

E-ARK standardisation for databases

Kuldar Aas
E-ARK3 Technical lead

What is E-ARK?

EC funded E-ARK projects

- 2014 – 2017: E-ARK
- 2018 – 2019: E-ARK4ALL
- 2019 – 2021: E-ARK3

eArchiving Building Block

- EC-owned building block that aims to support interoperability in digital archiving
- E-ARK3 as the current service provider

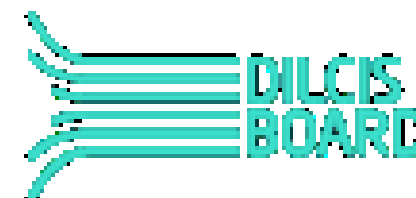
Digital Information LifeCycle Interoperability Standard (DILCIS) Board

- Maintains specifications developed in E-ARK projects ..
- .. and SIARD



eArchiving

Facilitate the preservation, migration, reuse and trust of your data



E-ARK vision (2014)

Vision: All digital preservation systems receive, store and provide access to information regardless of its size, type or format according to a set of agreed principles which allow systems to identify, verify and validate the information in a uniform way

Goal: Interoperability between data sources, archives and reuse environments is improved to a point where digital preservation tools can be reused across borders and institutions



E-ARK and databases

The idea of E-ARK was first developed among national archives

Focus still remains on 'records' in public sector and business

- Content in relational databases
- ERMS
- other



National Archives SIARD scenario

- Potentially 1000s on relational databases in the public sector
 - *About 95% of public records in Estonia assumed to be relational data
- 100s of databases of archival value
- Need to preserve data and service definitions (i.e. views, queries)
- Need for a simple, standardised preservation format (i.e. SIARD)

riha.ee/Avaleht

LOG IN

RIIGI INFOSÜSTEEMI HALDUSSÜSTEEM

[home page](#) RIHA catalog RIHA repository Help Center

search

Overview of the state information system

You can find descriptions of state information systems and data in the state information system administration system RIHA

900
An active institution and company

over 1300
registered information system and database

- Q** Browse the RIHA catalog
All information systems listed in RIHA can be found in the public RIHA catalog. The catalog provides information on how many different systems make up the Estonian state information system in general.
Browsing the information system in RIHA:
- P** Information system management
In order for others to find your information system, it must first be written down. Whether it is just being created or has been in use for years, describe it in RIHA.
The description of the information system in RIHA
- X** Joining the X-Road
The X-Road data exchange layer can be used for secure and interoperable data exchange with state information systems.
Joining the X-Road allows you to:
 - focus on the functionality of the information

DILCIS standardisation effort

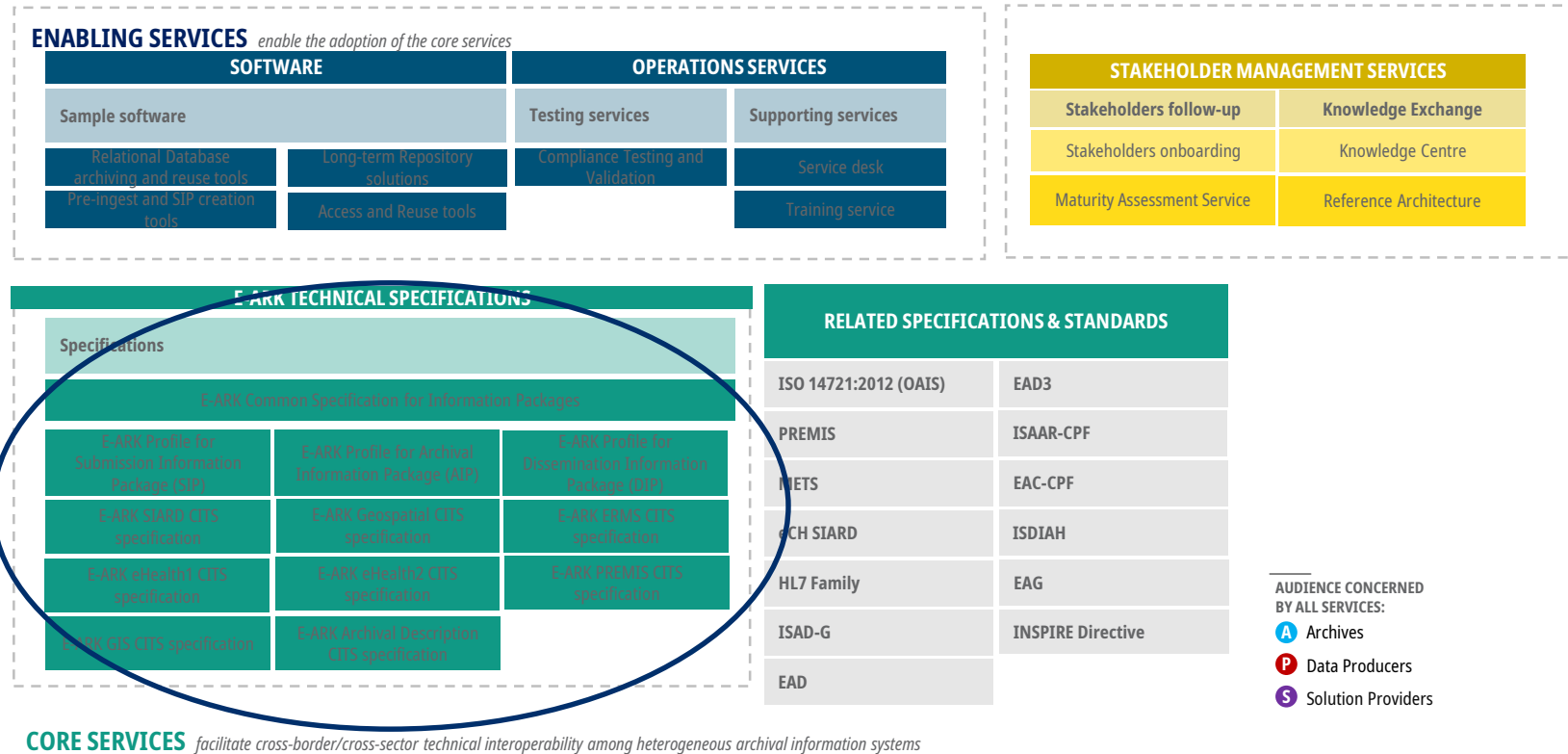
Generic Information Package Specifications

- CSIP, SIP, AIP, DIP

Content Information Type Specifications (CITS)

- SIARD, ERMS, Geodata, eHealth(x2)
- Archival description, PREMIS

Each specification has its own space and team



AUDIENCE CONCERNED BY ALL SERVICES:

- A** Archives
- P** Data Producers
- S** Solution Providers



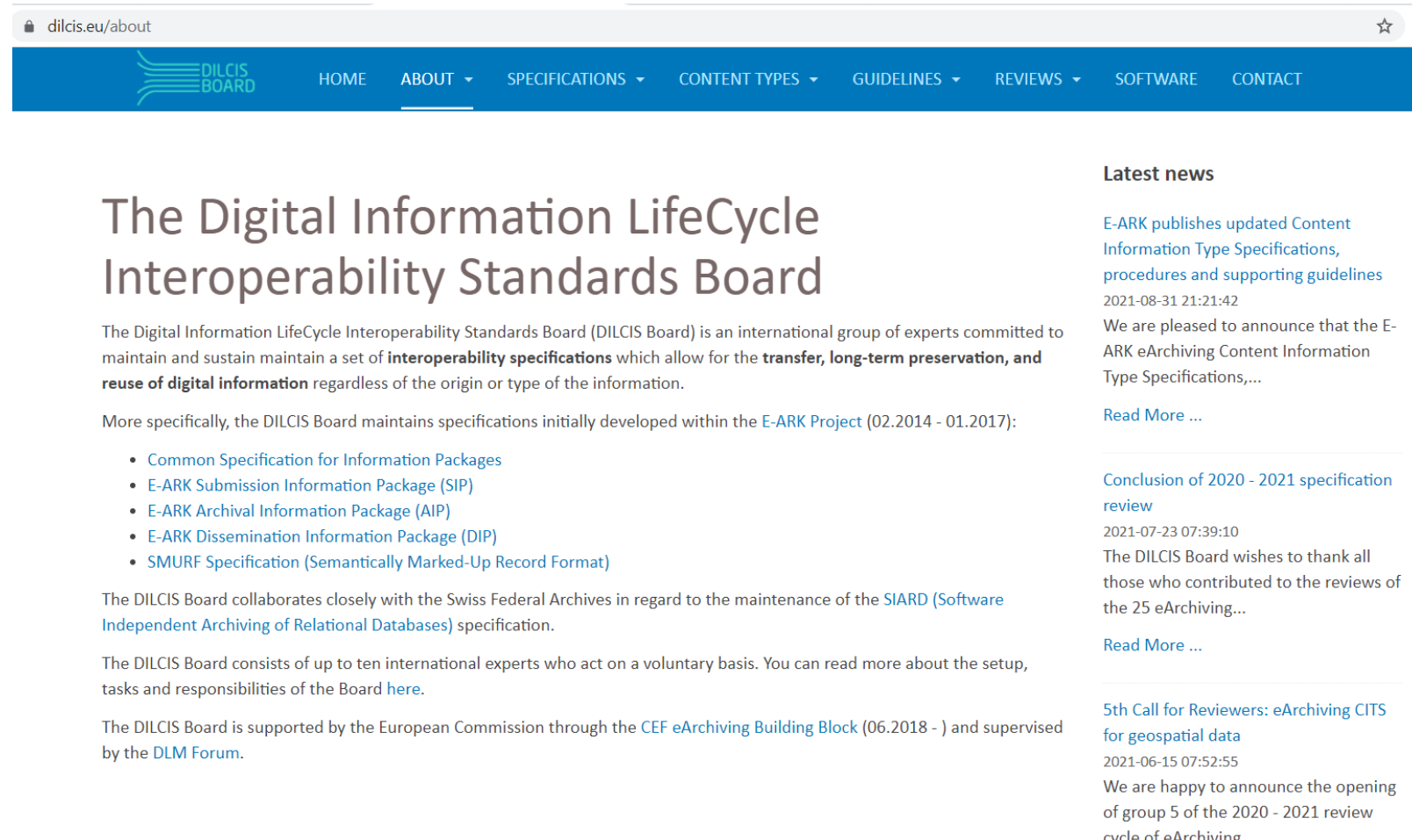
DILCIS standardisation support

Guidelines and procedures for CITS maintenance and development
<https://www.dilcis.eu/guidelines>

Dedicated landing pages for all specifications
<https://www.dilcis.eu/content-types/siard>

Open reviews
SIARD 2.2 RFC in 10.2020 – 01.2021

GitHub sites for spec development and issue handling
<https://github.com/DILCISBoard/SIARD>



The screenshot shows the website dilcis.eu/about. The navigation menu includes: HOME, ABOUT, SPECIFICATIONS, CONTENT TYPES, GUIDELINES, REVIEWS, SOFTWARE, CONTACT. The main heading is "The Digital Information LifeCycle Interoperability Standards Board". The introductory text states: "The Digital Information LifeCycle Interoperability Standards Board (DILCIS Board) is an international group of experts committed to maintain and sustain maintain a set of **interoperability specifications** which allow for the **transfer, long-term preservation, and reuse of digital information** regardless of the origin or type of the information." A list of specifications includes: Common Specification for Information Packages, E-ARK Submission Information Package (SIP), E-ARK Archival Information Package (AIP), E-ARK Dissemination Information Package (DIP), and SMURF Specification (Semantically Marked-Up Record Format). The sidebar contains "Latest news" with three items: "E-ARK publishes updated Content Information Type Specifications, procedures and supporting guidelines", "Conclusion of 2020 - 2021 specification review", and "5th Call for Reviewers: eArchiving CITS for geospatial data".



E-ARK and SIARD

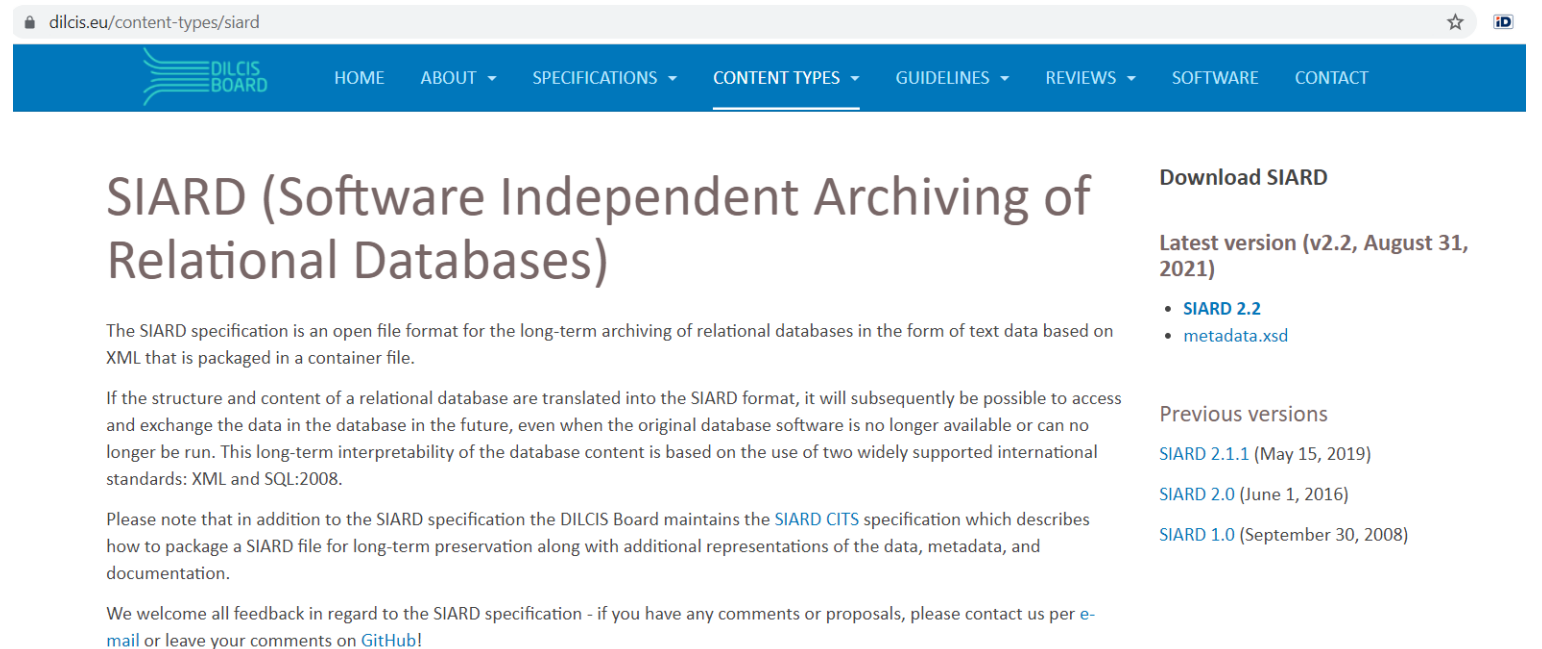
* SIARD is maintained in collaboration of DILCIS Board and Swiss Federal Archives

Common work on SIARD since 2017 (v2.0)

More people to work with SIARD = more errors and issues found

SIARD scalability as a special concern for E-ARK members (DNA, NAE)

- SIARD v2.2 includes support for LOBs outside SIARD



The screenshot shows the website dilcis.eu/content-types/siard. The page title is "SIARD (Software Independent Archiving of Relational Databases)". The navigation menu includes: HOME, ABOUT, SPECIFICATIONS, CONTENT TYPES, GUIDELINES, REVIEWS, SOFTWARE, and CONTACT. The main content area contains the following text:

The SIARD specification is an open file format for the long-term archiving of relational databases in the form of text data based on XML that is packaged in a container file.

If the structure and content of a relational database are translated into the SIARD format, it will subsequently be possible to access and exchange the data in the database in the future, even when the original database software is no longer available or can no longer be run. This long-term interpretability of the database content is based on the use of two widely supported international standards: XML and SQL:2008.

Please note that in addition to the SIARD specification the DILCIS Board maintains the [SIARD CITS](#) specification which describes how to package a SIARD file for long-term preservation along with additional representations of the data, metadata, and documentation.

We welcome all feedback in regard to the SIARD specification - if you have any comments or proposals, please contact us per [e-mail](#) or leave your comments on [GitHub](#)!

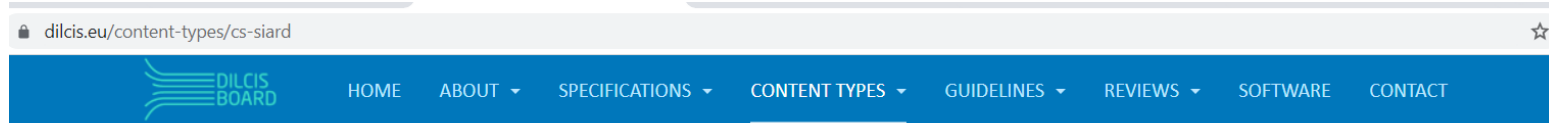
On the right side, there is a "Download SIARD" section with the text "Latest version (v2.2, August 31, 2021)" and a list of links: [SIARD 2.2](#) and [metadata.xsd](#). Below that is a "Previous versions" section listing: [SIARD 2.1.1](#) (May 15, 2019), [SIARD 2.0](#) (June 1, 2016), and [SIARD 1.0](#) (September 30, 2008).



E-ARK and CITS SIARD

A relational database transfer might include more than just a SIARD snapshot:

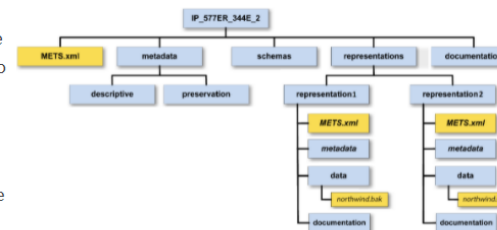
- Original dump (and application)
- Additional metadata (Dublin Core, EAD, ...)
- Documentation
- ...



CITS SIARD

The CITS SIARD (Content Information Type Specification for Relational Databases using SIARD) is a specification that describes how to package and preserve relational database content. This is primarily done by packaging SIARD files into information packages that conform to the Common Specification for Information Packages.

The specification helps you to apply a common way of storing multiple representations of a database (for example a proprietary backup and a SIARD snapshot) in a single package along with appropriate metadata and binary documentation of the dataset.



Download CITS SIARD

Latest version (v1.0.0, August 31, 2021)

- [CITS SIARD v1.0.0](#)
- [E-ARK-SIARD-ROOT.xml](#)
- [E-ARK-SIARD-REPRESENTATION.xml](#)
- [Guideline_CITS_SIARD_1_0_0.pdf](#)

We welcome all feedback in regard to the SIARD CITS specification - if you have any comments or proposals, please contact us per [e-mail](#) or leave your comments on [GitHub](#)!

How many ways are there to archive a database?

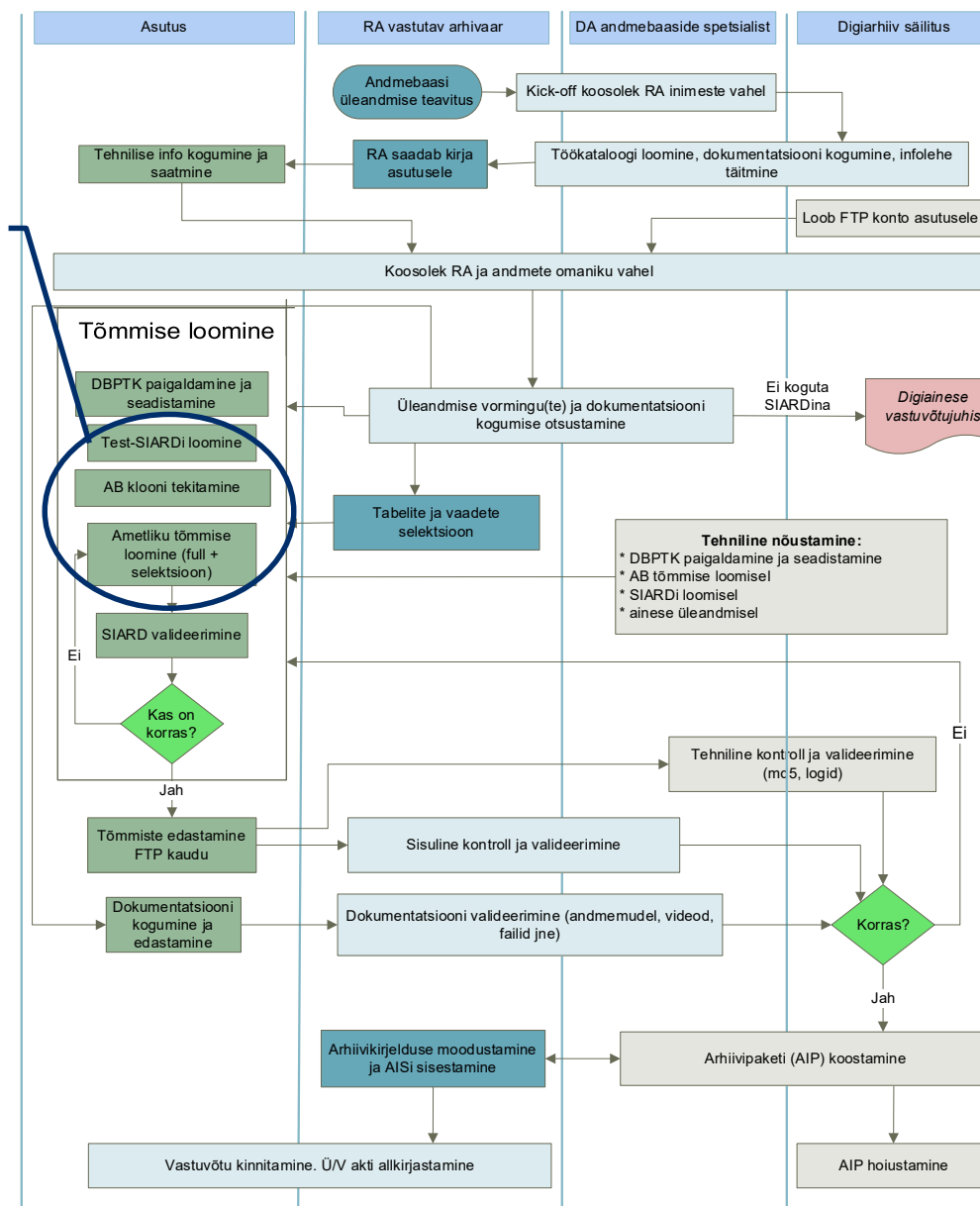
SIARD is just one of the options!

The whole archival process is much wider than just creating a snapshot

- Example: to archive the whole database or parts of it, relational data or materialised views / services?

SIARD software (DBPTK, SIARD Suite) behaves differently – which one to choose?

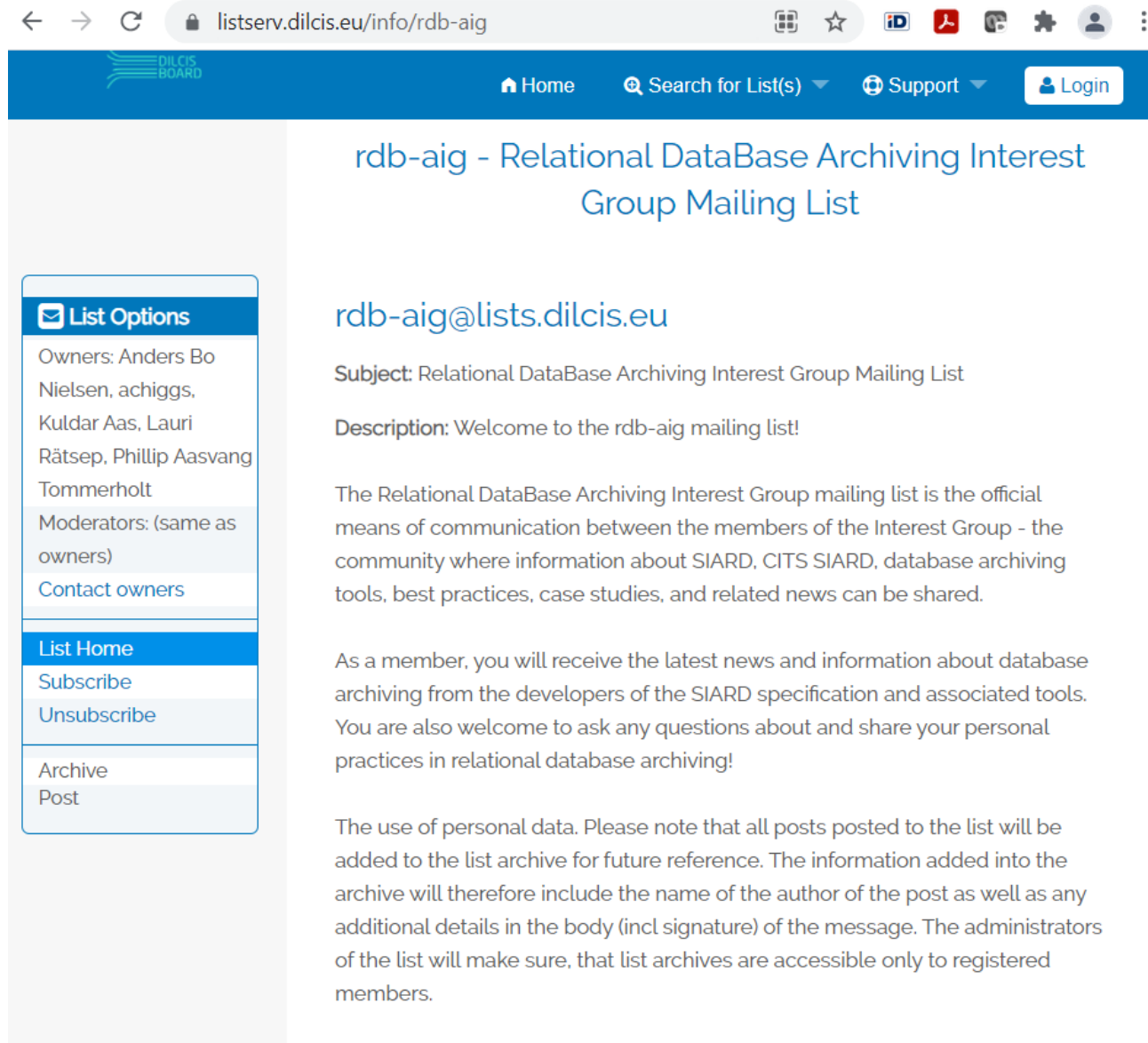
SIARD created here



How many ways are there to archive a database?

E-ARK thinks it's crucial to share our experiences

- Relational DataBase Archiving Interest Group (rdb-aig)
- Two case studies written in 2020
- NEW! rdb-aig mailing list
<https://listserv.dilcis.eu/info/rdb-aig>



The screenshot shows a web browser displaying the page for the 'rdb-aig - Relational DataBase Archiving Interest Group Mailing List'. The browser's address bar shows the URL 'listserv.dilcis.eu/info/rdb-aig'. The page features a blue header with the 'DILCIS BOARD' logo and navigation links for 'Home', 'Search for List(s)', 'Support', and 'Login'. The main content area is divided into two columns. The left column contains a 'List Options' sidebar with sections for 'List Options' (listing owners and moderators), 'List Home' (with 'Subscribe' and 'Unsubscribe' links), and 'Archive Post'. The right column contains the list's title, email address 'rdb-aig@lists.dilcis.eu', subject, description, and several paragraphs of text explaining the list's purpose and privacy policy.

← → ↻ listserv.dilcis.eu/info/rdb-aig [Grid] [Star] [ID] [PDF] [Globe] [User] [Menu]

DILCIS BOARD Home Search for List(s) Support Login

rdb-aig - Relational DataBase Archiving Interest Group Mailing List

rdb-aig@lists.dilcis.eu

Subject: Relational DataBase Archiving Interest Group Mailing List

Description: Welcome to the rdb-aig mailing list!

The Relational DataBase Archiving Interest Group mailing list is the official means of communication between the members of the Interest Group - the community where information about SIARD, CITS SIARD, database archiving tools, best practices, case studies, and related news can be shared.

As a member, you will receive the latest news and information about database archiving from the developers of the SIARD specification and associated tools. You are also welcome to ask any questions about and share your personal practices in relational database archiving!

The use of personal data. Please note that all posts posted to the list will be added to the list archive for future reference. The information added into the archive will therefore include the name of the author of the post as well as any additional details in the body (incl signature) of the message. The administrators of the list will make sure, that list archives are accessible only to registered members.

List Options

Owners: Anders Bo Nielsen, achiggs, Kuldar Aas, Lauri Rätsep, Phillip Aasvang Tommerholt

Moderators: (same as owners)

[Contact owners](#)

List Home

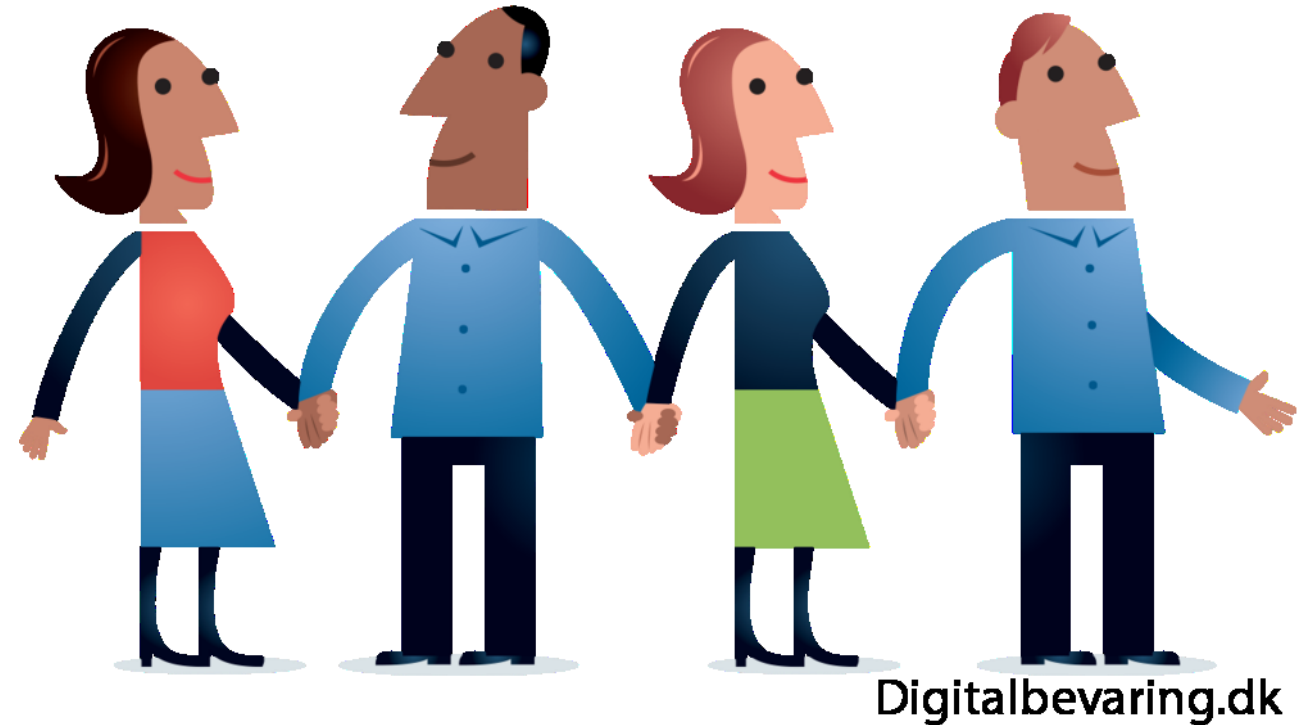
[Subscribe](#)

[Unsubscribe](#)

[Archive Post](#)

The future of ~~E-ARK~~ and SIARD and DILCIS

- DILCIS (and E-ARK) mindset is about being open and inclusive!
- DILCIS to support CITS development (as opposed to lead)
 - SIARD and CITS SIARD groups are currently led by DNA and SFA, with contributions by many others
 - Communication MUST be improved → 2021 SIARD v2.2 RFC got a total of three responses..
- SIARD validation and tool conformance as a serious issue
- rdb-aig mailing list to serve as the birthplace for new ideas!
- E-ARK to lobby with EC for funding



Questions?

Kuldar Aas

kuldar.aas@ra.ee

Ready to join DILCIS?

Find out more at:

<https://dilcis.eu/content-types/siard>

Join the rdb-aig mailing list at:

<https://listserv.dilcis.eu/info/rdb-aig>

