?lang_cell<http://siard.link#value> ?language.
  #Get Whether official or not from country_languages table
            ?row <http://siard.link#hasCell> ?official.
            ?official <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/country_languages/official>;
          <http://siard.link#value> ?official_status
}
```

Federated Query Capital

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
SELECT ?name_en ?country_name ?capital_name WHERE {
#Get language ID
            ?lang_cell<http://siard.link#value> "Urdu";
            <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/languages/language>.
            ?lang_row <http://siard.link#hasCell> ?lang_cell.
            ?lang_row <http://siard.link#hasCell> ?lang_id_cell.
            ?lang_id_cell <http://siard.link#value> ?lang_id;
    <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/languages/language_id>.
#Use language ID to get Country ID from country_languages table
            ?id_cell <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/country_languages/language_id>;
            <http://siard.link#value> ?lang_id.
            ?row <http://siard.link#hasCell> ?id_cell;
            <http://siard.link#hasCell> ?country_id_cell.
            ?country_id_cell <http://siard.link#hasColumnType>
<https://ld.admin.ch/(...)/nation/country_languages/country_id>;
            <http://siard.link#value> ?country_id.
#Use country ID to get country name from country table
            ?cell <http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/countries/country_id>;
            <http://siard.link#value> ?country_id.
            ?country_row <http://siard.link#hasCell> ?cell;
            <http://siard.link#hasCell> ?country_cell.
            ?country_cell<http://siard.link#hasColumnType> <https://ld.admin.ch/(...)/nation/countries/name>;
    <http://siard.link#value> ?country_name.
            BIND(STRLANG(?country_name, "en") as ?name_en)
#Use country name to get capital from wikidata:
SERVICE <https://query.wikidata.org/sparql> {
            ?country_URI wdt:P31 wd:Q3624078;
            rdfs:label ?name_en;
            wdt:P36 ?capital_URI.
            ?capital_URI rdfs:label ?capital_name
            FILTER(LANG(?capital_name) = "en")
            FILTER(LANG(?name_en) = "en")
}
} LIMIT 10
```

Advantages of SIARD as linked data

- Interoperability of data records
- Opportunity for automation

Getting started

- Releases · sfa-siard/siard-suite - GitHub
- SIARD-Suite-Getting-Started
- SIARD-Suite-User-Manual

Questions / Contact

- audun.lund@bar.admin.ch
- nora.wyler@bar.admin.ch
- DA-Support@bar.admin.ch
- Datenbankarchivierung: SIARD Suite (admin.ch)

Folgen Sie dem Bundesarchiv / Suivez les Archives fédérales /
Seguite l'Archivio federale / Follow the Federal Archives:

Thank you

for your attention!

Any

questions?