E-ARK standardisation efforts for databases

Kuldar Aas (National Archives Estonia, Tallin, Estonia)

Kuldar Aas is the Data Governance Programme Lead at the Ministry of Economic Affairs and Communication of Estonia. Until 2021, he was the deputy director of the Digital Archives of the National Archives of Estonia. In this position he was actively involved in developing national records management and cultural heritage metadata standards, creating requirements and guidelines for the ingest, description and preservation of national datasets and electronic records from EDRM systems. Aas has participated in a number of European collaborative projects (PROTAGE, APEx, YEAH, APIS), and was the initiator and technical coordinator of E-ARK projects in 2014-2021.

Leaving the world of current best practice, Aas covered standardisation more generally. He has been in digital preservation for almost 20 years and began this career looking at the preservation of databases.

This became a big endeavor when European Archival Records and Knowledge preservation (E-ARK) was launched as an EU funded project in 2014 and repeated twice until the E-ARK3 project whose outcomes are now promoted by the Digital Information Lifecycle Interoperability Standards (DILCIS) board. E-ARK and DILCIS aim to support interoperability. Aas described the relationship between the E-ARK project, the current eArchiving EU activities and the DILCIS board.

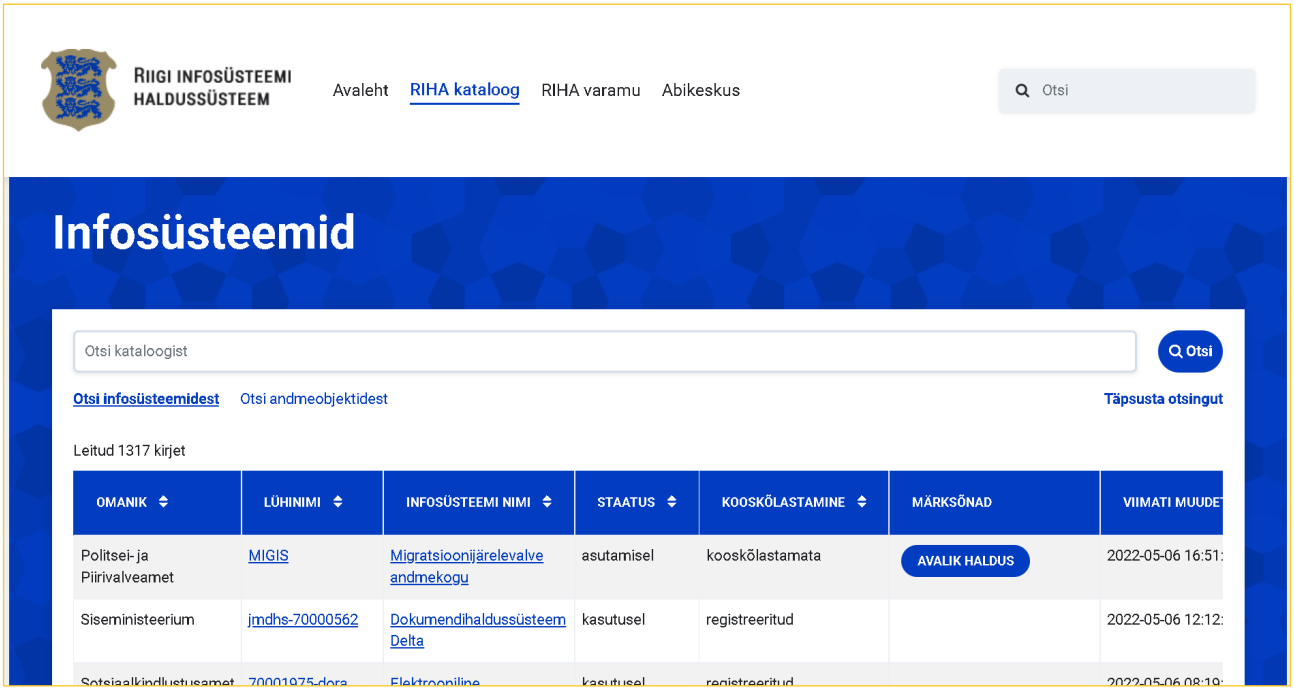

Figure 21: Estonia's government information system catalogue mentions 1317 entries

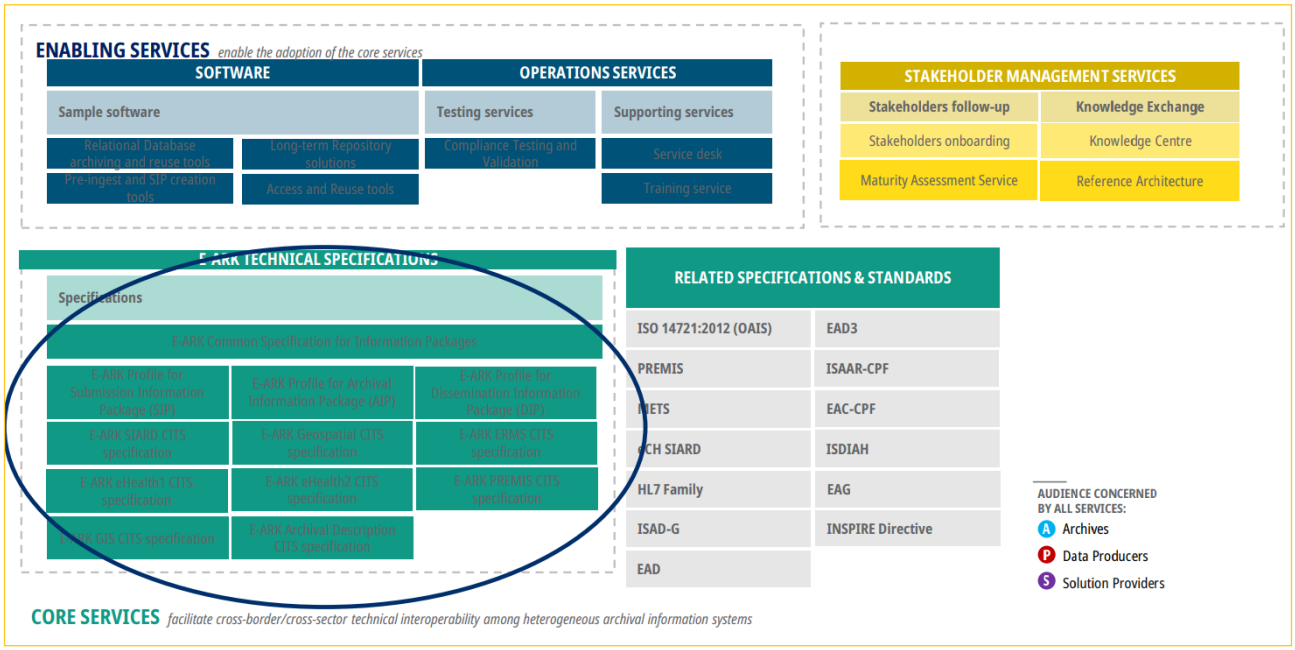

Figure 22: Overview on the DILCIS standardisation effort

The original E-ARK vision was that all digital preservation systems receive, store, and provide access to information regardless of its size, style or format according to a set of agreed principles, which allow systems to identify, verify and validate the information in a uniform way. This was aimed at interoperability between data source archives and re-use environments. E-ARK developed among national archives of some smaller European nations and the focus remains on records in public sector and businesses – for example content in relational databases, ERMS and other systems – more generally: any kind of information that has legal value.

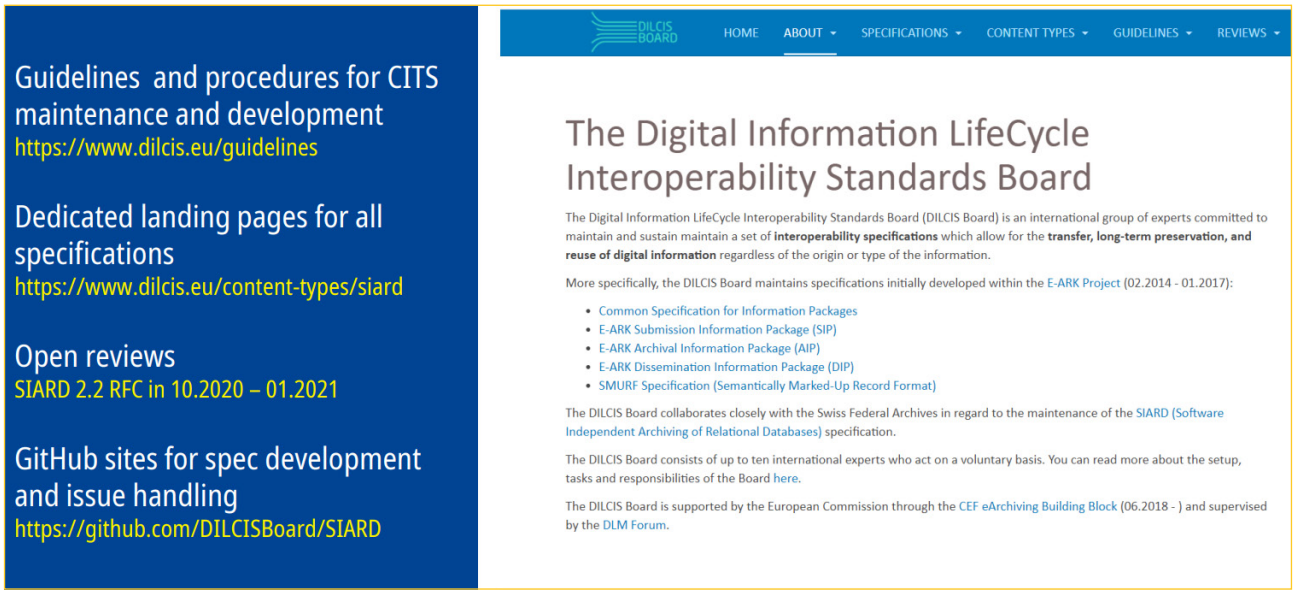

Figure 23: The DILCIS portfolio as of 2021 and its homepage

A specific problem at the Estonian National Archives is that the public sector might hand over potentially thousands of relational databases – 95% of public records in Estonia are relational data, existing on many different platforms (figure 21). It has been decided there needs to be a universal and standardised preservation format that could connect to all these different databases and hide the complexity from the end users.

Figure 24: The history of SIARD, taken from Guideline for CITS SIARD (2021), p. 12

The DILCIS standardisation effort has focused on generic information package specifications as well as content specifications such as SIARD. Lots of work happens on GitHub and this is where the community can get involved and leave feedback. The more people work with SIARD, the more errors and issues are found and fed into the development of the standard.

Scalability of SIARD is a key issue, but also documentation. A relational database transfer might include more than just a SIARD snapshot. It may come with additional metadata and documentation as well as a dump of the original DB and application. Therefore, the Content Information Type Specification (CITS) for SIARD was released in August 2021 together with accompanying Guidelines.

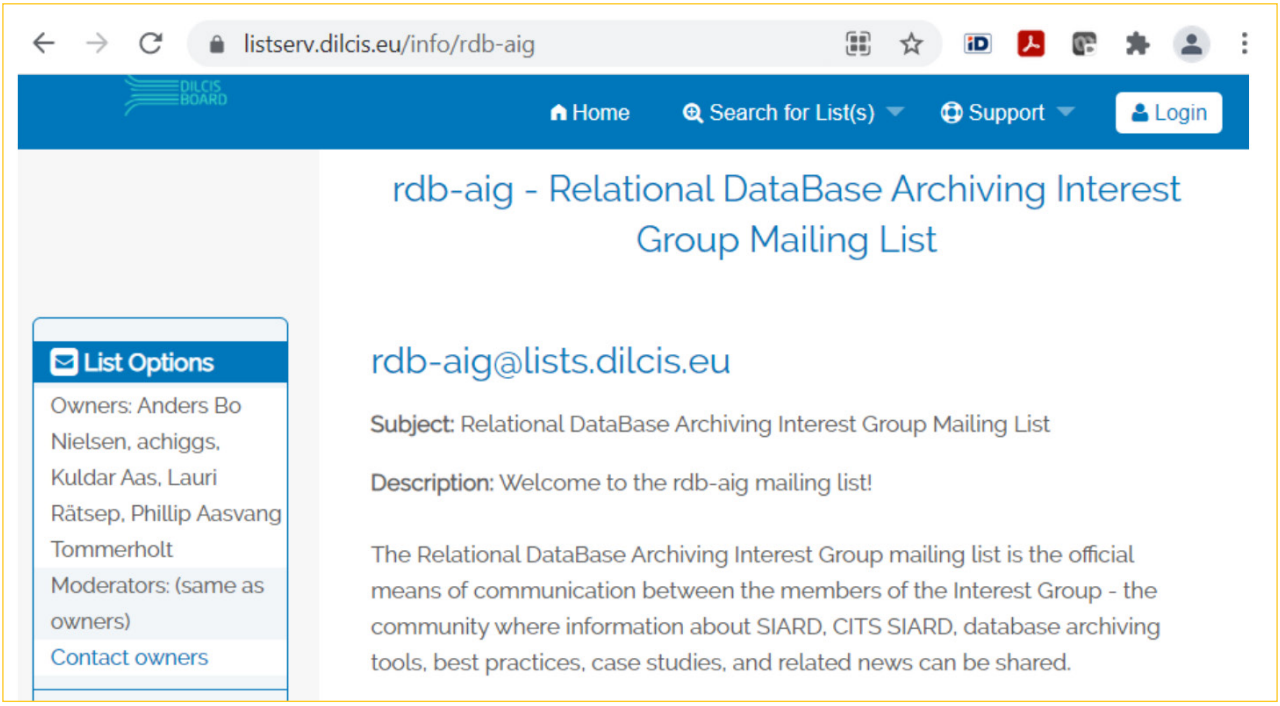

Figure 25: Listserver of the Relational Database Archiving Interest Group (rdb-aig) Mailing List

SIARD is only one option for archiving a database and only one step of the process – there needs to be much decision making around this related to the whole archival process. Should you archive the whole database or just parts of it? Should relational data be captured or materialised views/services? One also needs to choose appropriate software, noting that SIARD software behaves differently.

It is important to share experiences on database preservation, for instance in the DILCIS Relational Database Archiving Interest Group (see further). Note there are two case studies written in 2020 (Preserving databases, 2020; Preserving [...] Case Study 2, 2020). The first one covers many national implementations; the second is only about implementing Large Object database content.

DILCIS will take this work forward in an open and inclusive way. The DILCIS board has noted that communication must be improved as the 2021 SIARD v2.2 request for comment only had a limited number of responses but this workshop demonstrates that there are more people interested in the topic of database preservation. The hope is that E-ARK will be able to lobby with the European Commission for further funding.

Most people involved in E-ARK and DILCIS have an IT focus rather than being specialists at communication. This, Kuldar Aas said, needs to change.

Questions and discussion

- Which features would you ask to be incorporated in DILCIS and SIARD moving forward? Aas noted that he would like to turn that question around and ask for ideas from the community – already some excellent ideas are coming out of the workshop.
- Concerning SIARD: What about data that is misread or incorrectly read? How would you re-import the data multiple times? Faria referred that you can do partial exports from a database into SIARD, or you can load SIARD into a living DBMS, which does not need to be the same vendor as initially, change the data, and export back into SIARD. For using this technique over time, you might experience issues if the source database changes schemas, but you could do extra work to align them together, for example with an archival view.
- SIARD development – a few things on the list, support for bi-temporal databases included. Encouraged to use the mailing list to add further ideas (https://listserv.dilcis.eu/info/rdb-aig).
- Kai Naumann mentioned geodata. SIARD is moving in this direction. Any time Oracle stores geometry, these outlines can be stored as a GML (Geographic Markup Language) file. Aas mentioned they are trying to separate them in the specification. GML conversion may be incorporated into DBPTK. Faria reported they export Oracle into single GML file but add all other rows (related content) into the GML as attributes. The strategy within the tool is to use GML not SIARD for geodata as this format can be opened in other systems.