Efficient Data Management Systems for Research Data in Databases

Martin Weise [0000-0003-4216-302X]

Data Science, TU Wien

May 7rd, 2024
About me

PreDoc researcher at DS-IFS
MSc in “Software Engineering & Internet Computing” (TU Wien)
BSc in “Software & Information Engineering” (TU Wien)
About me (cont.)
Data Deluge

Research becomes increasingly data-driven

Vast amounts of information generated & analyzed, demands a combination of methods/technologies to store, process, share and reuse research data.
• Infrastructure supporting (implied) reuse of academic data transforms at snails pace
• Hindered knowledge discovery and (ultimately) innovation
• RDM practices ([meta-]data documentation, long-term preservation, sharing) exist
• Delayed data publications due to data stewardship overhead ~⇒ questionable benefit for researchers
• Traditional trail of quality control (isolated validation) on data publications before data can be published
• Databases as collections of **structured** data are very efficient data storage systems
• Wide spectrum:
  • In-process library written in C to e.g., store state information,
  • Graph to e.g., detect drug-drug interaction (adverse),
  • Neural network deep-learning.
• Some guidelines for efficient RDM in this context exist (mainly database design in 3NF) → normalization algorithms are widespread now
Use Case 1: Efficient data publications

- **RDM units lack skills** to deploy DBMS', import dataset
- Consider **user access privileges**, data security implications, make data accessible for legitimate interest
- **Decrease of utility** $\rightarrow$ unauthorized access, data breaches, -integrity, DoS
- Deposit **database dump** in a trusted repository/data mart $\rightarrow$ shifts burden to researchers who wants to use the data

Warrants advanced RDM practices and accompanying applied services to make research data available with machine-actionable interfaces to find, access, interoperate and reuse the data.
Use Case 2: Reproducible experiments

- **Complex data** in some disciplines (high-energy physics, genomics, material science)
- Analysis on a single machine **not possible** (lack of storage-, compute- or economic resources)
- Ubiquitous working environments for researchers to **analyze** their complex data
- Web-interfaces **tailored to their needs** (naming, workflows, etc.) $\Rightarrow$ VREs (e.g. Galaxy, CERN VRE)

Demand for machine-actionable interfaces for data in DBs, flexible integration into VREs (agnostic from research domain), efficient subsetting mechanisms, computing close to data.
Use Case 3: Environment for confidential data

- Protection/maintaining control over confidential data while allowing access by third parties → significant challenge
- Access data in a highly controlled and monitored environment, high security guarantees
- Security by obscurity prevalent, slow shift towards data visiting

Missing systematic knowledge exchange for these trusted research environments → documentation of architecture and processes needed
FAIR Guiding Principles


Contribution: guide, motivation

- Software framework and **data management system** for global academic federations currently handling 450PB\(^1\) globally
- Identities get a **namespace** (c.f. UNIX home directory), integration into LDAP systems
- Multiple protocols for addressing storage (HTTP, WebDAV, S3)
- User-controllable replication management

Contribution: concepts, architecture, operation experience

\(^1\)2019, likely way more currently

- **Reusable analyses** for future reuse/reproduction in a YAML file (assets, parameters, processes) i.e. expensive-to-gather particle physics data
- **Container**-based common-/batch-/DAG workflow
- Executes **workflows** in cloud-environment (Kubernetes)

Contribution: concepts, architecture, operation experience

Contribution: building blocks

Contribution: operation experience (and approachable by TUW)
Definition 1: Data Management Systems for Research Data in Databases (DMSfRDB) consist of a (i) data infrastructure (management), (ii) research data repository (metadata), and (iii) database management system (data).

**P1**: Missing systematic understanding of the characteristics and building blocks of DMSfRDB.

**Motivation**: General (high-level) building blocks of DMSfRDB are known for TREs on abstract degree, descriptions are missing a technological baseline and concrete technical blocks how to achieve that (with open-source software).
P2: Haphazard and manual privacy preserving data publishing for confidential data by researchers.

Motivation: privacy models exist for decades, tools exist too. Storage of secrets and audit information (fingerprinting metadata) within the infrastructure necessary. Introduction role data publisher.
P3: Limited utility of big-volume research data in DMSfRDB.

Motivation: FAIRness of data in databases is very poor, interfaces exist but cannot be used by e.g. aggregators, poor subset mechanisms. Level of trust needed to gain access is high.
Stakeholder interactions (edges) and interests (clouds)
“What are the key technical elements and their characteristics of DMSfRDB?”

- **RQ 1.1** “What are the key technical elements of DMSfRDB in a private cloud environment?”
- **RQ 1.2** “What are the key characteristics of DMSfRDB to setup and to what degree can they be measured?”

**Feedback?**
Scope too broad? Useful? Reasonable? → after talk pls
“What are prerequisites and methods for efficient publications of data in DMSfRDB?”

- **RQ 2.1** “To which degree can data subsets be published in DMSfRDB without significant decrease of data utility?”
- **RQ 2.2** “To which degree can confidential data be published in DMSfRDB in an academic context?”

Feedback?
Scope too broad? Useful? Reasonable? → after talk pls
“To what extent can findability, availability and interoperability be improved in the context of DMSfRDB?”

- **RQ 3.1** “To what extent can findability be improved in the context of DMSfRDB?”
- **RQ 3.2** “To what extent can availability be improved in the context of DMSfRDB?”
- **RQ 3.2** “To what extent can interoperability be improved in the context of DMSfRDB?”

Feedback?

Scope too broad? Useful? Reasonable? → after talk pls
Established **three cycle view** of design science research [4] within the *information systems* field based on the *information systems research framework* [?].
Design science is **pragmatic** (practical, relevant) in nature but not sufficient for good solution: need **strong contributions** in the relevance cycle and rigor cycle:

- Connecting establishing contextual environment with DMSfRDB (*Relevance Cycle*) ⇐
- Connecting design science activities of efficient DMSfRDB with the knowledge base of the contextual environment (*Rigor Cycle*)
- Building/evaluating the efficient DMSfRDB and processes (*Design Cycle*) ⇐
Main Methods

Narrow down the (high-level) groups, use the established design science methodology **engineering cycle** [15] for information systems/software engineering.
Overview of work plan

Figure: Planned work assignments (A1-6) and three cycle view of design science research.
Problem Investigation [15]

- Problem definition (A1) to establish pivot for the thesis
  - Limit scope for Security/Data Management/DBMS
  - Reasonable focused enough cross-section for components and formulation of generalized solution

- Perform a systematic mapping study [8] to structure a research topic area in early phases\(^2\)

- Perform a stakeholder analysis [1] (part of A2) to early identify them and analyze level of awareness & desire to carry out the improvement (desires vs. goals).

\(^2\)Compare systematic mapping study [6] with strong completeness demand
Use Case 1 (“Efficient Data Publications”):

Use Case 2 (“Reproducible experiments”):

Use Case 3 ("Environment for confidential data"): 


Treatment Design [15]

- Perform a **multiple case study** [10] (A3) to devise theoretical SDMfRDB building blocks for specific use cases of interest → strengths in systematic cross-case comparison³
- Perform a **systematic mapping study** [8] (A4) to identify and analyze⁴ the SoTA RDM approaches and methods

---

³Opposed to single-case study  
⁴Suitable method for weakly defined areas
Future Work

Finish the proposal & proficiency evaluation

- **Relevance cycle**: contributions sufficient ✓
- Design cycle: development ongoing (software engineering)
- **Rigor cycle**: publish method/experience paper(s): ACM/IMS JDS, ACM TMIS, …
Remarks?

- Methodology proper?
- Related work suggestions? Did I miss your paper?
- Research questions sufficiently small?
- Interested in contributing?
A Taxonomy of Stakeholders: Human Roles in System Development. 

Rucio: Scientific data management. 
*Computing and Software for Big Science*, 3(1), August 2019. 


Systematic literature reviews in software engineering – a systematic literature review.  

Planning Future Health: Developing Big Data and System Modelling Pipelines for Health System Research. 
doi:10.11128/sne.27.tn.10396.

The theory contribution of case study research designs. 
doi:10.1007/s40685-017-0045-z.

Trusted research environments: Analysis of characteristics and data availability. 
2024. 

doi:10.2218/ijdc.v17i1.825.
Design Science Methodology for Information Systems and Software Engineering.
2014.

The FAIR Guiding Principles for Scientific Data Management and Stewardship.
Scientific Data, 3(1), Mar 2016.
doi:10.1038/sdata.2016.18.
Reana: A system for reusable research data analyses. 
doi:10.1051/epjconf/201921406034.