Discussion group (A) on standards, software, deployment

Discussion group A was chaired by Kuldar Aas (National Archives Estonia) and Carl Wilson (OPF)

Carl Wilson leads all technical activities at Open Preservation Foundation (UK). An open source enthusiast, he is an experienced software engineer with a focus on software quality through testing. His professional interest is using virtualisation, automation and continuous delivery techniques to improve the software development process.

Biography of Kuldar Aas: see p. 28.

Lead questions standards:

- Is standardisation in the area scarce, sufficient, or exaggerated?

Standardisation has to be improved, beginning with SQL: there are differences between e.g., PostgreSQL and Oracle. It is really the extensions to the standard ISO SQL (references) that cause issues: TransactSQL (MS), PLSQL (Oracle). In addition, access to databases is not standardised: there is no standard or best practice on how to do it in legal or other administrative contexts.

There is also a lack of adoption for SIARD outside the archival sector. SIARD must evolve.

Participants discussed SQLlite as a storage format, instead or on top of SIARD, but did not get to a clear result. There are advantages in using a binary, well-documented format as regards storage costs. It saves the overhead incurred by XML tags, as can be seen in the success of the GeoPackage format in geoinformatics. It was contentious whether storage cost should influence archival format decisions.

Some thought SQLite is more prone to obsolescence than SIARD, because the latter is plain text and more interpretable. Most think SIARD cannot become unreadable in the future. There were claims that SIARD is designed for archiving and based on the ISO SQL:2008 standard, while SQLite does not fully comply with this standard. Others thought that SQLite could be kept perfectly interpretable if a specification survives along with the data, as with JPEG or TIFF.

Access should be viewed differently: preservation formats are not the same as access formats. One must distinguish between transfer formats, transformation formats and preservation formats.

- The discussion was all about relational databases. Does anyone preserve different data sources (e.g., data lakes, RDF, NoSQL data)?

Moved to Group B, fused with the NOSQL question (CROSSREF inside)

Lead questions software:

- Do we have complete software tool suites for the task?

The answer was no. It was agreed that more software is available than 10 years ago, but not enough. Slovenia requires a software to become certified before it can be used by agencies. Data “transparency” is one of the criteria (clear export format) for this process. One issue is the security of the software and sensitive information stored in the database. Another issue is the access to the data.

Authenticity also comes to mind and there is a lack of tools to preserve it. One example is the “sign off” of a dataset which is difficult to preserve. This should be assured in the initial software (e.g. by hash and log files). The archive should get all the information to preserve the authenticity. But this depends on the design of the database. For example, for every change in the database a hash is generated. Maybe this is a step towards preserving authenticity. Blockchain usually has too much overhead to accomplish this requirement.

Torbjørn Aasen reminded the audience of

- SIARD spec interoperability issues (https://github.com/DILCISBoard/SIARD/issues/43)
- a lack of practical examples (https://github.com/DILCISBoard/SIARD/issues/44), inconsistent file path reference for LOBs in DBPTK Desktop v2.5.4 (https://github.com/keeps/dbptk-developer/issues/476).

He hoped for improvements on the situation. A party must decide on what is within the spec and what is not within the spec, so the vendors can all act accordingly and increase common interoperability. An important tool to reach this aim would be best practice test cases of SIARD from the different usages and vendors, avoiding interoperability problems in real cases.

- Can we influence the big data industry?

The basic answer was no. Nevertheless, there were ideas on how to catalyse reactions from the industry (see also Group B).