Discussion group (B) on strategies, efficiency, documentation

Discussion group B was chaired by Annette Strauch-Davey (University Hildesheim) and Kai Naumann (Landesarchiv Baden-Württemberg)

Annette Strauch-Davey is European Ethnologist and "Digital Humanist" from the Georg-August University of Göttingen. She worked and lived in Wales for almost twenty years (Museum of Welsh Life, National Library of Wales). She also worked for the project bwFLA (EaaS) at the kiz, University of Ulm. She was working for the INF project in the Collaborative Research Center SFB 1187 before moving to Hildesheim in order to build up services for Research Data Management at the Hildesheim University Foundation, aiming to make the research process as efficient as possible and meet expectations and requirements of the university, research funders, and legislation.

For Kai Naumann's biography, see p. 6.

Lead questions efficiency:

- Do we need persistence for whole databases, or can we mostly rely on derived datasets?

Yes and no, depending on data types. In the case of social surveys, it is sometimes important to preserve views of information displayed to respondents (e.g. questions). GESIS is looking into SIARD, but is it suited for Social Media data, for behavioral sciences? Content is more important than user interfaces. It always depends on what the end users want from the database – this will be specific to the content/context and future use cases. There is a need for low-level preservation formats, but format diversity and the urge to innovate must be considered. An interesting example is the statistical microdata community that established the DDI (Data Documentation Initiative) creating an XML metadata standard mainly for the social and economic sciences.

- Do we need additions to intellectual property legislation regarding database archiving?

Limited budgets for archival interests do not match the prices charged by DBMS manufacturers (see Martin Rechtorik, cross-reference). This dims the prospects for emulation. National and EU legislation must be further relaxed for libraries and archives (US fair use clauses for education as a model). Software that is no more on the first market should be allowed to be used without licence fees. Andreas Lange (former Computerspiele Museum Berlin, current affiliation?) is lobbying for this on the EU level. (References https://www.softwarepreservationnetwork.org/dmca-rulemaking-reform/, https://efgamp.eu/2020/02/14/dsm-directive-first-efgamp-statement-submitted/).

Is it feasible to think about having licences for software products in the archives? Note that you also need to emulate the hardware as well as the software. There are many technical problems that will be encountered. Microsoft Access is still widely used. Oracle has been mentioned often in the workshop. It is important to lobby for long-term preservation formats.

Further dependencies were mentioned like Docker and GitHub. One needs to consider all elements otherwise there is no chance for reproducibility. The community does not yet know how to deal with these challenges. Docker images need to be exported to standard container formats, and GitHub to Git repositories in the long term.

Knowledge to access a database is important too. We need to store as much as we can afford but may lose data with every migration. Scientific databases prefer to keep original data when feasible, and migrated data for easy access.

Documentation is a hot topic. Presentations have not included much on access either. It is hard to imagine how to rebuild a very complex database with numerous tables and columns. It might be difficult to achieve this and to be certain afterwards that what has been rebuilt is authentic. For example, you might have to be able to convince a judge that the data has been created at the time and by the person indicated in its metadata.

Lead questions strategies:

- Are there multiple, technically different business cases like emulation vs. migration or only one basic business case that varies only in terms of IT ecosystems and DBMS types?

The business cases differ, it is not one basic business case. Sometimes it is a question of "belief". Often organisations tend to make their business case a special one.

- Are the emerging NOSQL technology, resource description frameworks (RDFs), and technically diverse Data Lakes even more difficult to preserve?

Of course, but solutions in this area vary much more than in the relational DB world. For example, the Cassandra DB is well-suited for frequently changing timelines, like on streaming servers. NoSQL can be stored in JSON (JavaScript Object Notation) or XML but to date there is no tool like DBPTK or Spectral Core Full Convert that would access all types of NoSQL DBMS.

"Archiving by design" should be a principle; systems should be designed with "mothballing" and archiving as first class use cases.

Most of the time, as a database archivist, you are documenting the data information systems and how they work. The documentation of the software itself is the bigger issue than preserving the data in the software.

When you have formatted pieces of data, the formatting of the data is an important metadata that is in fact often lost.

The focus must be on the future and the future use of the data.

- Can geodata (coordinates, polygons) be included in the standard database workflows?

Yes, there is a chance to preserve Oracle Spatial with DBPTK, but it is not possible to preserve PostgreSQL. It is still impossible to do so in SIARD, because simple Geographic Information System (GIS) features are not captured in tables. Geometry can be exported into GML. Negotiation with developers to incorporate simple GIS features into SIARD should begin. To date, the SIARD spec does not allow GML to be contained.