Databases for 2080

Workshop Proceedings

Edited by Kai Naumann

Landesarchiv
Baden-Württemberg

Dialog Digital
Landesarchiv Baden-Württemberg
Band 2

Herausgegeben vom
Landesarchiv Baden-Württemberg

Typeset by: BookDesigns, Heidesee

Landesarchiv
Baden-Württemberg

urn:nbn:de:101:1-2022071903

Contents

Preface

Sustainability of data is an aim often neglected if you look at the money and effort invested in digital solutions in research, politics, and industry. In the last 15 years, database technology has continually progressed, performance has scaled and conferences on database technology have flourished. In contrast, there were very few dedicated events on how to secure efficient and effective access to database content for longer periods. Three workshops took place in the first decade of this century, at Berne in 2003 (Erpaworkshop 2003), Edinburgh (PresDB 2007), and Stockholm (van Horn 2007). Discussions continued at preservation conferences like the Digital Curation Conference, iPRES, or Preservation and Archiving Special Interest Group (PASIG) meetings. But it was after a long gap that the "Databases for 2080" workshop of 2021 once again addressed all aspects of this subject.

Society depends not only on database technology, but also on the ability to trust. Trust between humans is not only indispensable across space, but also across time. The signal sent by a certain person in 2020 must be clearly and unambiguously interpretable in 2080. Archivists at different places in the world first started to preserve database content in the 1970s. At the Landesarchiv Baden-Württemberg, we started only in 2002 and in 19 years, we have gathered over 300 million items in about 40 data series, mostly of statistical nature. The subjects include, among many others, museums, criminal justice, and environmental data. We are merely a tiny player in the world of data archives.

All data has been ingested in the DIMAG software suite that we started to use in 2006. Since 2010, we have been developing it with a community of development partners that now comprises over 200 archives in German-speaking countries. DIMAG is suitable for all kinds of digital objects. We have versatile and robust ingest tools and established methods to ensure authenticity. We are working on preservation planning tools that enable controlled format migration. But are our holdings still coded in the most suitable database formats? Did we sufficiently document them? Do we provide efficient ways for re-using database content? Many of these questions have been asked by colleagues worldwide and may be answered.

This is the reason for Kai Naumann's activities, as explained in the opening chapter. I would like to thank all presenters, minute takers, and discussion hosts, especially Kai and all other people at the Landesarchiv and worldwide who helped in preparing, hosting, and reporting this workshop.

Gerald Maier
President of Landesarchiv Baden-Württemberg

Scope, content, and advice for readers

This proceedings ebook contains short reports on all presentations, short presenter biographies, excerpts of the prominent presentation slides and, for some presentations, manuscripts.

The presentations were recorded at the workshop and the videos, alongside with presentations, can be viewed at the Landesarchiv's website (https://www.landesarchiv-bw.de/de/aktuelles/termine/72973 or search "DBs for 2080").

The first chapters deal with options and strategies, the later chapters with standardisation and application. The ebook closes with references, discussion results, and a conclusion.

- Readers unfamiliar with terms and acronyms may want to consult the glossary (p. 64).
- Readers unfamiliar with concepts should read the introduction first, but might want to dive deeper and read (Naumann 2021).
- German speakers might want to consult the very instructive article by Däßler and Schwarz (2010).
- The references include not only every citation mentioned in the text, but also other works that appeared as relevant in the preparation process.

Biographies of people who helped editing this publication are at the end of this book (p. 63). Presenters' biographies are below the titles of their contributions.

Introduction

Kai Naumann (Landesarchiv Baden-Württemberg, Stuttgart, Germany)

Kai Naumann is a historian and archivist at the Landesarchiv Baden-Württemberg, Stuttgart, Germany. He works in the fields of appraisal and transfer of digital and paper records, access and use, and archival law. He has been working on database preservation for 15 years. He is member of committee of the German Conference of State Archive Directors (KLA). He teaches database preservation classes at the Potsdam University of Applied Sciences.

Naumann started with a definition of a database as presumed in the workshop, a modular view that includes graphical user interfaces, the business logic, the database management system and, if necessary, the storage layer. The underlying technologies have been developing since the 1970s and are now very diverse (RDBMS Genealogy 2018, figure 1).

In April 2020, Naumann publicised a challenge to the community. A variety of solutions came from this call. The solutions included the use of CSV, XML, Disk Imaging, Docker, Web Crawler and a hybrid solution (the latter explained by Brigitte Mathiak, p. 9 and 10). In order to estimate the economic consequences of the different choices, he has mapped the solutions to a fictitious timeline up to the year 2080 that also showed the expectable costs for each solution (figure 2).

Naumann also reported the research published so far and mentioned the rather slow progress the SIARD format has made since its inception in 2007 (more by Kuldar Aas, p. 28).

He introduced the participants and pointed out that the workshop was intended to gather the worldwide state of the art in the field, thus helping institutions to make informed decisions in this area. In order to create an atmosphere for debate, the number of participants had been limited to under 50 people chosen as representatives of different professions and research areas.

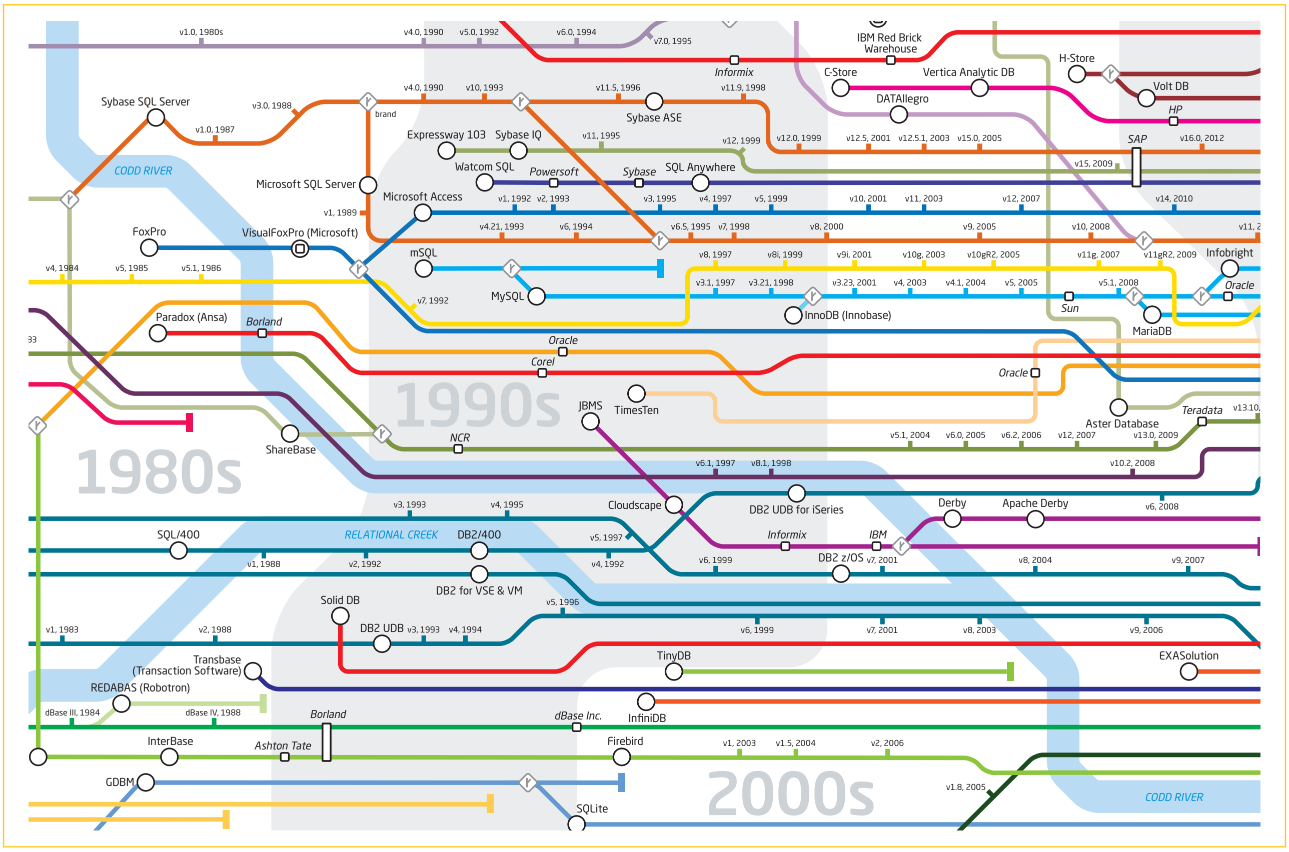
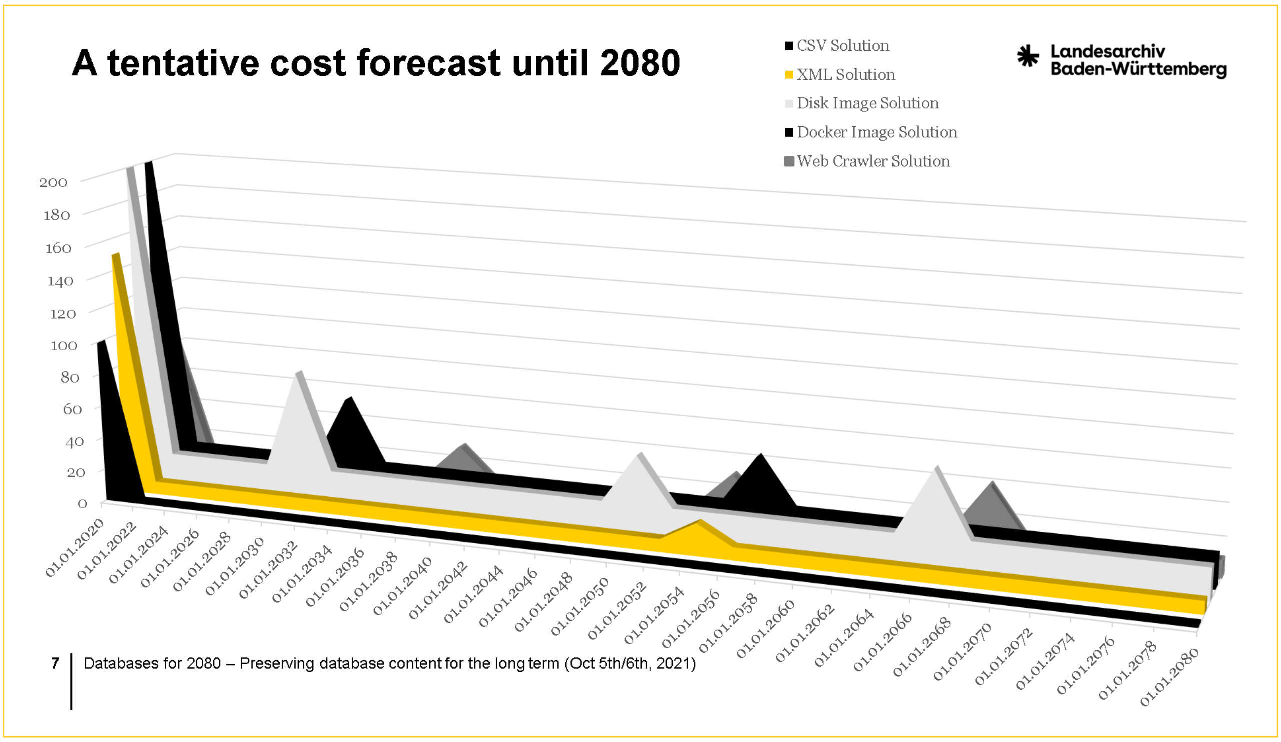


Figure 1: HPI Genealogy of Relational Database Management Systems (cutout)



A tentative cost forecast until 2080

■ CSV Solution
■ XML Solution
■ Disk Image Solution
■ Docker Image Solution
■ Web Crawler Solution

✱ Landesarchiv Baden-Württemberg

7 | Databases for 2080 – Preserving database content for the long term (Oct 5th/6th, 2021)

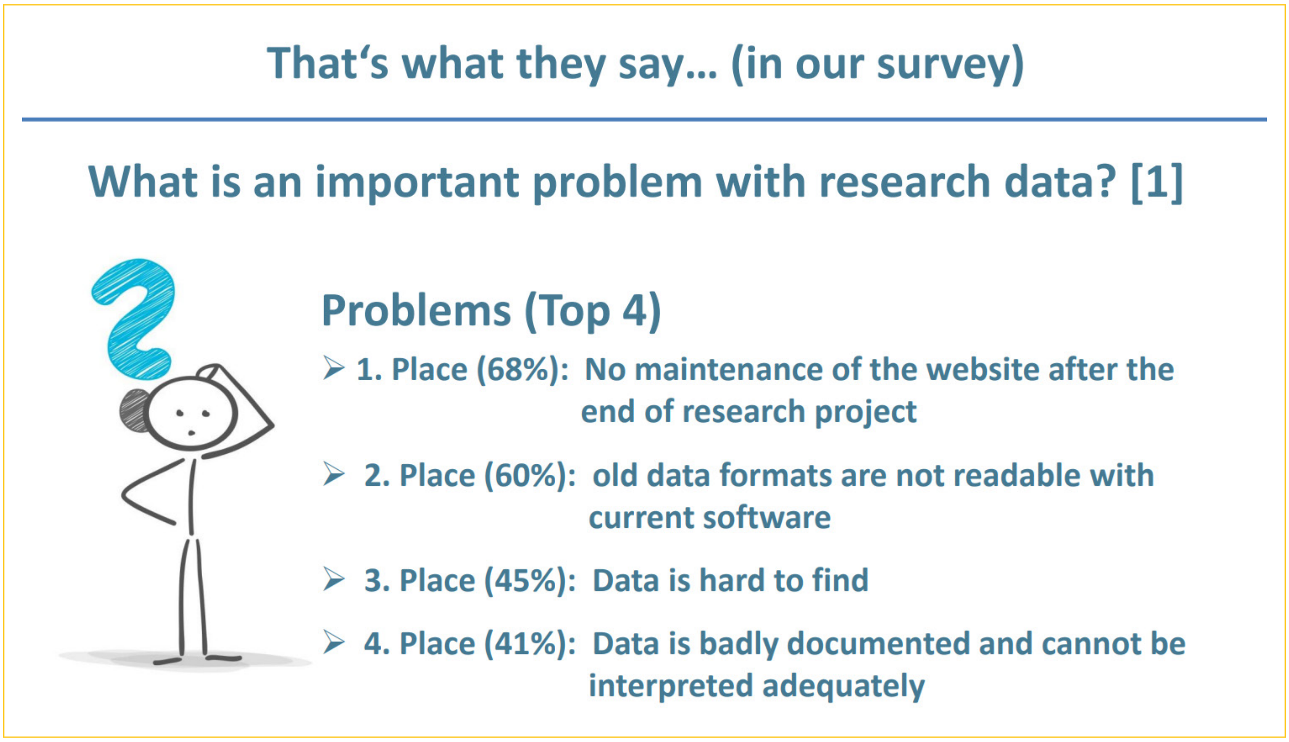
Figure 2: Cost forecast for different DB archiving solutions as estimated by Naumann (2021)

Sustainability strategies for digital humanities systems

Brigitte Mathiak (GESIS social science institute, Cologne, Germany)

Brigitte Mathiak is a senior scientist at GESIS (German institute for the social sciences). Before, she held a Junior Professorship for Digital Humanities at the University of Cologne and was speaker of the Data Center for the Humanities (DCH). In this position, she studied the sustainability of Digital Humanities projects as part of the DFG (Deutsche Forschungsgemeinschaft, German Research Foundation) project SustainLife. She holds a diploma in informatics.

“Why is sustainability an issue in digital humanities?” Mathiak and colleagues asked humanities scholars this question (Neuefeind 2020) and the most frequent answer was that there was no maintenance of websites at the end of their research projects. The average lifespan of digital scholarly editions is 8.5 years – in contrast to books, which survive for centuries without attention. But what is a digital scholarly edition? It is a book with annotations but online, usually text based. These websites often have a database or an XML repository to store the raw data, but interactive components such as search functions, text statistics and translations, and visualisation on top of the raw data are constructed in many different ways. All these requirements make interactive components extremely hard to preserve.



That‘s what they say… (in our survey)

What is an important problem with research data? [1]

Problems (Top 4)

- 1. Place (68%): No maintenance of the website after the end of research project
- 2. Place (60%): old data formats are not readable with current software
- 3. Place (45%): Data is hard to find
- 4. Place (41%): Data is badly documented and cannot be interpreted adequately

Figure 3: Results of the Kronenwett/Mathiak 2017 survey (see references)

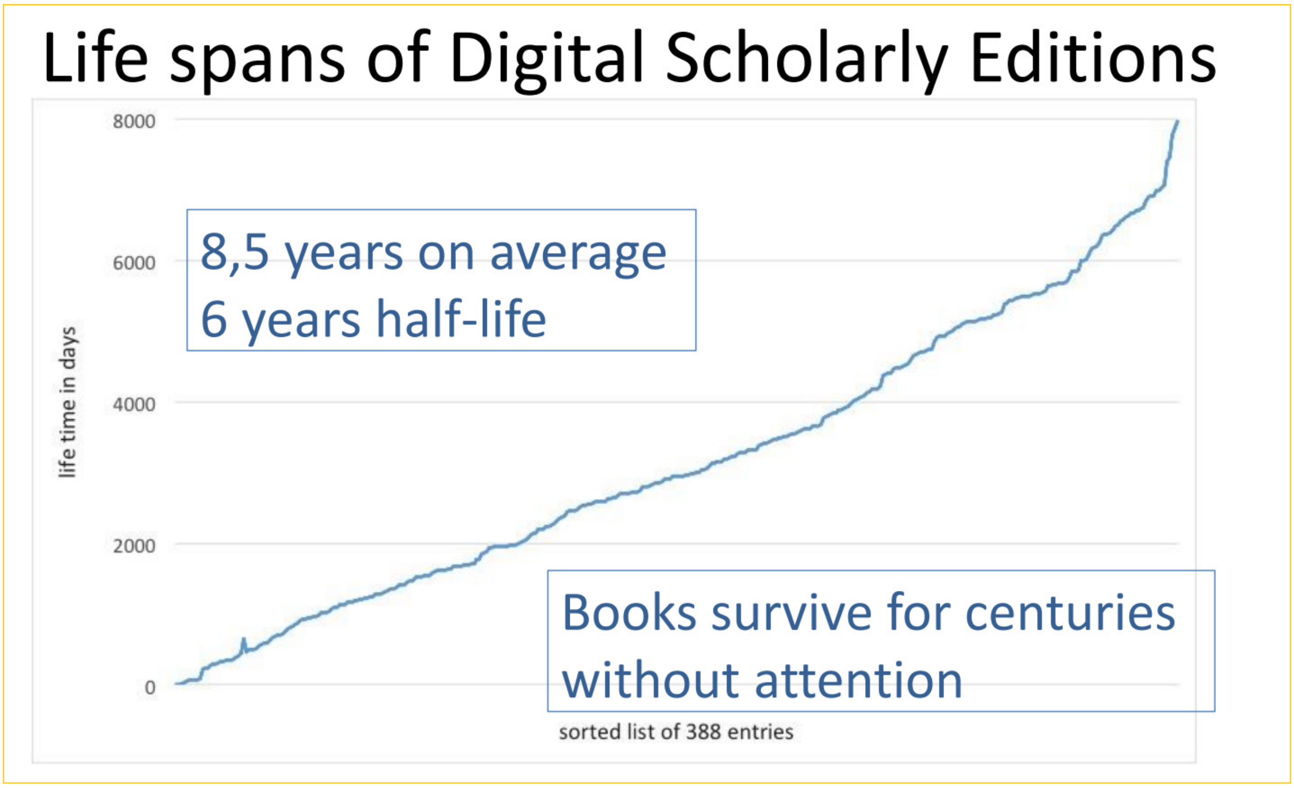


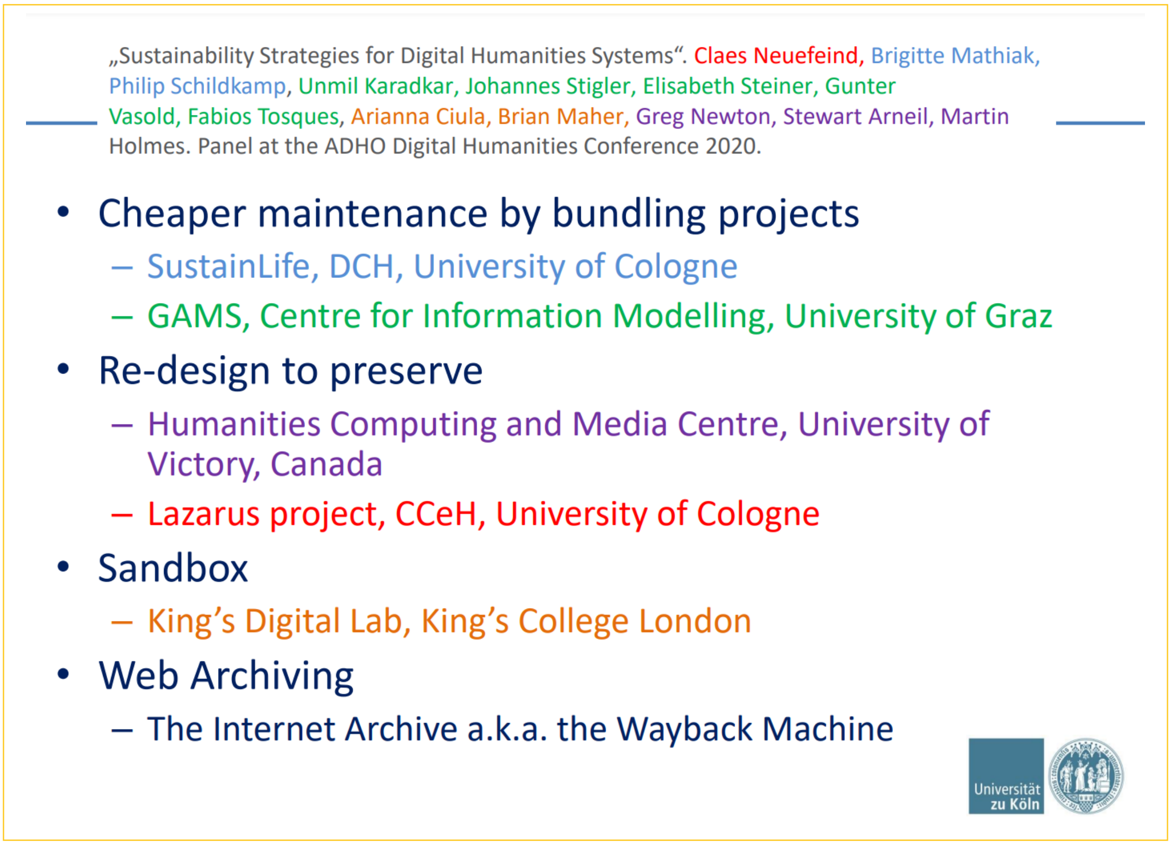
Figure 4: Life spans of Digital Scholarly Editions (Neuefeind 2020)

Mathiak outlined several projects that deal with this problem (see figure 5). Some try to sustain the scholarly editions as living systems, reduce them to a smaller application with only standard access methods or move the old systems to newer platforms. These methods can be used retroactively on legacy projects, they are relatively quick to learn and to apply. Their application does improve maintenance costs significantly but inevitably re-design or shutdown is necessary to reduce security risks.

Other projects try to steer scholars away from standalone custom software and toward more modular, standards-based software frameworks like the Austrian GAMS (see references) to implement their digital assets which then will be maintained through the central system.

An interesting crossover solution is having the Internet Archive (https://archive.org) harvest a scholarly edition with adapted settings. Simultaneously, the database management system's content is preserved, which makes it possible to tentatively reconstruct the whole application based on its data and the application in the future. This approach involves low costs in the beginning.

Sustainability of digital projects is a challenge. It requires adequate funding to keep them fully functional. Cheaper options are web archiving (may only cost 5 minutes of your time) or a static HTML page (with or without Javascript), but not all projects can be transformed without loss and findability of the resource suffering.



„Sustainability Strategies for Digital Humanities Systems“. Claes Neuefeind, Brigitte Mathiak, Philip Schildkamp, Unmil Karadkar, Johannes Stigler, Elisabeth Steiner, Gunter Vasold, Fabios Tosques, Arianna Ciula, Brian Maher, Greg Newton, Stewart Arneil, Martin Holmes. Panel at the ADHO Digital Humanities Conference 2020.

- Cheaper maintenance by bundling projects
  - SustainLife, DCH, University of Cologne
  - GAMS, Centre for Information Modelling, University of Graz
- Re-design to preserve
  - Humanities Computing and Media Centre, University of Victory, Canada
  - Lazarus project, CCeH, University of Cologne
- Sandbox
  - King’s Digital Lab, King’s College London
- Web Archiving
  - The Internet Archive a.k.a. the Wayback Machine

Figure 5: Mathiak’s comparison of sustainability strategies. Representatives for each project associated by colour coding.

It is possible to use a combination of measures including tombstone pages[1] and archiving of the data and the code to mitigate this. These methods can be used retroactively on legacy projects. They are relatively quick to learn and apply and thus reduce maintenance costs significantly. But inevitably re-design or shutdown is necessary to reduce security risks.

Mathiak concluded that a multi-layered approach (e.g. King’s Digital Lab, see references) and prefabricated environments like GAMS (see references) are good ideas. She also stressed that the people working in the projects are a key factor for preservation efforts and adequate funding is necessary. A good approach would be to think about individuals who might want to access the resource in 100 years’ time (at the point of creation/design). It can be more helpful to think about them rather than current users.

Questions and discussion

Audience commented that these approaches are very focused on web archiving, but it should be noted there are other ways of tackling database preservation.

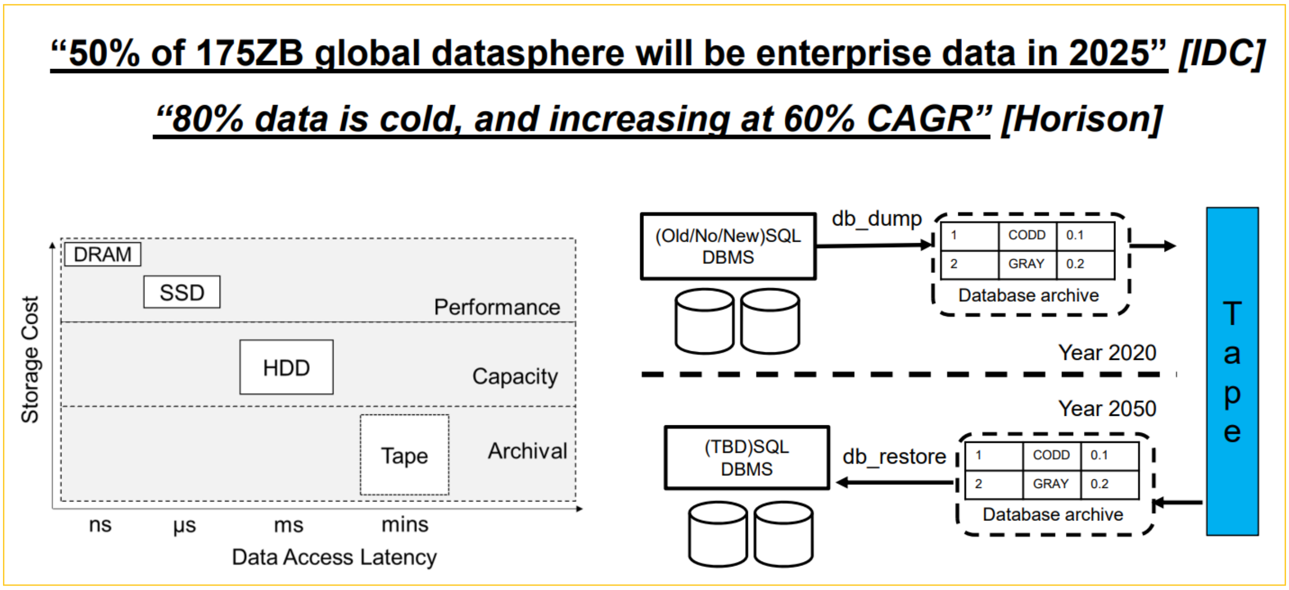
[1] The term is used metaphorically for URLs signaling a former presence of web content that has been removed or transferred to another URL.

Storing and reviving databases on DNA

Raja Appuswamy (EURECOM research institute, Sophia Antipolis, France)

Raja Appuswamy is currently working as an Assistant Professor at EURECOM. Previously, he worked as a Visiting Professor at EPFL, as a Visiting Researcher in the Systems and Networking group at Microsoft Research, Cambridge, and as a Software Development Engineer in the Windows 7 kernel team at Microsoft, Redmond. He received his PhD in Computer Science from the Vrije Universiteit, Amsterdam, where he worked under the guidance of Prof. Andrew S. Tanenbaum on designing and implementing a new storage stack for the MINIX 3 microkernel operating system. He also holds dual Masters degrees in Computer Science and Agricultural Engineering from the University of Florida.

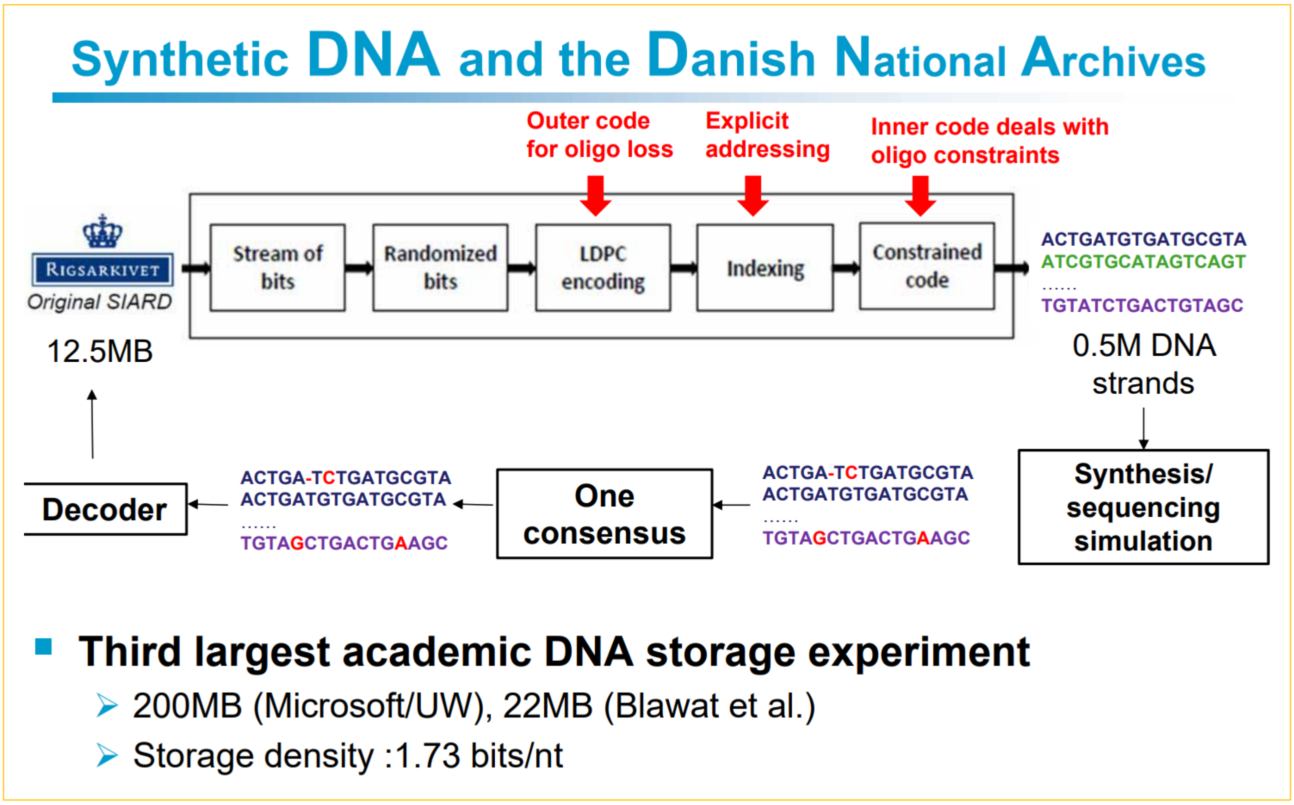
Appuswamy cited recent studies showing that nearly 80 percent of all stored data are ‘cold’ (infrequently accessed) and this number is increasing over time (Memishi et al. 2019). Data often needs to be stored and kept for legal compliance reasons, typically on magnetic tape. There are problems with these tapes, because tape vendors typically only support backwards compatibility for two generations. This poses problems if large collections need to be moved to newer tape formats, which hold especially true for audiovisual data (Perlmutter 2021).



“50% of 175ZB global datasphere will be enterprise data in 2025” [IDC]

“80% data is cold, and increasing at 60% CAGR” [Horison]

Figure 6: Scenarios on growth of data and proportion of “cold” data.



Synthetic DNA and the Danish National Archives

- **Third largest academic DNA storage experiment**
  - 200MB (Microsoft/UW), 22MB (Blawat et al.)
  - Storage density :1.73 bits/nt

Figure 7: The DNA4DNA collaboration

Magnetic media is only one possibility for storing data. DNA (desoxyribonucleic acid) is another option. It is a molecule made from four different nucleotides which can be sorted into one big string. DNA storage refers to single strand DNA manufactured and stored outside the body – no living beings are involved.

Sequencing technology to read the DNA is available. Storage density on DNA is much higher than on tape. DNA is also very durable, can survive for a long time (thousands of years) and in harsh conditions. The project Oligoarchive focuses on using DNA as an intelligent storage medium. The project is looking at how DNA can be used to store files.

There are economic limits today, as writing DNA is very costly, while reading devices are already becoming affordable. In spite of this, the application of the concept to databases has been tested. The Danish National Archives have encoded a 12.5 megabyte SIARD database file to DNA and reread it successfully (Marinelli et al. 2021).

The OligoArchives project cooperates with the company EUPALIA that has long experience with emulation and traditional persistent data carriers like microfilm.

There are some complications.

- The need to encode the data. The data in the DNA needs to store metadata too. The metadata and the data can be decoded into the data for the database. DNA synthesis is very expensive (10^7 times more expensive than tape) currently. DNA does not solve media or format obsolescence.
- The need to preserve the decoder (and its mechanisms) by self-explaining instructions.
- For decoding a DNA data carrier, you have to inject a liquid and by that destroy it. DNA today is a write-once, read-once medium. This shortcoming can be overcome by keeping several samples of a data object (tiny physical particles) in one data container and by re-writing new copies at certain intervals.

Questions and discussion

- The strong need for a comprehensive way of declaring the content and meaning of an arbitrary byte sequence extracted from DNA might induce the Big Data industry to work on improvements in the database archiving sector.
- Having “bootstrap routines” that help reviving data out of its own metadata might become commercially important.
- DNA storage researchers thus welcome academic and industry people who want to get involved.

Database preservation for industry customers

Florian Hartl (CSP GmbH & Co KG Company, Munich, Germany)

Florian Hartl is a software developer. He has been doing archiving projects for the automotive industry for over 8 years now. He is working with CHRONOS, a software from his employer CSP. Currently he is working as a key account manager and does project management for archiving projects.

Hartl did not bring a slide presentation but talked and showed a specimen of network connected automotive production tool: a screwdriver. Equipment used in the automotive industry produces a lot of data. This data may need to be kept for a lifespan of usually 30 years. The standard retention period for compliance reasons in this business is 25 years, but it is also good to have a time buffer of five additional years. The industry does not necessarily know what they need to do with the data in the future, but it is needed for legal reasons in case there is an accountability problem with a car in the future. What is important to them?

- Safe storage – stored not just in one place but all over the world
- Revision security – write once, read many – no feature for changing or deleting the files. Important that they can guarantee they have not changed the data.
- Documentation and knowledge – they must keep everything together in the documentation so that a future user has all the knowledge they need to analyse and understand the data.

Why is knowledge so important? Hartl has worked with old COBOL files and it was challenging, as the knowledge was not preserved with the data. Nobody could tell him anything about the data and it was created 10 years before he was born.

The Diesel Emissions Scandal triggered more demand for proving that data collected during the production period had not been manipulated at any later stage. Hartl's company CSP has thus grown from 30 to 120 people due to increasing demand from the automotive industry. CSP expects there will be more data in the future. The data that the depicted screwdriver type generates in a 15-year lifetime is about 15 terabyte. In the next five years, they expect screwdrivers to generate more than 15 terabyte per year.

Archiving with CHRONOS progresses in four steps:

- The first step is knowledge – gather producers and IT people and find out what they need. You must talk to multiple people in factories and the IT sector to find out which data is important to be collected and stored.
- If you have done a good identification phase, the archiving runs smoothly. CHRONOS uses a format that is similar to SIARD (Fitzgerald 2013). It would be simple to map between them.

- Then the data in the DBMS that were archived in the process are being replaced by links to the archive.
- After 30 years, long-term preserved data is deleted completely from the archives.

Data are split up into data files and metadata. Hartl said his company was not keen on encrypting data, as it is not easy to use the data without the ability to decrypt. Plain text files are best for storing the data. Hash values are also kept seeing if the file changes. CSP has a system that allows staff to check this on the fly.

CHRONOS is also used in the financial sector and other industries. Hartl concluded with a side-glance on the storage systems behind CHRONOS. The company uses a storage system called EfficientNodes that prevents changes on the hardware side based on hash values.

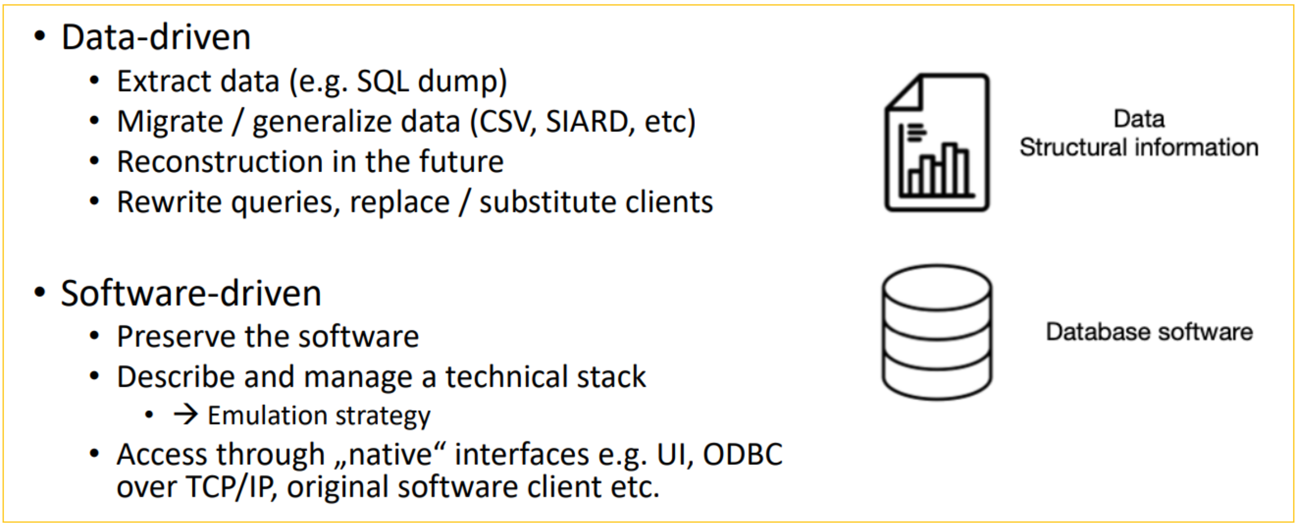
Figure 8: Industry screwdriver as used in the automotive industry.

Emulation options for legacy databases

Klaus Rechert (University Freiburg i.Br., Germany)

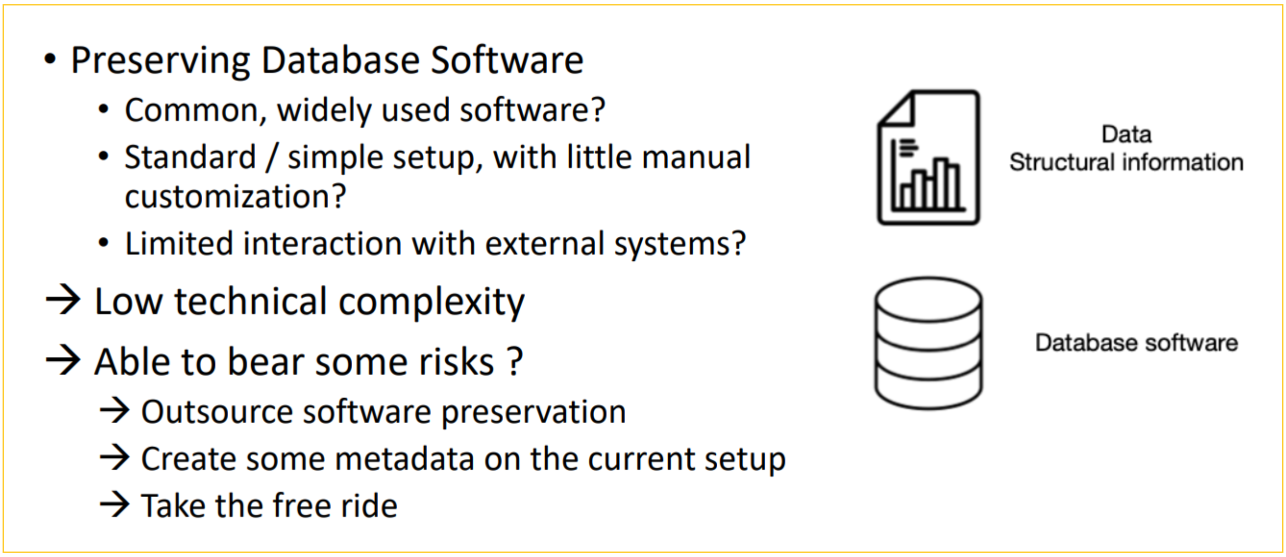
Klaus Rechert recently became professor at the University of Applied Sciences Kehl. In October 2021, he was researcher and professor at the University of Freiburg. He was the principle investigator of bwFLA (Baden-Württemberg Functional Long-Term Archiving and Access) and has been the architect behind Emulation as a Service. He is involved in multiple national and international projects related to digital preservation, reproducible science and research data management. He was awarded a Diploma in Computer Science in 2005, and a doctoral degree in 2013 from the University of Freiburg.

Rechert started his presentation by citing the challenge of a safe and economic preservation over many decades. He reminded the audience of a „Database" being a weakly defined term, from a Computer Science perspective (hence a definition given by Naumann, p. 6). Is it possible to plan this far ahead and what are the risks? There are two basic strategies: either data-driven – focused on extracting the data, or software-driven – focused on the technical stack. There is also a middle ground between the two.



- Data-driven
  - Extract data (e.g. SQL dump)
  - Migrate / generalize data (CSV, SIARD, etc)
  - Reconstruction in the future
  - Rewrite queries, replace / substitute clients
- Software-driven
  - Preserve the software
  - Describe and manage a technical stack
    - → Emulation strategy
  - Access through „native" interfaces e.g. UI, ODBC over TCP/IP, original software client etc.

Figure 9: The data-driven and the software-driven strategy



- Preserving Database Software
  - Common, widely used software?
  - Standard / simple setup, with little manual customization?
  - Limited interaction with external systems?

→ Low technical complexity

→ Able to bear some risks ?

→ Outsource software preservation

→ Create some metadata on the current setup

→ Take the free ride

Data
Structural information
Database software

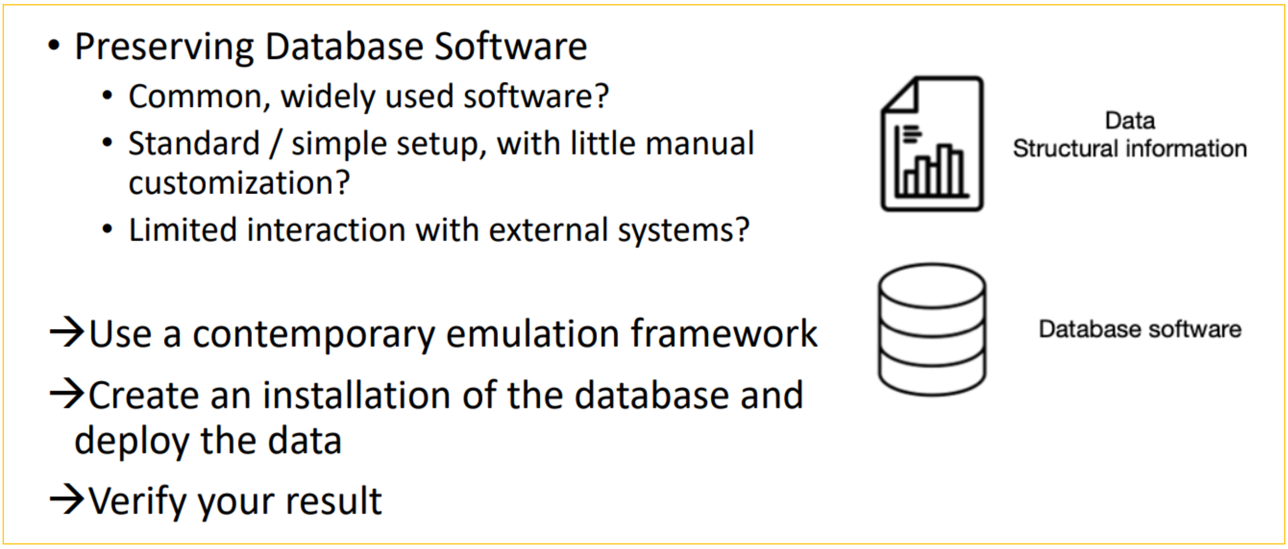
Figure 10: The (lazy) strategy

Rechert referred to a software-driven strategy that mostly relies on emulation. There are several subtypes:

- A lazy strategy – if you have a widely used database software already set up with little manual customization, you can create metadata around the dump and take a free ride, hoping for an emulation solution to be around on the day the data are revived. However, this is risky.
- A less lazy strategy is to preserve the database software and use a contemporary emulation framework. This is a better way of ensuring that it will work in the future.

The technical stack for the less lazy strategy (figure 11) comprises the:

- data,
- database software,
- configured environment,
- virtual hardware, and
- contemporary CPUs and connectors to machines and users.



- Preserving Database Software
  - Common, widely used software?
  - Standard / simple setup, with little manual customization?
  - Limited interaction with external systems?

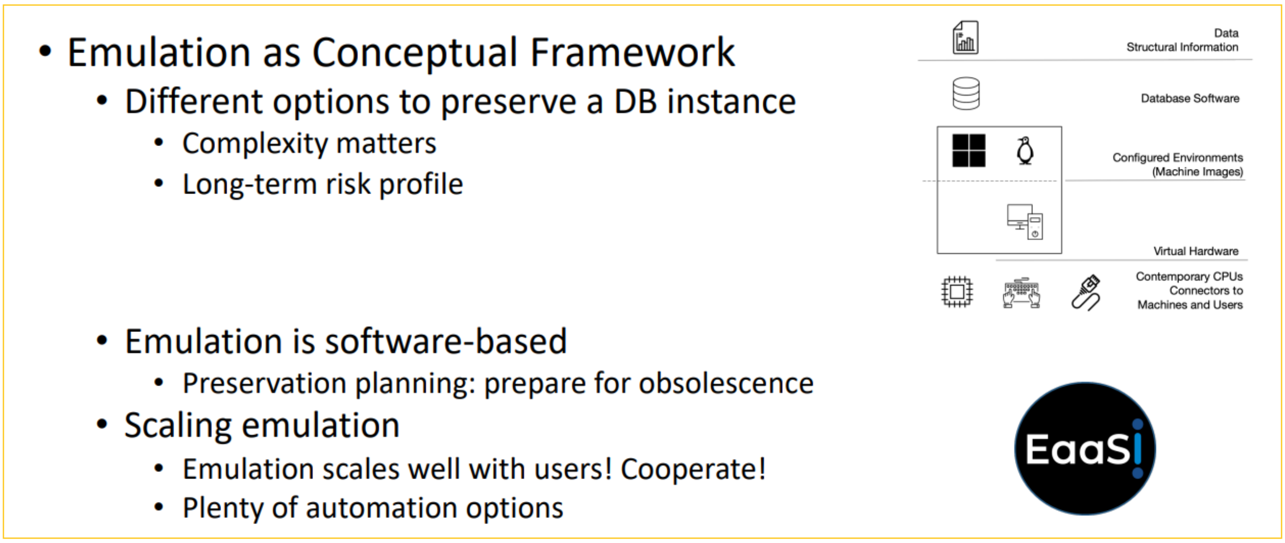
→Use a contemporary emulation framework

→Create an installation of the database and deploy the data

→Verify your result

Data
Structural information
Database software

Figure 11: The (less lazy) strategy



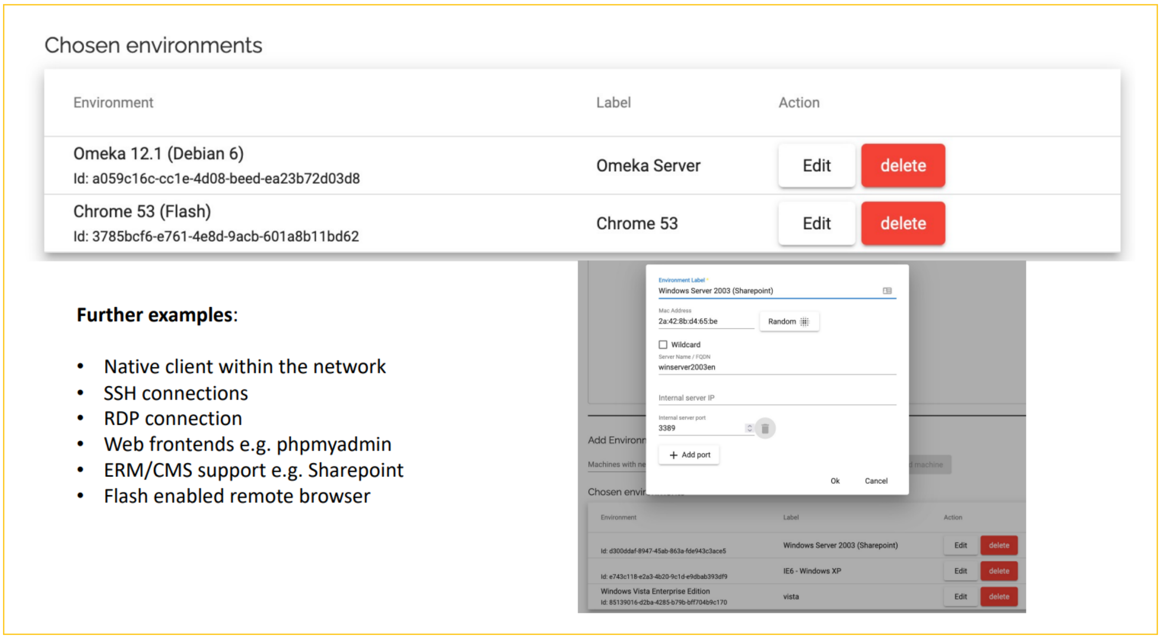
- Emulation as Conceptual Framework
  - Different options to preserve a DB instance
    - Complexity matters
    - Long-term risk profile
  - Emulation is software-based
    - Preservation planning: prepare for obsolescence
  - Scaling emulation
    - Emulation scales well with users! Cooperate!
    - Plenty of automation options

Data
Structural Information
Database Software
Configured Environments
(Machine Images)
Virtual Hardware
Contemporary CPUs
Connectors to
Machines and Users
EaaSI

Figure 12: Conclusions on emulation

Complex setups may require manual or installation steps. If you have Microsoft Office Sharepoint and you know the version of Sharepoint, you can create a simple solution using a template, as this is a common scenario. Suppliers only need to ask the user a few questions. This works well if you can reuse the stacks. You can automate to run emulation at scale – not hard for someone within the community to create scripts.

Rechert pointed out the importance of keeping the data separate from the image. A very complex setup may need full system preservation, which means to preserve the full virtual machine with a disk image. One needs to think about dependencies. In this case, his company may also be able to rebuild a machine from a file system or backup.



Chosen environments

Environment	Label	Action
Omeka 12.1 (Debian 6) Id: a059c16c-cc1e-4d08-beed-ea23b72d03d8	Omeka Server	Edit delete
Chrome 53 (Flash) Id: 3785bcf6-e761-4e8d-9acb-601a8b11bd62	Chrome 53	Edit delete

Further examples:

- Native client within the network
- SSH connections
- RDP connection
- Web frontends e.g. phpmyadmin
- ERM/CMS support e.g. Sharepoint
- Flash enabled remote browser

Environment Label *
Windows Server 2003 (Sharepoint)

Mac Address
2a:42:8b:d4:65:be

□ Wildcard

Server Name / FQDN
winserver2003en

Internal server IP

Internal server port
3389

Environment	Label	Action
Id: d300ddaf-8947-45ab-863a-fde943c3ace5	Windows Server 2003 (Sharepoint)	Edit delete
Id: e743c118-e2a3-4b20-9c1d-e9dbab393df9	IE6 - Windows XP	Edit delete
Windows Vista Enterprise Edition Id: 85139016-d2ba-4285-b79b-bff704b9c170	vista	Edit delete

Figure 13: Options for emulated network environments with Stabilize.

He saw several options of how to preserve the software stack. If you follow an emulation strategy for preservation, you must prepare for obsolescence. The good thing is that everyone has the same problem so it is a safe bet that we will see other emulators in the future. A key issue is to preserve the images. Another important observation is around complexity. The complexity and risk profile should guide any option that you take, whether a template-based or full system emulation is deployed. Emulation scales well. The more users, the more people are available to cooperate with (see EaaSI in references).

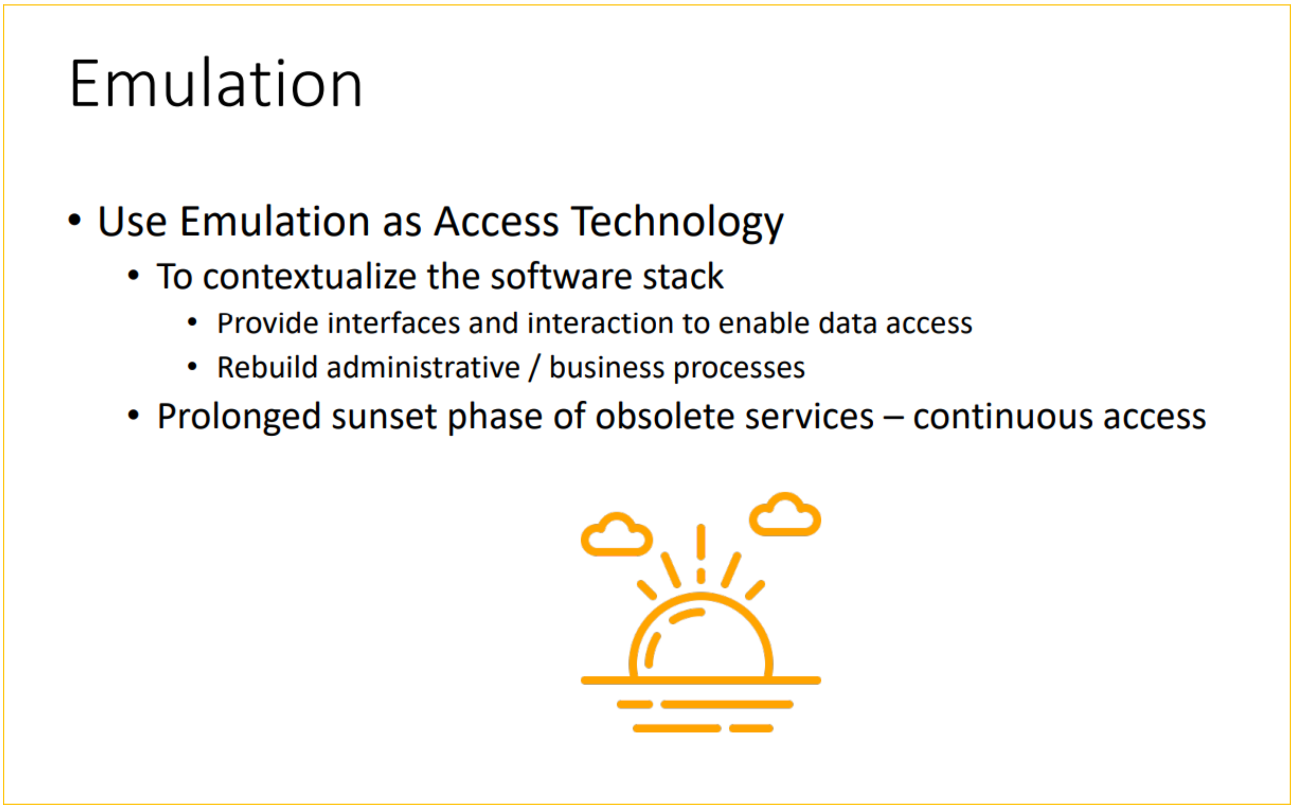
Emulation is also useful as an access technology. Here, Rechert introduced Stabilize (see references). Emulation allows you to prolong the sunset phase of services. Once you have a software stack you can set up an emulated network. Users need to provide IP address and port info as appropriate, and can “fire up” a machine. An example shown was running on an SQL Server.

Emulation is a software-focused strategy – the complexity of the setup and anticipated risks are key to implementation. Emulation is a tool to offer a prolonged (endless) sunset phase of obsolete services. Operating obsolete software systems remains a huge non-technical, especially legal, problem but there are some technical solutions possible.

Questions and discussion

- Is emulating databases running on mainframes feasible – for example COBOL? How to access those mainframes? Where is the extracted data stored? – Rechert thought mainframe technology mostly didn’t need emulation since the data contained was mostly quite simply organised and could be best preserved via format migration.

- Will Oracle or their successors care for users of their obsolete DBMS versions in 60 years' time? Will they allow users to use their intellectual property at all? – Rechert said this shouldn't be too much of a problem if companies realise that they have no market to sell that product anymore.
- Emulating the whole database is only available for a very small subset of database engines.
- How do you even connect to the old database? New official database drivers refuse to connect to the old versions. The same holds true for ODBC (32-bit version only, for example).
- How do you emulate specific versions of mainframe DB2, for example? Or Oracle? Or Amazon Web Services (AWS) DynamoDB?
- Will someone have data in the same virtual machine? What if there will be a cluster?
- What about the licensing? What about the cost? What if the price models change?
- Databases vary wildly even between versions of the same databases (for example hashing functions change in MySQL).
- Rechert was also uncertain about the survival rate of concrete DBMS versions – people must prepare for obsolescence. If emulation solutions will be around, it would solve the problem for the next 10-20 years. Afterwards, there need to be people who will build the stack to wrap this into a new framework. This would have to be done every so many years.
- One conclusion was that emulation is a very interesting idea, but often hard to implement.



Emulation

- Use Emulation as Access Technology
  - To contextualize the software stack
    - Provide interfaces and interaction to enable data access
    - Rebuild administrative / business processes
  - Prolonged sunset phase of obsolete services – continuous access

Figure 14: Emulation as contextualising and prolonging technology

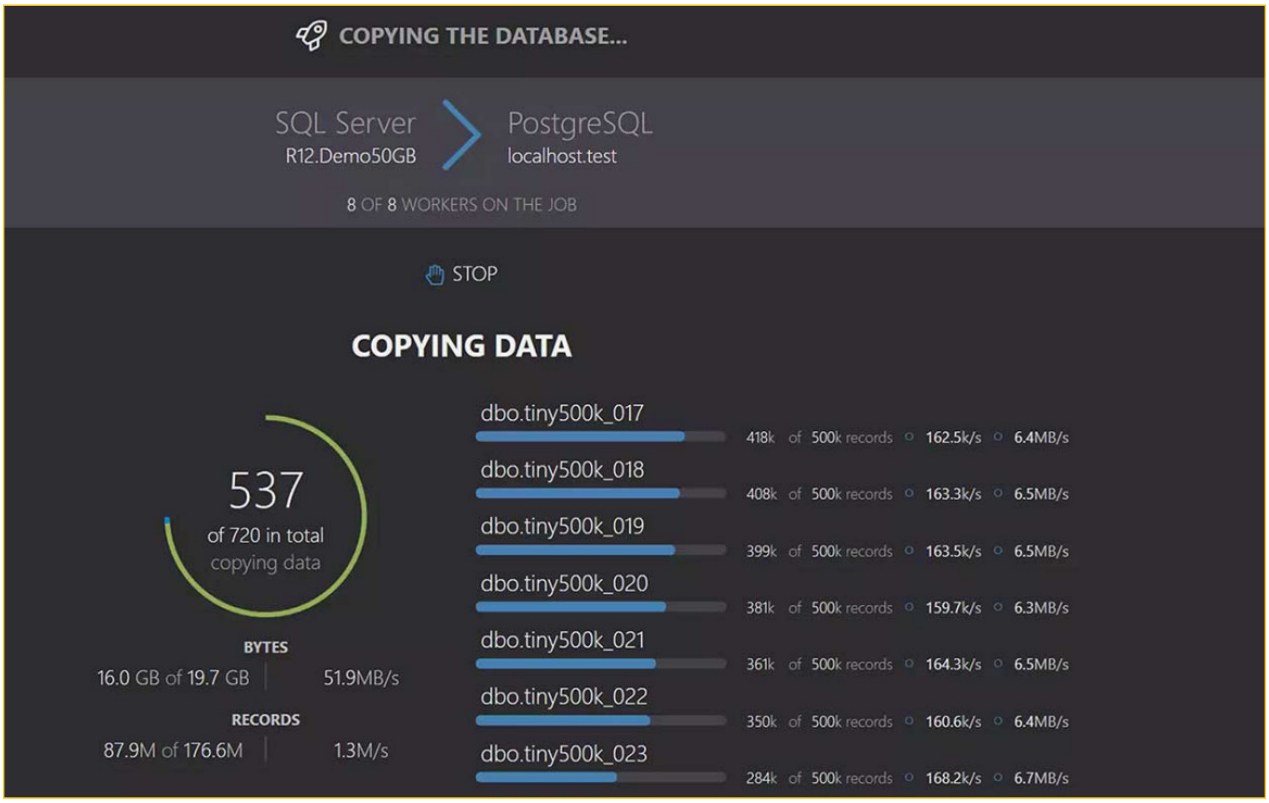
Database preservation through conversion technology

Damir Bulic (Spectral Core Ltd Company, Dublin, Ireland)

Damir Bulic has 30 years of experience building software in a wide variety of languages, frameworks, and environments. He is the owner of Spectral Core, a company specializing in database migrations with customers in more than 90 countries. For the past 20 years, his focus has been on database tooling. His interest is in removing vendor lock-in, SQL parsing, analysis, and transformation.

Bulic's talk continued the topic of best current practice. His company helps other companies in copying data from one place to another. They support around 40 database formats and big data, migration to the cloud or data lakes. The conversion software is called "Full Convert". It supports, among many other formats, transformation into SIARD from 40 different database management systems.

The enterprise level product is called "Omni Loader" and is a distributed migration cluster. This is an ideal solution for migrating databases on premises to the cloud. It handles hundreds of terabytes similarly to "Full Convert" on a single computer for the basic use case.



SQL Server
R12.Demo50GB
PostgreSQL
localhost.test

8 OF 8 WORKERS ON THE JOB

COPYING DATA

537
of 720 in total
copying data

BYTES
16.0 GB of 19.7 GB | 51.9MB/s

RECORDS
87.9M of 176.6M | 1.3M/s

dbo.tiny500k_017 418k of 500k records ○ 162.5k/s ○ 6.4MB/s
dbo.tiny500k_018 408k of 500k records ○ 163.3k/s ○ 6.5MB/s
dbo.tiny500k_019 399k of 500k records ○ 163.5k/s ○ 6.5MB/s
dbo.tiny500k_020 381k of 500k records ○ 159.7k/s ○ 6.3MB/s
dbo.tiny500k_021 361k of 500k records ○ 164.3k/s ○ 6.5MB/s
dbo.tiny500k_022 350k of 500k records ○ 160.6k/s ○ 6.4MB/s
dbo.tiny500k_023 284k of 500k records ○ 168.2k/s ○ 6.7MB/s

Figure 15: Spectral Core Full Convert software at work

Another Spectral Core product to mention is “SQL Tran” which allows for data definition language (DDL) translations of very complex schemas. SQL Tran extracts complex data from storage containers and copies it somewhere where you have full control. This feature is also important to prepare for obsolescence. Also, there is “Documenter”, a tool for database schema documentation. Spectral Core feels their software for working with SIARD is very effective.

Bulic then shifted to a critical review of SIARD. According to him, SIARD is useful now but will not be useful in five years’ time, as databases are getting larger and more complex. He named the following challenges for SIARD:

- It is XML and zipped
- A single file that cannot be serialised
- Hierarchical text format
- Verbose and clunky
- Useful only for small datasets
- Cannot handle much of what we see in the real world today. Datasets are growing at fast pace and SIARD will be less applicable in future.
- Full convert can handle up to 16 connections, but ZIP writing cannot be paralleled.
- SIARD cannot be supported in Omni Loader (most powerful database migration software they sell).

Looking to the future, Bulic predicted the end of the concept of vertical scaling, and Moore’s law (observation that the number of transistors in a microchip doubled about every two years, see references) being no longer applicable. According to him, horizontal scaling is the answer, meaning a division of calculation tasks in separate workloads that are executed on multiple machines and reunited for the result. The cloud is all about horizontal on-demand scaling and this trend will continue. Workflows are moving away from personal computers.

A new possible format Bulic proposed for archival data should have a separate structure for data and schema, highly compressed chunks of columnar data (10 times better compressed), and extensible data types. A perfect archival format should allow for searches and should not have to be all in one file. It could in fact be distributed across the world. It could probably be based on SQLite as this is well used everywhere. SQLite can support billions of records.

Questions and discussion

- Faria noted that the SIARD standard is already developing in some of the areas mentioned, for example how it could be segmented into different files. SIARD supports several Terabytes currently. He saw an importance to view and understand the tradeoff between performance and suitability for long-term preservation. There are benefits in having a self-documenting file without reliance on a technology stack. Performance may decrease but the benefits are around interoperability, simplicity, and the ability to use it in different contexts.

- Rechert once again got back to the purpose of preservation. Sometimes you have logic encoded into the front end. With complex databases and systems, you need to spend some effort to reconstruct it so that future users get the same results when they run the same queries etc. Every approach has its merits, but he still reminds us to think about what is needed 60 years from now. Bulic noted that there needs to be some sort of ISO format that is readable and efficient for the future.
- Appuswamy compared the discussions on Bulic's SIARD successor to the format used in DNA storage – it needs to be self-describing and compressed. He suggested the term "binary SIARD". Faria noted we have a binary SIARD once it is zipped. There may be room for improvement, Appuswamy thought.
- Incremental backup is very common in databases, Appuswamy noted. How do we handle this with emulation and other solutions?

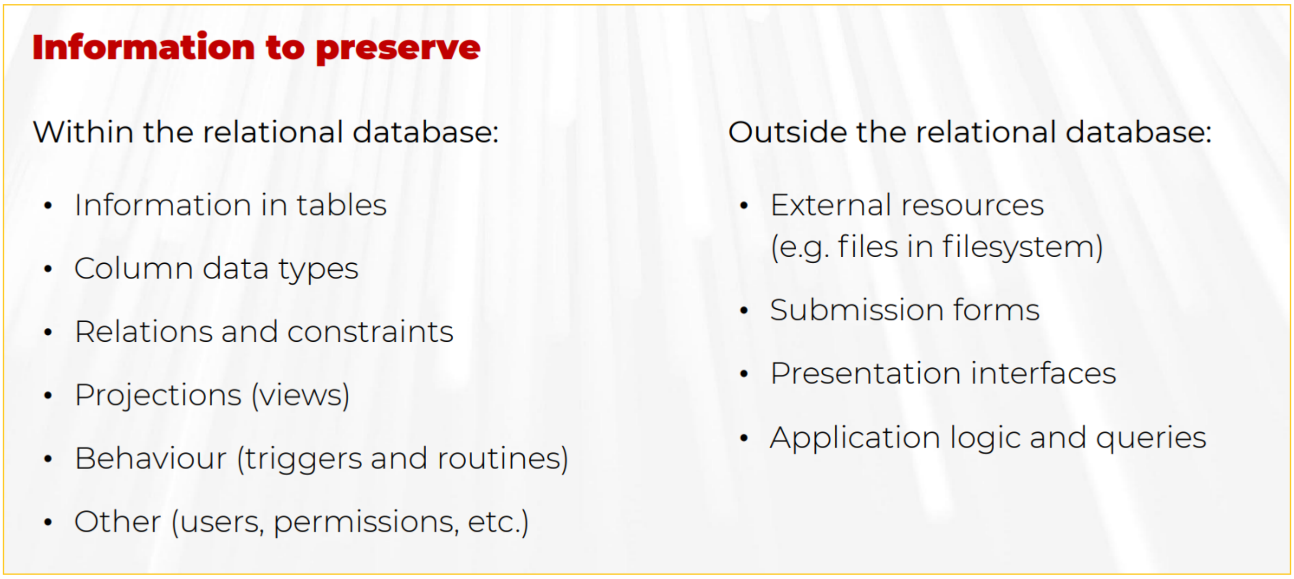
Preservation through standardisation with DBPTK

Luis Faria (KEEP Solutions LDA Company, Porto, Portugal)

Luis Faria is Research and Innovation Director at KEEP SOLUTIONS, working in research and development of solutions for digital preservation and information management since 2005. He holds a PhD in Computer Science with specialization in Digital Preservation by the University of Minho, has done his degree in Computer Science at the same University in 2005. Has participated in several research and development projects in the area of digital preservation, such as SCAPE, E-ARK, 4C and VeraPDF. He is co-author of preservation formats specifications SIARD 2 and E-ARK IP, and is manager of the open-source project RODA and Database Preservation Toolkit (DBPTK). Faria has been working on the challenge of database preservation for the last eight years, particularly looking at preservation through standardisation.

There is a lot of different information that may need to be preserved (see figure 16). Every database preservation strategy has advantages and disadvantages.

- Hardware and software museums have reproduction accuracy but are difficult to maintain
- Emulation also has good accuracy and an advantage of not needing to maintain hardware but is also difficult to maintain and set up and needs users to understand old systems.



Information to preserve

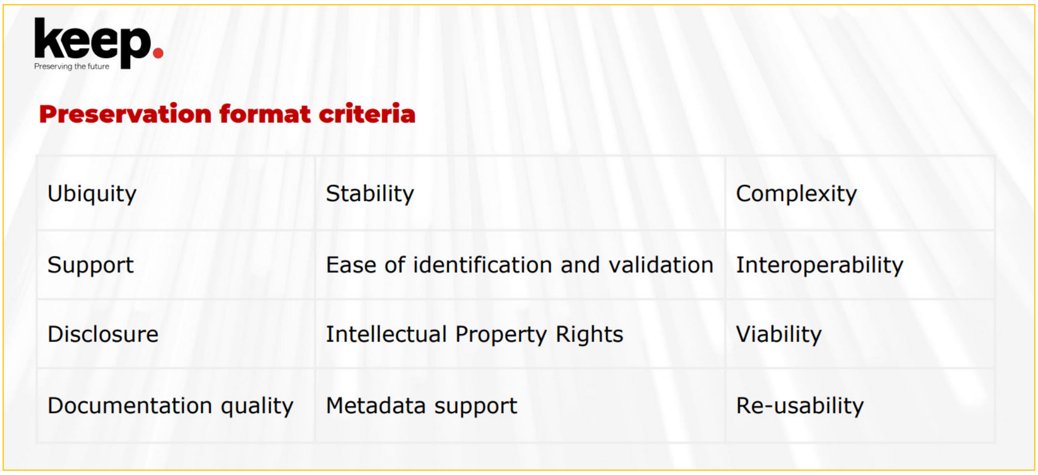
Within the relational database:

- 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

Figure 16: Information to preserve within and outside a database management system (DBMS)



keep.
Preserving the future

Preservation format criteria

Ubiquity	Stability	Complexity
Support	Ease of identification and validation	Interoperability
Disclosure	Intellectual Property Rights	Viability
Documentation quality	Metadata support	Re-usability

Figure 17: Preservation format criteria according to Brown (2008) as presented by L. Faria

- File format migration makes it easier to use and reuse information, and there is no need to maintain hardware or software. The risk is that information may be lost in migration.
- Encapsulation keeps files together with all necessary documentation. This can postpone costly actions and means no need to keep hardware and software. Disadvantages are a huge cost for timely access to information, difficulties to gather documentation on all formats and to ensure quality.

When looking for a preservation format, Faria followed Adrian Brown (2008, see figure 17). SIARD scores highly on most of these criteria. It is based on international standards and is for database data, structure and behavior. A simple database archive flow would be that a producer submits a database file to the archive in SIARD formats. This is validated, put into a simple catalogue and may be viewed by the end user with the database viewer. A fuller workflow would include Keep's repository software RODA and more robust processes by the producer to document and capture the database.

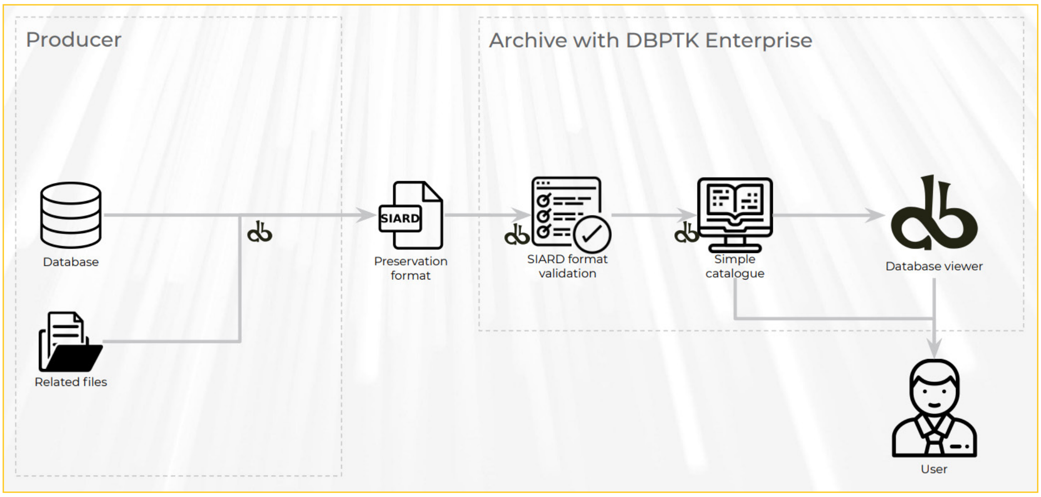


Figure 18: DBPTK simple database archive flow

The Database Preservation Toolkit (DBPTK) can be downloaded and used by anyone. There are three separate tools that can be used – DBPTK Desktop, Enterprise and Developer. The Desktop version features a connection to a database, stores the content in SIARD, and provides some problem-solving help. It can work with several formats. Users can create a migration report which notes migration changes and losses in the export. The user can then edit the SIARD metadata and enrich it. This will also allow for a full validation of the SIARD data.

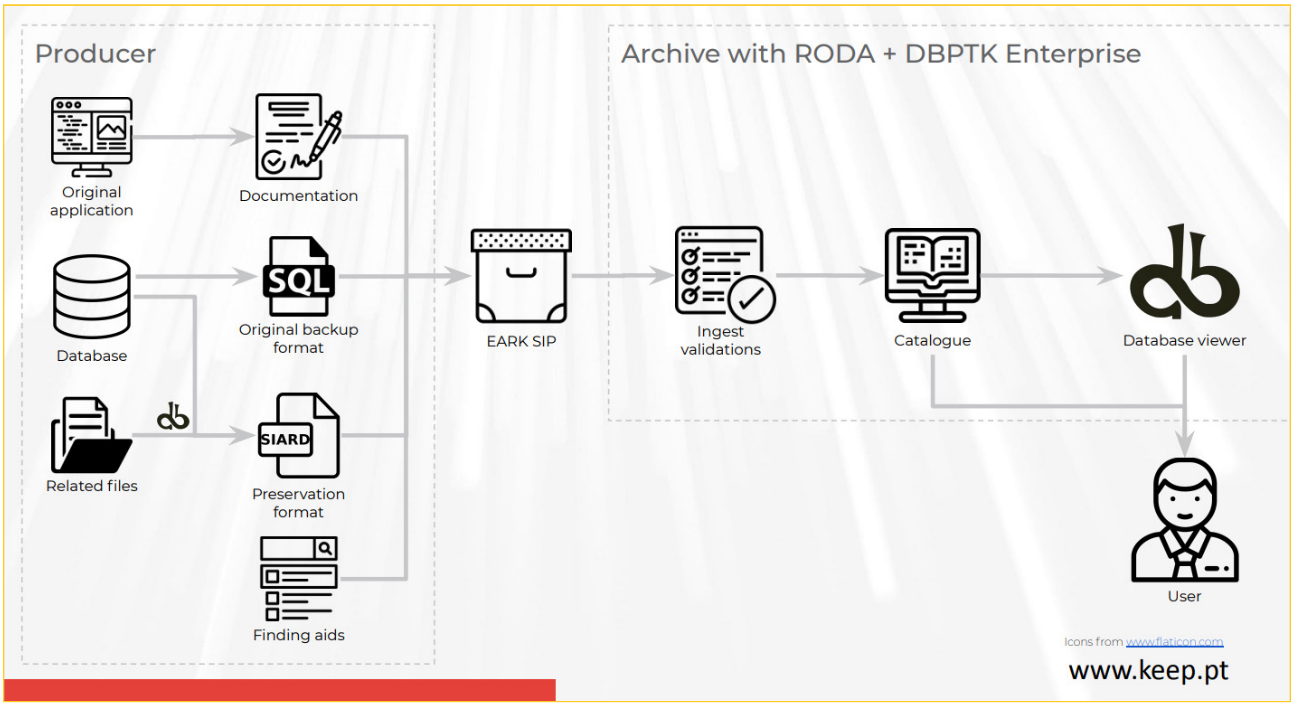
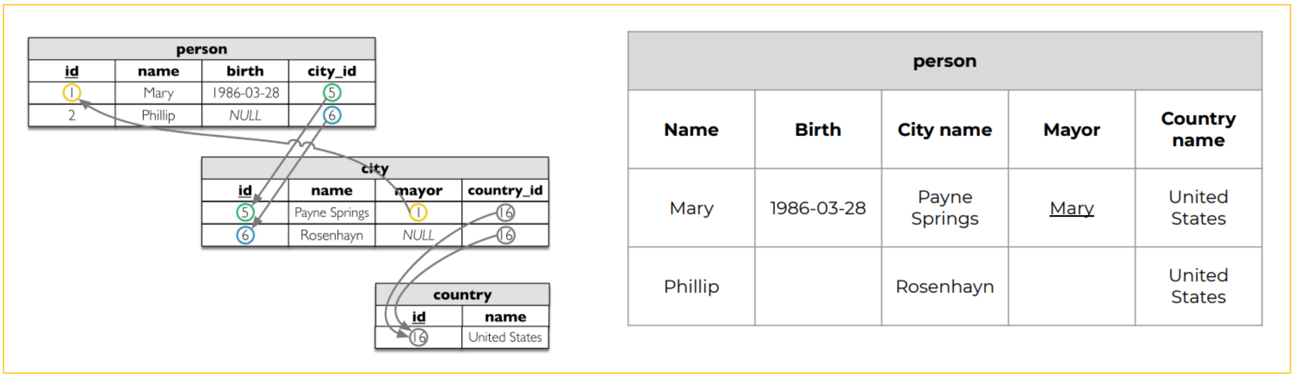


Figure 19: DBPTK full database archival flow

The Enterprise version aims at larger institutions with multiple users. These institutions can do data transformation (de-normalisation) and can provide access straight to their users through the web, allowing them to browse and search the database content. It also allows them to put SIARD back into a database including an activity log and supports multiple languages.

The Developer version features a command line and a java library. It is open source for custom development and allows specialised support for new or legacy database systems. Includes many other features for archiving databases and accessing archived databases.



person			
id	name	birth	city_id
1	Mary	1986-03-28	5
2	Phillip	NULL	6

city			
id	name	mayor	country_id
5	Payne Springs	1	16
6	Rosenhayn	NULL	16

country	
id	name
16	United States

person				
Name	Birth	City name	Mayor	Country name
Mary	1986-03-28	Payne Springs	Mary	United States
Phillip		Rosenhayn		United States

Figure 20: Data transformation (aka denormalisation) feature of DBPTK Enterprise

Faria brought two case studies: First, hospital legacy databases to support specific hospital use cases. Lacking sufficient documentation, the information is exported to SIARD where expert analysis creates the missing documentation. Second, a European taxation and customs union: trader messages archive – this is a new European Union (EU) service that will provide a centralised interface with customs authorities. All transactional messages must be archived for a decade. The productive system exports parts of the database to SIARD every hour, through RODA services, in a continuous extract/archive/validation workflow.

Questions and discussion

- Kai Naumann asked where DBPTK and RODA are installed – they mostly sit at the computing centres chosen by the agencies they support but can be deployed elsewhere if networks allow.
- Torbjørn Aasen asked how customers could create a migration log and preserve it alongside the data. – Currently, DBPTK has no way to include the report within the SIARD file as it is produced after the SIARD file has been created. The report can be taken separately but could use an information package such as the E-ARK SIP to package it up with the database content.
- Boris Domajnko asked which log would be best to keep with the SIARD file. Who will look at these after 10 years? What information is relevant to future users? Faria pointed out that the reports should also be subject to some standardisation. There is an execution log that is not intended as a report but will be useful if something goes wrong. He said you should keep almost everything (just in case) but some things are easier to keep than others.
- Faria mentioned Merkle Hash Trees (see references) which allow you to create checksums for every row and column that can be joined hierarchically allowing easy validation against manipulations.

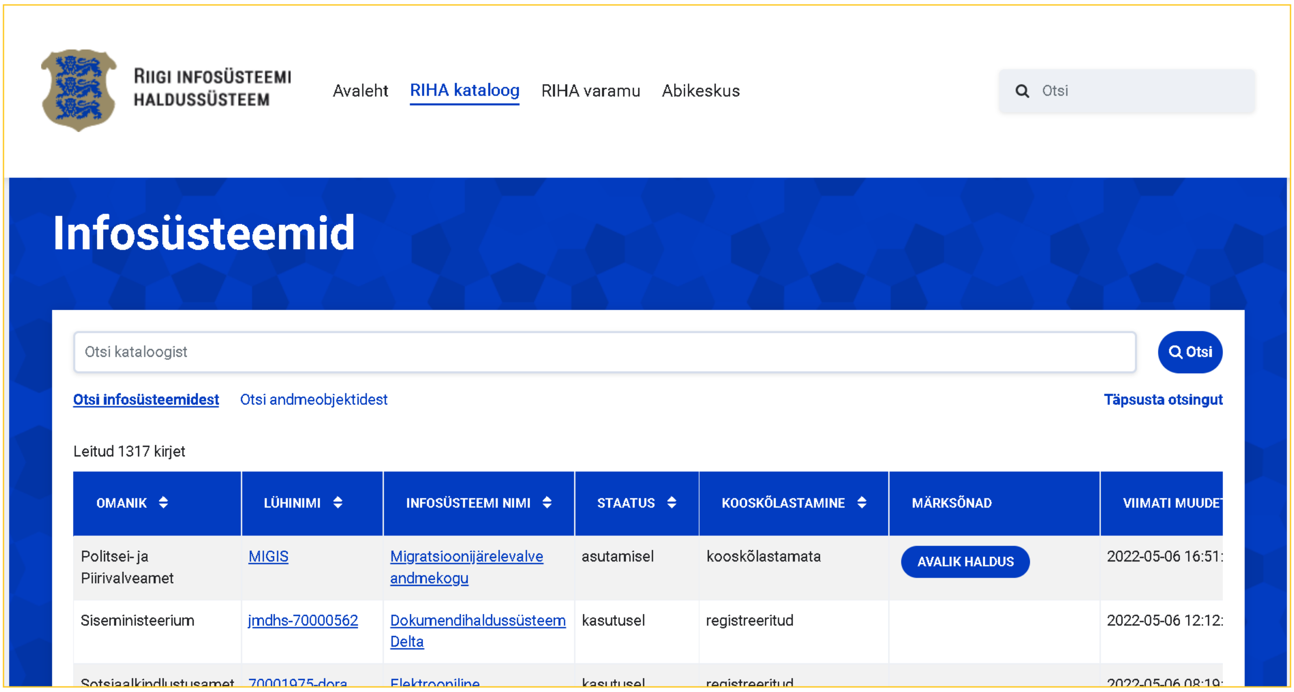
E-ARK standardisation efforts for databases

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.



Infosüsteemid

Leitud 1317 kirjet

OMANIK	LÜHINIMI	INFOSÜSTEEMI NIMI	STAATUS	KOOSKÕLASTAMINE	MÄRKSÕNAD	VIIMATI MUUDE
Politsei- ja Piirivalveamet	MIGIS	Migratsioonijärelevalve andmekogu	asutamisel	kooskõlastamata	AVALIK HALDUS	2022-05-06 16:51:
Siseministeerium	jmdhs-70000562	Dokumendihaldussüsteem Delta	kasutusel	registreeritud		2022-05-06 12:12:
Sotsiaalkindlustusamet	70001975-dora	Elektrooniline	kasutusel	registreeritud		2022-05-06 08:19:

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

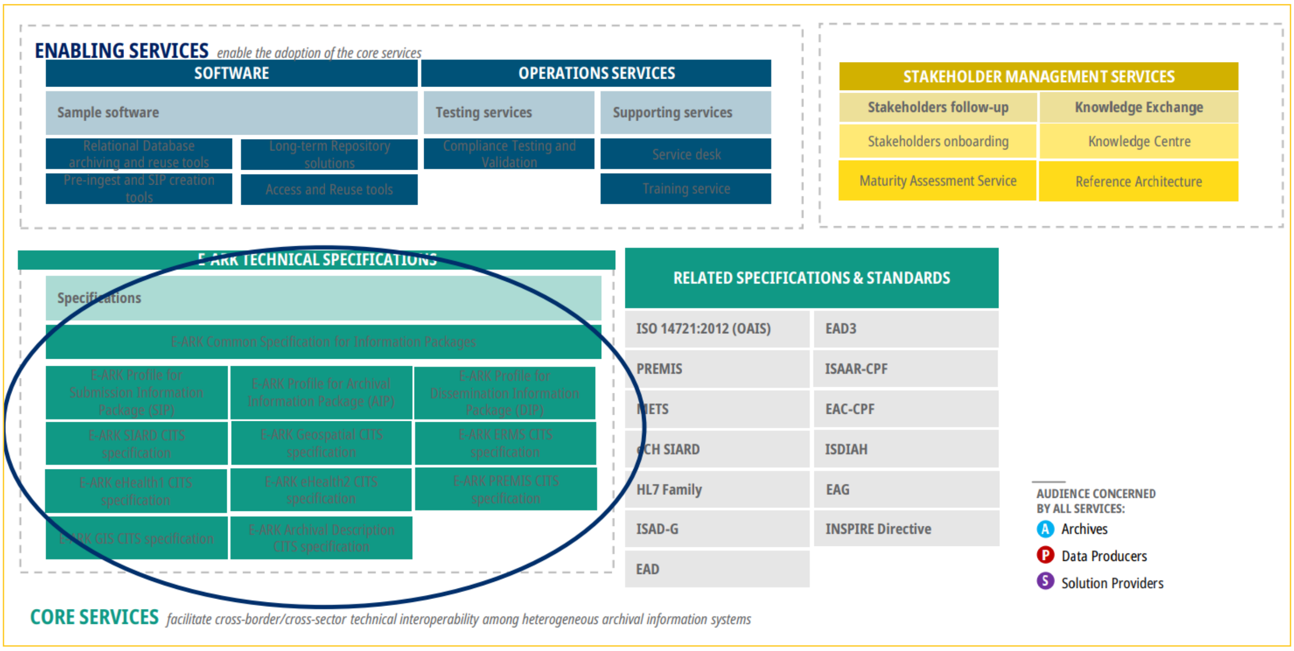
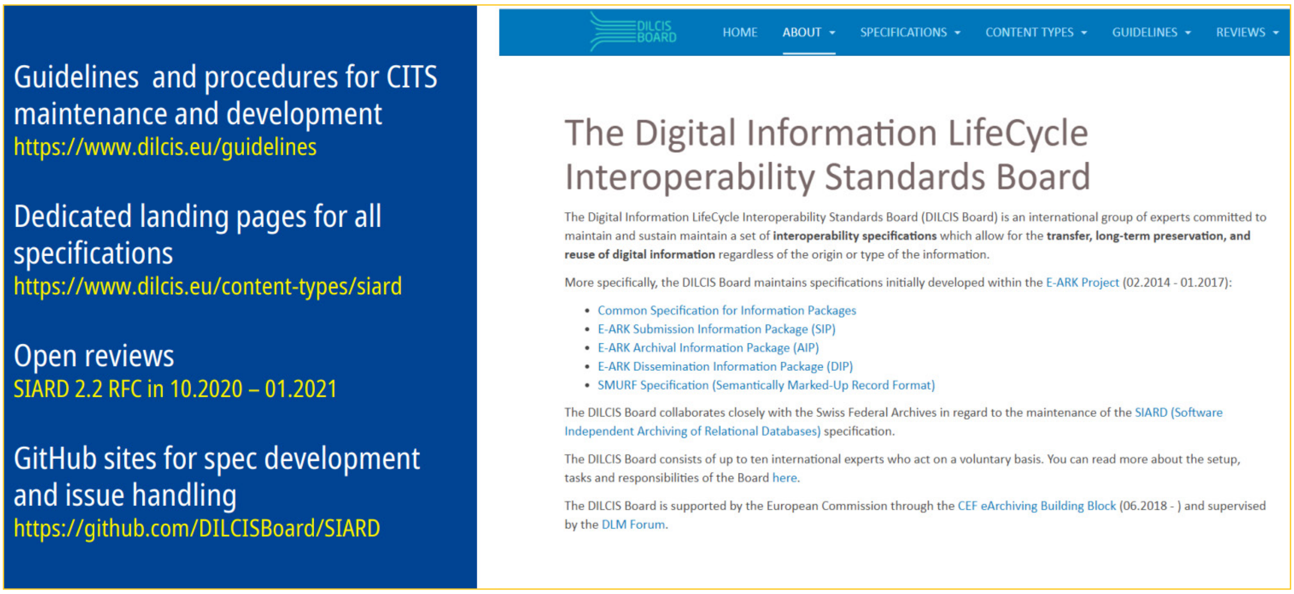

ENABLING SERVICES enable the adoption of the core services
SOFTWARE
Sample software
Relational Database archiving and reuse tools
Long-term Repository solutions
Pre-ingest and SIP creation tools
Access and Reuse tools
OPERATIONS SERVICES
Testing services
Compliance Testing and Validation
Supporting services
Service desk
Training service
STAKEHOLDER MANAGEMENT SERVICES
Stakeholders follow-up
Knowledge Exchange
Stakeholders onboarding
Knowledge Centre
Maturity Assessment Service
Reference Architecture
E-ARK TECHNICAL SPECIFICATIONS
Specifications
E-ARK Common Specification for Information Packages
E-ARK Profile for Submission Information Package (SIP)
E-ARK Profile for Archival Information Package (AIP)
E-ARK Profile for Dissemination Information Package (DIP)
E-ARK SIARD CITS specification
E-ARK Geospatial CITS specification
E-ARK ERMS CITS specification
E-ARK eHealth1 CITS specification
E-ARK eHealth2 CITS specification
E-ARK PREMIS CITS specification
E-ARK GIS CITS specification
E-ARK Archival Description CITS specification
RELATED SPECIFICATIONS & STANDARDS
ISO 14721:2012 (OAIS)
EAD3
PREMIS
ISAAR-CPF
METS
EAC-CPF
eCH SIARD
ISDIAH
HL7 Family
EAG
ISAD-G
INSPIRE Directive
EAD
AUDIENCE CONCERNED BY ALL SERVICES:
A Archives
P Data Producers
S Solution Providers
CORE SERVICES facilitate cross-border/cross-sector technical interoperability among heterogeneous archival information systems

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.


Guidelines and procedures for CITS maintenance and development
https://www.dilcis.eu/guidelines

Dedicated landing pages for all specifications
https://www.dilcis.eu/content-types/siard

Open reviews
SIARD 2.2 RFC in 10.2020 – 01.2021

GitHub sites for spec development and issue handling
https://github.com/DILCISBoard/SIARD

The Digital Information LifeCycle Interoperability Standards Board

The Digital Information LifeCycle Interoperability Standards Board (DILCIS Board) is an international group of experts committed to maintain and sustain maintain a set of **interoperability specifications** which allow for the **transfer, long-term preservation, and reuse of digital information** regardless of the origin or type of the information.

More specifically, the DILCIS Board maintains specifications initially developed within the E-ARK Project (02.2014 - 01.2017):

- Common Specification for Information Packages
- E-ARK Submission Information Package (SIP)
- E-ARK Archival Information Package (AIP)
- E-ARK Dissemination Information Package (DIP)
- SMURF Specification (Semantically Marked-Up Record Format)

The DILCIS Board collaborates closely with the Swiss Federal Archives in regard to the maintenance of the SIARD (Software Independent Archiving of Relational Databases) specification.

The DILCIS Board consists of up to ten international experts who act on a voluntary basis. You can read more about the setup, tasks and responsibilities of the Board here.

The DILCIS Board is supported by the European Commission through the CEF eArchiving Building Block (06.2018 -) and supervised by the DLM Forum.

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.

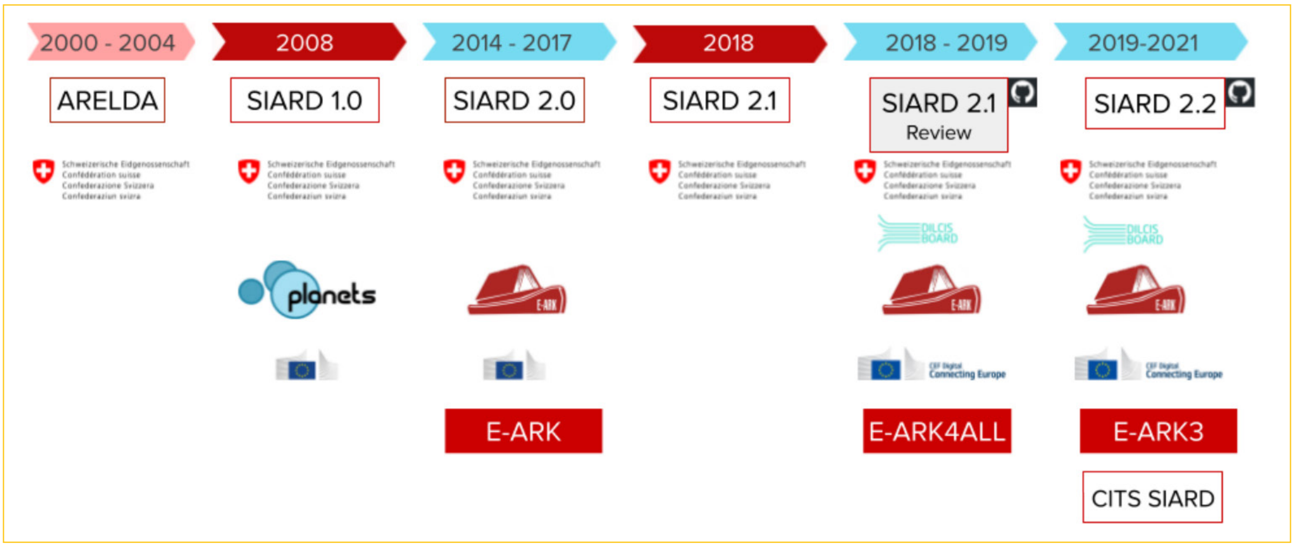
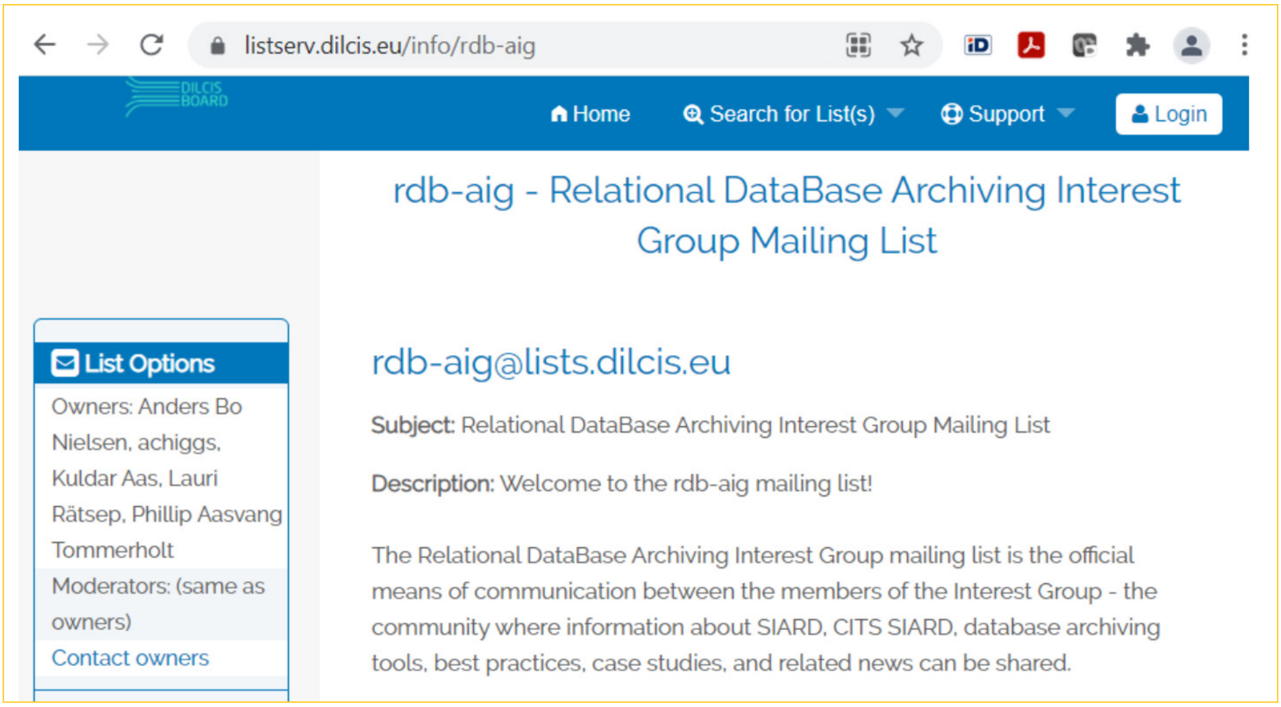


Figure 24: The history of SIARD, taken from Guideline for CITS SIARD (2021), p. 12

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.



rdb-aig - Relational DataBase Archiving Interest Group Mailing List

Owners: Anders Bo Nielsen, achiggs, Kuldar Aas, Lauri Rätsep, Phillip Aasvang Tommerholt

Moderators: (same as owners)

rdb-aig@lists.dilcis.eu

Subject: Relational DataBase Archiving Interest Group Mailing List

Description: Welcome to the rdb-aig mailing list!

The Relational DataBase Archiving Interest Group mailing list is the official means of communication between the members of the Interest Group - the community where information about SIARD, CITS SIARD, database archiving tools, best practices, case studies, and related news can be shared.

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.

Transfer and Preservation of Databases at the National Archives of Australia: Problems and Directions

James Doig (National Archives Australia (NAA), Canberra, Australia)

James Doig has worked at National Archives of Australia (NAA) for 20 years. He is a historian and archivist. He has worked in a number of roles at NAA mainly relating to digital preservation, archival skills development and collection management. In 2016-17, he was project manager for the relocation of 115 shelf kilometres of records into NAA's newly built storage and preservation facility in Canberra.

Doig said he felt like Australia was playing catch up sometimes compared with European approaches. Like many archives, NAA began receiving digital records early on (since 1970) – came mainly on magnetic tape and stored in an off-line storage environment. They were retrieved and accessed from computers in reading rooms when publicly requested. However, this was not sustainable because of obsolescence – a classic digital preservation problem. In the 1990s, when it turned out that 'do nothing' was the wrong approach, NAA adopted a distributed custody model from 1996-2000. Agencies then managed their archival records under a management regime worked out by NAA. This was when the NAA was developing early standards for record keeping metadata, which informed the management regime for digital records.

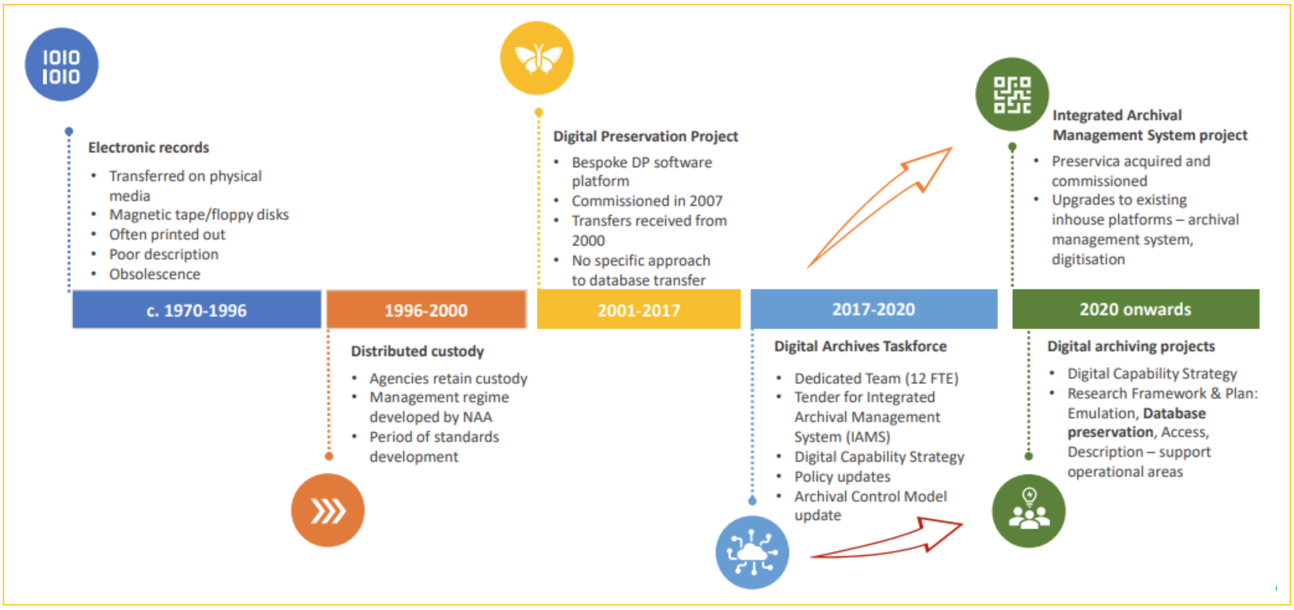

1010
1010
Electronic records
• Transferred on physical media
• Magnetic tape/floppy disks
• Often printed out
• Poor description
• Obsolescence
c. 1970-1996
1996-2000
Distributed custody
• Agencies retain custody
• Management regime developed by NAA
• Period of standards development
Digital Preservation Project
• Bespoke DP software platform
• Commissioned in 2007
• Transfers received from 2000
• No specific approach to database transfer
2001-2017
2017-2020
Digital Archives Taskforce
• Dedicated Team (12 FTE)
• Tender for Integrated Archival Management System (IAMS)
• Digital Capability Strategy
• Policy updates
• Archival Control Model update
Integrated Archival Management System project
• Preservica acquired and commissioned
• Upgrades to existing inhouse platforms – archival management system, digitisation
2020 onwards
Digital archiving projects
• Digital Capability Strategy
• Research Framework & Plan: Emulation, **Database preservation**, Access, Description – support operational areas

Figure 26: A brief history of digital preservation at NAA

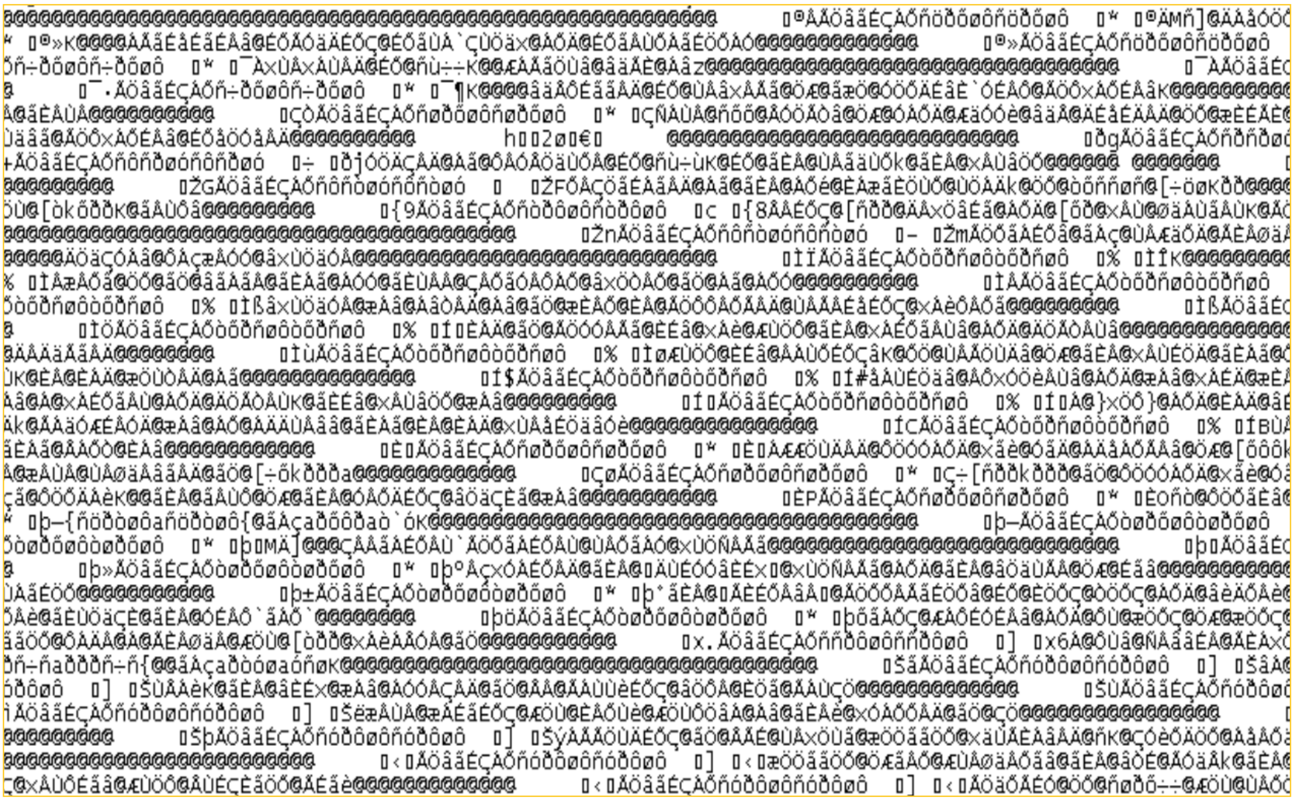
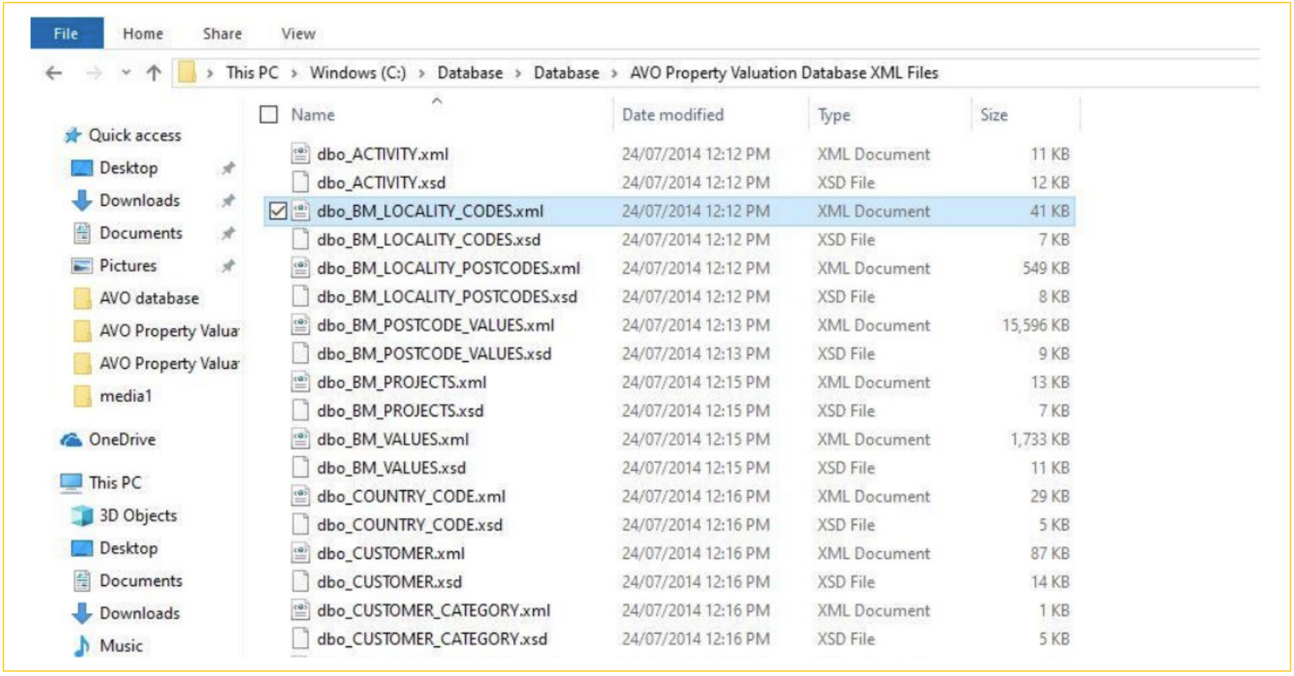


Figure 27: Recovered data stream without interpretation

In the early 2000’s, NAA started carrying out data recovery on some of those early disks. One of the data recovery projects was an early 1980’s public enquiry that used a database system to manage all the material of the inquiry, such as evidence, interviews, submissions etc. The data recovered from this inquiry would provide good insights into early computing practices. However, it is a challenge to interpret the content of the recovered data, e.g. the character encoding and other technical aspects. One of the problems relates to lack of information about the computer system and the software used. In the 1980’s, no one thought this information would be important. Emulation might be a solution to effectively accessing this material.

In 2000, NAA started accepting digital records again, and has been ingesting digital records into a digital preservation system since 2007. Born digital content consists of about 10 TB (does not include AV or surrogates). In 2020, they got Preservica to replace their bespoke digital preservation platform. NAA has not received many transfers from purely database systems. They do get records from EDRMS, document management systems and other business systems, but export and manage the records, rather than treat them as databases. They have not received the number and quantity of digital records e.g., from EDRMS electronic document and records management systems (EDRMS) as they would expect. The vast majority of large transfers have been from closed agencies and short-term agencies (e.g., public enquiries). This is because there is a disconnect between the development of disposal schedules, when and what to transfer, and the difficulty agencies have in sentencing and appraising their records managed in business systems. The agencies decide what and when they will transfer.

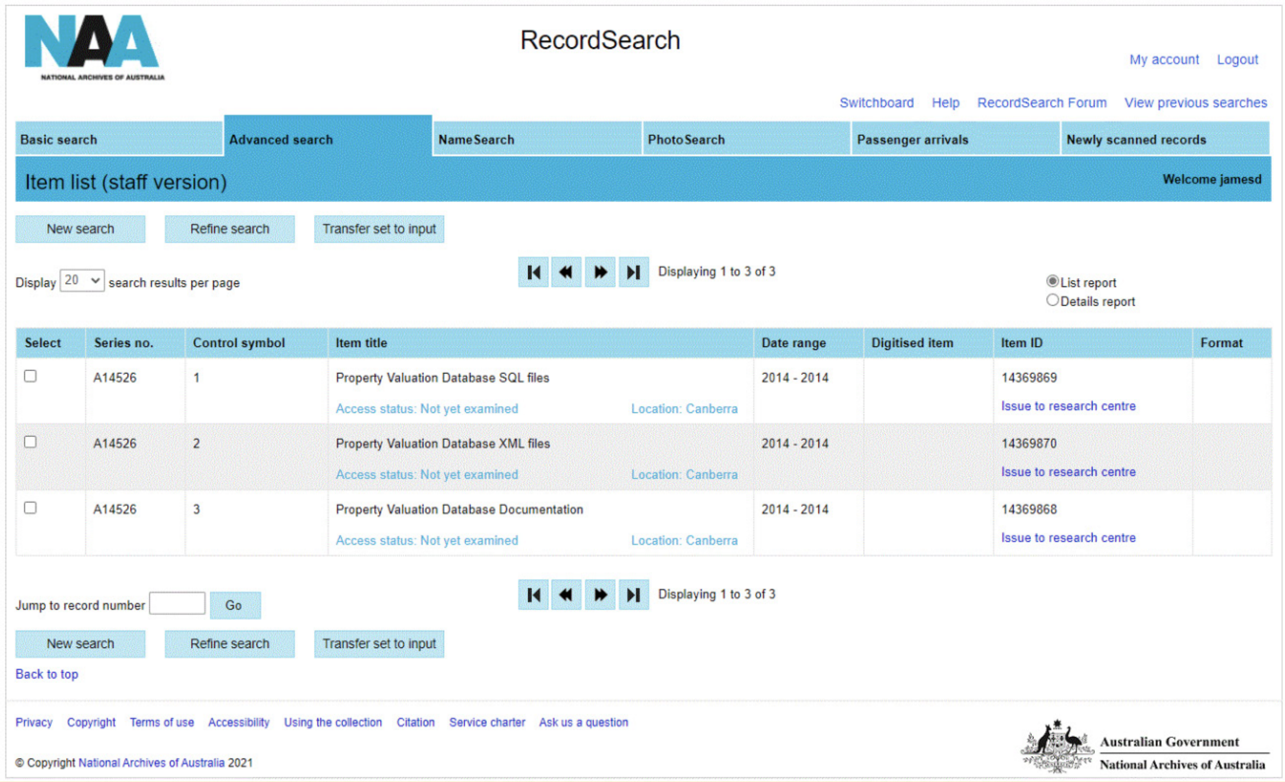


□	Name	Date modified	Type	Size
	dbo_ACTIVITY.xml	24/07/2014 12:12 PM	XML Document	11 KB
	dbo_ACTIVITY.xsd	24/07/2014 12:12 PM	XSD File	12 KB
☑	dbo_BM_LOCALITY_CODES.xml	24/07/2014 12:12 PM	XML Document	41 KB
	dbo_BM_LOCALITY_CODES.xsd	24/07/2014 12:12 PM	XSD File	7 KB
	dbo_BM_LOCALITY_POSTCODES.xml	24/07/2014 12:12 PM	XML Document	549 KB
	dbo_BM_LOCALITY_POSTCODES.xsd	24/07/2014 12:12 PM	XSD File	8 KB
	dbo_BM_POSTCODE_VALUES.xml	24/07/2014 12:13 PM	XML Document	15,596 KB
	dbo_BM_POSTCODE_VALUES.xsd	24/07/2014 12:13 PM	XSD File	9 KB
	dbo_BM_PROJECTS.xml	24/07/2014 12:15 PM	XML Document	13 KB
	dbo_BM_PROJECTS.xsd	24/07/2014 12:15 PM	XSD File	7 KB
	dbo_BM_VALUES.xml	24/07/2014 12:15 PM	XML Document	1,733 KB
	dbo_BM_VALUES.xsd	24/07/2014 12:15 PM	XSD File	11 KB
	dbo_COUNTRY_CODE.xml	24/07/2014 12:16 PM	XML Document	29 KB
	dbo_COUNTRY_CODE.xsd	24/07/2014 12:16 PM	XSD File	5 KB
	dbo_CUSTOMER.xml	24/07/2014 12:16 PM	XML Document	87 KB
	dbo_CUSTOMER.xsd	24/07/2014 12:16 PM	XSD File	14 KB
	dbo_CUSTOMER_CATEGORY.xml	24/07/2014 12:16 PM	XML Document	1 KB
	dbo_CUSTOMER_CATEGORY.xsd	24/07/2014 12:16 PM	XSD File	5 KB

Figure 28: XML files of database tables of VOIS (see text)

Doig said we learn the most from former bad practices, which highlights what needs to happen to improve practices and approaches. For example, data from 2007 election results. These data are in the public domain. The disposal schedule states what should have been received. NAA received the data in CSV files and other text files, packaged in ZIP files. Transferred data corresponds with the terms of the disposal schedule. However, when the systems used in the electoral process are considered, there is much potential functionality that has been lost. In addition, technical metadata has not been captured at transfer. The series registration is a very basic record lacking sufficient information to describe the systems in which the information was captured and analysed – it lacks technical information about the business systems and how it worked. There was very little information about the transferred CSV files. The whole dataset is registered on the archival control system (i.e. the catalogue) as a single item – this could turn out quite cumbersome when someone requests it. Several problems are evident – description, lack of technical documentation (big problem going forward), raw data format (CSV) and no explanation of the relationships between all the data in the transfer. Raw data does not preserve any functionality of the originating system.

Another example presented was a property valuation database used by the Australian Valuation Office, called VOIS, the Valuation Office Information System. It was transferred to NAA in 2014 when the agency was abolished and its functions moved to the Australian Taxation Office. In transfer they received the data from an MS SQL Server. As well as the native SQL data files, the transfer included an export of the database tables in XML. The files have been ingested into Preservica. The transfer included an export of database tables in XML. An advantage with XML is that it is both machine and human readable – but not that easy to import it back into a database. The transfer included a data dictionary and screen shots showing how it was used. Once again, the documentation obtained for the database was not as extensive as we would have liked, but at least it met standard expectations, e.g. the E-R model and data dictionary were transferred.



Item list (staff version)

Welcome jamesd

Display 20 search results per page

Displaying 1 to 3 of 3

● List report
○ Details report

Select	Series no.	Control symbol	Item title	Date range	Digitised item	Item ID	Format
□	A14526	1	Property Valuation Database SQL files Access status: Not yet examined Location: Canberra	2014 - 2014		14369869 Issue to research centre	
□	A14526	2	Property Valuation Database XML files Access status: Not yet examined Location: Canberra	2014 - 2014		14369870 Issue to research centre	
□	A14526	3	Property Valuation Database Documentation Access status: Not yet examined Location: Canberra	2014 - 2014		14369868 Issue to research centre	

Jump to record number

Displaying 1 to 3 of 3

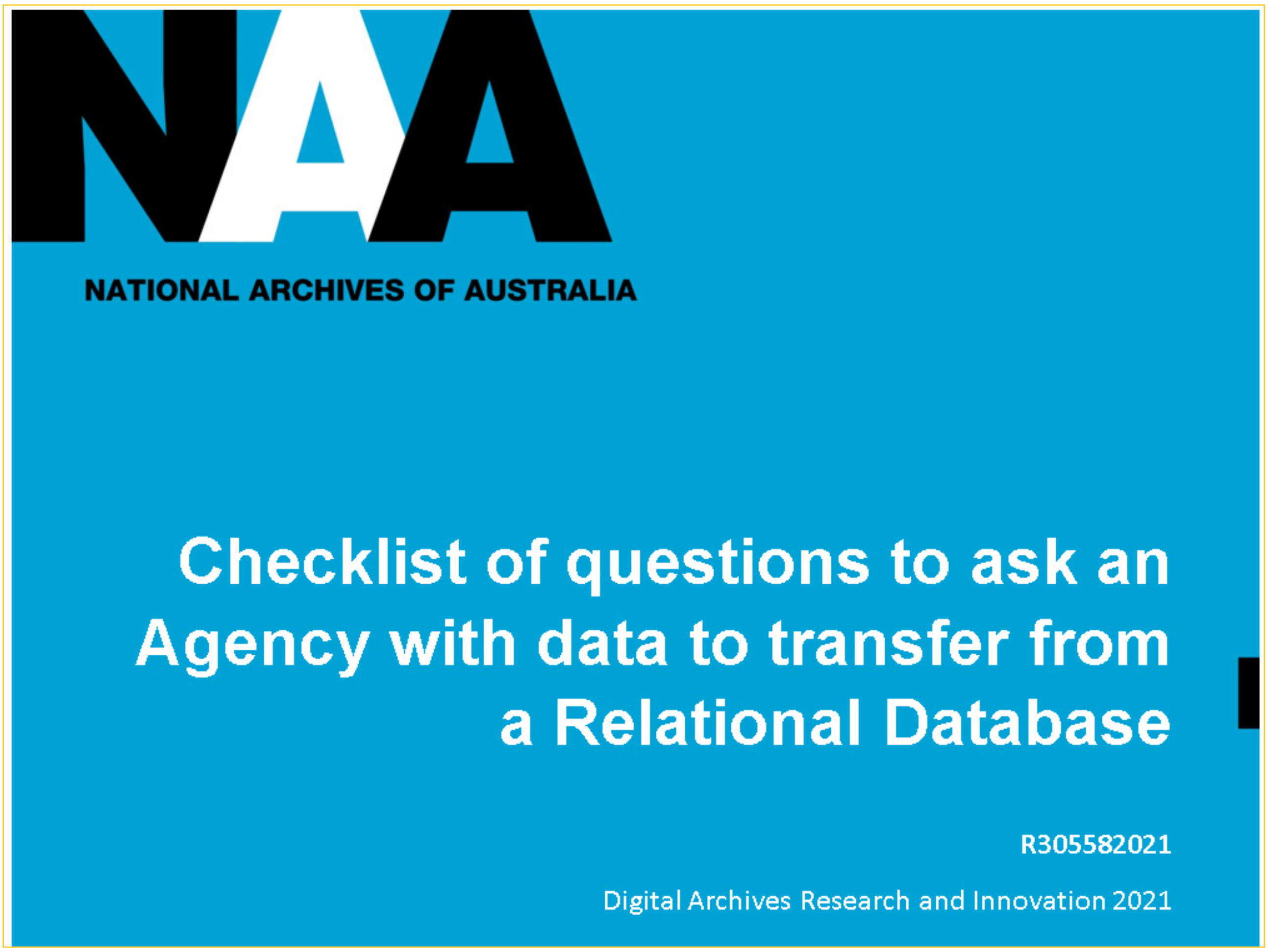
© Copyright National Archives of Australia 2021

Figure 29: Catalogue entries on the archived VOIS database

Doig identified problems in transfers of databases:

- Transfer decisions about what to transfer, how and when are not defined early enough, and disposal schedule and transfer are not well connected.
- We have not decided what we need to preserve to ensure meaningful access in the future.
- The raw data may not be enough: a native SQL format is software dependent. A flat file format (e.g. CSV) has limited usability. Both need technical documentation to understand data.
- Are there any characteristics of the database that we need to preserve to ensure meaningful accessibility and usability?
- There is not be a one-size-fits-all approach to database transfer and preservation.

And in order to resolve these problems, NAA commenced a database preservation project which ran from December 2020 to July 2021, and drew on the expertise of a Reference Group that was drawn from different business areas across the National Archives. It looked at DBPTK, which was easy to deploy and use. The approach was to create a SIARD file at NAA and import it into Preservica. The staff used the native SQL dump of the AVO database mentioned earlier. Of course, the DBPTK must connect to a live version of the database, so the native SQL files were imported to an instance of SQL Server, so that DBPTK could create the SIARD file. This process worked fine on a relatively small database.



NAA

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

Figure 30: NAA checklist for transfer of relational databases

Doig believed that, if archives adopt the DBPTK, creating SIARD files at the archive would be the norm, as cybersecurity concerns and rules mean that government agencies are reluctant to use third party software that has not been tested and accredited. NAA had a similar situation with a previous SIP creation tool – agencies were reluctant to download and deploy it because of cybersecurity concerns. This is a key point for government archives: preservation software that we expect records creators to use need to be ‘white listed’ to encourage use.

Another result of the project was a checklist of questions to ask an agency with data to transfer from a relational database – as staff guidance. A third product was guidance for determining options for transfer. In some cases, according to the guidance, it is appropriate to seek an export of raw data in a simple structured data format like CSV or plain text, or an export of reports, from a database, rather than to try to preserve database functionality.

This guidance also contains other advice, for example

- advice about frequency of transfer, which can be dependent on a number of different factors
- guidance for determining options for transfer
- transfer process maps
- a document describing the standard metadata requirements of the Australian Series System, as additional metadata elements mapped to other standards or products like PREMIS, the Australian Government Recordkeeping Metadata Standard and the Software Metadata Recommended Format Guide (Christophersen 2022).

The NAA has adopted a more flexible and hopefully a more sophisticated approach to database transfer. They still need to interpret the disposal class description, analyse the system in which records are held and so on. The outcome of that process is that there is not a one-size-fits-all approach. A database transfer could consist of a combination of these things, possibly all of them. But what is always needed is a full suite of technical documentation defining the data properties, and a full suite of descriptive archival metadata.

Doig concluded by stating that NAA has only begun to embark on a longer implementation process. The NAA has a few quite challenging database transfers in the pipeline (including Microsoft Dynamics and Lotus Notes) – but carrying out the processes is key to turning database transfers into business as usual, and to develop staff knowledge and capability and to continually improve the products they have developed. And perhaps most importantly, Doig sees a need to redevelop NAA's approach to creating Records Authorities or disposal schedules, so that we can embed transfer decisions and standards up front at the point of creation.

Questions and discussion

- Aas commented that indeed the current series-based arrangement logic does not work well for database transfers, where an item is not within a series but rather includes (multiple) series. Maybe the standard Records in Context (RiC) will provide a solution.
- James Doig responded that such issues have been looked into by TNA in their development of a new catalogue model (Catalogue Model proposal 2020), particularly in the section on Multiple Arrangement.
- Jenny Mitcham asked if NAA's checklists are available to look at. James replied that they are available on request, but that he is not sure how useful they will be as they were written for the Australian government environment and they still need to be thoroughly tested with agencies.
- Appuswamy asked what information is needed to document a database – e.g., the relationships. He wants to think about whether this information can be embedded within the database itself. Aas answered this question by pointing to the CITS SIARD (see references) by E-ARK.

A Generalised XML Client for the Bavarian State Archives

Markus Schmalzl (Directorate General of State Archives Bavaria, Germany)

Markus Schmalzl is an archivist and works at the Directorate-General of the Bavarian State Archives, previously at the State Archive for Upper Bavaria and the Bavarian Main State Archive at Munich. He is responsible for the transfer of digital data from EDRMS, geodatabases and other collections to archives and for developing interfaces for transfer processes. He is member of committee of the German Conference of State Archive Directors (KLA). He teaches at the Bavarian Archives School, the Bavarian University of Applied Sciences for Public Service and the University of Regensburg.

Schmalzl presented a one-size-fits-all approach for a standardised way of exporting data from government information systems that has been developed in the last few years: a new tool for database archiving.

Databases are widely used by state agencies. Not all the data are important for the long term, but some will need preservation. There is valuable information in high volumes with potential for scientific reuse and other purposes. The State Archives need interfaces for ingesting data from Bavarian government agencies. Data are sometimes volatile and can include Binary Large Object (BLOB) content. Typically, there are no managed archival interfaces for handling this type of information. The State Archives did not want to use emulation but instead applied an approach based on format migration.

Schmalzl's employer wanted to create a standard solution which would work for many systems. The key issue was to choose in which format this information was to be extracted. To close the gap the Bavarian State Archives created a generalised XML extraction tool. It was set up in a test environment October 2021. It is an archival client solution that transfers data in a fully automated way.

The project came to life in 2017 as a specialised client for transferring Bavarian government staff records from the SAP Human Resources (HR) software as PDF files, accompanied by XML metadata. As a user of this client system, you can receive, validate, and save the data and confirm receipt. The interface took a long time to develop. To use the client flexibly for other data sources, the 2017 project was extended to extend use to XML data objects of various schemas that conform to the standards of the different government branches. A mapping tool is the central component. The mapping tool can define how to validate data, whether an appraisal is required and what metadata are presented in the graphical user interface for appraisal. The XML data can be accompanied by text and images.

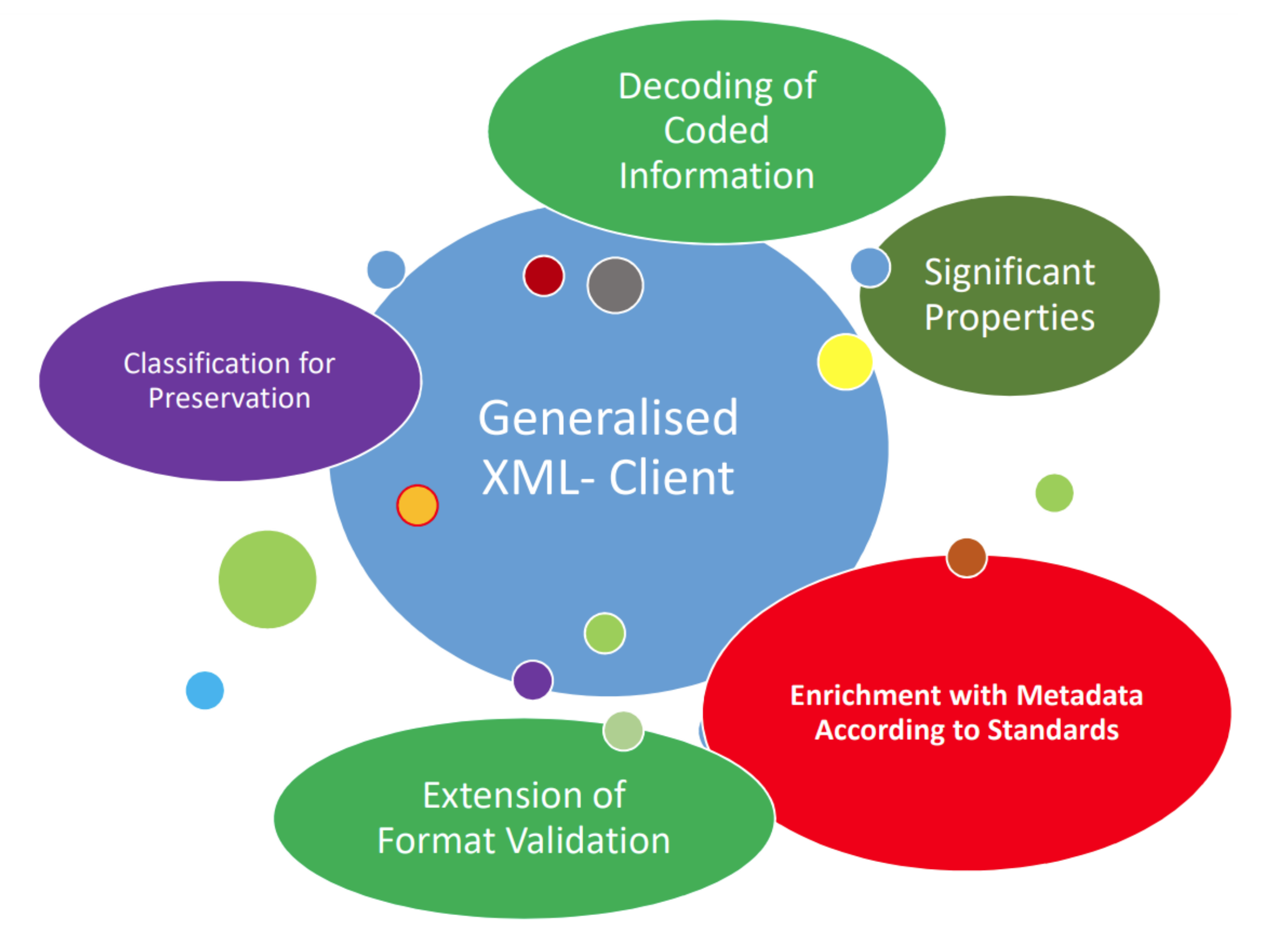
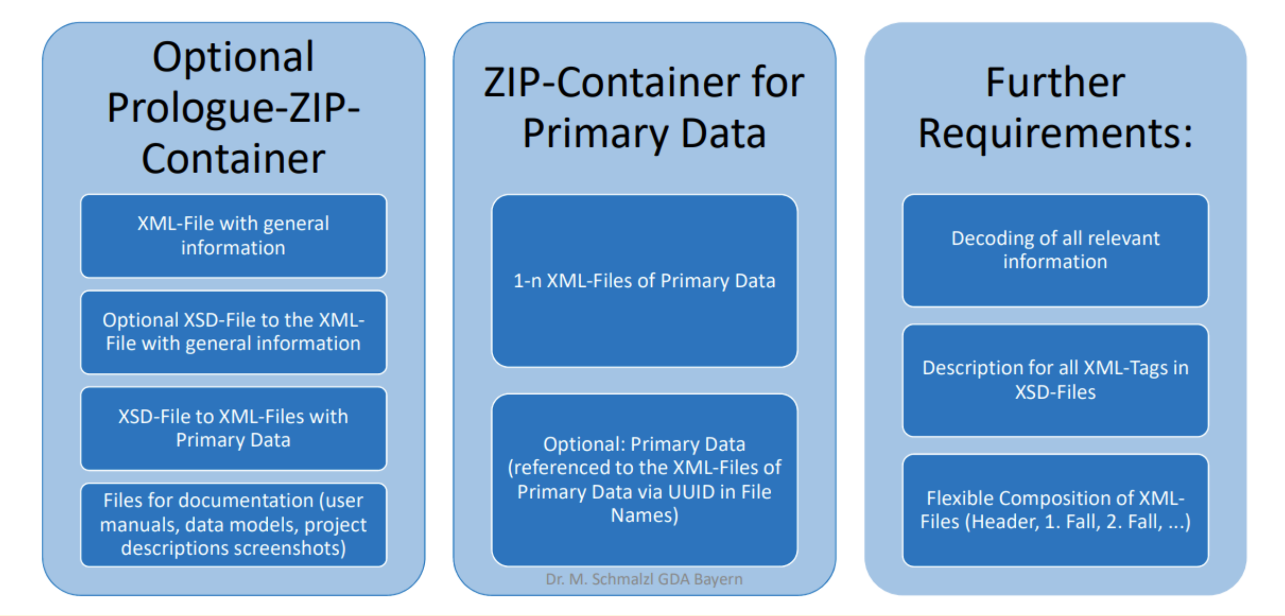

Decoding of Coded Information
Significant Properties
Classification for Preservation
Generalised XML- Client
Enrichment with Metadata According to Standards
Extension of Format Validation

Figure 31: Options for further development of the Generalised XML Client

After configuring the mapping tool, a semi-automated workflow follows. Data are checked continually. Packages are restructured as needed by the mapping tool. The SIP is ingested into archival data storage and structured into AIPs that can be ingested. Attribute descriptions are incorporated into the AIP. Once data has been processed the institution gets a receipt. Hash values are controlled. All archival working steps are recorded: from the reception of the data to appraisal decisions, and finally to successful storage in the archive. The system is automated but flexible. A standardised structure is also maintained for deliveries.

Data are archived in a structured way in an appropriate long term archiving format which facilitates further processing and use. The State Archives do not archive the database management systems themselves. The rather apply this approach to data management systems of government and science.

Further development is planned. Issues relate to significant properties, enrichment with metadata and classification for preservation. Ultimately, the system allows the processing of large data volumes.



Optional Prologue-ZIP-Container

XML-File with general information

Optional XSD-File to the XML-File with general information

XSD-File to XML-Files with Primary Data

Files for documentation (user manuals, data models, project descriptions screenshots)

ZIP-Container for Primary Data

1-n XML-Files of Primary Data

Optional: Primary Data (referenced to the XML-Files of Primary Data via UUID in File Names)

Dr. M. Schmalzl GDA Bayern

Further Requirements:

Decoding of all relevant information

Description for all XML-Tags in XSD-Files

Flexible Composition of XML-Files (Header, 1. Fall, 2. Fall, ...)

Figure 32: Packaging structure for deliveries through the Generalised XML Client

Questions and discussion

- Will the Bavarian State Archives need a standardised format like SIARD going forward or will they use department standards? The Archives are not using the individual standards of departments currently as they contain partially coded information. It is important to extract certain data for reuse and appraisal. They are open to SIARD but believe this will not guarantee the reusability of the data secured in the databases. The Archives think it might not be the right way to go for the long term but might be convinced of this in the future.
- Kai Naumann explained about coded values in German government databases. Names of court or police stations are coded as values within the XML and the up-to-date key to the code is maintained elsewhere at a special agency at Bremen, another state of the German Federal Republic. Data are well structured, but some tags are not de-coded and this is a challenge when carrying out preservation. Markus confirmed that this is the problem with some German government XML standards.

Database ingest and use at public archives with the SIARD format

Torbjørn Aasen (Intercommunal Archive Møre og Romsdal, *Ålesund*, Norway)

Torbjørn Aasen is an IT specialist who has previously worked in systems development.

IKAMR is a Norwegian Municipal Archive in Middle Norway. It is a rural place with fjords and mountains. The organization has 28 employees including two IT archivists. IKAMR relies on the KDRS, a regional digital resource centre with 18 Municipal archives as members. They offer an OAIS based digital repository to members. Database archiving is done by routine. The IKAMR started XML table extractions in 2005, then was refined by using SIARD in 2014 and the national Noark standard in 2015.

- From 2005 onwards, table extractions were done by taking out each table and placing it into an individual XML file. Metadata on the DB structures were included as an ADDML.xml file (national standard). This way, IKAMR preserved 50 database systems between 2005 and 2011. They started by preserving three tables then added five more, finally ingesting full database in the archival work area.
- In 2014, IKAMR first ingested SIARD packages and has so far ingested 100 SIARD extractions. A few of them were extracted using SIARD Suite v1 and 2, some were using DBPTK 2, but most of them were using Spectral Core Full Convert, with SIARD 2.1 as the target format.
- Noark 5 extractions are the current Norwegian Archive Standard for recordkeeping systems used in public administration. The extractions are well-ordered, comply to the standard and can be validated against it. Fifty databases have been ingested since 2015 on this basis.

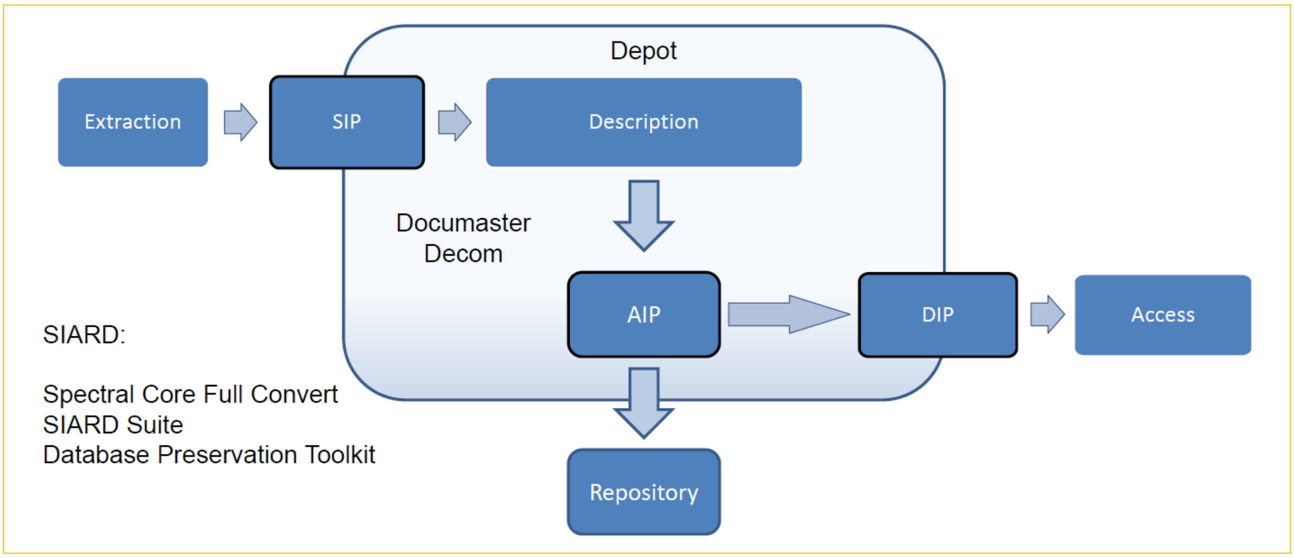


Figure 33: Production line for digital archives at Norwegian municipal archives

A SIARD extraction is imported into the tool Decom. The KDRS Decom server holds over 60 system description templates that are used to handle the unordered extractions (e.g. describing a particular childcare database system). This means that extractions from the same system over time are quicker as IKAMR can reuse the previous template. LOBS are transformed as part of the process, for example from office formats to PDF/A.

Regarding access to preserved SIARD extractions, the system for visualisation of preserved SIARD extractions is under evaluation. IKAMR is looking at DBPTK desktop enterprise and Asta 7 from Stiftelsen Asta. There are not many requests for access now, but this is likely to increase. Currently, Aasen migrates a SIARD file into a database for users and helps them use SQL queries to search for relevant content.

The challenges Torbjørn Aasen saw in the process are lack of time, especially for finding and prioritising the databases, loss of knowledge from staff who have left, scarce technical documentation, and in some cases a front-end logic that has not survived.

Challenges of using SIARD (but there are solutions and workarounds noted):

- Some old Oracle system slices files into 32 kb bits.
- Some systems have encrypted documents.
- Some old systems store binary files in CLOB text column and forces it into UTF-8.

Aasen gave positive feedback on DBPTK, an excellent solution including the SIARD validator. The validator needs to be tuned for interoperability. The viewer is also excellent. He also recommended Spectral Core Full Convert as a high-quality migration tool with high success rates. He has experienced some interoperability issues which are reported on GitHub (https://github.com/DILCISBoard/SIARD/issues). He is looking forward to exchanging experiences with the community for the common goal.

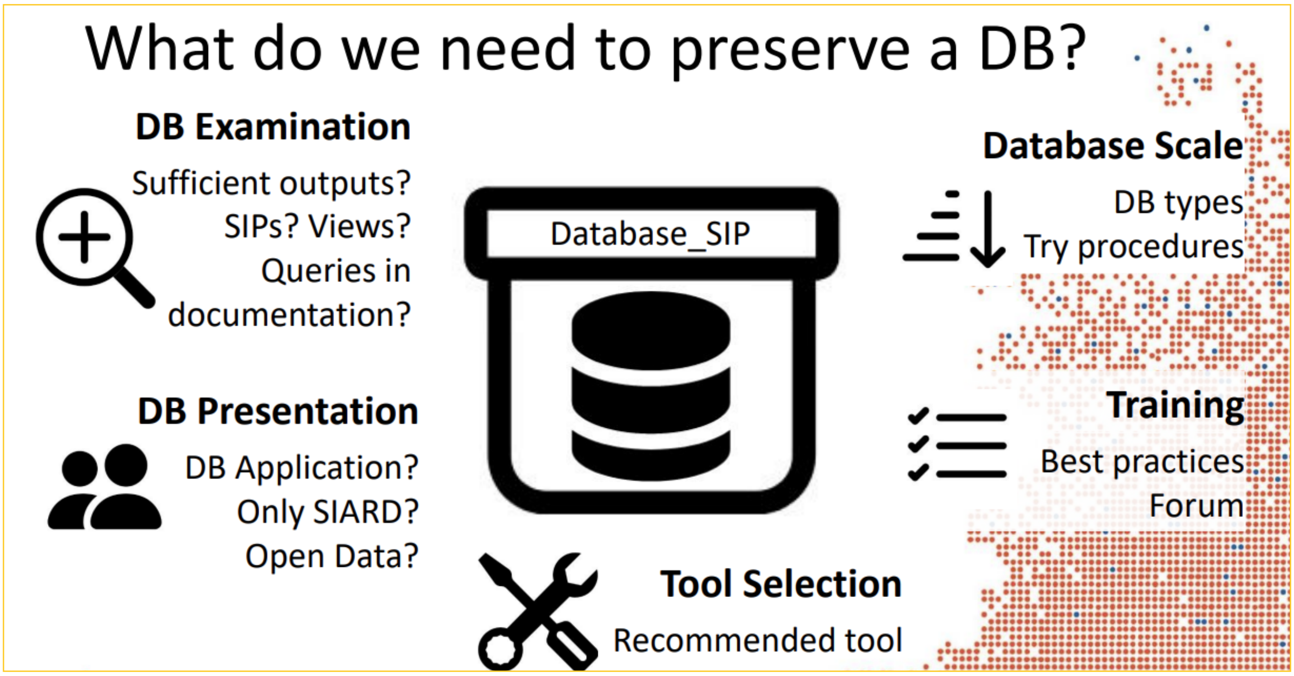
DB archiving and use at Czech authorities

Martin Rechtorik (National Archives Czech Republic, Prague)

Martin Rechtorik is historian and archivist and has been employed in the National Archives of the Czech Republic for the last four years. He is member of the digital archives methodological team and focuses on database preservation. He has almost 10 years of experience as record management specialist in private sector too.

Rechtorik's presentation had three parts – theory, sharing a practical experience and then looking at how to present data to users. On theory, he asked four questions: What does it mean to preserve a database? Which database do we choose to be preserved? How do we preserve it and at what stage (sometimes periodically across DB lifetime)? How do we present the data, depending on the legal question of who is allowed to access it (open or closed data)?

He then applied these questions (cf. figure 34: What do we need to preserve a DB?). What output is necessary for the future? Do we need queries/views? What is of real value and importance for the future? The presentation of databases is important. Can we present SIARD files? Rechtorik saw a need to select the right tool for presentation.



What do we need to preserve a DB?

DB Examination
Sufficient outputs?
SIPs? Views?
Queries in documentation?

DB Presentation
DB Application?
Only SIARD?
Open Data?

Database_SIP

Database Scale
DB types
Try procedures

Training
Best practices
Forum

Tool Selection
Recommended tool

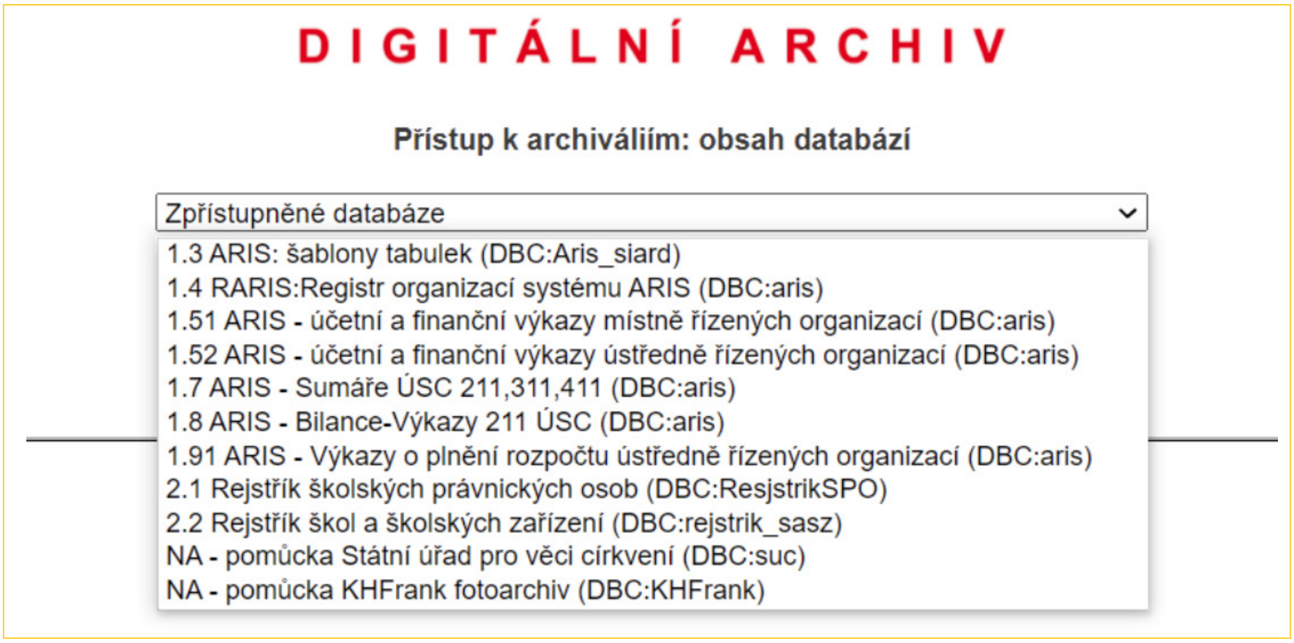
Figure 34: What do we need to preserve a DB?

Archives will have to do this at scale and to cover different types of databases – at least those in use in the public sector. They must also share this knowledge with colleagues across borders. The European Union wants member states to establish databases with similar outputs to be sent to EU authorities – this should be seen as an incentive to member states to cooperate on solving those problems.

An example is the ARIS (Automatic Budget Information System) of the Czech central government. It contains accounting and financial data of all state institutions (from the presidential office, ministries, municipalities to kindergartens) and covers the years 1997-2009. ARIS and its sibling systems RARIS and ARISweb were intended as paperless – an all-digital reference system. The contents are mostly open data. The Czech National Archives wanted to emulate the whole system because of the ARISWeb interface, which was used for public access, provided a lot of functionality. This plan turned out as too optimistic. The point of failure was licences for the Informix DBMS that were too costly. The National Archives resorted to a SIARD normalisation approach but also transferred CSV files for easier access. The Ministry of Finance first provided the database files as PostgreSQL dumps, while the company Keep LDA was engaged to help with the SIARD conversion. ARIS/ARISweb, an interface showing all financial data sheets, was shut down in September 2021.

Rechtorik also explained the way in which the ARIS data is now used at the National Archives. To present data in their reading room, the National Archives use the dbDIPview tool (p. 47). ARIS, ARISweb, and RARIS are now consultable through the following services:

- one part of the data is available as CSV exports divided into seven browsing packages
- the other part, a SIARD representation of the RARIS module was exported using the table filter feature from original SIARD file.

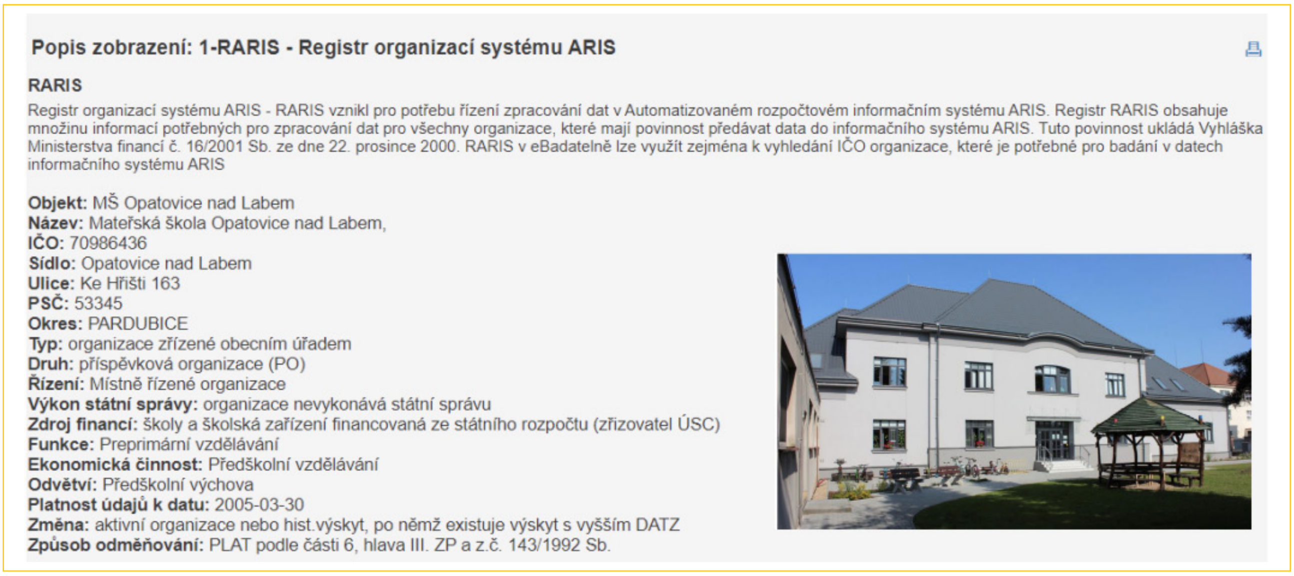

DIGITÁLNÍ ARCHIV

Přístup k archiváliím: obsah databází

Zpřístupněné databáze

1.3 ARIS: šablony tabulek (DBC:Aris_siard)
1.4 RARIS:Registr organizací systému ARIS (DBC:aris)
1.51 ARIS - účetní a finanční výkazy místně řízených organizací (DBC:aris)
1.52 ARIS - účetní a finanční výkazy ústředně řízených organizací (DBC:aris)
1.7 ARIS - Sumáře ÚSC 211,311,411 (DBC:aris)
1.8 ARIS - Bilance-Výkazy 211 ÚSC (DBC:aris)
1.91 ARIS - Výkazy o plnění rozpočtu ústředně řízených organizací (DBC:aris)
2.1 Rejstřík školských právnických osob (DBC:ResjstrikSPO)
2.2 Rejstřík škol a školských zařízení (DBC:rejstrik_sasz)
NA - pomůcka Státní úřad pro věci církevní (DBC:suc)
NA - pomůcka KHFrank fotoarchiv (DBC:KHFrank)

Figure 35: Service selection screen at the digital reading room at NACR



Popis zobrazení: 1-RARIS - Registr organizací systému ARIS

RARIS

Registr organizací systému ARIS - RARIS vznikl pro potřebu řízení zpracování dat v Automatizovaném rozpočtovém informačním systému ARIS. Registr RARIS obsahuje množinu informací potřebných pro zpracování dat pro všechny organizace, které mají povinnost předávat data do informačního systému ARIS. Tuto povinnost ukládá Vyhláška Ministerstva financí č. 16/2001 Sb. ze dne 22. prosince 2000. RARIS v eBadatelně lze využít zejména k vyhledání IČO organizace, které je potřebné pro badání v datech informačního systému ARIS

Objekt: MŠ Opatovice nad Labem
Název: Mateřská škola Opatovice nad Labem,
IČO: 70986436
Sídlo: Opatovice nad Labem
Ulice: Ke Hřišti 163
PSČ: 53345
Okres: PARDUBICE
Typ: organizace zřízené obecním úřadem
Druh: příspěvková organizace (PO)
Řizení: Místně řízené organizace
Výkon státní správy: organizace nevykonává státní správu
Zdroj financí: školy a školská zařízení financovaná ze státního rozpočtu (zřizovatel ÚSC)
Funkce: Preprimární vzdělávání
Ekonomická činnost: Předškolní vzdělávání
Odvětví: Předškolní výchova
Platnost údajů k datu: 2005-03-30
Změna: aktivní organizace nebo hist.výskyt, po němž existuje výskyt s vyšším DATZ
Způsob odměňování: PLAT podle části 6, hlava III. ZP a z.č. 143/1992 Sb.

Figure 36: RARIS database record of a kindergarten in its archival rendition at NACR (the picture being an illustration only)

The packages are enhanced with several simple tables to better understand the meaning of the data. RARIS can be used to find the necessary identification numbers of reports. The dbDIPview tool can handle CLOB (variable-length character large object string) or BLOB data easily regardless of whether they are inside or outside the SIARD file.

Rechtorik‘s talk was about a proof of concept for standardising the DB transfer process. The NACR has also accepted other DB systems in CSV and ACCDB formats, but now wants to deploy preservation and presentation of databases in a scalable and standardised way. An eminent goal is to provide means of periodic comparison of certain datasheets and variables over time.

To conclude, Rechtorik pointed out some doubts. DB preservation needs a plan for future generations. Will a description of columns be enough in 20 years? Will government agencies be satisfied with a SIARD file without a working environment? He was skeptical, seeing SIARD as a great standard but feeling that it must be extended to fit into more diverse use cases. He saw the necessity to share experiences with creation of SIARD and presentation tools for SIARD packages and archived DBs in general.

Questions and discussion

Kai Naumann named an approach that had not been mentioned during the workshop: to hop from database instance to instance over decades or longer. If an Informix environment is outdated, a service could move the data to Oracle, for which a new working environment is to be deployed. From there, two decades later, the data might travel on to MariaDB and another user interface. If there were mechanisms to prove that the jump from, e.g., Oracle to MariaDB was controlled and that nothing significant was lost in the transition, the gap of 60 years could be bridged, too.

Database archiving with dbDIPview: a brief overview

Boris Domajnko (Archives of the Republic of Slovenia, Ljubljana)

Boris Domajnko (M.Sc. of Computer Science) has been working at the Department of electronic archives and IT support since 2008. From the beginning, he was intrigued by the problem of preserving databases and making them accessible. He has used his experience in areas of software engineering from his previous work in the private sector and developed a solution that grew into a full production process.

dbDIPview (an acronym for database DIP viewer) is a tool and approach for long-term archiving of databases. It helps us make their information easily accessible for a non-technical user, be it an archivist or a reading room visitor. Domajnko developed it for the first database ingest in 2010 and continues to improve it whenever a new production scenario for ingest or access emerges. The archiving projects have been done with public registries based on dBase, its successors, and also with data from modern DBMSs. The tools are used both for ingest and for access scenarios.

Generally, in AIP, a content information package consists of a data object and representation description information. dbDIPview is used for creating, packing, and using the latter. The data object remains unchanged and can preserve its authenticity. It contains SIARD or CSV files or their combinations.

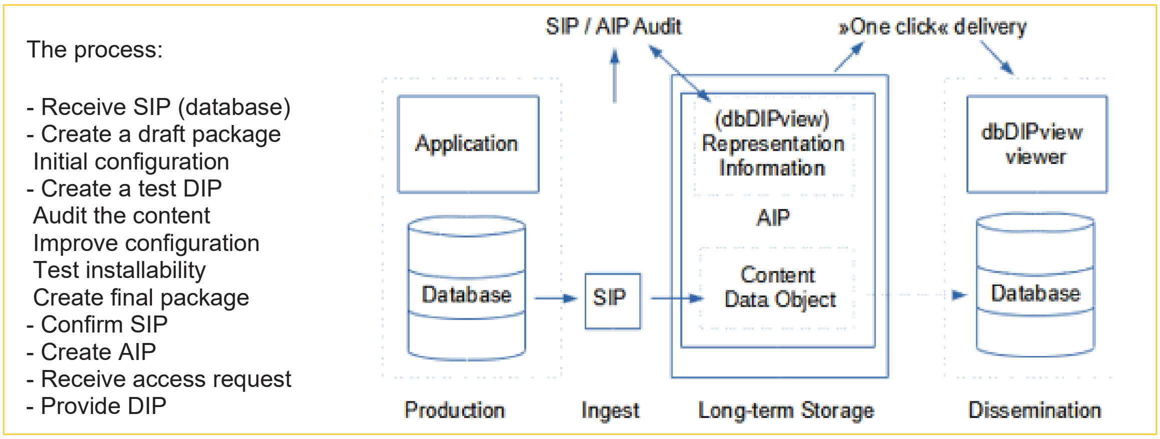
In the pre-ingest phase, a SIP with a data object is transformed into a preliminary DIP on a dbDIPview server. The archivist now configures the viewer details and examines the content, and finally rejects or accepts the SIP. At this stage, database knowledge is still needed as we try to imitate the search and reporting functionality of the original application by preparing expected use cases in XML. Therefore, the output of old and new applications will be compared.

With this approach, three risks are covered by ensuring the following:

- SIP content is checked before the archive confirms the ingest,
- the access is possible when the pre-ingest is done, and
- the documentation received is adequate.

In the end, the viewer configuration is ready, and the representation information object is built with a single command. This object will accompany the data object. An AIP can be set up and ingested. It is now available for automated deployment.

When an end user requests access to data, the archivists can deploy the database with a few commands in the admin menu, or in an automated way, based on an “order” XML file with needed information, including possible redaction. The end user can then get and use the data ordered – not as a database administrator, but as a user of an application that mimics output of the original application. Nested tree view menus are used for searching and pre-configured hyperlinks between reports are possible for further drilling down on data. Additional on-screen descriptions help to understand the data shown.

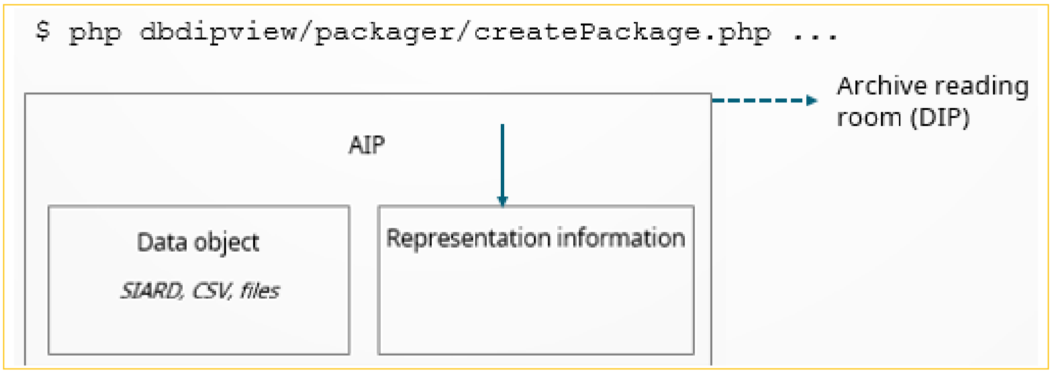


The process:

- Receive SIP (database)
- Create a draft package
 Initial configuration
- Create a test DIP
 Audit the content
 Improve configuration
 Test installability
 Create final package
- Confirm SIP
- Create AIP
- Receive access request
- Provide DIP

Figure 37: A risk minimizing process steps for DB archiving as proposed by Domajnko

Command line tools are used to create, install and uninstall packages. This example will validate the files and create a package that can contain CSV files or use external SIARD files:

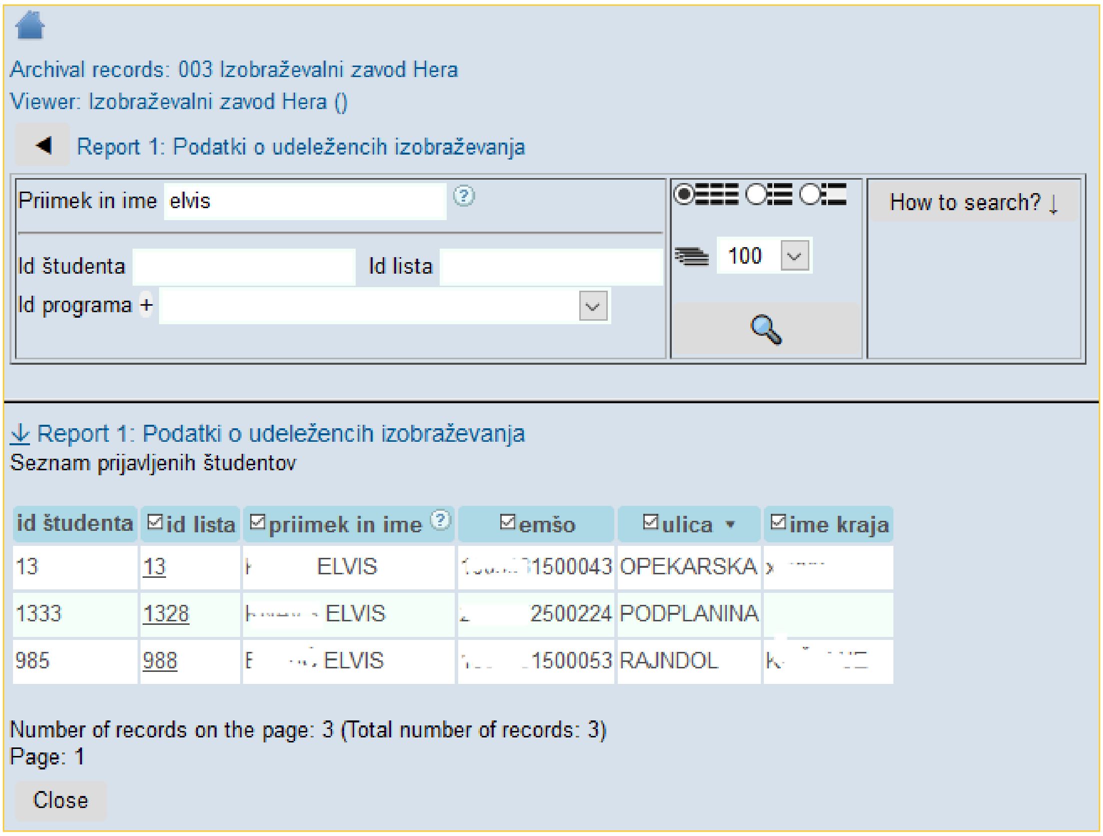


```
$ php dbdipview/packager/createPackage.php ...
```

Figure 38: The dbDIPview command for creating an AIP with representation information

The usability of displayed data can further be increased by allowing for longer text in the headers to explain the details about a certain table or report. External attachments and BLOB content can be shown. The solution is targeting end users who need friendly easy-to-use advanced search functionality. Future researchers could still get direct access to the tables, e.g. for statistical analysis. Nevertheless, the typical use of dbDIPview is getting an unambiguous concrete piece of information.

Every database ingest and use is a bit different. However, all required scenarios have been successfully covered so far.



Archival records: 003 Izobraževalni zavod Hera
Viewer: Izobraževalni zavod Hera ()

◀ Report 1: Podatki o udeležencih izobraževanja

Priimek in ime: elvis

Id študenta:

Id lista:

Id programa +:

100

↓ Report 1: Podatki o udeležencih izobraževanja
Seznam prijavljenih študentov

id študenta	☑id lista	☑priimek in ime	☑emšo	☑ulica ▾	☑ime kraja
13	13	[redacted] ELVIS	[redacted]1500043	OPEKARSKA	[redacted]
1333	1328	[redacted] ELVIS	[redacted]2500224	PODPLANINA	
985	988	[redacted] ELVIS	[redacted]1500053	RAJNDOL	[redacted]

Number of records on the page: 3 (Total number of records: 3)
Page: 1

Figure 39: A typical dbDIPview user interface screen with hyperlinks

Questions and discussion

- Is dbDIPview open source? – Yes, it is on GitHub: https://github.com/dbdipview/dbdipview/wiki, and our institution can also be contacted in case any additional information is needed.
- Raja Appuswamy asked if a SIARD reader is the right tool – couldn’t you move the SIARD file to a common database format instead of using the dbDIPview framework? This would allow use of all kinds of analysis tools that work on databases. Boris Domajnko replied that authenticity is also important. The typical archival use case is someone addressing the archive and asking for a document. The archivist must be able to use the authentic DB data and find the right information at once, without extra time for tackling the tables. People need a clear output from the tool. Of course, researchers who want to make in-depth studies can get the SIARD package and perform their analyses on any platform they desire.
- Torbjørn Aasen noted that it is important to work out what should be kept and what should be thrown away. It should be the aim to describe as much as feasible so users can later analyse it.
- James Doig remarked that resourcing is an issue. A normalisation approach such as SIARD standardises format makes life much easier for an archive going forward.

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).

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.

Plenary with outcomes of parallel groups, discussion of contentious issues, outlook

Group A started discussing standards and this included SIARD. Most participants regard SIARD as suitable and reasonable, despite the objections raised by Damir Bulic (p. 22). Clearly, SIARD needs more testing, more use cases, and examples. The group discussed tools and methods to assess authenticity of an export and found that there were little tools or methods available for this task.

In addition, the idea emerged that SQLite could serve as an option for future SIARD-like formats.

Important interoperability issues on SIARD were also discussed. When encountered, these issues mean extra work. Improved tool quality would be beneficial to all. All tools are very good but need to be made better

Group B noted differences between emulation and migration-based approaches. Migration appeared as the most popular way because it is mostly sufficient to derive a dataset from a database and not to preserve the whole database. Also, licencing issues and costs of solutions were discussed. NoSQL technology was identified as a field that SIARD doesn't cover. All deplored the lack of one-size-fits-all solution. Geodata can extract some geometries into GML alongside SIARD but needs to ensure the connections between the files/elements are present; otherwise it would be impossible to make sense of the information. Documentation of a database and its environment was discussed with the E-ARK Content Information Type Specification (CITS) for databases as an option for better templates.

How to meet again and keep discussion going?

Faria claimed that archiving by design would be a smooth solution.

Anders Bo Nielsen (Danish National Archives) got back to questions regarding similarities between GML and SIARD, confirming they contain similar issues.

Once again, the https://listserv.dilcis.eu/info/rdb-aig mailing list was advertised.

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.

References

This bibliography does not only show the cited works but every item that seemed of interest during the investigations for the workshop.

Anderson B., Braxton S., Imker H., Popp T. (2018), The Art of Preserving Scientific Data: Building Collaboration into the Preservation of a Legacy Database (iPRES 2018 Paper), https://doi.org/10.17605/OSF.IO/RH9SU.

Bewertung elektronischer Fachverfahren (2015), Arbeitspapier des VdA-Arbeitskreises Archivische Bewertung. In: Archivar 2015(1), p. 90-92, https://www.archive.nrw.de/sites/default/files/media/files/Archivar_1_10.pdf.

Barats C., Schafer V., Fickers A. (2020), Fading Away … The challenge of sustainability in digital studies. In: Digital Humanities Quarterly vol. 14 Nr 3, http://www.digitalhumanities.org/dhq/vol/14/3/000484/000484.html.

Barats C., Dister A., Gambette P., Leblanc J-M., Peres-Leblanc M. (2016), Analyser des pétitions en ligne: potentialités et limites d'un dispositif d'étude pluridisciplinaire. In: JADT 2016, Actes des Journées internationales d'Analyse statistique des Données Textuelles (2016). http://jadt2016.sciencesconf.org.

Brown A. (2008), The National Archives UK Digital Preservation Guidance Note 1: Selecting file formats for long-term preservation, https://cdn.nationalarchives.gov.uk/documents/selecting-file-formats.pdf.

Catalogue Model Proposal [of National Archives UK] (2020) (TNA-CMP20/1 Technical Paper), https://cdn.nationalarchives.gov.uk/documents/omega-catalogue-model-proposal.pdf.

Cha S.-J., Choi Y. J., Lee K.-C. (2015), Development of Preservation Format and Archiving Tool for the Long-Term Preservation of the Database (IMCOM 2015 Paper), https://dl.acm.org/doi/10.1145/2701126.2701192.

Cochrane E., Suchodoletz D., Crouch M. (2013), Database Preservation Using Emulation – a Case Study. In: Archifacts 2013, p. 80-95.

Cochrane E. (2012), Migrating a Windows 2000 Database Server to Virtualized and Emulated Hardware (Open Preservation Foundation Blogpost), https://openpreservation.org/blogs/migrating-windows-2000-database-server-virtualized-and-emulated-hardware/ – The process documentation formerly attached is to be found at https://yale.box.com/s/uro2wzbz9k149lulew3xzi4toxbgviof.

Christophersen A. et al. (2022). Software Metadata Recommended Formats Guide. Software Preservation Network, https://www.softwarepreservationnetwork.org/smrf-guide/.

Christophides V., Buneman P. (2007), Report on the First International Workshop on Database Preservation (PresDB'07), In: ACM SIGMOD Record, Vol. 36, Issue 3, pp. 55-58, September 2007, https://sigmodrecord.org/issues/sigmod-record-september-2007, https://doi.org/10.1145/1324185.1324197.

CITS SIARD. E-ARK Content Information Type Specification for Relational Databases using SIARD, Version 1.0.0, DILCIS Board, https://citssiard.dilcis.eu/specification/CITS_SIARD_version1_0_0.pdf.

Däßler R., Schwarz K. (2010), Archivierung und dauerhafte Nutzung von Datenbanken aus Fachverfahren – eine neue Herausforderung für die digitale Archivierung. In: Archivar 2010(1), p. 6-18. https://www.archive.nrw.de/sites/default/files/media/files/Archivar_1_10.pdf.

Dekeyser K. (2012), User story: archiving a FoxPro database (Open Preservation Foundation Blogpost), https://openpreservation.org/blogs/user-story-archiving-foxpro-database/.

Dorendorf S. (2007), Kosten und Nutzen von Datenbankreorganisationen: Grundlagen, Modelle, Leistungsuntersuchungen. In: Kemper, A., Schöning, H., Rose, T., Jarke, M., Seidl, T., Quix, C. & Brochhaus, C. (Hrsg.), Datenbanksysteme in Business, Technologie und Web (BTW 2007) – 12. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme" (DBIS). Bonn: Gesellschaft für Informatik e. V. (S. 397-416). https://dl.gi.de/handle/20.500.12116/31812.

EaaSI, Software Preservation Network Emulation-as-a-Service Infrastructure program, https://www.softwarepreservationnetwork.org/emulation-as-a-service-infrastructure/.

Erpaworkshop (2003), The Long Term Preservation of Databases, Bern April 9-11, https://www.yumpu.com/en/document/view/41511421/erpaworkshop-the-long-term-preservation-of-databases-erpanet/13, also https://docplayer.net/12064179-Erpaworkshop-the-long-term-preservation-of-databases-erpanet-workshop-report-bern-april-9-11-2003.html.

Faria L., Büchler M., Aas K. (2016), Workshop on Relational Database Preservation Standards and Tools (iPRES 2016 Paper), https://doi.org/11353/10.502816.

Ferreira B., Faria L., Ferreira M., Ramalho J. C. (2016), Database Preservation Toolkit. A relational database conversion and normalisation tool (iPRES 2016 Paper). https://hdl.handle.net/11353/10.503182.

Fitzgerald N. (2013), Using data archiving tools to preserve archival records in business systems – a case study (iPRES 2013 Paper), https://hdl.handle.net/11353/10.378094.

GAMS: Geisteswissenschaftliches [i.e. Humanities] Asset Management System, https://gams.uni-graz.at.

Gladney H. M., Lorie R. A. (2004), Trustworthy 100-Year Digital Objects: Durable Encoding for When It‘s Too Late to Ask. In: Computing Research Repository (CoRR), https://arxiv.org/abs/cs/0411092.

Guideline for CITS SIARD (2021), Guideline for the E-ARK Content Information Type Specification for Relational Databases using SIARD, Version 1.0.0, DILCIS Board, https://citssiard.dilcis.eu/guideline/Guideline_CITS_SIARD_1_0_0.pdf.

Heuscher S., Jährmann J., Keller-Marxer P., Möhle F. (2004), Providing authentic long-term archival access to complex relational data. In: European Space Agency Symposium ”Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data”, October, Frascati, Italy, 2004.

Van Horn J., van Pelt J. (2007), 1st INCF Workshop on Sustainability of Neuroscience Databases. In: Nature Precedings (2008), https://doi.org/10.1038/npre.2008.1983.1.

Keitel C. (2004), Erste Erfahrungen mit der Langzeitarchivierung von Datenbanken. Ein Werkstattbericht. In: Hering, R., Schäfer, U. (Ed.), Digitales Verwalten – Digitales Archivieren. 8. Tagung des Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“. https://dx.doi.org/10.15460/HUP.STAHH.19.82.

King’s Digital Lab (2019), Archiving and Sustainability. KDL‘s pragmatic approach to managing 100 Digital Humanities projects, and more … https://www.kdl.kcl.ac.uk/our-work/archiving-sustainability/.

Klopprogge M. R., Lockemann P. C. (1983), Modelling Information Preserving Databases: Consequences of the Concept of Time, in: VLDB ‚83: Proceedings of the 9th International Conference on Very Large Data Bases, S. 399–416.

Kronenwett S., Mathiak B. (2017), A Survey on Research Data at the Faculty of Arts and Humanities of the University of Cologne, in: Digital Humanities 2017 Conference Abstracts, 294-298, https://dh2017.adho.org/program-2/abstracts/.

Lindley A. (2013), Database Preservation Evaluation Report – SIARD vs. CHRONOS. Preserving complex structures as databases through a record centric approach? (iPRES 2013 Proceedings), https://dblp1.uni-trier.de/db/conf/ipres/ipres2013.html.

Marinelli E., Ghabach E., Bolbroe T., Sella O., Heinis T., Appuswamy R. (2021), DNA4DNA: Preserving Culturally Significant Digital Data with Synthetic DNA (iPRES 2021 paper), https://osf.io/z6yx3/.

Memishi B., Appuswamy R., and Paradies M. (2019), Cold Storage Data Archives: More Than Just a Bunch of Tapes, in: Proceedings of the 15th International Workshop on Data Management on New Hardware (DaMoN‘19). Association for Computing Machinery, New York, NY, USA, Article 1, 1–7, https://doi.org/10.1145/3329785.3329921.

Merkle signature scheme (Wikipedia Article), https://en.wikipedia.org/wiki/Merkle_signature_scheme.

Moore R. W. (2018), Archiving Experimental Data. Encyclopedia of Database Systems (2nd ed.) 2018.

Moore's Law (Wikipedia Article), https://en.wikipedia.org/wiki/Moore%27s_law.

Müller H. (2009), Archiving and Maintaining Curated Databases, In: 21. Workshop Grundlagen von Datenbanken : 02.-05. Juni 2009, Rostock-Warnemünde : Proceedings, p. 135-139, https://doi.org/10.18453/rosdok_id00002203.

Naumann K. (2021), 125 databases for the Year 2080. In: Proceedings of the 2020 Web Archiving and Digital Libraries Workshop, https://vtechworks.lib.vt.edu/handle/10919/99569.

Neuefeind C., Schildkamp P., Mathiak B., Karadkar U., Stigler J., Steiner E., Vasold G., Tosques F., Ciula A., Maher B., Newton G., Arneil G., Holmes M. (2020), Sustainability Strategies for Digital Humanities Systems, DH2020 Panel Paper, https://dh2020.adho.org/wp-content/uploads/2020/07/565_SustainabilityStrategiesforDigitalHumanitiesSystems.html.

Ohnesorge K. W., Aas K., Delve J., Lux Z., Tømmerholt P. M., Nielsen A. B., Büchler M. (2016), Tutorial on Relational Database Preservation (iPRES 2016 Paper), https://doi.org/11353/10.502822.

Olson J. E. (2010), Database archiving: how to keep lots of data for a very long time. ISBN 978-0123747204.

Olson, J. E. (2011), Data Quality and Database Archiving: The Intersection of Two Important Data Management Functions (5. MIT Information Quality Industry Symposium 2011 Presentation), http://mitiq.mit.edu/IQIS/Documents/CDOIQS_201177/Papers/01_05_T2B_Olson.pdf.

Perlmutter M. (2021), The Lost Picture Show: Hollywood Archivists Can't Outpace Obsolescence. In: Green Reporter, https://green-reporter.com/the-lost-picture-show-hollywood-archivists-cant-outpace-obsolescence/.

Peterson M., Zasman G., Mojica P., Porter J. (2007), 100 Year Archive Requirements Survey, https://www.snia.org/sites/default/orig/100YrATF_Archive-Requirements-Survey_20070619.pdf.

Potthoff J. (2012), Beweiswerterhaltendes Datenmanagement im elektronischen Forschungsumfeld. In: Müller P., Neumair B., Reiser H. & Rodosek G. D. (Ed.), 5. DFN-Forum Kommunikationstechnologien – Verteilte Systeme im Wissenschaftsbereich. Bonn: Gesellschaft für Informatik e.V.. (S. 109-118). https://dl.gi.de/handle/20.500.12116/18172.

Preserving databases using SIARD (2020), Experiences with workflows and documentation practices. RDB SIARD, V. 1.0, https://dilcis.eu/images/2020review/9_Draft_SIARD_Case_Study_1.pdf.

Preserving databases using SIARD. Case Study 2 (2020). Experiences with workflows and documentation practices. RDB SIARD, V. 1.0 Review version, https://dilcis.eu/images/2020review/10_Draft_SIARD_Case_Study_2.pdf.

Raselli D. (2014), Verfahren zur Langzeitarchivierung von Datenbankinhalten aus Fachanwendungen und die Dokumentation dazugehöriger Prozessvorgänge (Master thesis), Berne University, http://nbn-resolving.de/urn:nbn:de:0008-2017080108.

RDBMS Genealogy (2018), HPI Genealogy of Relational Database Management Systems v6, https://hpi.de/en/naumann/projects/rdbms-genealogy.html.

Rumianek M. (2013), Archiving and Recovering Database-driven Websites. D-Lib Mag. 19(1/2) (2013) http://www.dlib.org/dlib/january13/rumianek/01rumianek.html.

SIARD-2.1.1 Format Specification (2019), https://www.bar.admin.ch/dam/bar/en/dokumente/kundeninformation/siard_formatbeschreibung.pdf.download.pdf/siard_format_descriptioning.pdf.

Stabilize. Solving digital infrastructure obsolescence, https://www.stabilize.app/.

Steinke T., Padberg F., Schoger A., Rechert K. (2016), Project EmiL – Emulation of Multimedia Objects (iPRES 2016 Paper), https://hdl.handle.net/11353/10.503170.

Ur Rahman A., David G., Ribeiro C. (2010), Model Migration Approach for Database Preservation, 12th International Conference on Asia-Pacific Digital Libraries ICADL, Proceedings, https://dx.doi.org/10.1007/978-3-642-13654-2_10.

Weise M., Staudinger M., Michlits C., Gergely E., Stytsenko K., Ganguly R., Rauber A. (2022). DBRepo: A Semantic Digital Repository for Relational Databases (IDCC22 Presentation), https://doi.org/10.5281/zenodo.6637333.

Whitt R. S. (2017), „Through A Glass, Darkly“ Technical, Policy, and Financial Actions to Avert the Coming Digital Dark Ages, 33 Santa Clara High Tech. L.J. 117 (2017). https://digitalcommons.law.scu.edu/chtlj/vol33/iss2/1.

Zeller B., Herbst A., Kemper A. (2003), XML-Archivierung betriebswirtschaftlicher Datenbank-Objekte. In: Weikum, G., Schöning, H. & Rahm, E. (Hrsg.), BTW 2003 – Datenbanksysteme für Business, Technologie und Web, Tagungsband der 10. BTW Konferenz. Bonn: Gesellschaft für Informatik e.V.. (S. 127-146). https://dl.gi.de/handle/20.500.12116/30094.

Supporting editors' biographies

These proceedings were edited with the support of:

Francesco Gelati *(ORCID 0000-0002-6066-1308) is Head of Section for Digital Services and Internal Consultancy at the Hamburg University Archives (Germany). He studied history and archival science in Strasbourg and Venice. He worked previously in Brussels at the Belgian State Archives for EHRI (European Holocaust Research Infrastructure) as data import manager and in Munich at the Leibniz Institute for Contemporary History as archivist in charge of research data. He is a member of the DARIAH-EU (Digital Research Infrastructure for the Arts and Humanities) Research Data Management Working Group. His research fields are: archival metadata standards, FAIR data and data quality, linked data.*

Kevin A. McMahon *retired from Sandia National Laboratories (USA) in November 2019 after 31 years where he developed and managed numerous relational database systems. He was also an Adjunct Professor teaching Statistics and Systems Analysis at the Master's level for Webster University. Kevin holds a bachelors in Chemistry and an MBA in Management Information Systems.*

Oliver Watteler *is a senior researcher at GESIS – Leibniz Institute for the Social Sciences at Cologne (Germany), where he works in the department Data Services for the Social Sciences (DSS). He holds a master‘s degree in history and political science and is responsible for data acquisition. Together with colleagues he advises and trains on topics of research data management. His focus is on legal conditions of RDM, especially data protection, on which he also publishes. He is a member of the Leibniz Association‘s Data Protection Working Group and vice member of the GESIS Ethics Committee.*

Glossary

This glossary serves primarily for acronyms but also for names of companies and institutions.

APEx	Archives Portal Europe network of excellence
ARIS	Automated Budget Information System (Czech Republic)
AWS	Amazon Web Services
BLOB	Binary Large Object
CESSDA	Council of European Social Science Data Archives
CHRONOS	Software for database archival
CITS	Content Information Type Specification (part of E-ARK standards, see references)
CLOB	Variable-Length Character Large Object string
COBOL	Common Business Oriented Language
COPIM	Community-led Open Publication Infrastructures for Monographs
CPU	Central Processing Unit
CSP GmbH	German company
CSV	Comma Separated Values file
DB	Database
DB2	Database 2, IBM Relational DB products
DBMS	Database Management System
DBPTK	Database Preservation Toolkit by Keep LDA
DCC	Digital Curation Centre
DCH	Data Center for Humanities
DDI	Data Documentation Initiative
DDL	Data Definition Language
DFG	The German Research Foundation
DILCIS	Digital Information Lifecycle Interoperability Standards
DIMAG	German development community for digital repository software
DNA	Desoxribonucleic Acid
E-ARK	European Archival Records and Knowledge preservation
EDRMS	Electronic Document and Records Management System
E-R	Entity Relationship
EUPALIA	French company that specialises in digital longevity
EURECOM	French research centre in digital sciences
GESIS	German institute for the social sciences
GIS	Geographic Information System
GML	Geographic Markup Language
GUI	Graphical User Interface

HR	Human Resources
HTML	Hyper Text Markup Language
IKAMR	Norwegian Municipal Archive
ISO	International Standards Organisation
IT	Information Technology
iPRES	A series of international conferences on digital preservation
JPEG	Joint Photographic Experts Group image file format
JSON	JavaScript Object Notation
KDRS	Norwegian data centre for municipalities
KLA	German Conference of State Archive‘s Directors
Keep LDA	Portuguese company
MySQL	Open-source relational database system
NAA	National Archives Australia
NACR	National Archives Czech Republic
NoSQL	Not only Structured Query Language
OAIS	Open Archival Information System
ODBC	Open Data Base Connectivity
OPERAS	Open Scholarly Communication in the European Research Area for Social Sciences and Humanities
PASIG	Preservation and Archiving Special Interest Group
PDF	Portable Document Format
PROTAGE	EU research project, 2007-2011
RARIS	See ARIS
RDF	Resource Description Framework
RiC	Records in Contexts
RODA	Long term digital repository solution by Portuguese company Keep
SAP	Systems, Applications, and Products in Data Processing
SIARD	Software Independent Archival of Relational Databases
SIP	Submission Information Package
SQL	Structured Query Language
SQLite	Database format optimised for portability
TB	Terabyte
TIFF	Tagged Image File Format
TNA	The National Archives (UK)
VOIS	Valuation Office Information System (Australia)
XML	Extension Markup Language