Welcome to this live webinar on
The large object handling capabilities of SIARD with the Database Preservation Toolkit

Start 10:00

15 June 2023

Audience notes for the Live Webinar

Your cameras have been turned off and microphones muted.

If you have any technical issues during the event, please use the chat function.

Please use the Q&A for questions to speakers. These will be addressed at the end of the event.

Please note that this webinar will be recorded. No attendee personal information will be captured in these recordings.

Agenda

10:00 – 10:05

eArchiving Initiative welcome

Jaime Kaminski – eArchiving Initiative training activity lead

10:05 – 10:50

The large object handling capabilities of SIARD with the Database Preservation Toolkit

István Alföldi – Poliphon

10:50 – 11:00

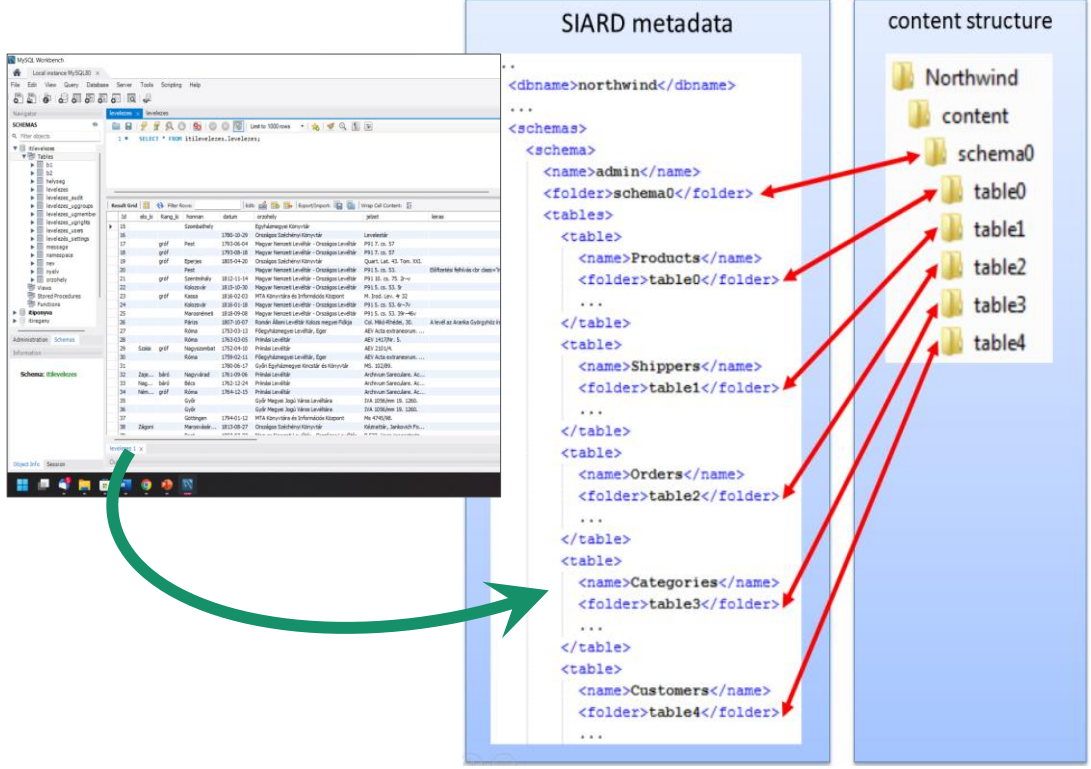
Q&A

Large object handling capabilities of SIARD with Database Preservation Toolkit

István Alföldi, Poliphon (alfi@poliphon.hu)

eArchiving Initiative Training Webinar

Content



- Introduction to DB archiving and SIARD (non-technical)
- Pilot at DH-Lab, Budapest (non-technical)
- LOB handling with SIARD (more technical)

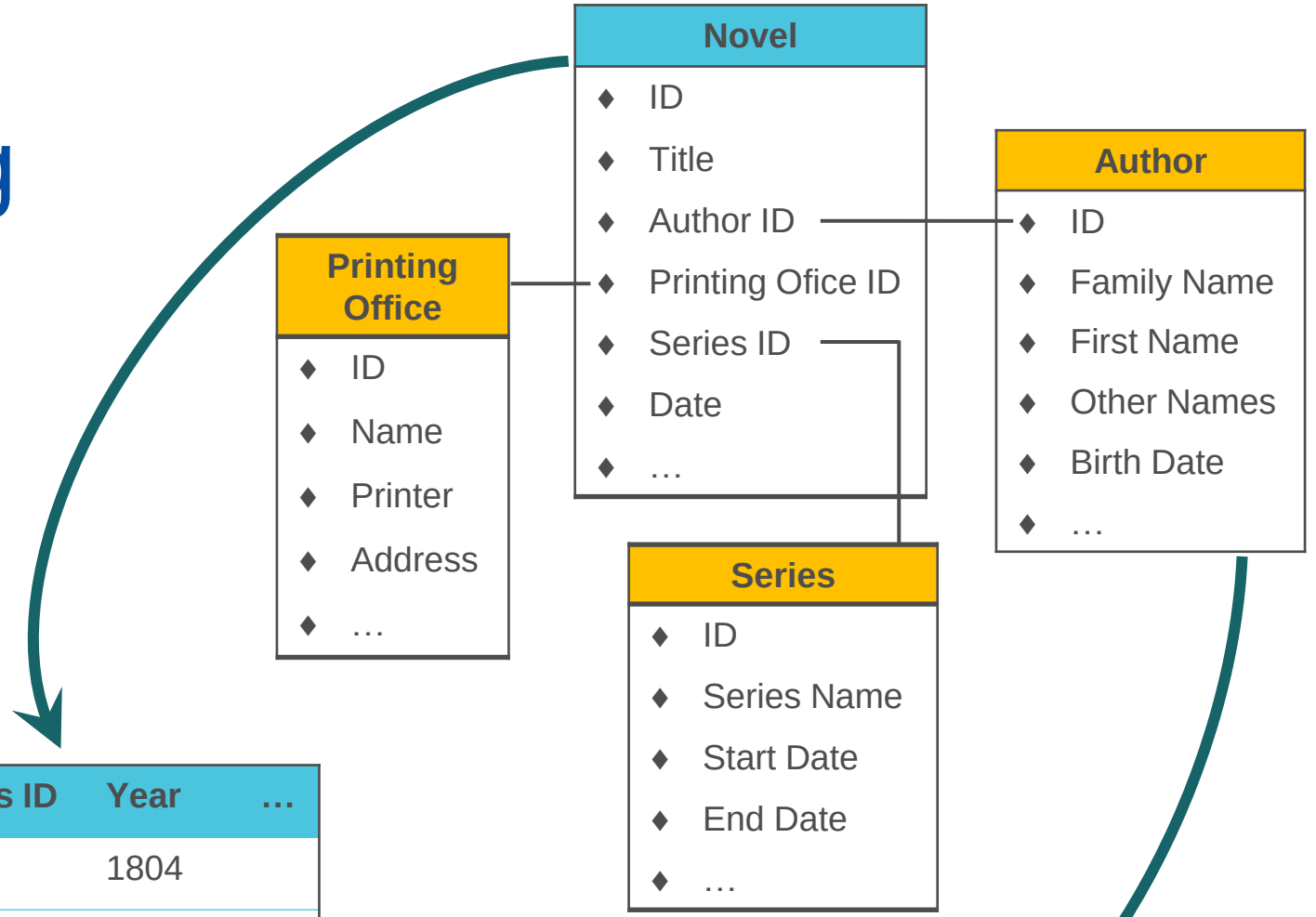
European Commission

Introduction to DB archiving and SIARD

European Commission

Database archiving

- Relational database
- Tables, Columns and Rows



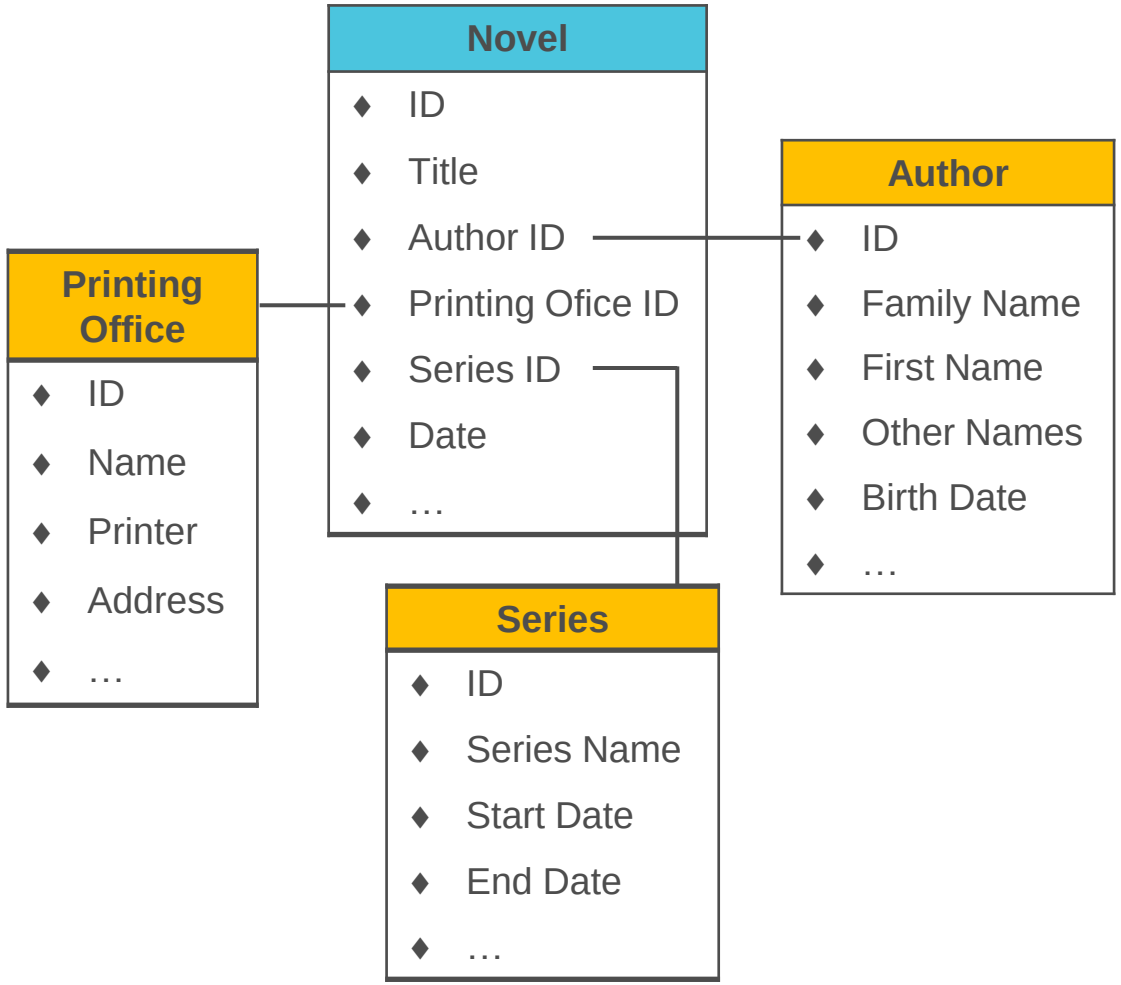
ID	Title	Author ID	PH ID	Series ID	Year	…
101	Kámor	503	2	14	1804	
102	Ottília	201	2	11	1786	
103	A Bölcsek köve	145	7	9	1793	

ID	Family Name	First name	Other Names	…
503	Campe	Joachim	Heinrich	
504	Eötvös	Károly		

Database archiving

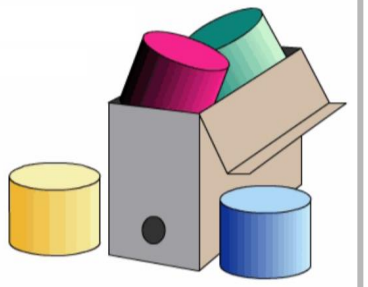
Other DB elements

- Schemas
- Views
- Keys
- Triggers
- Constraints
- Users, groups, rights
- Etc.



SIARD

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra



E-ARK

- Software Independent Archival of Relational Databases (SIARD)
- Open file format for the long-term archiving of relational databases
- Capable of archiving practically all types of DB elements
- Text data based on XML, packaged in a ZIP container files
- Based on standards including Unicode, XML, SQL:2008, URI, ZIP
- Originally developed by the Swiss Federal Archives, later versions in cooperation with the E-ARK projects.
- Now at version 2.2

European Commission

SIARD

3.1 Use of standards

To ensure that the contents of a database remain interpretable over a long period, the SIARD format is essentially based on two ISO standards: XML and SQL:2008.

ID	Description of requirement	M/O
G_3.1-1	All database content is stored in a collection of files in XML 1.0 format[4] that conform to schema definitions according to XML Schema 1.0[5]. The schema definitions and SQL code must in each case conform to SQL:2008 in accordance with ISO/IEC 9075. The only exceptions are BLOB and CLOB data (Binary Large Objects and Character Large Objects) which are stored in separate binary and text files but are referenced in the XML files.	M

3.2 Databases as documents

A relational database is treated as a single document to be archived, so that the references between the data in individual tables are preserved.

ID	Description of requirement	M/O
G_3.2-1	A relational database is archived in a single SIARD file. This file may reference externally stored Large Objects that belong to the database in a larger sense. In rare cases a SIARD file may need to be segmented due to size.	M

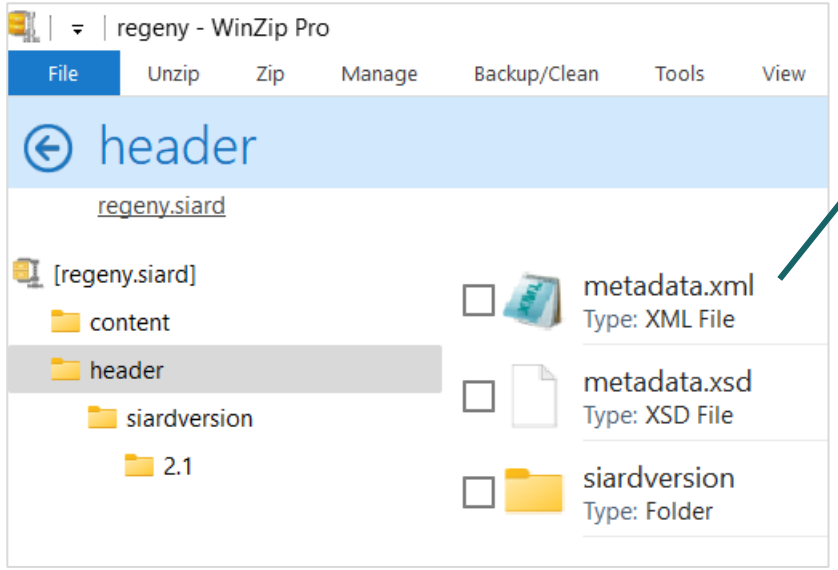
4.1 Construction of the SIARD archive file

The SIARD archive file is realised as a ZIP archive.

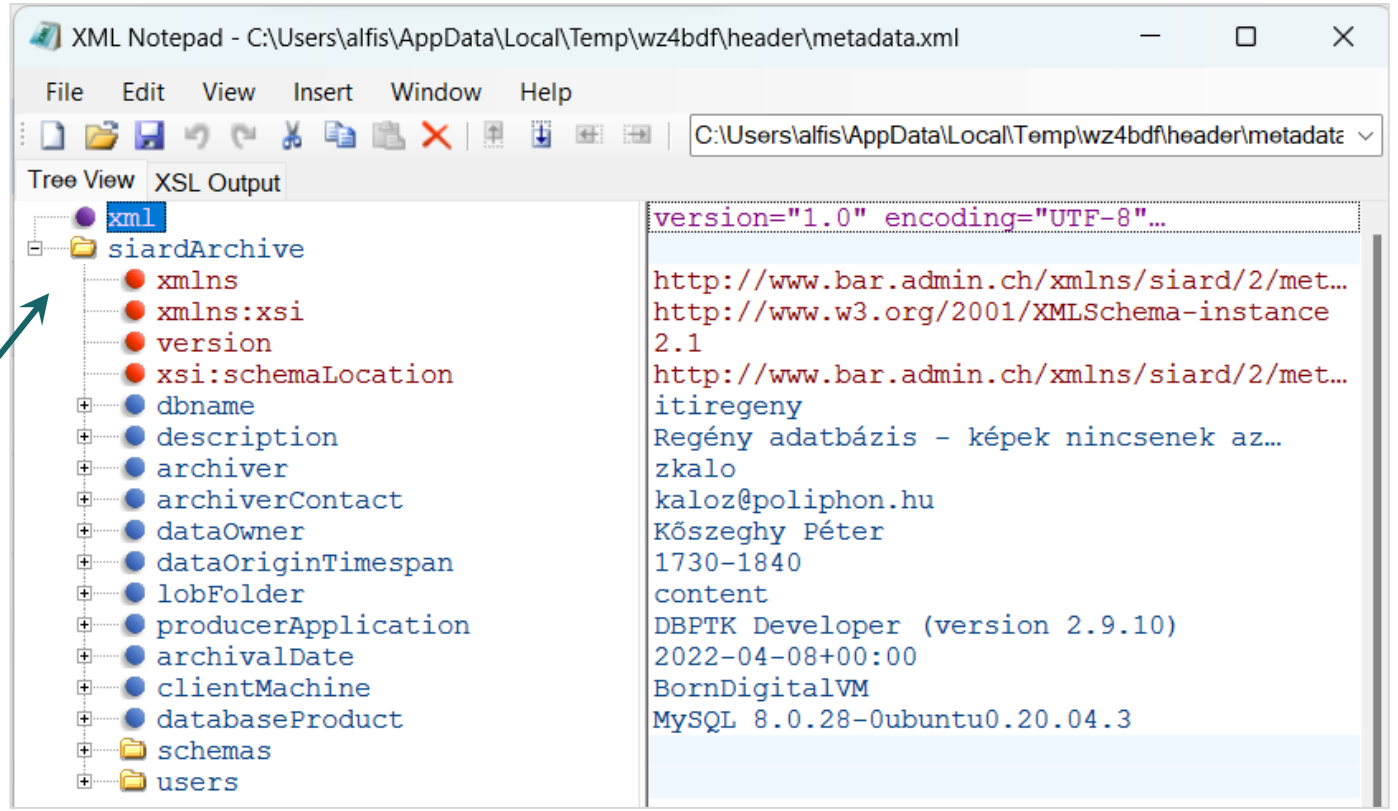
ID	Description of requirement	M/O
G_4.1-1	The SIARD file is stored as a single, ZIP archive in accordance with the specification published by the company PkWare, version 6.3.2[8].	M

SIARD – Header

ZIP file \header



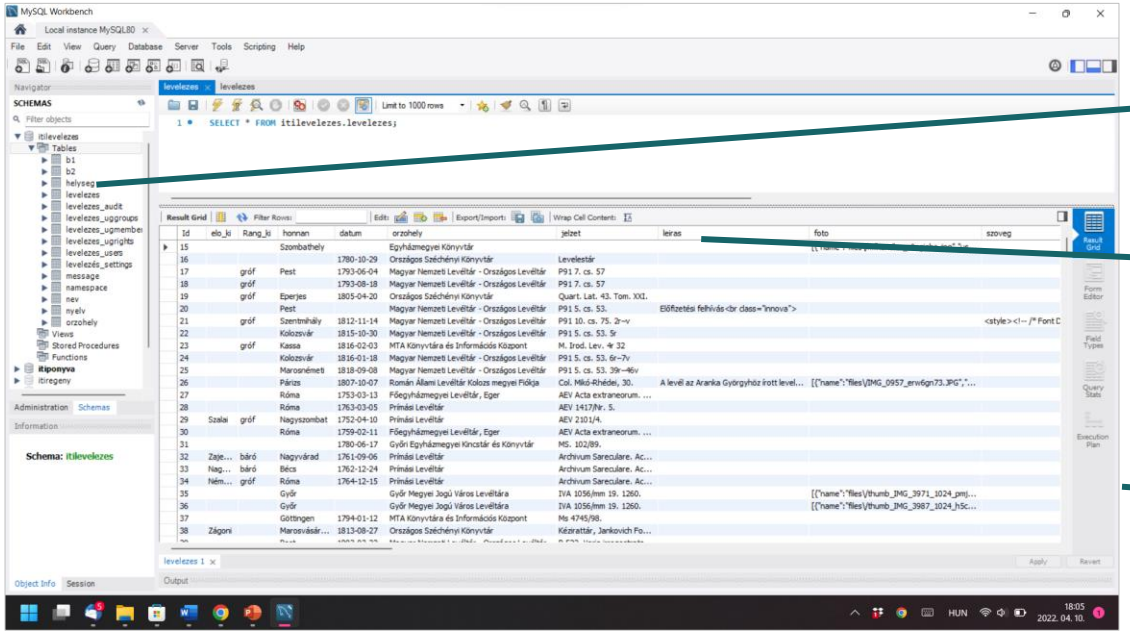
Metadata.xml



SIARD – Mapping

metadata.xml
Type: XML File

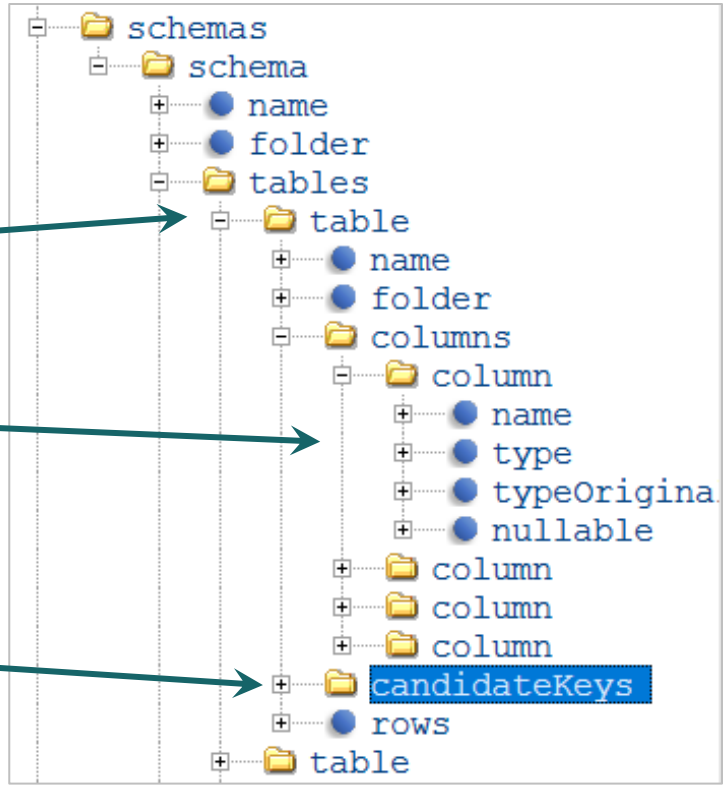
Database



Table

Column

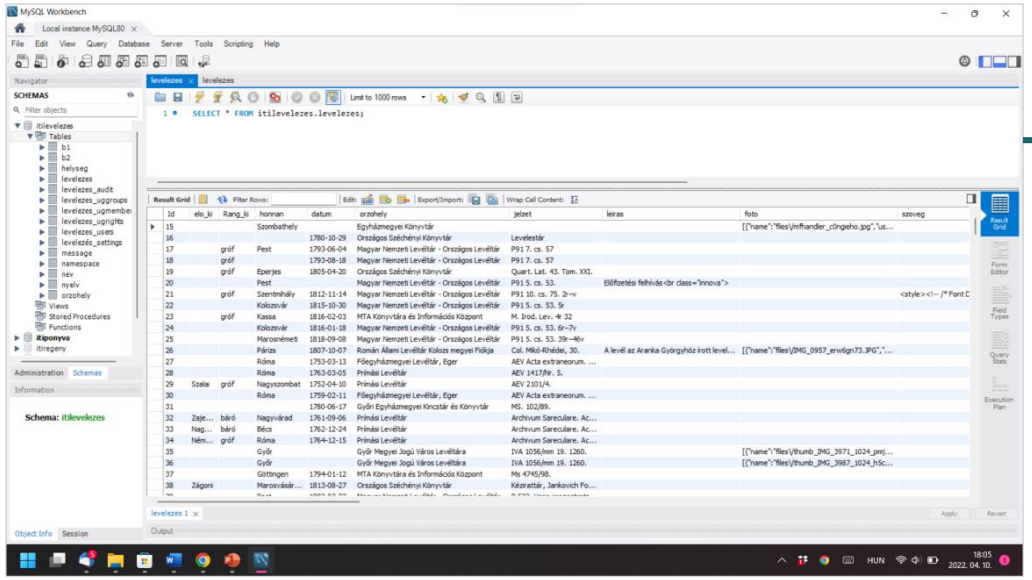
Keys



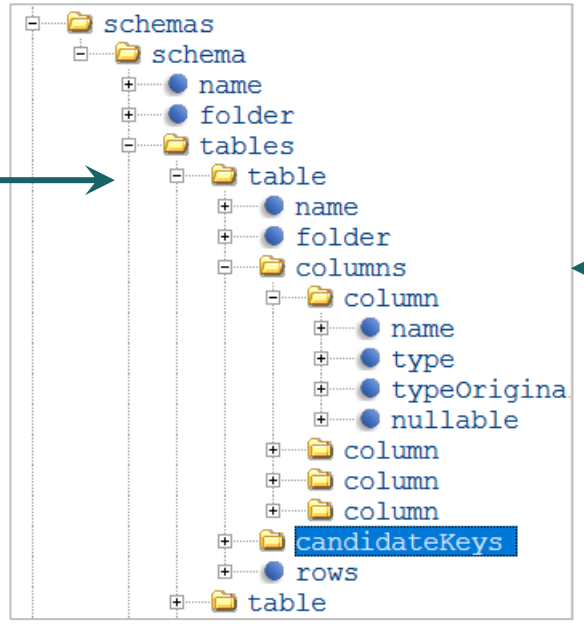
European Commission

SIARD – Mapping

Database



metadata.xml
Type: XML File



SIARD specification

5 Requirements for metadata

- 5.1 Database level metadata
- 5.2 Schema level metadata
- 5.3 Type level metadata
- 5.4 Attribute level metadata
- 5.5 Table level metadata
- 5.6 Column level metadata
- 5.7 Field metadata
- 5.8 Primary key metadata
- 5.9 Foreign key metadata
- 5.10 Reference metadata
- 5.11 Candidate key metadata
- 5.12 Check constraint metadata
- 5.13 Trigger level metadata
- 5.14 View level metadata
- 5.15 Routine level metadata
- 5.16 Parameter metadata
- 5.17 User level metadata
- 5.18 Role level metadata
- 5.19 Privilege level metadata

SIARD – Content

ZIP file \content

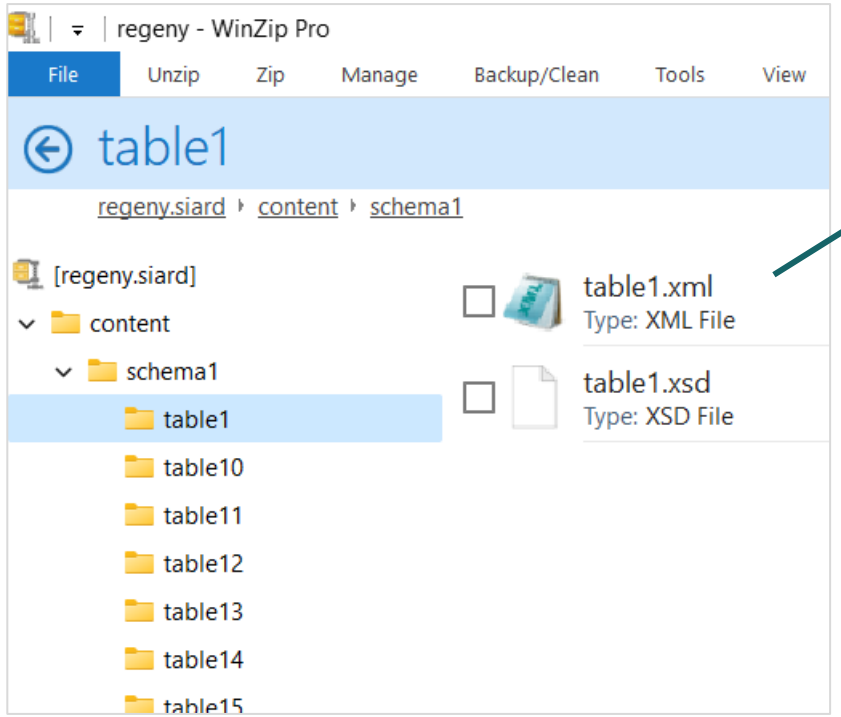
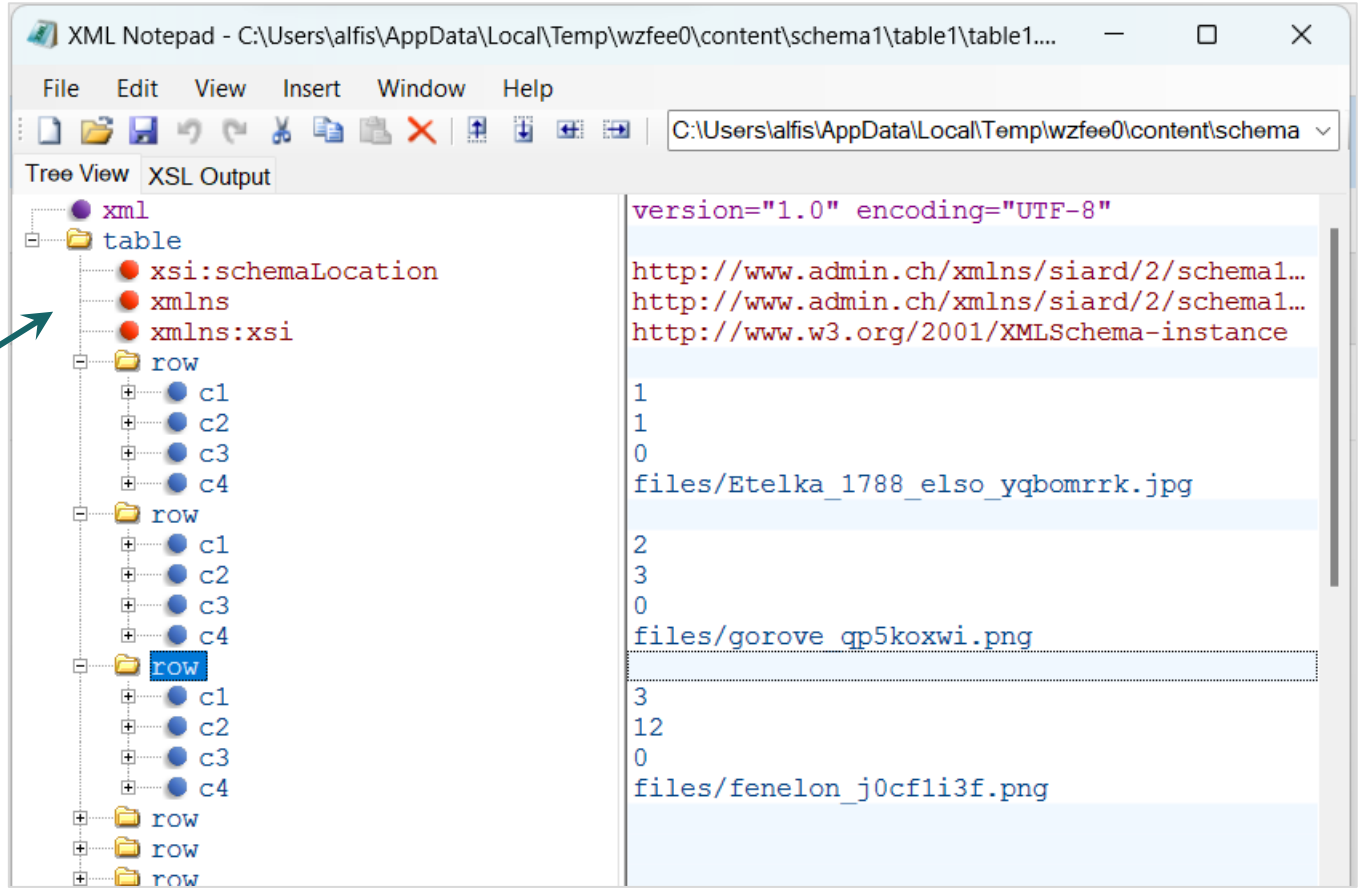
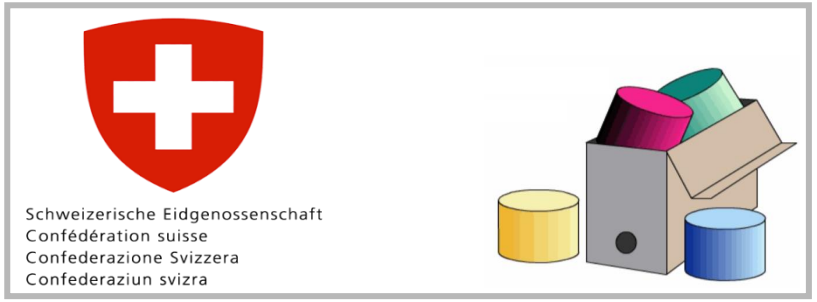


Table1.xml



SIARD



4 Introduction to Database Archival

1. Make sure you know which parts of the database need to be archived. If needed, get in touch with the responsible personnel (e.g. someone from the archive responsible for appraisal).
2. Prepare the database for archival: create a new user on the database system that only has read permissions to objects that need to be archived. If needed, create a copy of your database (or certain tables/parts of it) or create views. The database may not be changed during the archival process; otherwise, extraction with SIARD Suite will fail. Never archive from a live system.
3. Download the database using SIARD Suite.
4. Quality control: check the SIARD file to make sure that everything needed is included, spot check some entries to ensure everything went well.
5. Advanced quality control: load SIARD file into a database system again. Run some defined queries on the original database as well as on the archived one and compare results.
6. Supplement SIARD file with metadata.
7. Define which external documentation needs to be archived together with the SIARD file to ensure comprehensibility of the data (e.g. code tables, system documentation, Entity-Relationship-Diagram, …).

(Manual SIARD-Suite 2.2, Swiss Federal Archives)

European Commission

Pilot at DH-Lab, Budapest

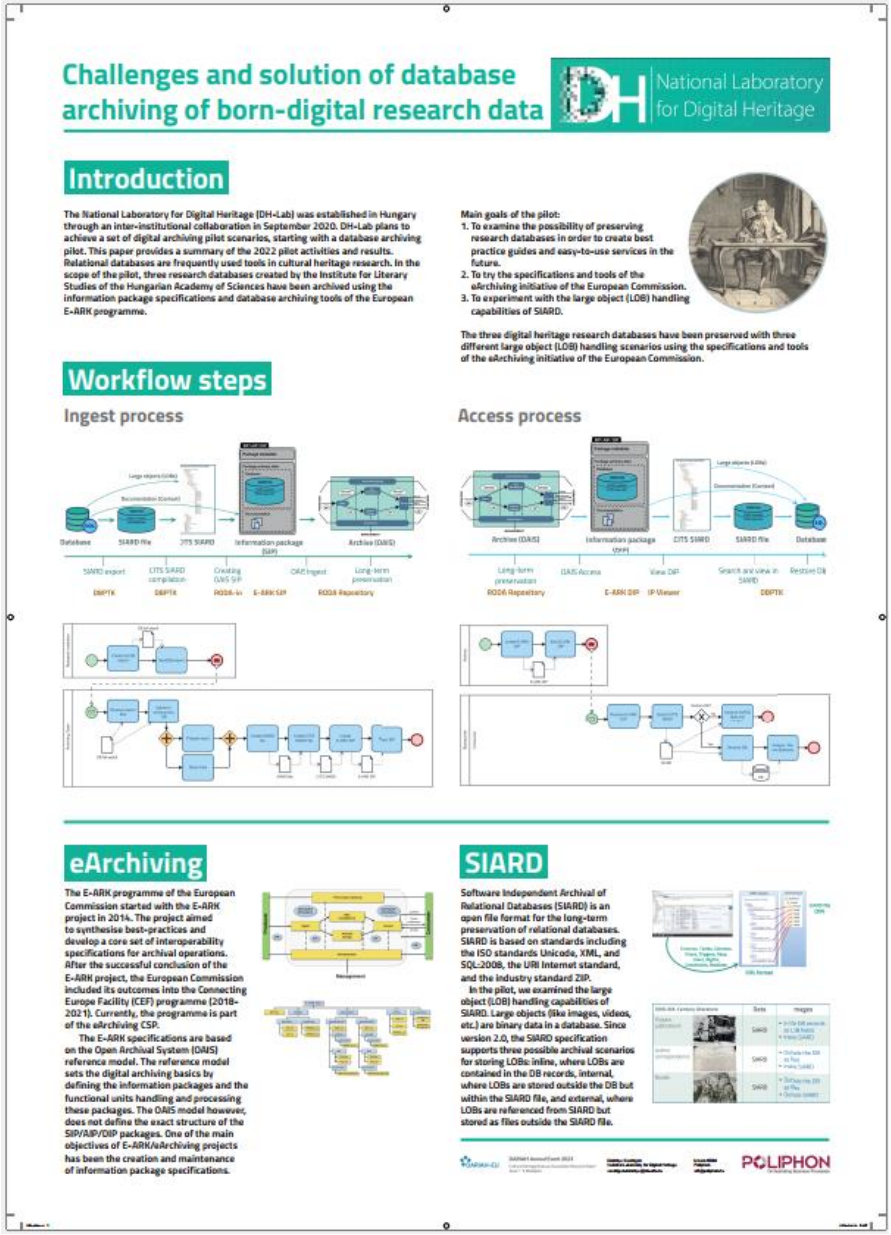
DH-Laboratory, ELTE

- National Laboratory for Digital Heritage (DH-Lab)
- Born Digital project
- Archiving research data

European Commission

DH-Lab pilot

- Research database archiving pilot
- Institute of Literary Studies
- DARIAH 2023 Annual Event, at Budapest
- White paper (coming soon)



European Commission

DH-Lab pilot

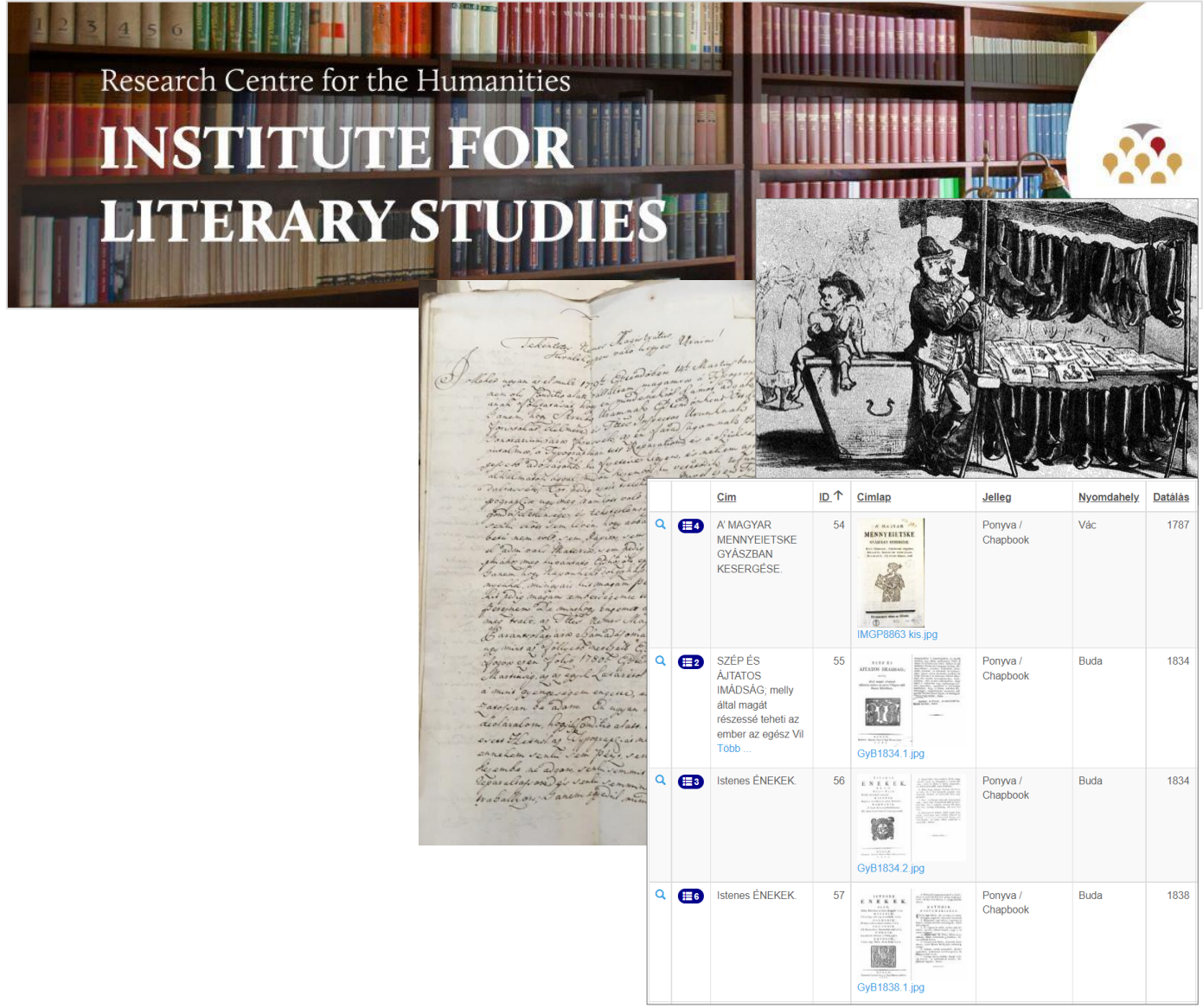
		Cím	ID	Címlap	Jelleg	Nyomdahely	Datálás
	4	A' MAGYAR MENNYEIETSKE GYÁSZBAN KESERGÉSE.	54	IMGP8863 kis.jpg	Ponyva / Chapbook	Vác	1787
	2	SZÉP ÉS ÁJTATOS IMÁDSÁG; melly által magát részessé teheti az ember az egész Vil Több ...	55	GyB1834.1.jpg	Ponyva / Chapbook	Buda	1834
	3	Istenes ÉNEKEK.	56	GyB1834.2.jpg	Ponyva / Chapbook	Buda	1834
	6	Istenes ÉNEKEK.	57	GyB1838.1.jpg	Ponyva / Chapbook	Buda	1838

Literature in Western Hungary in 17–19th century

- Novels in Hungary (1730–1836)
- Popular prints in 17th–19th century Hungary
- Literary and scholarly correspondence

European Commission

DH-Lab pilot

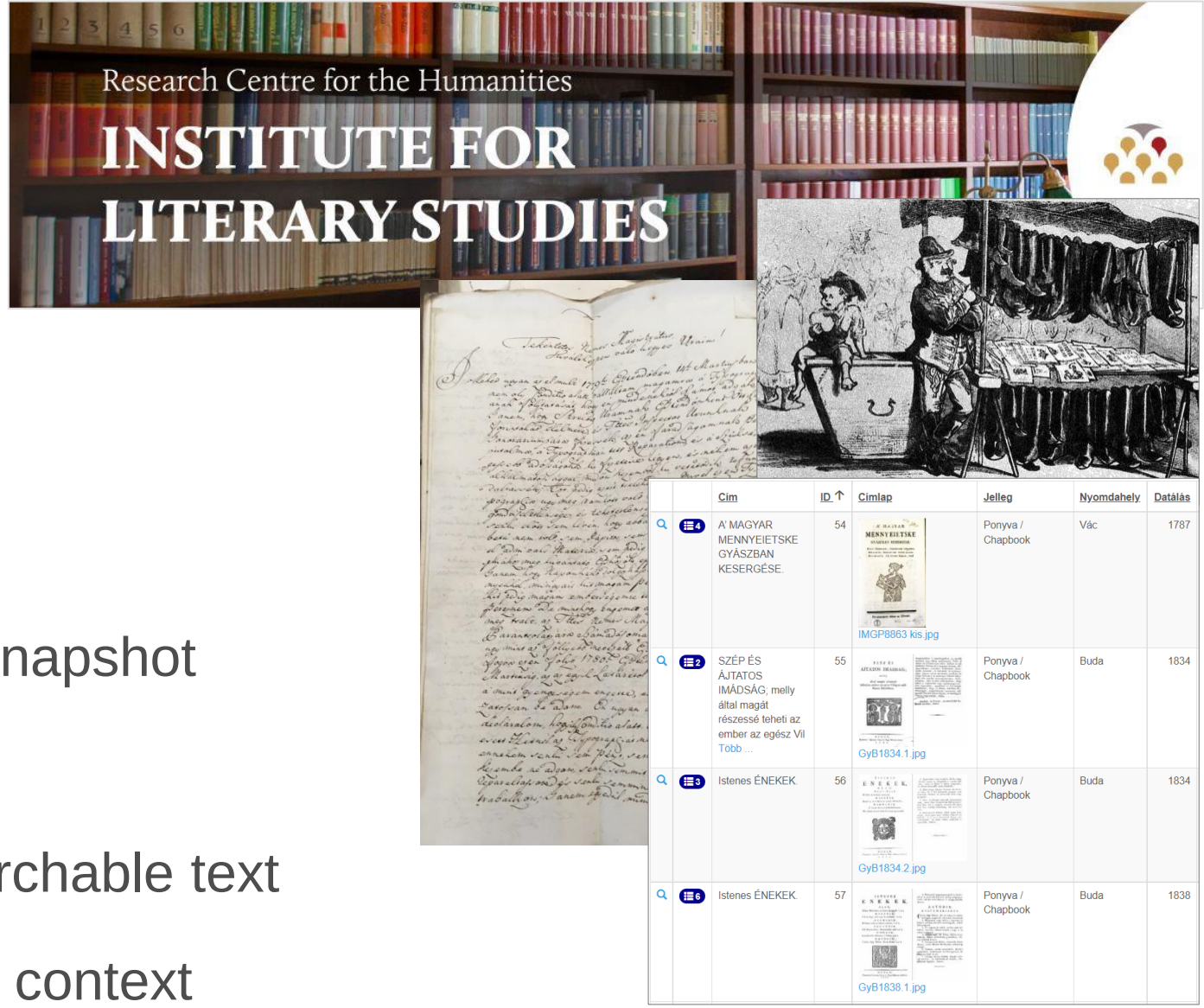


Large Objects in the DB

- Images
- Texts
- Formatted HTML pages

European Commission

DH-Lab pilot

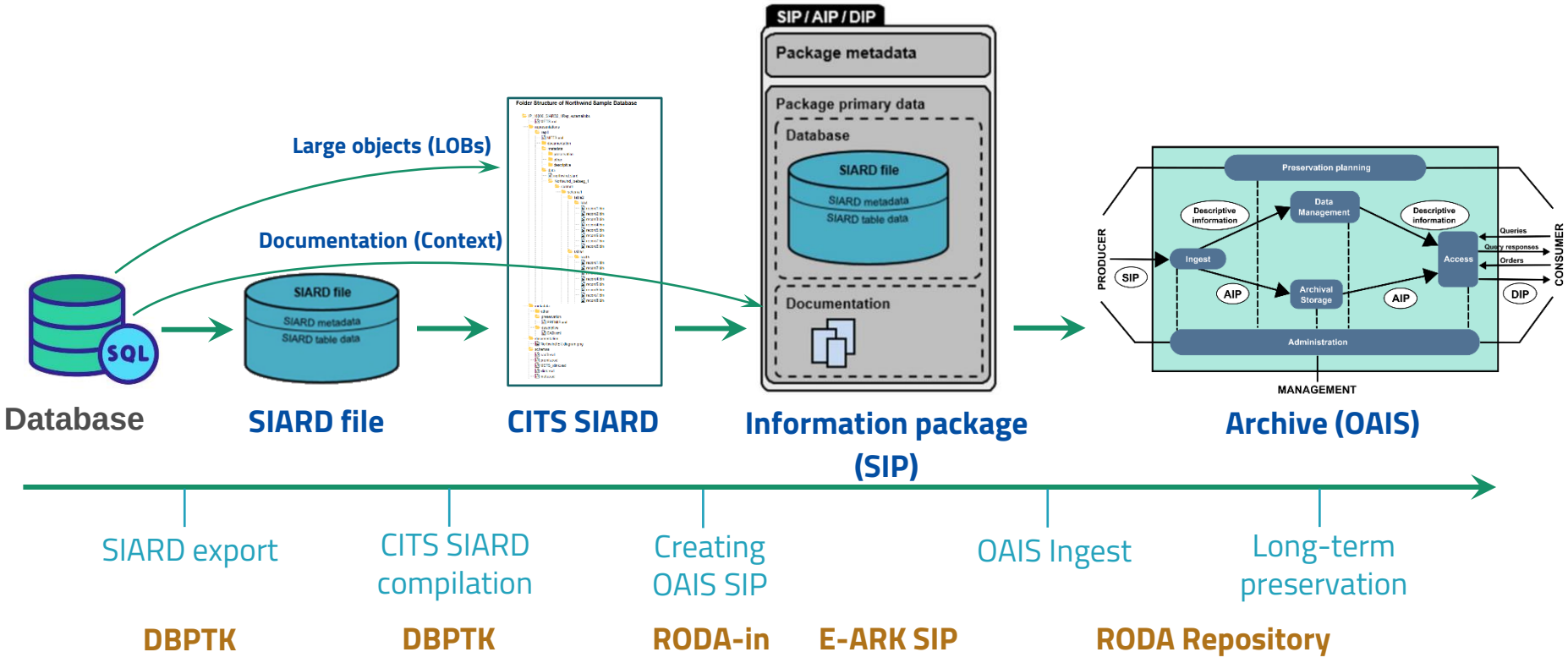


Data preparations

- Active research databases → Snapshot
- → Original SIP
- Formatted HTML pages → Searchable text
- PHP user surface → Images as context
- Considered the SIP for the Pilot

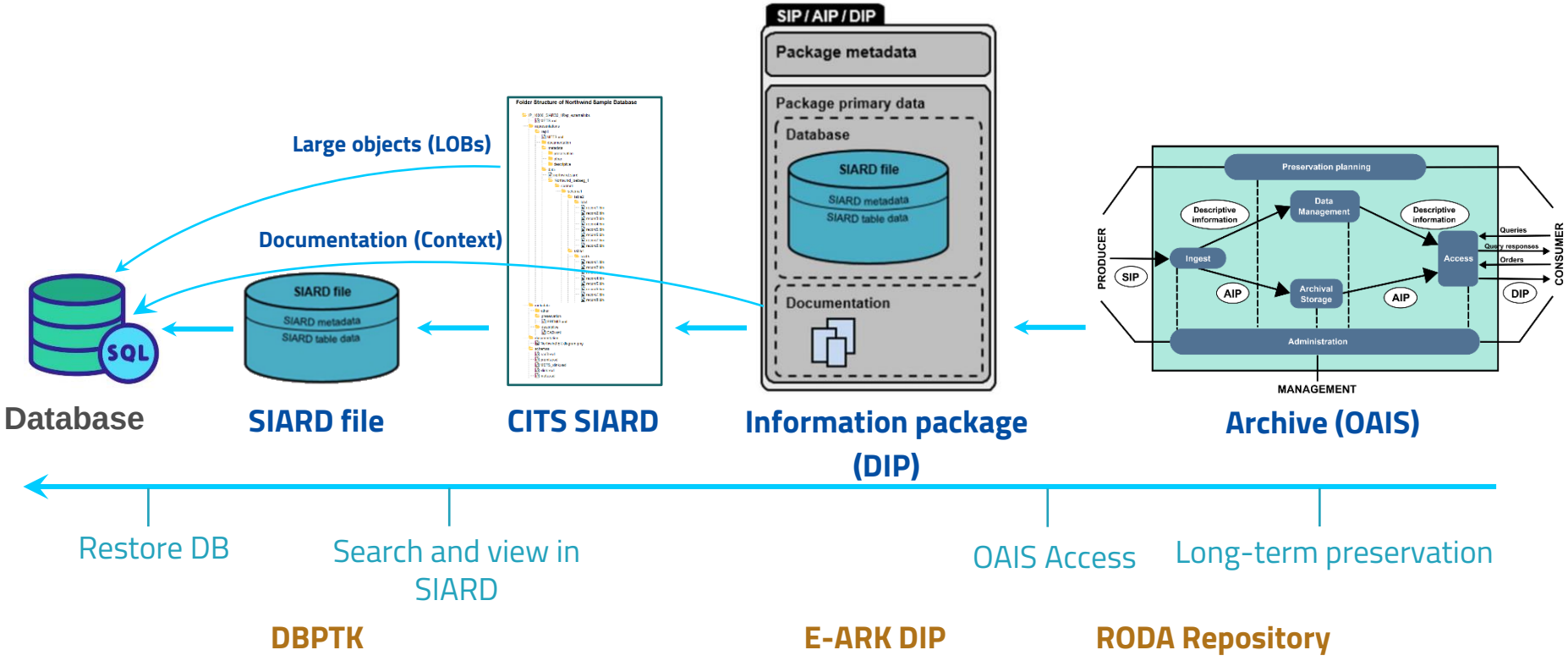
European Commission

The Workflow



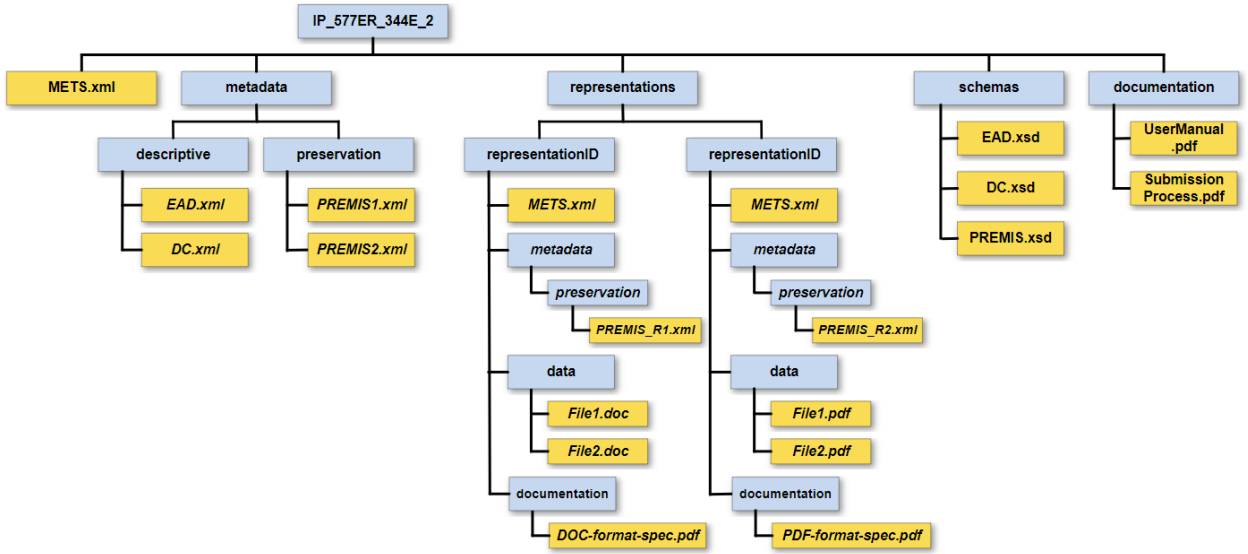
European Commission

The Workflow



Used eArchiving components

- SIARD 2.1
- E-ARK SIP/DIP
- CITS SIARD
- Database Preservation Toolkit
- RODA-in
- RODA Repository
- IP Viewer

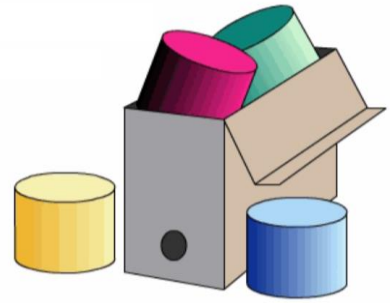


European Commission

Large Object (LOB) handling with SIARD

SIARD version history

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra



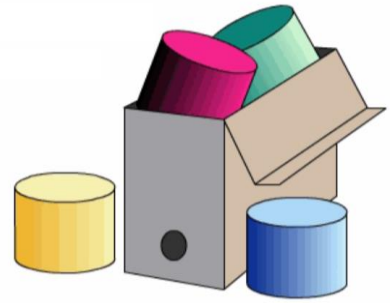
E-ARK

SIARD v1.0	Original version by the Swiss Federal Archives. Replaced by version 2.1.1: Version 1.0 has been replaced by a more recent and current version. Its use remains possible, but it is recommended to use the present version.
SIARD v2.0	Abrogated: Version 2.0 has been displaced because of errors and ambiguities that might have led to practical problems in the long term. This old issue must no longer be used, and either the present version or version 1.0 must be used.
SIARD v2.1	Replaced by version 2.1.1: Version 2.1 corrects errors and ambiguities in version 2.1. It has been developed by the eCH Fachgruppe on Digital Preservation but is no official standards by the eCH.
SIARD v2.1.1	Replaced by version 2.2: Version 2.1.1 documents the current state of the SIARD file format. It is identical to version 2.1 save for a few precisions in the wording.
SIARD v2.2	Version 2.2 adds support for files outside the database according to part 9 of SQL:2008 (ISO/IEC 9075-9:2008 – SQL/MED) as well as scalability supporting large objects stored outside the SIARD file. Apart from this it is identical to version 2.1.1. It has been developed by the DILCIS Board during the E-ARK3 project.

European Commission

LOBs in SIARD

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra



E-ARK

Large objects in general are binary data in a database. Binary data is mostly referred to as a binary large object (BLOB) and large character-based data are named character large objects (CLOB).

From version 2.0 SIARD specification supports three possible archival scenarios for storing LOBs:

1.) Inline - LOBs are contained in the DB records.

2.) Internal - LOBs are stored outside the DB records but within the SIARD file.

3.) External - LOBs are referenced from SIARD but stored as files outside the SIARD file.

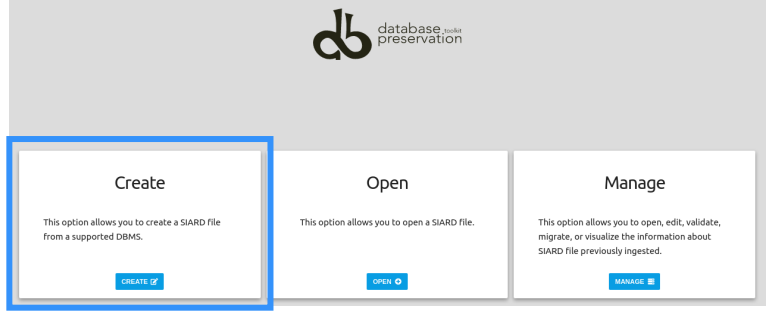
European Commission

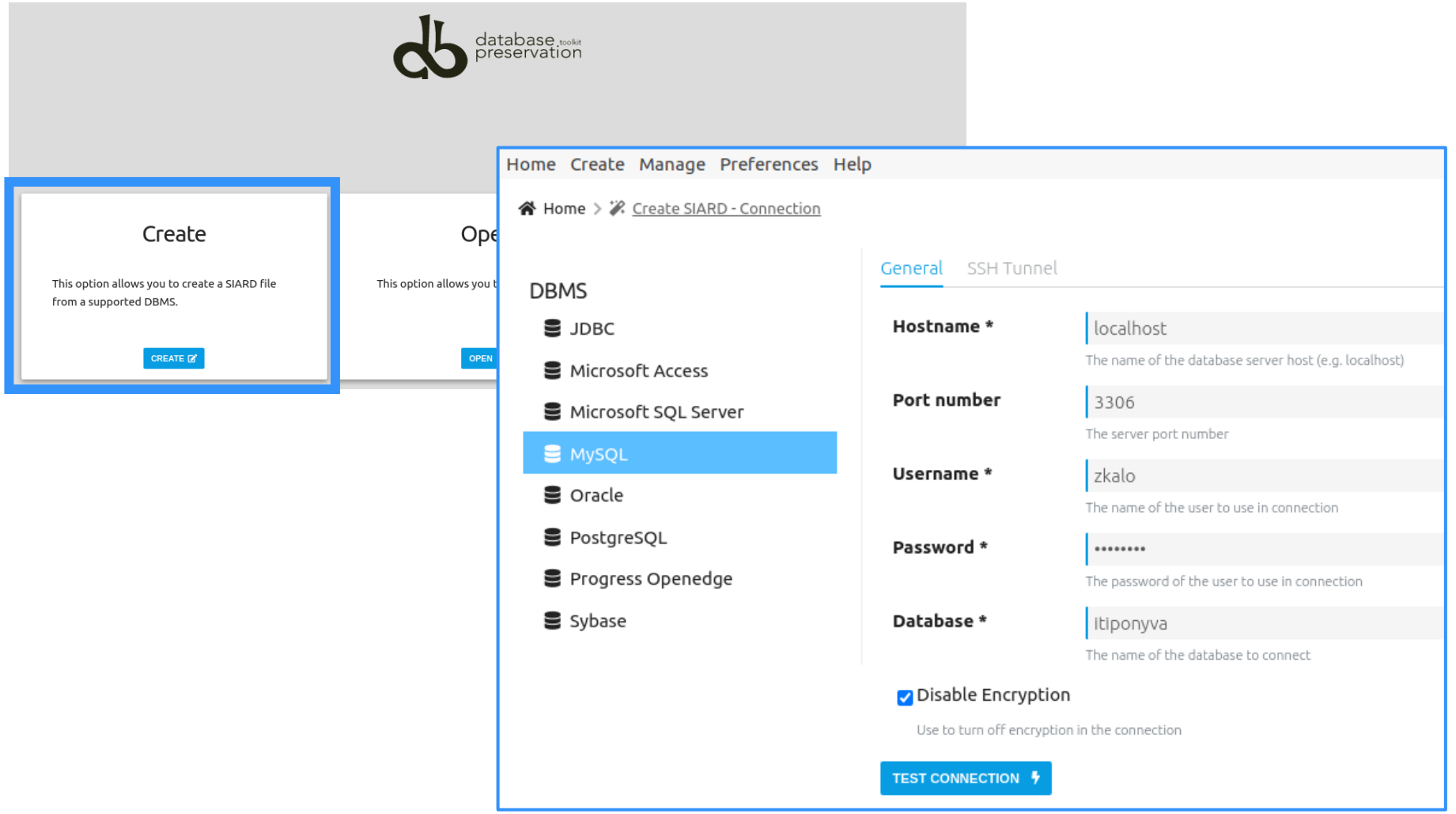
DH-Lab archiving scenarios

XVIII-XIX. Century Literature		Data	Images
Popular publications		SIARD	• Inline • In the DB records as LOB fields
Novels		SIARD	• Internal • Outside the DB as files
Author correspondence		SIARD	• External • Outside the DB as files

European Commission

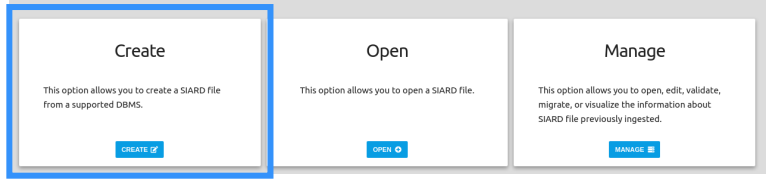
DBPTK – Create SIARD file

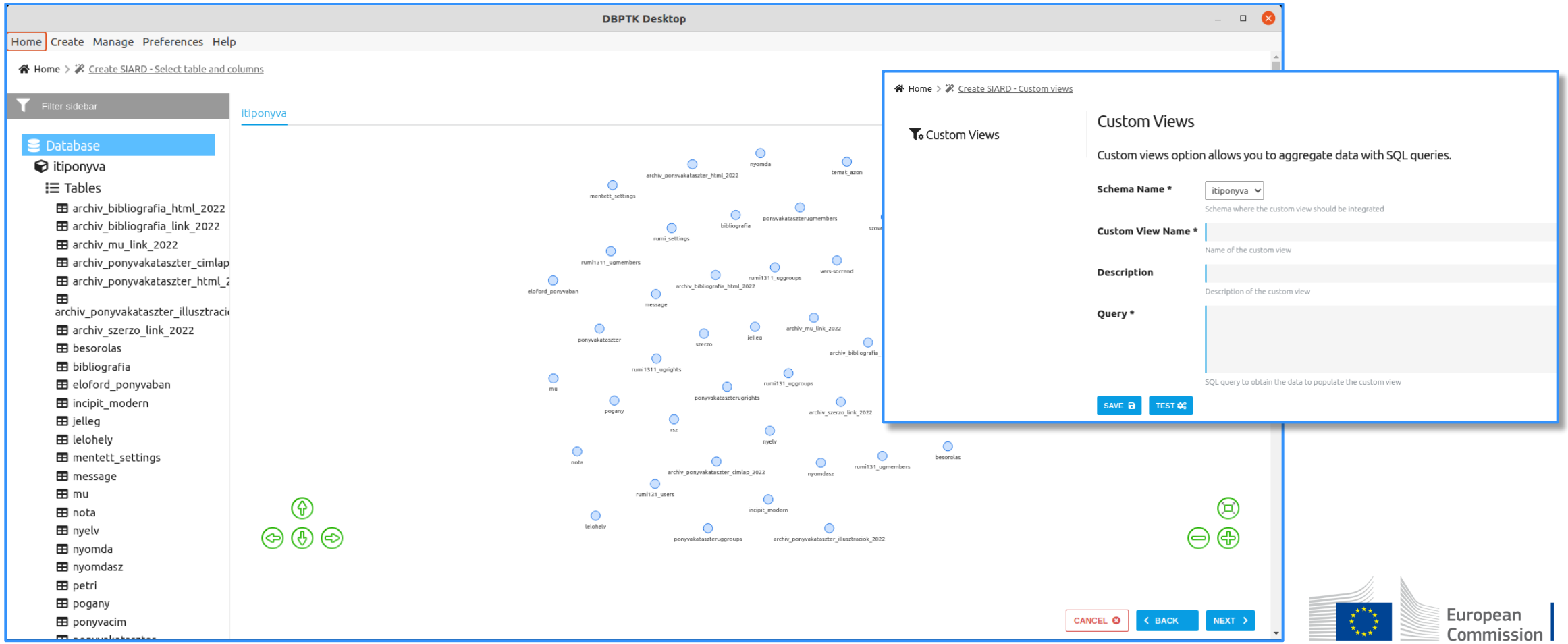




European Commission

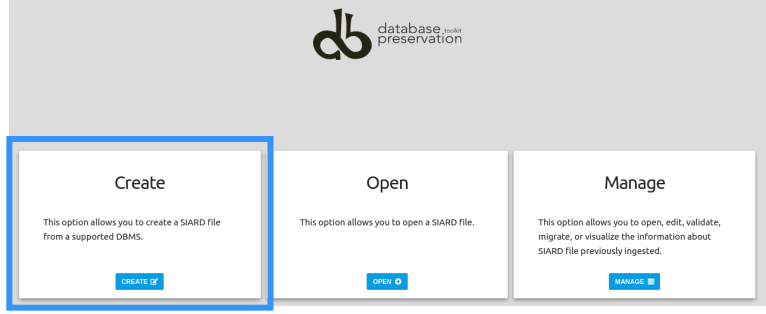
DBPTK – Create SIARD file





European Commission

DBPTK – Create SIARD file



Home > Create SIARD - SIARD export options

SIARD Export Options

Resumes the different options available to create a SIARD file.

SIARD Version SIARD 2

Destination Folder * BROWSE /home/zkalo/BornDigital/Pilot/siard/test.siard

Path to SIARD file

☐ Compress

Compress the SIARD file with deflate method

☐ Format XML

Write human-readable XML

☑ Save LOBs outside SIARD file

External LOBs per Folder 1000 Files

The maximum number of files present in an external LOB folder. Default: 1000 files

External LOBs Folder Size 0 Megabytes

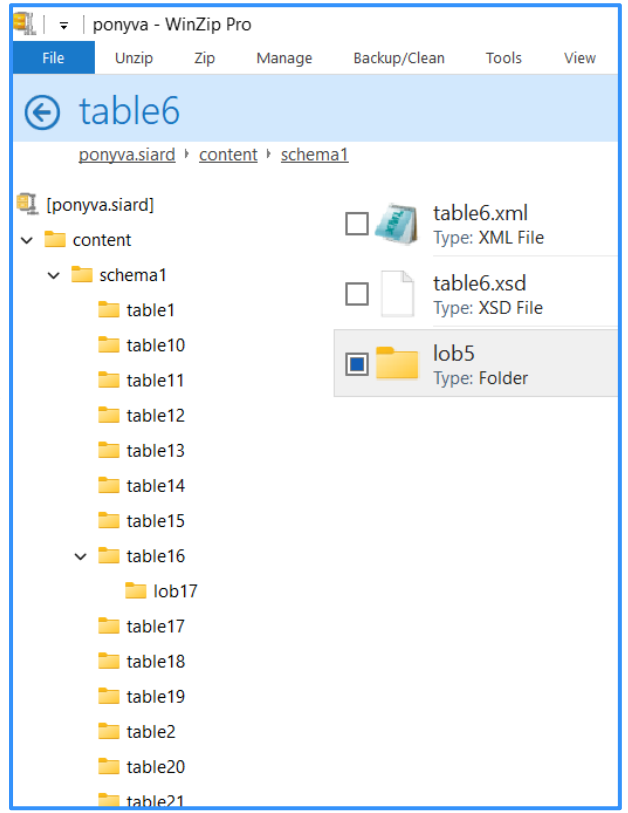
Divide LOBs across multiple external folders with (approximately) the specified maximum size (in Megabytes). Default: do not divide

Digest SHA-256

Define the type of integrity information. Supported types are: MD5, SHA-1 and SHA-256.

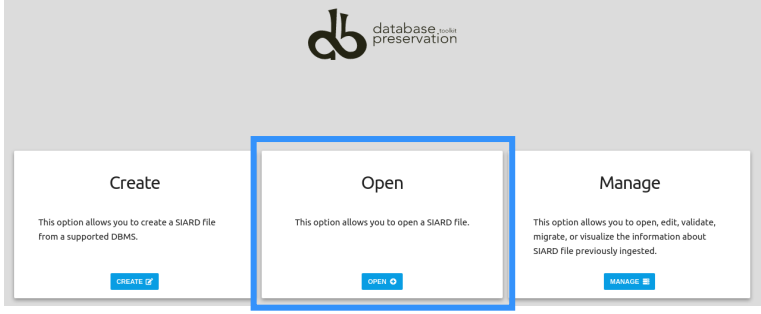
Font Case Lower case

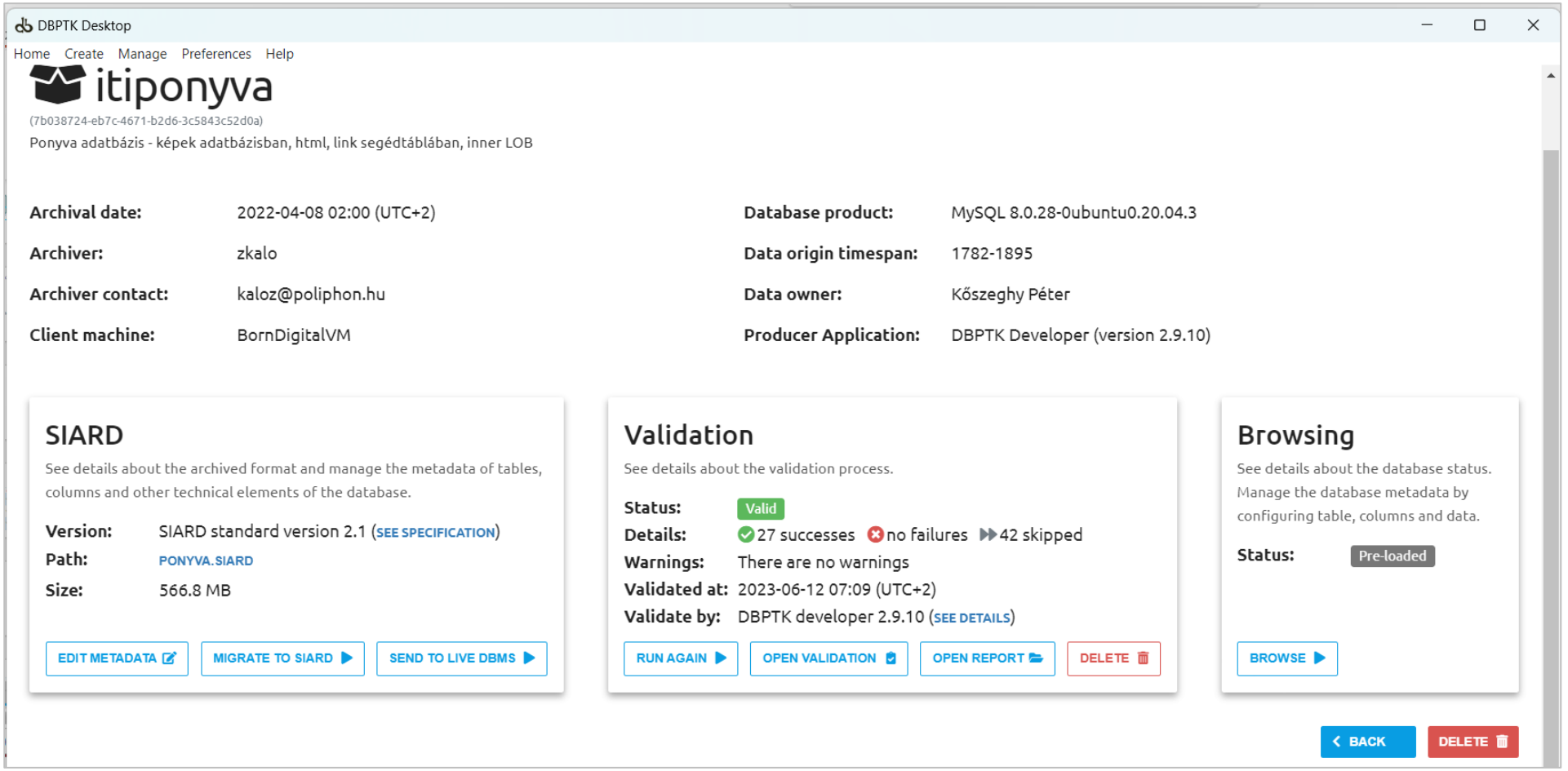
Define the type of font case for the message digest. Supported font case are: upper case and lower case.



European Commission

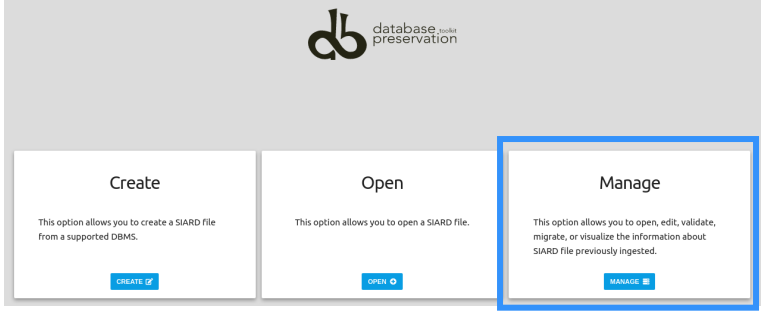
DBPTK – View SIARD file

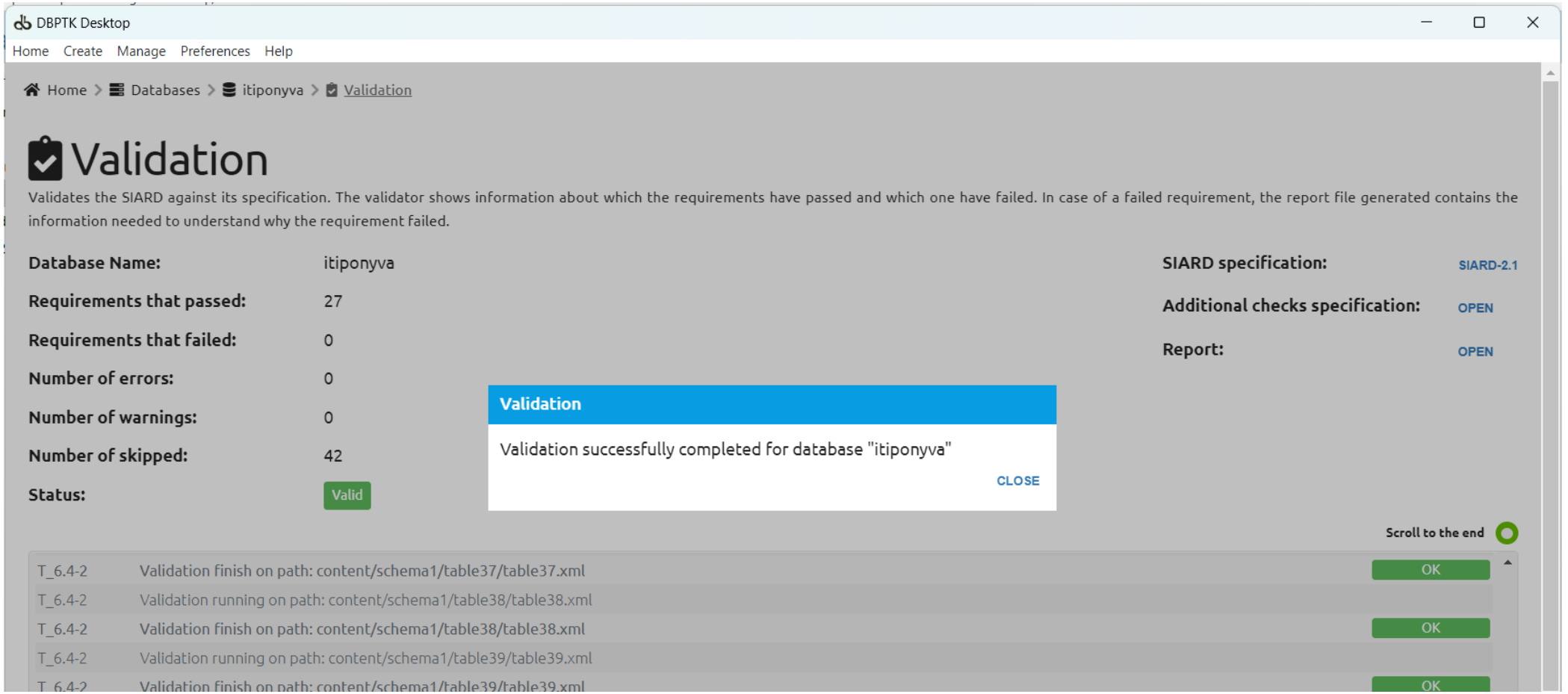




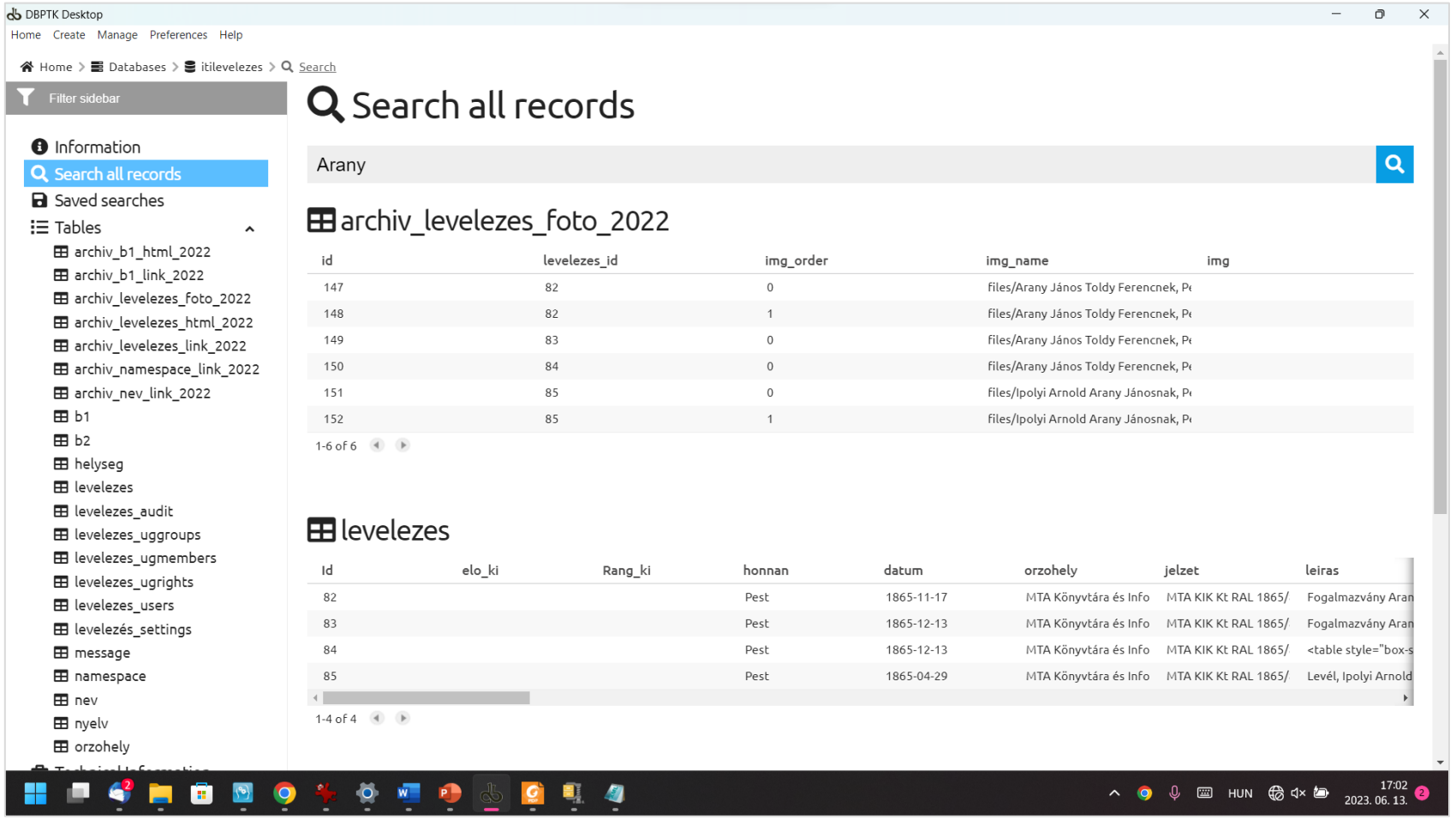
European Commission

DBPTK – Validate SIARD file



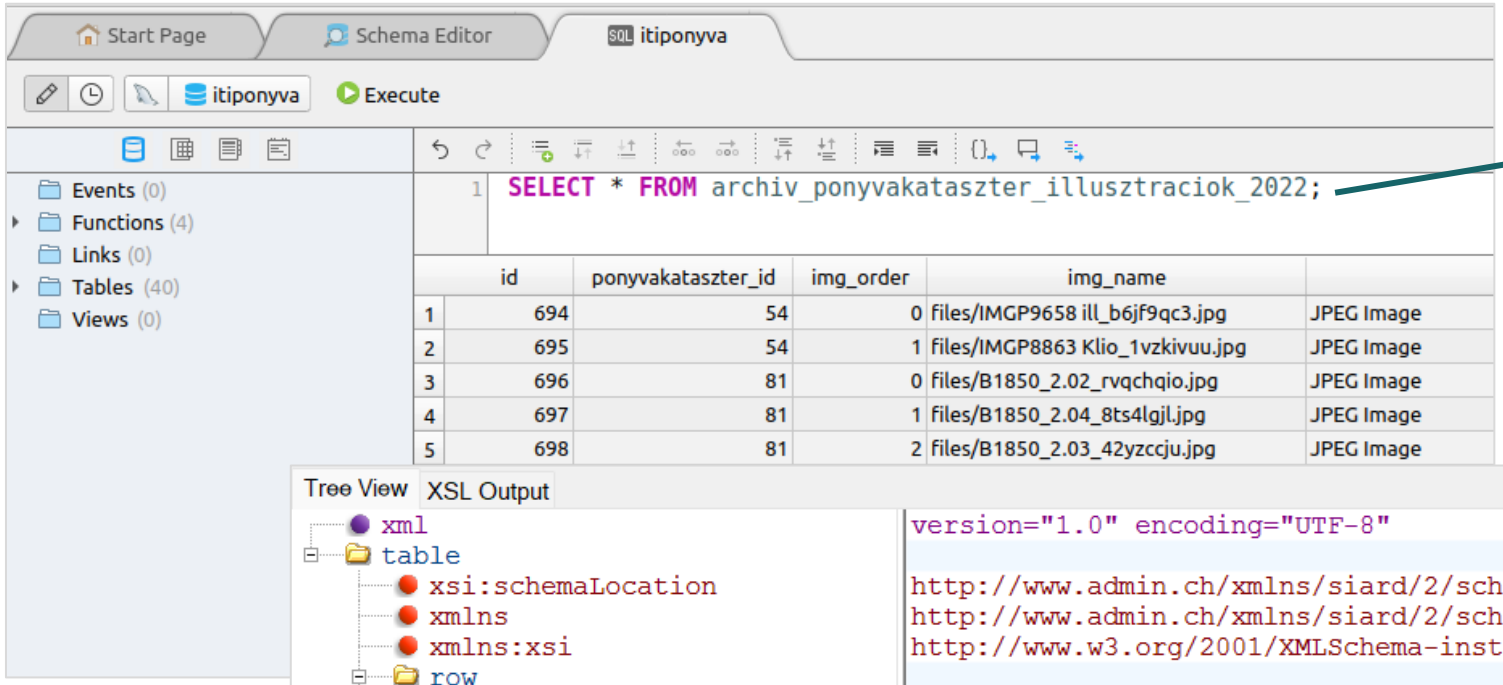


DBPTK – Search in SIARD file

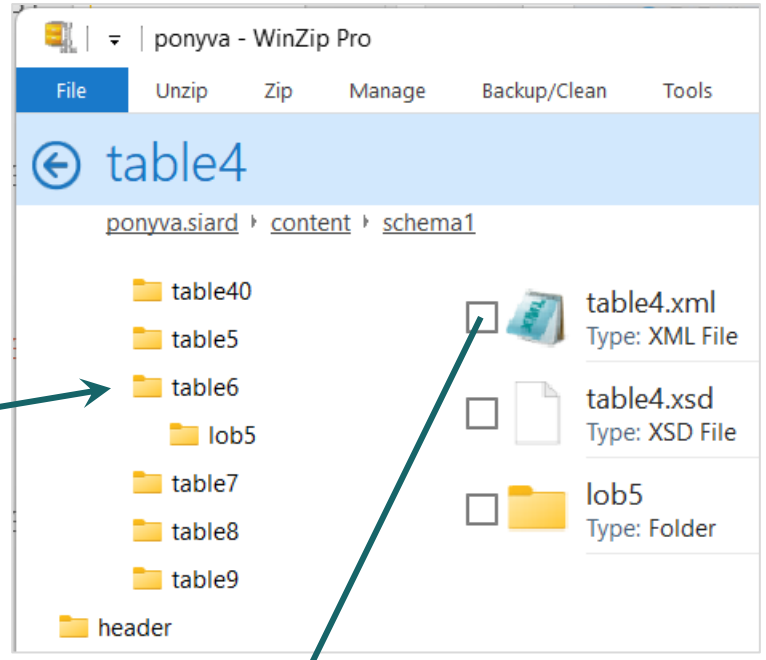


European Commission

Large Objects - Internal

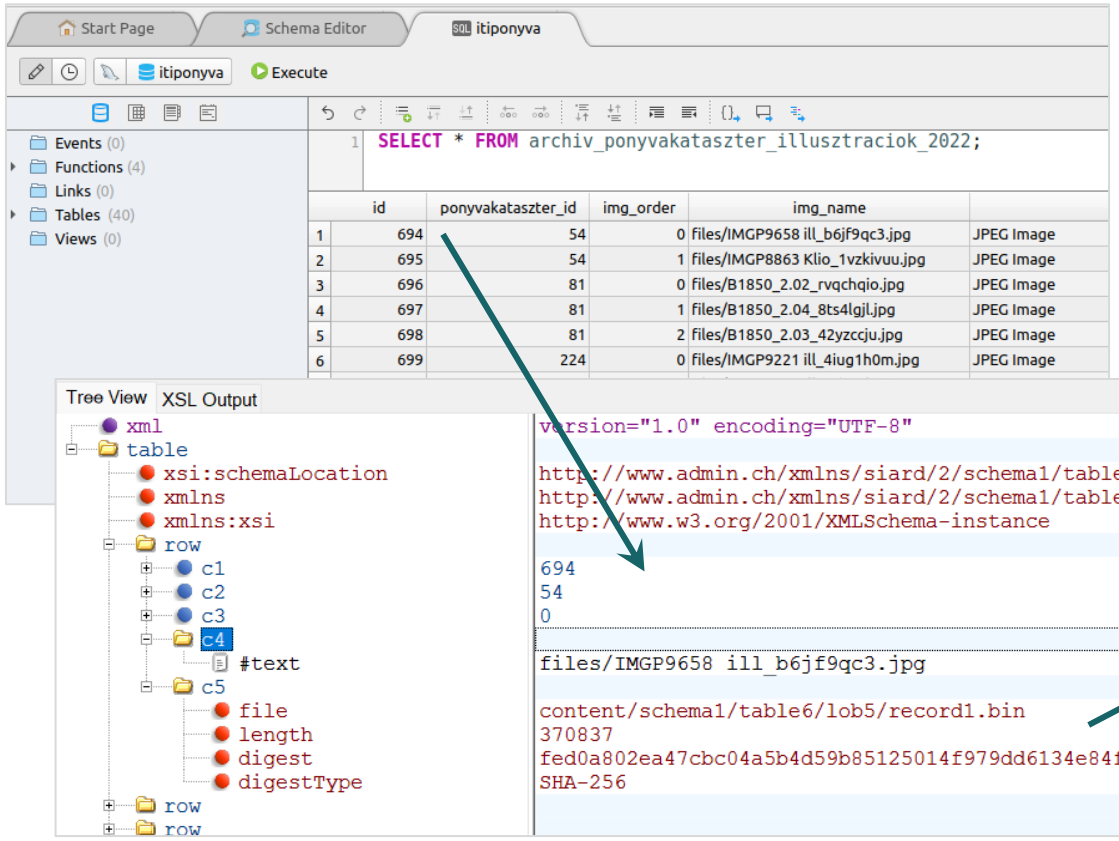


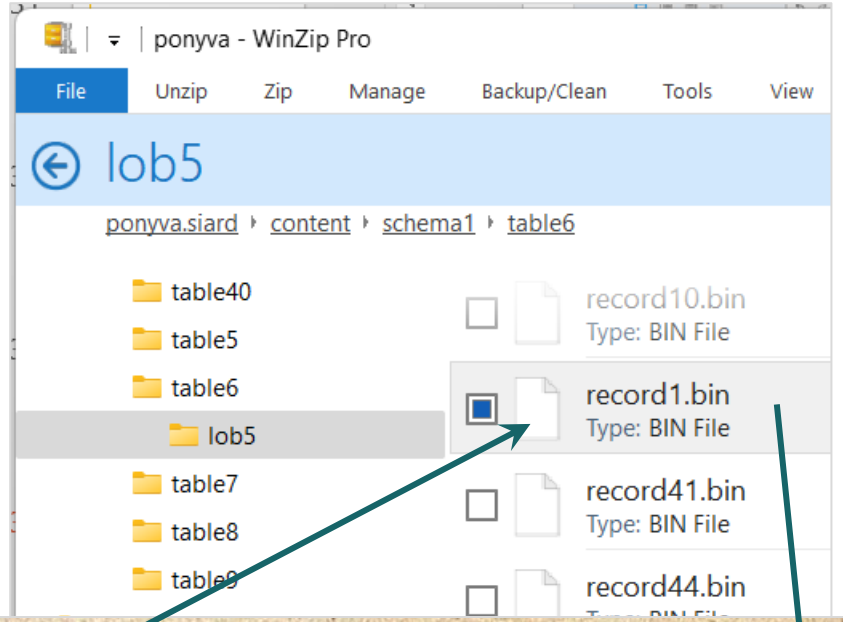
	id	ponyvakataszter_id	img_order	img_name	
1	694	54	0	files/IMGP9658 ill_b6jf9qc3.jpg	JPEG Image
2	695	54	1	files/IMGP8863 Klio_1vzkivuu.jpg	JPEG Image
3	696	81	0	files/B1850_2.02_rvqchqio.jpg	JPEG Image
4	697	81	1	files/B1850_2.04_8ts4lgjl.jpg	JPEG Image
5	698	81	2	files/B1850_2.03_42yzccju.jpg	JPEG Image



European Commission

Large Objects - Internal





European Commission

Large Objects – Internal

Start Page | Schema Editor | itiponyva

itiponyva | Execute

- Events (0)
- Functions (4)
- Links (0)
- Tables (40)
- Views (0)

```
1  SELECT * FROM archiv_ponyvakataszter_illusztraciok_2022;
```

	id	ponyvakataszter_id	img_order	img_name	
1	694	54	0	files/IMGP9658 ill_b6jf9qc3.jpg	JPEG Image
2	695	54	1	files/IMGP8863 Klio_1vzkivuu.jpg	JPEG Image
3	696	81	0	files/B1850_2.02_rvqchqio.jpg	JPEG Image
4	697	81	1	files/B1850_2.04_8ts4lgjl.jpg	JPEG Image
5	698	81	2	files/B1850_2.03_42yzccju.jpg	JPEG Image
6	699	224	0	files/IMGP9221 ill_4iug1h0m.jpg	JPEG Image
7	700	224	1	files/IMGP9222 ill_r8huolwf.jpg	JPEG Image
8	701	224	2	files/IMGP9227 ill_0kjjyiix.jpg	JPEG Image
9	702	225	0	files/IMGP8830 ill_r7b4ndxe.jpg	JPEG Image
10	703				
11	704				

DBPTK Desktop

Home | Create | Manage | Preferences | Help

Home > Databases > itiponyva > archiv_ponyvakataszter_illusztraciok_2022

Filter sidebar

- Information
- Search all records
- Saved searches
- Tables
  - archiv_bibliografia_html_2022
  - archiv_bibliografia_link_2022
  - archiv_mu_link_2022
  - archiv_ponyvakataszter_cimlap_2
  - archiv_ponyvakataszter_html_202
  - archiv_ponyvakataszter_illusztraciok
  - archiv_szerzo_link_2022
  - besorolas
  - bibliografia
  - eloford_ponyvaban
  - incipit_modern

archiv_ponyvakataszter_illusztraciok_2022

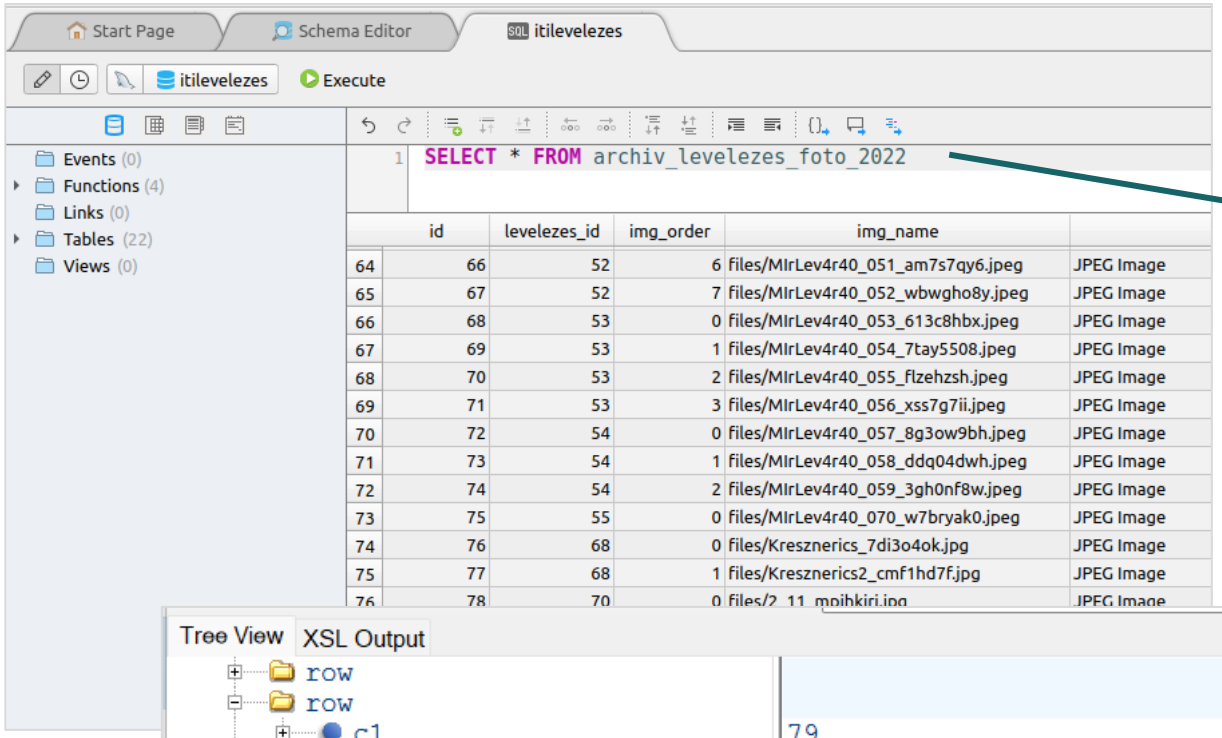
MANAGE ARCHIV_PONYVAKATASZTER_ILLUSZTRACIOK_2022 OPTIONS

Search... advanced

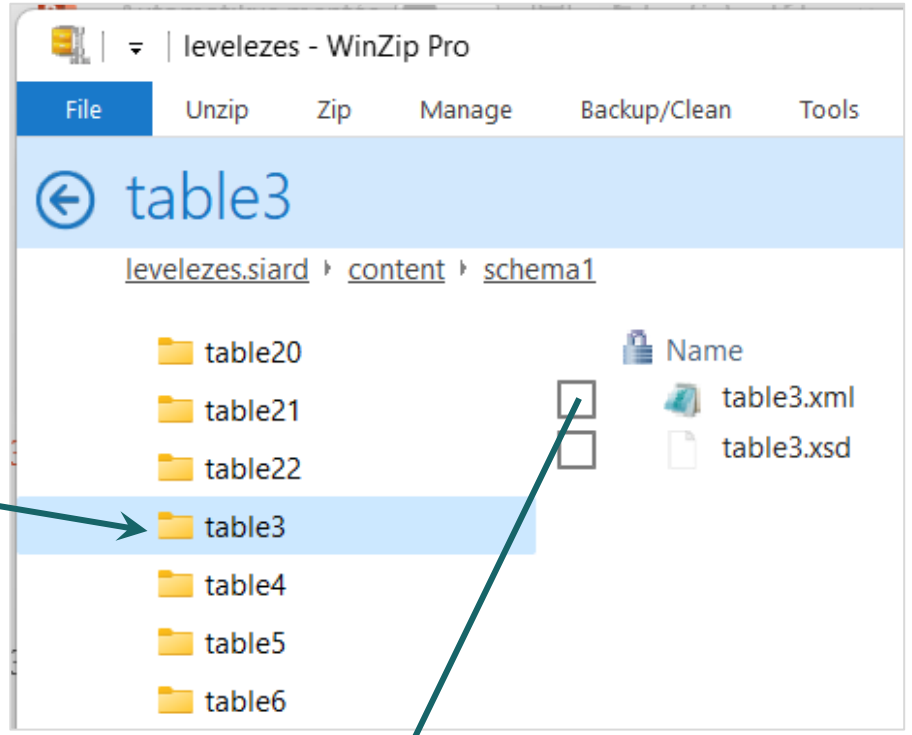
id	ponyvakataszter_id	img_order	img_name	img
694	54	0	files/IMGP9658 ill_b6jf9q	Download
695	54	1	files/IMGP8863 Klio_1vzk	Download
696	81	0	files/B1850_2.02_rvqchqi	Download
697	81	1	files/B1850_2.04_8ts4lgjl	Download
698	81	2	files/B1850_2.03_42yzccj	Download
699	224	0	files/IMGP9221 ill_4iug1h	Download
700	224	1	files/IMGP9222 ill_r8huol	Download
701	224	2	files/IMGP9227 ill_0kjjyii	Download

European Commission

Large Objects - External

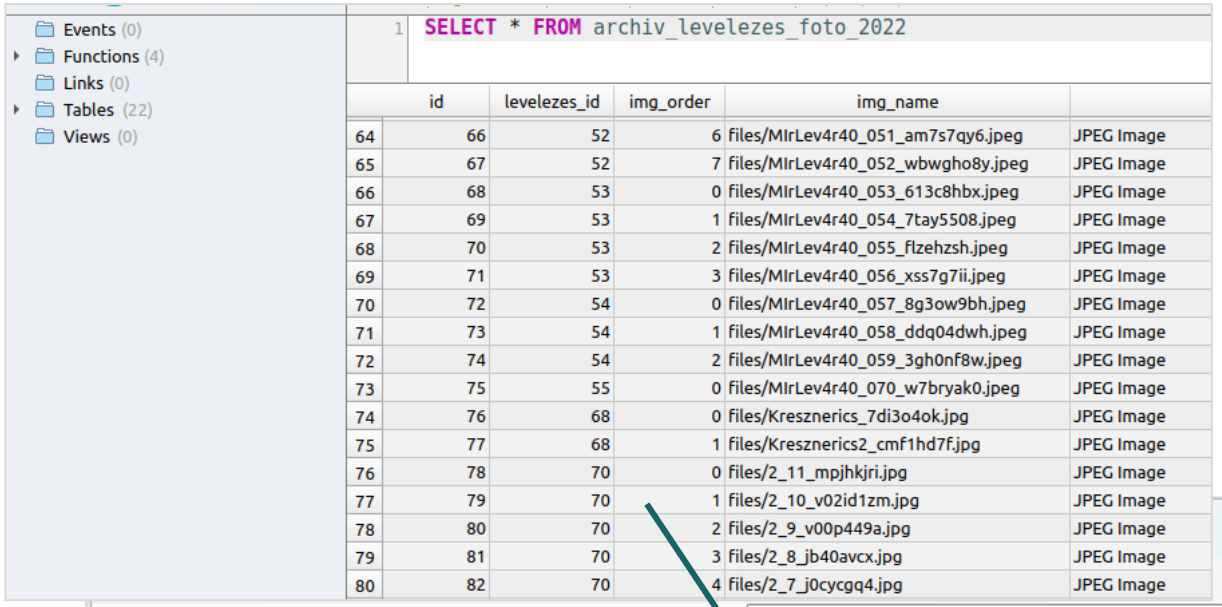


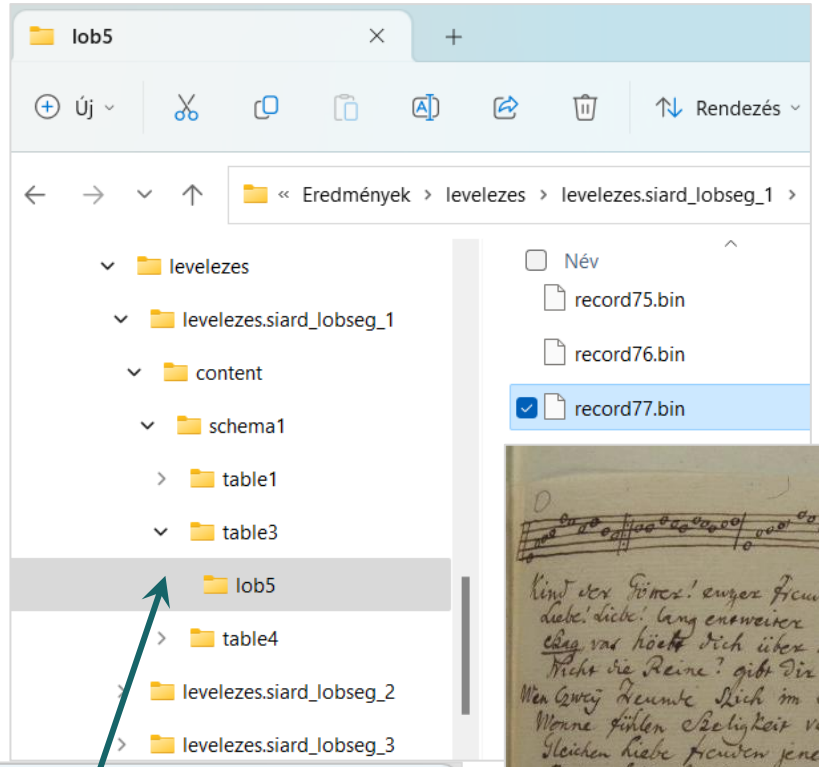
	id	levelezes_id	img_order	img_name	
64	66	52	6	files/MIrLev4r40_051_am7s7qy6.jpeg	JPEG Image
65	67	52	7	files/MIrLev4r40_052_wbwgho8y.jpeg	JPEG Image
66	68	53	0	files/MIrLev4r40_053_613c8hbx.jpeg	JPEG Image
67	69	53	1	files/MIrLev4r40_054_7tay5508.jpeg	JPEG Image
68	70	53	2	files/MIrLev4r40_055_flzehzsh.jpeg	JPEG Image
69	71	53	3	files/MIrLev4r40_056_xss7g7ii.jpeg	JPEG Image
70	72	54	0	files/MIrLev4r40_057_8g3ow9bh.jpeg	JPEG Image
71	73	54	1	files/MIrLev4r40_058_ddq04dwh.jpeg	JPEG Image
72	74	54	2	files/MIrLev4r40_059_3gh0nf8w.jpeg	JPEG Image
73	75	55	0	files/MIrLev4r40_070_w7bryak0.jpeg	JPEG Image
74	76	68	0	files/Kresznerics_7di3o4ok.jpg	JPEG Image
75	77	68	1	files/Kresznerics2_cmf1hd7f.jpg	JPEG Image
76	78	70	0	files/2_11_mojhkiri.jpg	JPEG Image

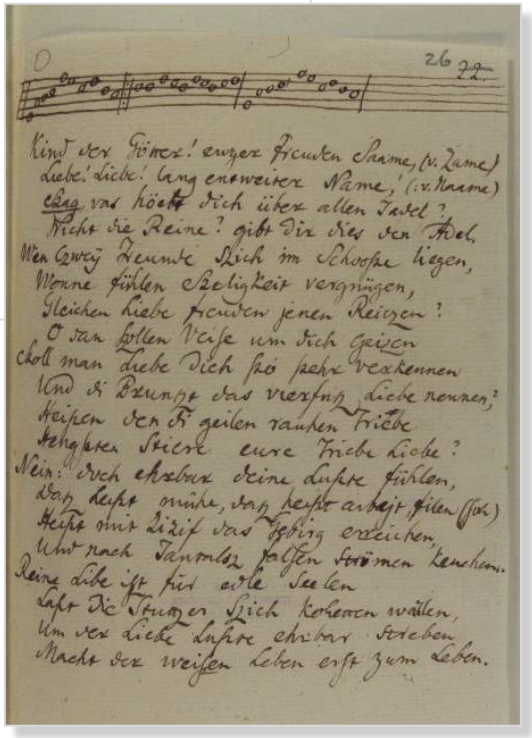


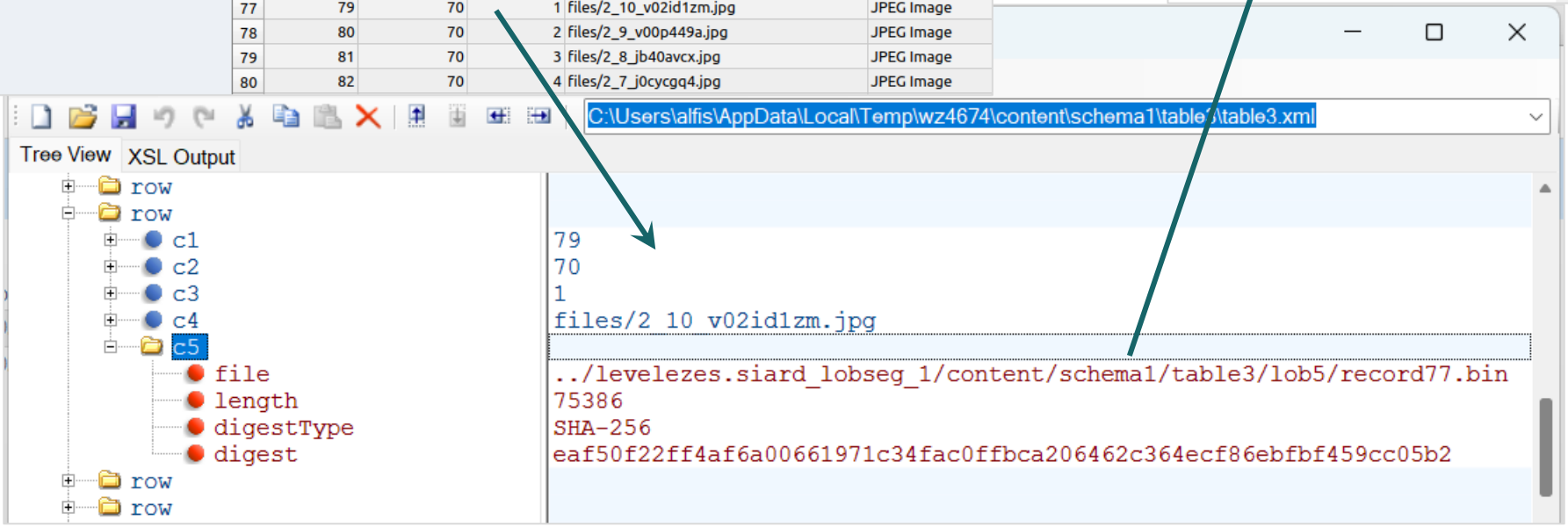
European Commission

Large Objects - External



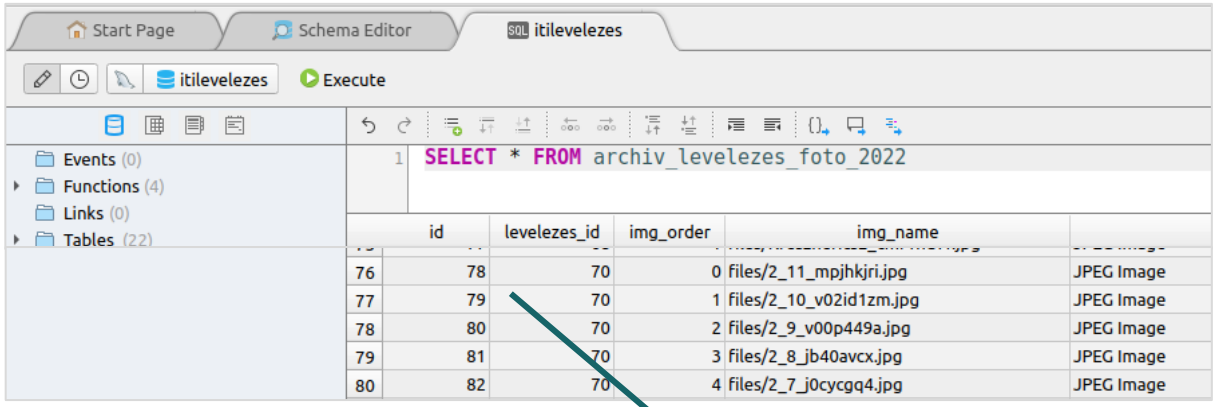


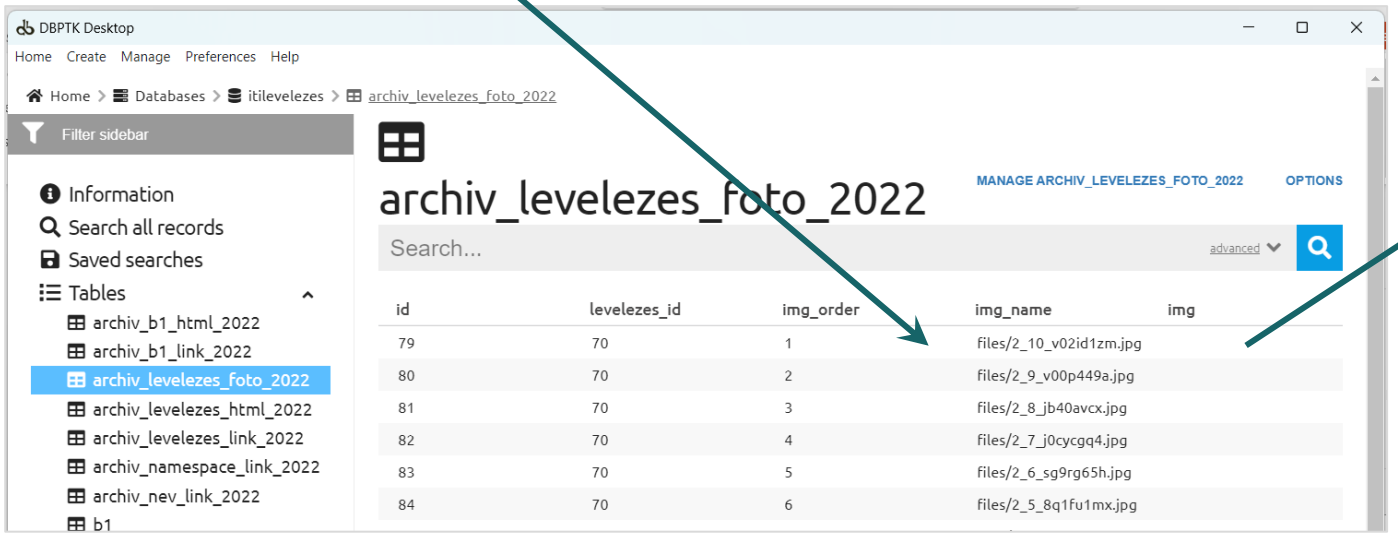


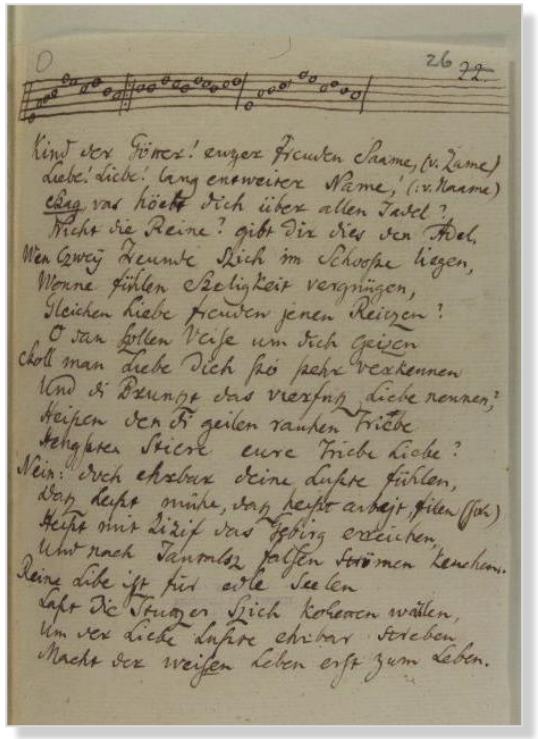


European Commission

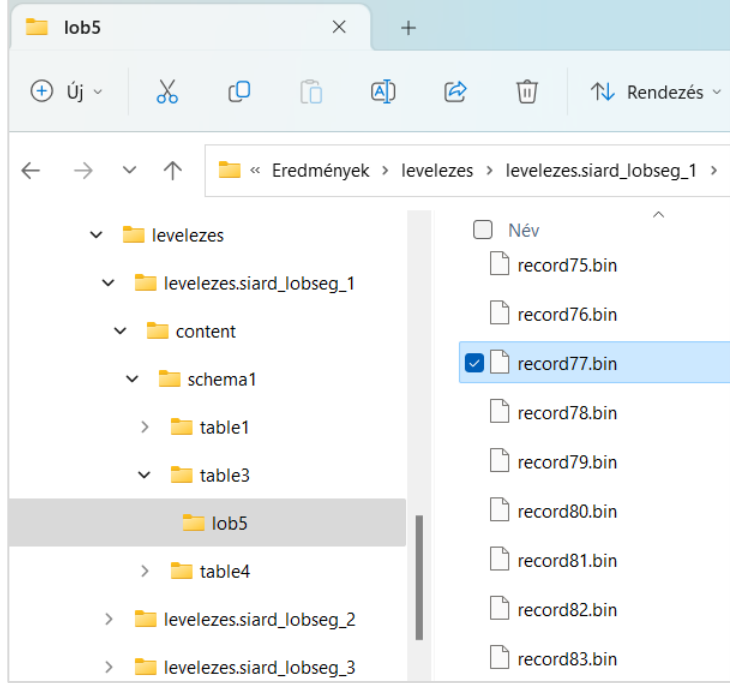
Large Objects – External







External file structure



- SIARD 2.1 does not define the file structure for External LOBs
- There is a recommendation from the E-ARK project: *Recommendation for storing large objects outside the SIARD file (E-ARK project, 2017)*
- SIARD 2.2 defines the structure for External LOBs (but DBPTK has not implemented it yet)

External file structure

DBPTK v2.5.10 (E-ARK recommendation)

5.2.2 Illustration of the folder structure for the example

```
Northwind.siard <!-- packaged as a ZIP file ->
        content/
        header/
                metadata.xml
                metadata.xsd
Northwind_lobseg_0/
        content/schema0/table2/lob4/record0.bin <!-- row 1, col 4 has record0.bin ->
        content/schema0/table2/lob4/record1.bin
        content/schema0/table2/lob4/record2.bin
        content/schema0/table2/lob4/record3.bin <!-- amount limit reached -->
Northwind_lobseg_1/
        content/schema0/table2/lob4/record4.bin
        content/schema0/table2/lob4/record5.bin
        content/schema0/table2/lob4/record6.bin <!-- size limit reached -->
Northwind_lobseg_2/
        content/schema0/table2/lob4/record7.bin
```

5.2.3 Extract of metadata.xml for the example

```
<dbname>Northwind</dbname>
<dataOwner>...</dataOwner>
<dataOriginTimespan>2015</dataOriginTimespan>
<lobFolder>file:///Archives/Northwind/</lobFolder>
```

SIARD 2.2 (Requirement L_7.1-0)

Example

```
Northwind.siard
Northwind_lobs/
    s0_t2_c4/
        seg_0/
            t2_c4_r1.bin
            t2_c4_r2.bin
            t2_c4_r3.bin
            t2_c4_r4.bin
            t2_c4_r5.bin
            t2_c4_r6.bin
            t2_c4_r7.bin
            t2_c4_r8.bin
    s0_t2_c8/                   <!-- new column -->
        seg_0/
            t2_c8_r3.bin
    s0_t11_c6/                  <!-- new column -->
        seg_0/
            t11_c6_r7.bin
```

metadata.xml

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>...
<siardArchive>...<lobFolder>./Northwind_lobs/</lobFolder>
..
<column>...<lobFolder>s0_t2_c4/</lobFolder>...</column>
<column>...<lobFolder>s0_t2_c8/</lobFolder>...</column>
..
<column>...<lobFolder>s0_t11_c6/</lobFolder>...</column>
```

DATALINK type

- SIARD 2.2 also supports the DATALINK type
- DATALINK type
  - DATALINK is a special SQL type intended to store URLs in database according to SQL:2008 part 9 SQL/MED (ISO/IEC 9075-9:2008).
  - It contains a reference to a LOB in a system outside of the RDB, but partly controlled by the RDBMS. In SIARD it is treated as a LOB with information about the original path.
- Rarely used due to limited support by major DB vendors in the past.

Keep in touch

ec.europa.eu/

europa.eu/

@EU_Commission

@EuropeanCommission

European Commission

europeancommission

@EuropeanCommission

EUTube

EU Spotify

European Commission

Thank you

Contact

https://e-ark4all.eu/

info@e-ark-foundation.com

@EU_eArchiving

https://www.linkedin.com/company/eu-earchiving-initiative

https://www.youtube.com/@e-ark