



# Preserving databases for the future

- Dr. Kai Naumann, November 15, 2023
- DOI [10.5281/zenodo.10405758](https://doi.org/10.5281/zenodo.10405758)

# My employer at a glance

- knowledge centre about the past of the state of Baden-Württemberg
- key research infrastructure for the past
- saves records of all kinds as cultural heritage
- provides transparency of decision-making
- founding member of the DIMAG software development group (with >200 other partners from Austria, Germany, and Switzerland)
- 9 sites throughout the country
- 11 million EUR overall budget
- 1210 years: oldest dated charter
- 21.878.815 pages of scanned documents
- 294.403.593 dataset rows





Challenge, made up in the first COVID weeks, 2020.

How do you preserve 125 databases of diverse origin for future use from the year 2080 onwards?

Prepare them in such a way that they can be used in as many ways as possible in 2080.

In the following 60 years

- a) no cost should be incurred apart from secure storage
- b) the database contents must not be publicly accessible.

# Why this challenge?

- Aside email and records management systems, where does evidence for decisions in society, economy, politics reside?
  - line of business systems
  - geographic information systems
  - data warehouses
  - ...
- At what cost are these data constantly migrated to successor systems?
- How to assure that these data are migrated without a bias or loss?



# The 1961 census of Baden-Württemberg

- executed in 1961 on rented IBM machines
- 6 million individual punched cards destroyed in 1968 by a flooding
- Surviving part: calculated sums on ca. 1,592,821 punched cards
- migrated
  - to magnetic tape in the 1960s
  - from unknown code to ASCII (maybe via EBCDIC)
  - to CD-ROM in the 1990s
- Transferred to the State Archives in 2006
- **Can we do better for the 60 years to come?**

5  
hlf bitte mit

Volkszählung  
1961

Kai Naumann, DB Archiving, 11/2023

Auch auf Deine genauen Angaben kommt es an

LABW StAL E 258 II Bü 214

<http://www.landesarchiv-bw.de/plink/?f=2-335336>

# The challenge and how it evolved

## 2020

- challenge spread via email
- presented at Joint Conference on Digital Libraries  
<https://vtechworks.lib.vt.edu/handle/10919/99569>
- and presented at WeMissiPres2020

## 2021

- Virtual Workshop Databases for 2080  
<https://www.landesarchiv-bw.de/de/aktuelles/termine/72973>

## 2022

- Workshop Proceedings  
<https://doi.org/10.53458/books.109>
- Follow-up iPRES2022 Workshop „Eternalize DBs“  
<https://www.dpconline.org/docs/miscellaneous/events/2022-events/2791-ipres-2022-proceedings/file>



# Highlights of the DBs for 2080 workshop

## Databases for 2080

Workshop Proceedings

Edited by Kai Naumann

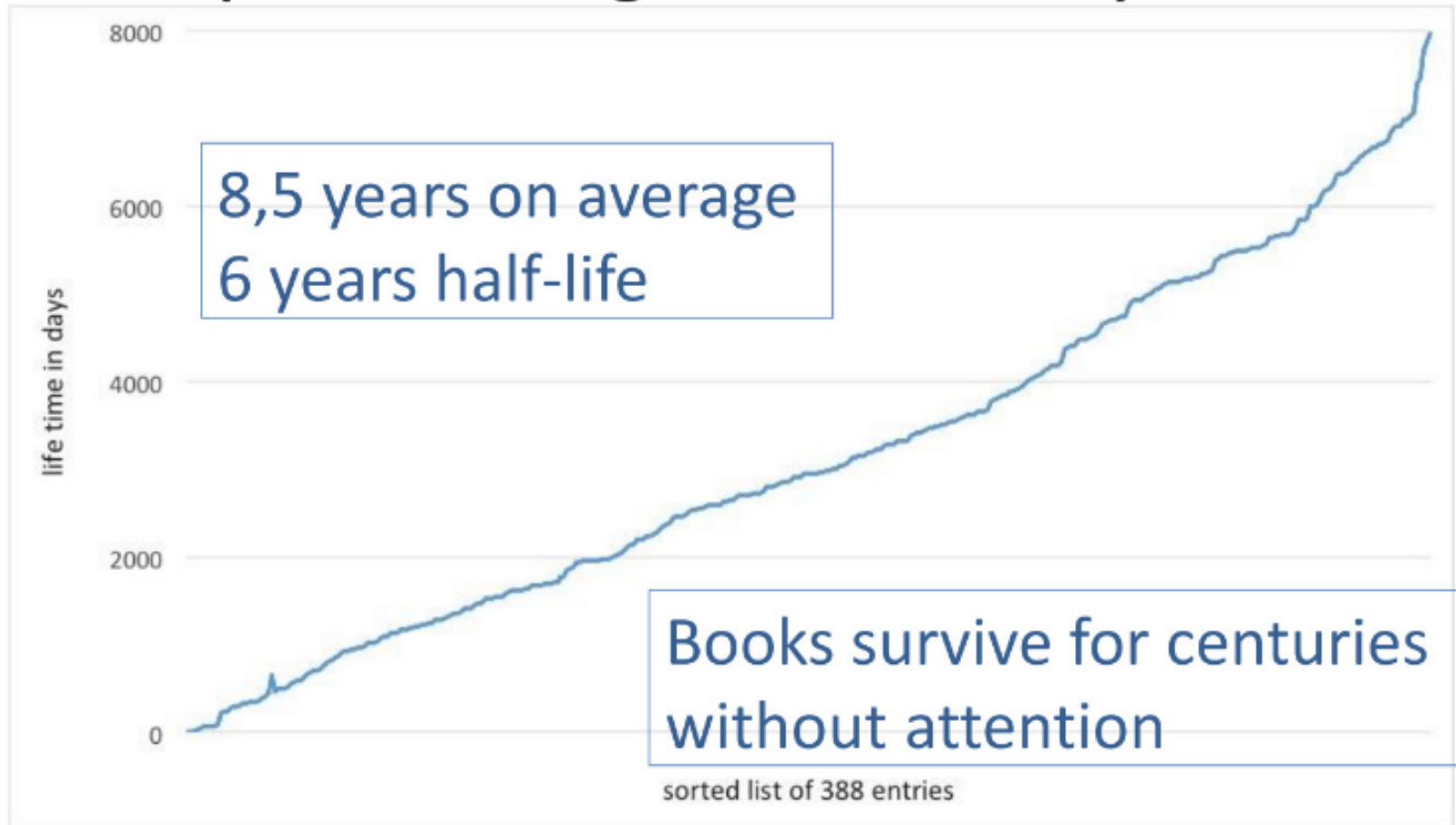
[https://nbn-resolving.org/  
urn:nbn:de:101:1-2022071903](https://nbn-resolving.org/urn:nbn:de:101:1-2022071903)



# Workshop outcomes (1): B. Mathiak



## Life spans of Digital Scholarly Editions





## Information to preserve

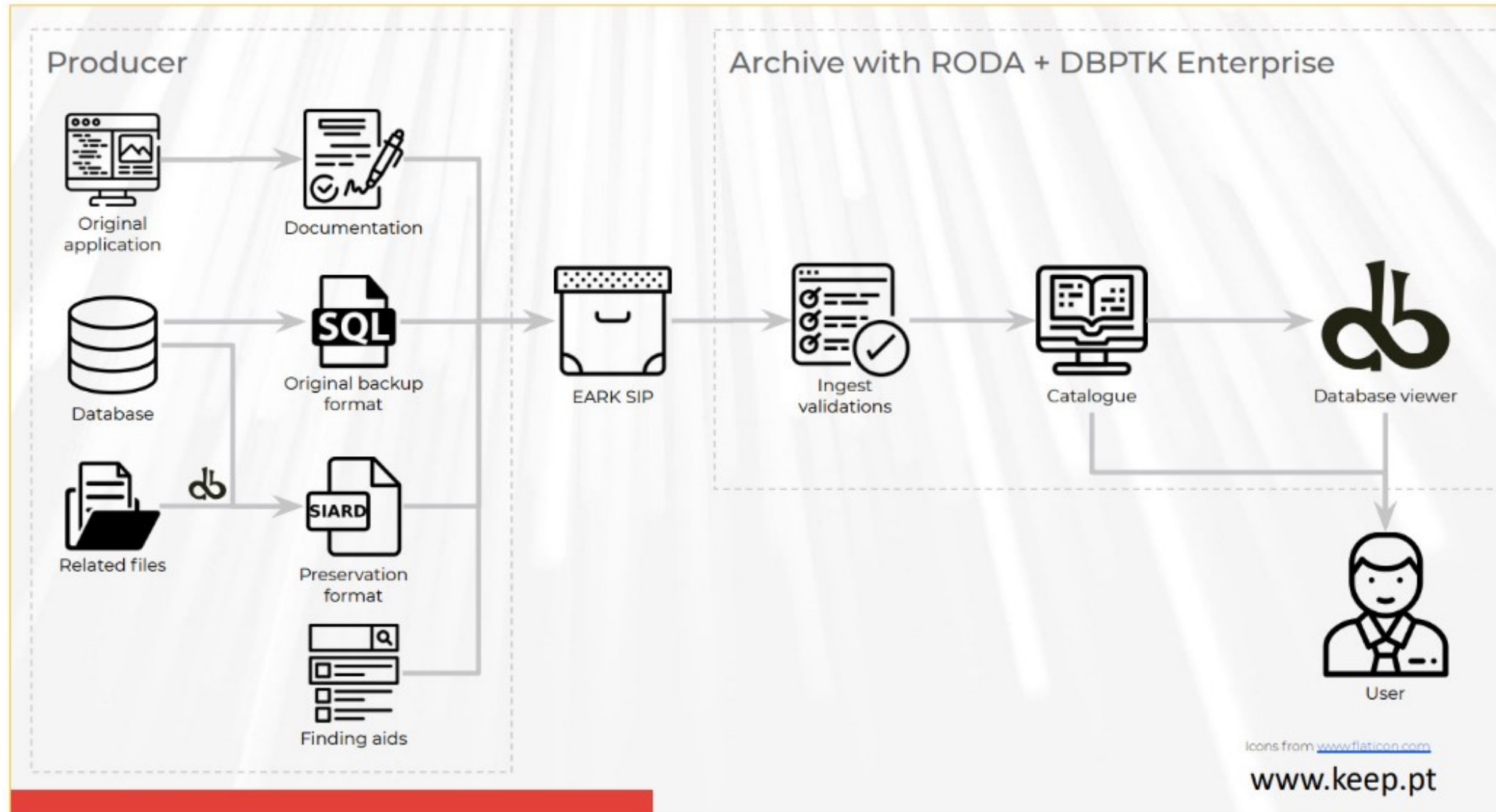
Within the relational database:

- Information in tables
- Column data types
- Relations and constraints
- Projections (views)
- Behaviour (triggers and routines)
- Other (users, permissions, etc.)

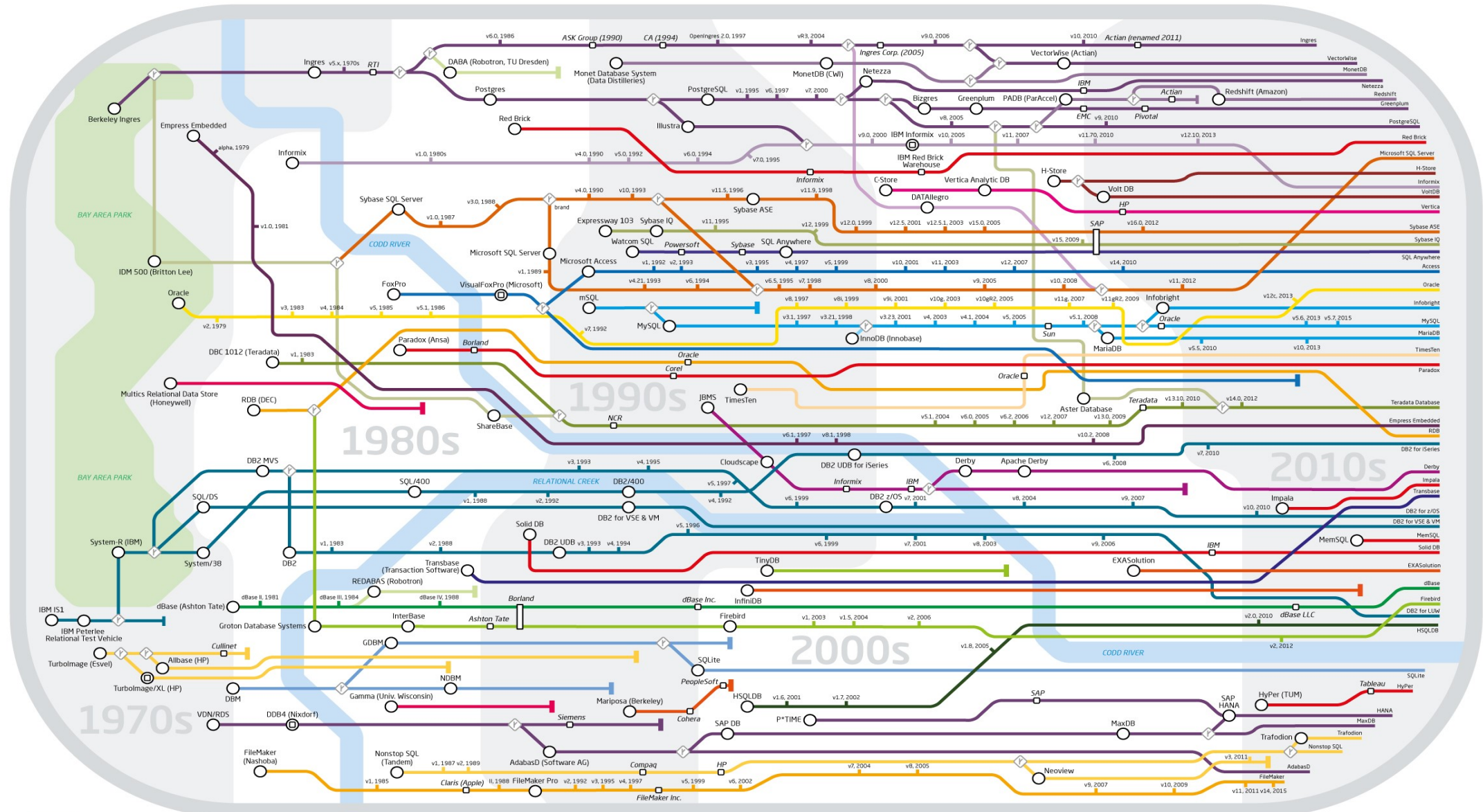
Outside the relational database:

- External resources (e.g. files in filesystem)
- Submission forms
- Presentation interfaces
- Application logic and queries

# Workshop outcomes (3): L. Faria

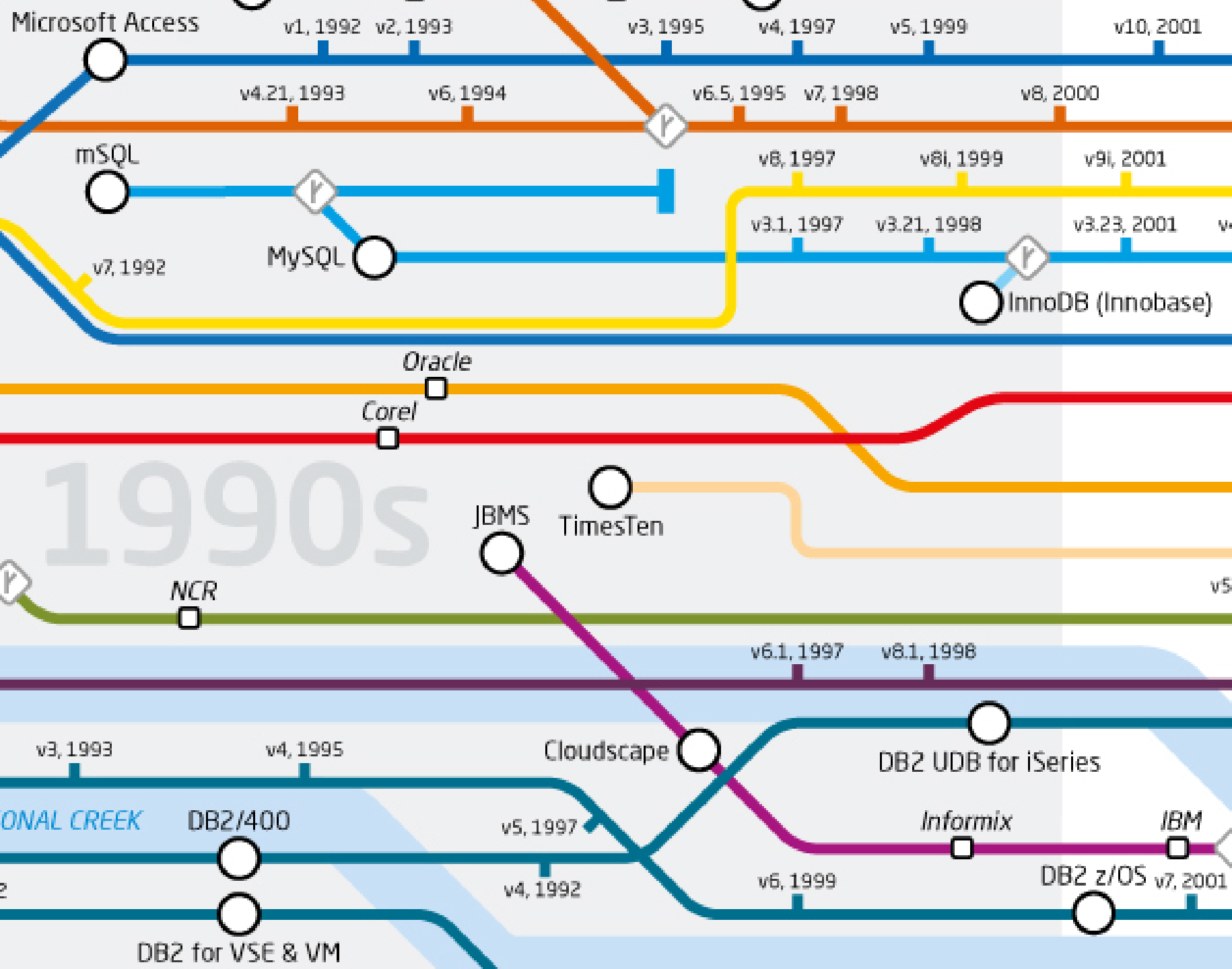


# Genealogy of Relational Database Management Systems



## Key to lines and symbols

- DBMS name (Company)
- Acquisition
- v9.2006 Versions
- ⏏ Discontinued
- ◇ Branch (Intellectual and/or code)
- Crossing lines have no special semantics



Felix Naumann et al.,  
Genealogy of  
Relational Database  
Management  
Systems, v6.0, 2018,

<https://hpi.de/naumann/projects/rdbms-genealogy.html>


(Detail)



# Workshop outcomes (4): K. Aas



# Workshop outcomes (5): J. Doig



**NATIONAL ARCHIVES OF AUSTRALIA**

**Checklist of questions to ask an  
Agency with data to transfer from  
a Relational Database**

R305582021

Digital Archives Research and Innovation 2021



# Recent developments and hypotheses

# Outcomes

- SIARD conceived around 2000, but supposedly insufficient for the large scale
- SQLite as lightweight and robust DB storage emerged out of a damage control system for battleships in 2000 and used in **+10.000 products**
- SIARD an EU standard candidate
- **three products** (DBPTK, SIARD Suite, Spectral Core Full Convert) produce SIARD files

# and Assumptions (1)

- the demand was first voiced by professionals (archivists)
- but only large-scale needs of society and industry will make the demand viable



US Navy, Public domain, via Wikimedia Commons

# Outcomes

- still no industry-wide DB storage format
- 2020ies: DNA and nanolayer ceramic storage will change data centers
- 2020ies: emulation of ancient software becoming a business niche as two EU companies offer services

# and Assumptions (2)

- long-term storage media will lead to long-term formats
- the need for migrating old formats may become less urgent



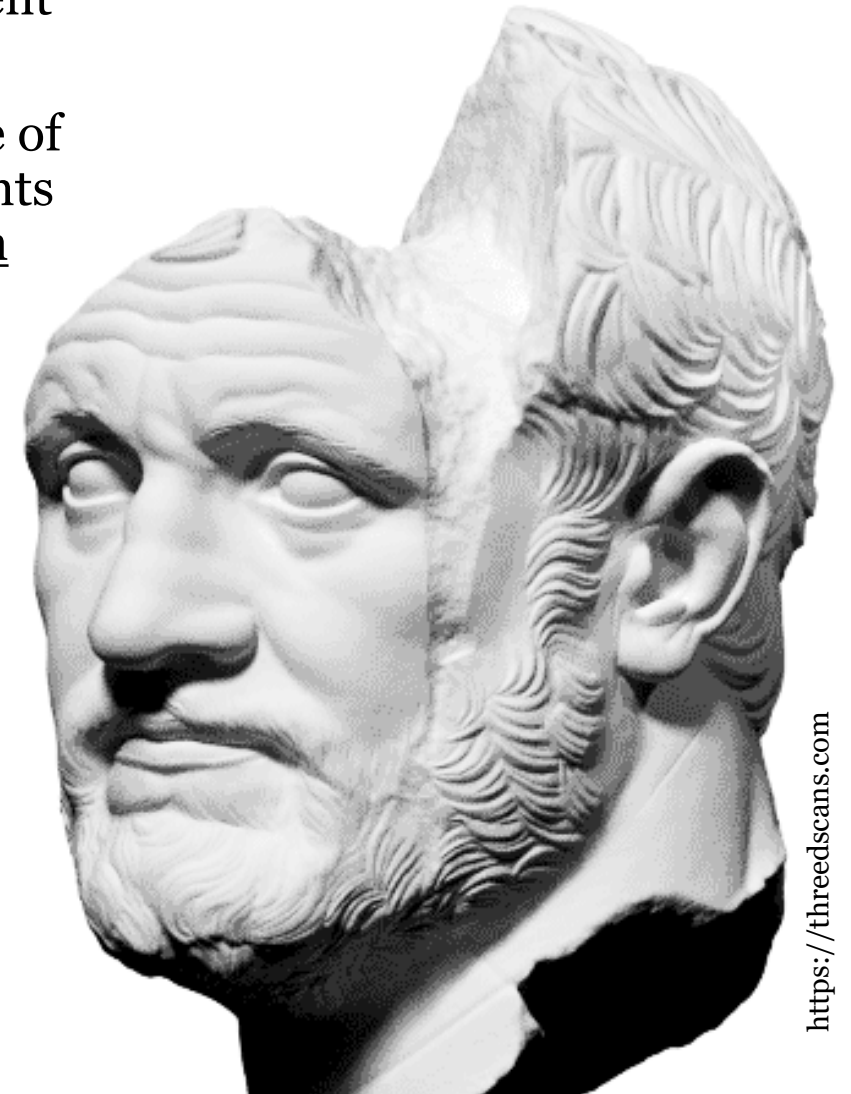


# Outcomes

- 2022 ChatGPT shows AI has enormous potential but is a black box issuing near-human statements
- AI relies on its training data (ground truth)

# and Assumptions (3)

- there is a need for persistent evidence on ground truth
- a society based on the rule of law and responsibility wants to know the base on which humans or AI made decisions – also over the long term



# Outcomes

- 2022 DIW Berlin proposal for an Open Data format for datasets replacing proprietary formats (STATA, SPSS, R)  
<https://git.soep.de/opensda/specification>
- 2023: SIARD format better adapted to Binary Large Objects (BLOBs)

# and Assumptions (4)

- improvement will come in small steps alongside technical evolution





**Two small steps**



















# Add-on for the 1961 census in 2022

## Structure navigation

Tektonics:

- D: Digital archive
  - A: Archives
    - State Archives of Ludwigsburg
      - Upper and Central Authorities since 1945
        - Ministry of Finance
          - Statistical State Office Baden-Württemberg:...
            - Census census

Go to AID:

 Signature/AID	 Title	Details
	Insert here:  representation	
219-1965	 Protocol to 2-1915280 Data from sum cards	
D A R 1	  Fixed width format	 <input type="checkbox"/>
D A R 2	  CSV format (partial representation)	 <input type="checkbox"/>
D A R 3	  CSV format with decrypted sorting features	 <input type="checkbox"/>
D A R 4	  SQLite	 <input type="checkbox"/>

# Simple manual on DB archiving

1. Document the original environment
  - Gather manuals, specifications, screenshots or screencasts. Make reviving the database a self-explaining “bootstrap” routine.
2. Continue with options:
  - Option A: Select essential entities
    - You can use existing reporting data.
    - You can create specific archival reports
  - Option B: Preserve DBMS performance
    - This can be done with proprietary dumps (e.g. Oracle, or SIARD, SQL 2008, or SQLite dumps).
  - 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.
    - good advice: use option C with A or B in parallel.
3. From then on, you keep the package, wait, and monitor it regularly.





**Thinking big**

# An alliance for database preservation?

## – Who joins?

- developers of long-term storage (DNA, nanomaterials)
- manufacturers of long-term goods (e.g. buildings, aerospace, defense)
- guardians of long-term structures (e.g. nuclear waste disposal agencies)
- scientists using long-term data (e.g. economics, geo-engineering, climate research)

## – Who facilitates?

- United Nations, China, EU, US?

## – Who are allies?

- Digital Preservation Coalition  
<https://www.dpconline.org/>
- Open Preservation Foundation  
<https://openpreservation.org/>
- Data Documentation Initiative  
<https://ddialliance.org/>
- European Open Science Cloud
- many national research data infrastructures
- large libraries and archives

## – When does it start?

- Depends on you ☺.

# And now?

- Engage!
- Read <https://doi.org/10.53458/books.109> and maybe watch the recordings mentioned therein
- Subscribe to mailing list  
<https://listserv.dilcis.eu/info/rdb-aig>

**Thank you!**

**Dr. Kai Naumann**

**Landesarchiv Baden-  
Württemberg**

Abteilung Archivischer Grundsatz

Urbanstr. 31A

70182 Stuttgart, Germany

Phone: +49 711 212-4284

Email: [kai.naumann@la-bw.de](mailto:kai.naumann@la-bw.de)

[@kai\\_naumann@mastodon.social](https://mstdn.social/@kai_naumann)

<https://orcid.org/0000-0002-2799-1030>

Web: [www.landesarchiv-bw.de](http://www.landesarchiv-bw.de)