Optimistic epilogue (seven months later)

Kai Naumann

When I did the wrap-up in October 2021, I reminded the audience of the most common obstacles to database preservation that are outside of academic research: understaffed DB admins, too little standards and software support, proprietary extensions, problems with handling large XML files that do not parse well on most computers (same for CSV files with one million rows not opening in Excel), as well as accidents and disasters.

For reasons of time, I withheld some slides with a simple manual for database preservation:

1. **Document the original environment**

Gather manuals, specifications, screenshots or screencasts of user interfaces, maybe follow standards like Data Documentation Initiative (DDI), PREMIS, or the CITS SIARD (see references) from E-ARK.

In short: try to make reviving the database a self-explaining “bootstrap” routine.

2. **Continue with options:**

Option A: Select essential entities

You can only use existing reporting data (for accounting, statistics, OLAP processing) and renounce on the full database content.

You can create specific archival reports, also by modifying existing reports, in various formats like CSV, SQLite, SIARD, or other XML or JSON.

Option B: Preserve DBMS performance

This can be done with proprietary dumps (e.g. Oracle, or SIARD, SQL 2008, or SQLite dumps. For the latter cases, you might miss some parts. You can make the informed decision to do so, but avoid losing performance without notice.

Option C: Preserve complete performance

This can be done by calling for emulation services. Packaging standards for OS, DBMS, and GUI might emerge in the next years, also specialized companies might help. It may be a good advice to use less intricate strategies as a benchmark in parallel.

3. **From then on, you keep the package in persistent, replicated, secure storage, keep access methods alive, wait, and monitor the data regularly.**

Getting back to the obstacles: during the preparation of the final text version of this workshop documentation, I decided that outlooks are a little better than I first estimated. As Christine Barats et al. (2020) have pointed out, nearly all areas of science set up platforms similar to the ones depicted by Mathiak (p. 8), for example OPERAS and COPIM for scholarly communication, and Dataverse and CESSDA for social science data, let alone the large and standardised data stores for data in chemistry, biology, medicine, physics, many other academic disciplines, and also in government agencies. The University of Vienna (Weise 2022) has demonstrated a tool at the International Digital Curation Conference that facilitates documentation of relational DB content.

Also, science and society both are getting impatient waiting for interoperability because it becomes an economic imperative. Regarding privacy legislation, stable and trustworthy ways of controlled ways for accessing outdated information are also vital for the rule of law. “Security by obscurity” is an outdated concept in the age of data crawlers (see e.g. Barats et al. 2016). The better old data systems are accessible, the better they can be protected against unlawful access. It will be legal and economic forces that drive the vision of this book forward.