RESEARCH DATA ALLIANCE

Data Citation Working Group Mtg @ P19 June 23 2022, Seoul

Andreas Rauber, Mark Parsons

research data sharing without barriers

rd-alliance.org

Agenda

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Harvard Data Science Review Paper
- New Ref Implementations:
  - RDF
- New Pilots:
  - DBRepo: Open source database repository system
  - OSSDIP: Secure data visiting platform for sensitive data
- “New" directions:
  - Information Retrieval Systems
  - AI online learning systems
- Other issues, next steps

Welcome and Intro

Welcome!

to the maintenance meeting

of the

WGDC

Agenda

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Harvard Data Science Review Paper
- New Ref Implementations:
  - RDF
- New Pilots:
  - DBRepo: Open source database repository system
  - OSSDIP: Secure data visiting platform for sensitive data
- “New" directions:
  - Information Retrieval Systems
  - AI online learning systems
- Other issues, next steps

Challenge: States of Dynamic Data

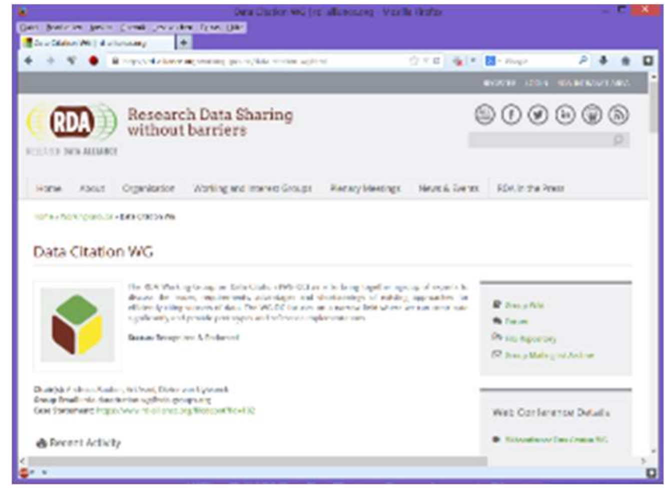
- Usually, datasets have to be static
  - Fixed set of data, no changes:
    no corrections to errors, no new data being added
- But: (research) data is **dynamic**
  - Adding new data, correcting errors, enhancing data quality, …
  - Changes sometimes highly dynamic, at irregular intervals
- Current approaches
  - Identifying entire data stream, without any versioning
  - Using “accessed at” date
  - “Artificial” versioning by identifying batches of data (e.g. annual), aggregating changes into releases (time-delayed!)
- Would like to identify precisely the **data as it existed at a specific point in time**

Challenge: Granularity of Subsets

- What about the **granularity** of data to be identified?
  - Enormous amounts of CSV data
  - Researchers use specific subsets of data
  - Need to identify precisely the subset used
- Current approaches
  - Storing a copy of subset as used in study -> scalability
  - Citing entire dataset, providing textual description of subset -> imprecise (ambiguity)
  - Storing list of record identifiers in subset -> scalability, not for arbitrary subsets (e.g. when not entire record selected)
- Would like to be able to identify precisely the **subset of (dynamic) data used** in a process

RDA WG Data Citation

- Research Data Alliance
- WG on **Data Citation: Making Dynamic Data Citeable**
- March 2014 – September 2015
  - Concentrating on the problems of **large, dynamic (changing) datasets**



- Final version presented Sep 2015 at P7 in Paris, France
- Endorsed September 2016 at P8 in Denver, CO

- Since: support for take-up/adoption, lessons-learned

https://www.rd-alliance.org/groups/data-citation-wg.html

Dynamic Data Identification and Citation

We have: Data + Means-of-access (“query”)

Dynamic Data Identification and Citation

We have: Data + Means-of-access ("query")

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Dynamic Data Identification and Citation

We have: Data + Means-of-access ("query")

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Steps:

1. Data → versioned (history, with time-stamps)

Dynamic Data Identification and Citation

We have: Data + Means-of-access (“query”)

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Steps:

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

Dynamic Data Identification and Citation

We have: Data + Means-of-access ("query")

Dynamic Data Citation:
Cite (dynamic) data dynamically via query!

Steps:

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

2. Access → **store & assign PID to "QUERY"**, enhanced with
   – **Time-stamping** for re-execution against versioned DB
   – **Re-writing** for normalization, unique-sort, mapping to history
   – **Hashing** result-set: verifying identity/correctness

   leading to landing page

Data Citation – Deployment

- Researcher uses workbench to identify subset of data
- Upon executing selection („download“) user gets
  - Data (package, access API, …)
  - PID (e.g. DOI) (Query is time-stamped and stored)
  - Hash value computed over the data for local storage
  - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

Note: query string provides excellent provenance information on the data set!

- ubset of data
- er gets
  - Data (package, acce s API, …)
  - PID (e.g. DOI) (Que is time-stamped and stored)
  - Hash value compute over the data for local storage
  - Recommended citati n text (e.g. BibTeX)
- PID resolves to land g page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

Note: query string provides excellent provenance information on the data set!

- ... ubset of data
- ... er gets
  - Data (pac...
  - PID (e.g. ...
  - Hash valu...
  - Recommended citati... text (e.g. BibTeX)

This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!

- PID resolves to land... g page
  - Provides detailed metadata, link ... o parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

Note: query string provides excellent provenance information on the data set!

This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!

Identify which parts of the data are used. If data changes, identify which queries (studies) are affected

- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citation text (e.g. BibTeX)
- PID resolves
  - Provides det
  - Option to ret
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Recommendations

Preparing Data & Query Store
- R1 – Data Versioning
- R2 – Timestamping
- R3 – Query Store

When Resolving a PID
- R11 – Landing Page
- R12 – Machine Actionability

When Data should be persisted
- R4 – Query Uniqueness
- R5 – Stable Sorting
- R6 – Result Set Verification
- R7 – Query Timestamping
- R8 – Query PID
- R9 – Store Query
- R10 – Citation Text

Upon Modifications to the Data Infrastructure
- R13 – Technology Migration
- R14 – Migration Verification

Data Citation – Output

- **14 Recommendations**
  grouped into 4 phases:
- **2-page flyer**
  https://rd-alliance.org/recommendations-working-group-data-citation-revision-oct-20-2015.html
- **Detailed report: Bulletin of IEEE TCDL 2016**
  http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf
- **Adopter's reports, webinars**
  https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html
- **Review / Lessons Learned**
  Andreas Rauber et al., Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data Harvard Data Science Review, 3(4), 2021.
  DOI [10.1162/99608f92.be565013](10.1162/99608f92.be565013).

HDSR Paper: From Principles to Adoption

Andreas Rauber, Bernhard Gößwein, Carlo Maria Zwölf, Chris Schubert, Florian Wörister, James Duncan, Katharina Flicker, Koji Zettsu, Kristof Meixner, Leslie D. McIntosh, Reyna Jenkyns, Stefan Pröll, Tomasz Miksa, and Mark A. Parsons: **Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data.** Harvard Data Science Review (HDSR), 3(4), 2021.
DOI [10.1162/99608f92.be565013]

- Principles
- 4 Reference implementations
- 8 Adoptions as Case Studies
- **Lessons Learned**

HDSR
Volume 3 Issue 4
DOI: 10.1162/99608f92.be565013
ISSN: 2644-2353

Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data

Andreas Rauber[1], Bernhard Gößwein[1,3], Carlo Maria Zwölf[4], Chris Schubert[5], Florina Wörister[1], James Duncan[6], Katharina Flicker[1], Koji Zettsu[7], Kristof Meixner[1], Leslie D. McIntosh[8], Reyna Jenkyns[9], Stefan Pröll[10], Tomasz Miksa[1,11] Mark A. Parsons[2]

1 TU Wien, Vienna, Austria
2 University of Alabama in Huntsville, AL, USA
3 Earth Observation Data Centre, Vienna, Austria
4 LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne University, UPMC Univ Paris, Meudon, France
5 Climate Change Centre Austria, Vienna, Austria
6 Forest Ecosystem Monitoring Cooperative, University of Vermont, Burlington, VT, USA
7 National Institute of Information and Communications Technology, Tokyo, Japan
8 Ripeta, Saint Louis, MO, USA
9 Ocean Networks Canada, University of Victoria, Victoria, BC, Canada
10 Cropster, Innsbruck, Austria
11 SBA Research, Austria

Abstract

Precisely identifying arbitrary subsets of data so that these can be reproduced is a daunting challenge in data-driven science, the more so if the underlying data source is dynamically evolving. Yet, an increasing number of settings exhibit exactly those characteristics: larger amounts of data being continuously ingested from a range of sources (be it sensor values, [online] questionnaires, documents, etc.), with error correction and quality improvement processes adding to the dynamics. Yet, for studies to be reproducible, for decision-making to be transparent, and for meta studies to be performed conveniently, having a precise identification mechanism to reference, retrieve, and work with such data is essential. The Research Data Alliance (RDA) Working Group on Dynamic Data Citation has published 14 recommendations that are centered around time-stamping and versioning evolving data sources and identifying subsets dynamically via persistent identifiers that are assigned to the queries selecting the respective subsets. These principles are generic and work for virtually any kind of data. In the past few years numerous repositories around the globe have implemented these recommendations and deployed solutions. We provide an overview of the recommendations, reference implementations, and pilot systems deployed and then analyze lessons learned from these implementations. This article provides a basis for institutions and data stewards considering adding this functionality to their data systems.

1 Introduction

Accountability and transparency in automated decisions (ACM US Public Policy Council, 2017) have important implications on the way we perform studies, analyze data, and prepare the basis for data-driven decision making. Specifically, reproducibility in various forms, that is, the ability to recompute analyses and arrive at the same conclusions or insights is gaining importance. This has impact on the way analyses are being performed, requiring processes to be documented and code to be shared. More critically, data–being the basis of such analyses and thus likely the most relevant ingredienable in any data-driven, decision-making process–needs to be findable and accessible if any result is to be verified. Yet, identifying precisely which data were used in a specific analysis is a nontrivial challenge in most settings: Rather than relying on static, archived data collected and frozen in time for analysis, today's decision-making processes rely increasingly on continuous data streams that should be available and usable on a continuous basis. Working on last year's (or last week's) data is not an acceptable alternative in many settings. Data undergo complex preprocessing routines, are recalibrated, and data quality is continually improved by correcting error. Thus, data are often in a constant state of flux.

Additionally, data are getting 'big': Enormous volumes of data are being collected, of which specific subsets are selected for analysis, be they a small number of individual values to massive subsets of even bigger data sets. Describing which subset was actually being used–and trying to re-create the exact same subset later based on that description–may constitute a daunting challenge due to the complexity of subset selection processes (such

1

Large Number of Adoptions

- **Standards / Reference Guidelines / Specifications:**
  - Joint Declaration of Data Citation Principles:
    Principle 7: Specificity and Verifiability (https://www.force11.org/datacitation)
  - ESIP:Data Citation Guidelines for Earth Science Data Vers. 2 (P14)
  - ISO 690, Information and documentation - Guidelines for bibliographic references and citations to information resources (P13)
  - EC ICT TS5 Technical Specification (pending) (P12)
  - DataCite Considerations (P8)
- **Reference Implementations**
  - MySQL/Postgres (P5, P6)
  - CSV files: MySQL, Git (P5, P6, P8, Webinar)
  - XML (P5)
  - CKAN Data Repository (P13)
  - SPARQL (P17)

Large Number of Adoptions

- **Early pilot implementations, use cases**
  - DEXHELPP: Social Security Records (P6)
  - NERC: ARGO Global Array (P6)
  - LNEC: River dam monitoring (P5)
  - CLARIN: Linguistic resources, XML (P5)
  - MSD: Million Song Database (P5)
  - many further individual ones discussed …

Large Number of Adoptions

- **Adoptions deployed**
  - CBMI: Center for Biomedical Informatics, WUSTL (P8, Webinar)
  - VMC: Vermont Monitoring Cooperative (P8, Webinar)
  - CCCA: Climate Change Center Austria (P10/P11/P12, Webinar)
  - EODC: Earth Observation Data Center (P14, Webinar)
  - VAMDC: Virtual Atomic and Molecular Data Center (P8/P10/P12, Webinar)
  - Ocean Networks Canada (P12, Webinar)
- **In progress**
  - NICT Smart Data Platform (P10/P14)
  - Dendro System (P13)
  - Deep Carbon Observatory (P12)

Lessons Learned as an FAQ (1 of 2)

- **Do the recommendations work for any kind of data?** Yes, it appears so.
- **Do all updates need to be versioned?** Ideally, yes. In practice, probably not.
- **May data be deleted?** Yes, with caution and documentation.
- **What types of queries are permitted?** Any that a repository can support over time.
- **Does the system need to store every query?** No, just the relevant queries.
- **Which PID system should be used?** The one that works best for your situation.
- **When multiple distributed repositories are queried, do we need complex time synchronization protocols?** No, not if the local repositories maintain timestamps.

Lessons Learned as an FAQ (2 of 2)

- **How does this support giving credit and attribution?** By including a reference to the overall data set as well as the subset.
- **How does this support reproducibility and science?** By providing a reference to the exact data used in a study.
- **Does this data citation imply that the underlying data is publicly accessible and shared?** No.
- **Why should timestamps be used instead of semantic versioning concepts?** Because there is no standard mechanism for determining what constitutes a 'version.'
- **How complex is it to implement the recommendations?** It depends on the setting.
- **Why should I implement this solutions if my researchers are not asking for it or are not citing data?** Because it's the right thing for science.

Takeaways from the paper

- It works and it’s not as hard as it seems.
  - Not all Recommendations need to be implemented or at least not at once.
- All found value in adopting even a subset of the Recommendations because it improved services or workflows or archive practices.
- Technical migration still somewhat untested but a fact of life for archives.
- It’s not really about credit.
- It’s the way of the future.

WGDC Webinar Series

- https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html
  - Implementation of the RDA Data Citation Recommendations by **Ocean Networks Canada (ONC)**
  - Implementation of the RDA Data Citation Recommendations the **Earth Observation Data Center (EODC) for the openEO platformby**
  - **Automatically generating citation text from queries for RDBMS and XML data sources**
  - Implementing of the RDA Data Citation Recommendations by the **Climate Change Centre Austria (CCCA) for a repository of NetCDF files**
  - Implementing the RDA Data Citation Recommendations for **Long-Tail Research Data / CSV files**
  - Implementing the RDA Data Citation Recommendations in the **Distributed Infrastructure of the Virtual and Atomic Molecular Data Center (VAMDC)**
  - Implementation of Dynamic Data Citation at the **Vermont Monitoring Cooperative**
  - Adoption of the RDA Data Citation of Evolving Data Recommendation to **Electronic Health Records**

RDA WGDC Recommendations - Summary

- ***Benefits***
  - Allows **identifying, retrieving and citing the precise data subset** with minimal storage overhead by only storing the versioned data and the queries used for extracting it
  - Allows retrieving the data both **as it existed** at a given point in time as well as the **current view** on it, by re-executing the same query with the stored or current timestamp
  - It allows to identify and cite even an **empty set**!
  - The query stored for identifying data subsets provides valuable **provenance data**
  - Query store collects **information on data usage**, offering a basis for data management decisions
  - **Metadata** such as checksums support the verification of the correctness and **authenticity** of data sets retrieved
  - The same principles work for **all types of data**

Agenda

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Harvard Data Science Review Paper
- New Ref Implementations:
  - RDF
- New Pilots:
  - DBRepo: Open source database repository system
  - OSSDIP: Secure data visiting platform for sensitive data
- “New" directions:
  - Information Retrieval Systems
  - AI online learning systems
- Other issues, next steps

Q&A

Any questions?

Any issues identified?

Anybody in the progress of (planning to) implement the recommendations?

Adoption Stories or Plans

- Let us know if you are (planning to) implement (part of) the recommendations

- Submit your adoption story to the RDA Webpage:

https://www.rd-alliance.org/recommendations-outputs/adoption-stories

Agenda

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Harvard Data Science Review Paper
- New Reference Implementation:
  - RDF
- New Deployments:
  - DBRepo: Open source database repository system
  - OSSDIP: Secure data visiting platform for sensitive data
- “New" directions:
  - Information Retrieval Systems
  - AI specifically online learning systems
- Other issues, next steps

Implementing the RDA WGDC Recommendations to Make RDF Data Citeable

Filip Kovacevic, TU Wien

WGDC Implementation for RDF Data

How to enable timestamp-based statement-level versioning for RDF based data representations via RDF-star?

Datasets … … must have *nested quoted triples in subject position** with a deletion timestamp as object at the first nesting level and a creation timestamp as object at the second nesting level.

```
<<
<<:Bob :occupation :Cook>>
vers:valid_from "2021-04-07T12:00:00.000+00:00" ^^xsd:date
>> vers:valid_until "9999-12-31T00:00:00.000+00:00" ^^xsd:date.
```

Queries and Update statements … … must use the creation and deletion timestamp properties at *Basic Graph Pattern (BGP)* level.

Who has „cook“ as occupation now?

```
Select ?s {
   << << ?s :occupation :Cook >>
     vers:valid_from ?valid_from >>
   vers:valid_until ?valid_until .
   filter(?valid_form
        <= "2022-06-20T12:00:00.000+00:00"
        < ?valid_until)
}
```

Triple stores … … must support RDF* and SPARQL* with multi-level nesting

E.g. GraphDB, Jena TDB

* https://w3c.github.io/rdf-star/tests/trig/syntax/manifest.html#trig-star-nested-1

WGDC Implementation for RDF Data

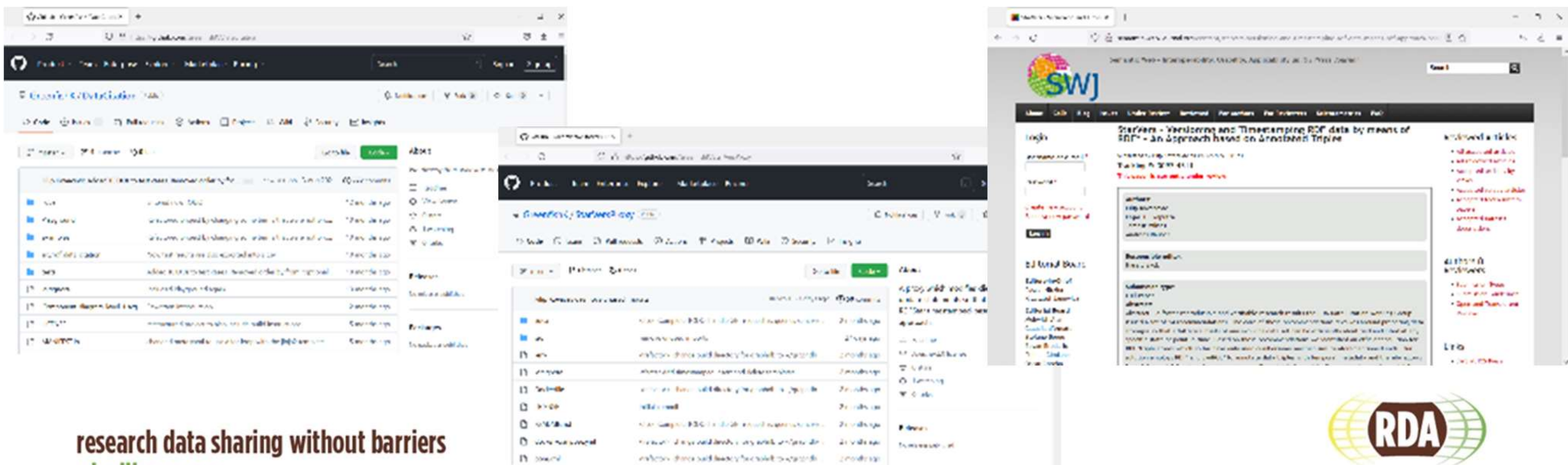
Two reference implementations:

- Python API
  - User provides SPARQL endpoints, update statements and queries.
  - API adds timestamps and filters to statement and query bodies.
  - Executes them against the provided SPARQL endpoints
- Proxy server
  - Abstracts versioning/timestamping from user
  - User simply sends SPARQL update statements and queries via an arbitrary interface to the proxy
  - Proxy modifies requests to add versioning and forwards them to the configured SPARQL endpoint
- Evaluated with Jena TDB and GraphDB
- Essential for tracking evolution in ontologies!

WGDC Implementation for RDF Data

Further Reading

- API: https://github.com/GreenfishK/DataCitation
- Proxy: https://github.com/GreenfishK/StarVersProxy
- Paper (under review): http://semantic-web-journal.org/content/starvers-versioning-and-timestamping-rdf-data-means-rdf-approach-based-annotated-triples



Agenda

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Harvard Data Science Review Paper
- New Ref Implementations:
  - RDF
- New Pilots:
  - DBRepo: Open source database repository system
  - OSSDIP: Secure data visiting platform for sensitive data
- “New" directions:
  - Information Retrieval Systems
  - AI online learning systems
- Other issues, next steps

RESEARCH DATA ALLIANCE

DBRepo – A Repository System Hosting Relational Databases

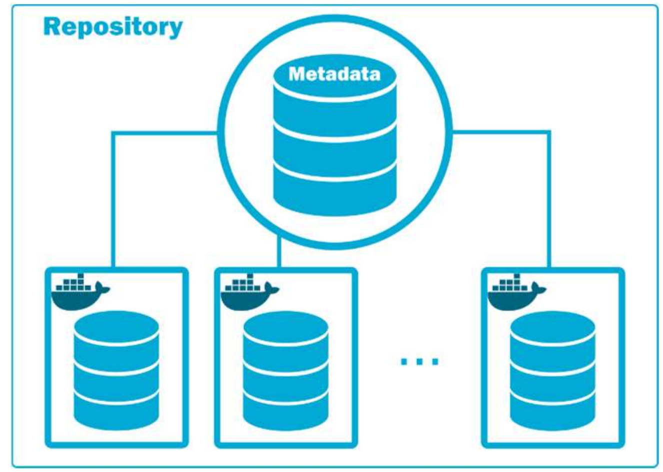
Martin Weise, TU Wien

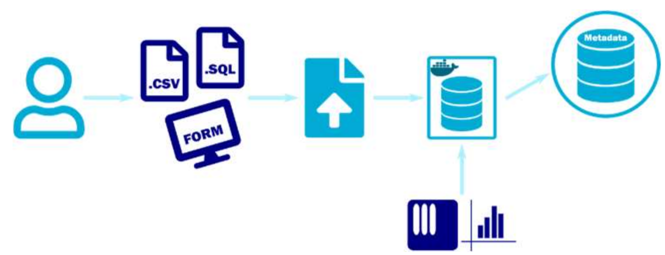
research data sharing without barriers

rd-alliance.org

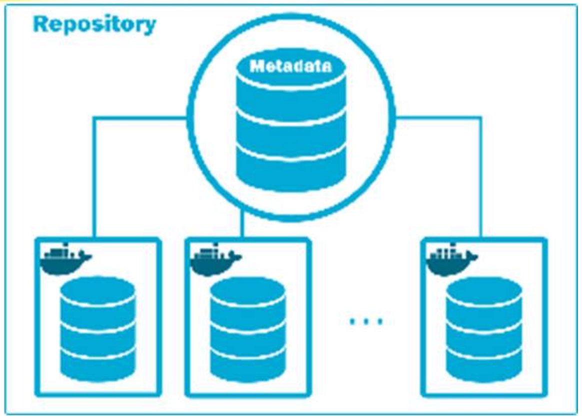
DBRepo – A Database Repository

- Cloud hosted repository for structured research data
- Supports data **versioning** and
- **FAIR** principles
- Guarantees reproducibility
- Data is **cite-able**
- Different levels of SQL-knowledge
- Microservice Architecture
- Each database encapsulated in a Docker container





DBRepo Principles



- Each database is encapsulated in a Docker container: flexible, scalable
- Metadatabase makes databases findable
  - DB description, data license, …
  - Table names
  - Attribute names
  - Measurement units
  - Mapped to controlled vocabularies
  - Search by statistical properties
  - „*List databases that contain temperature measurements in the range of 100-250 degrees Kelvin that are accessible for researchers at ACONet member institutions or public*“

Persistent Identification of Arbitrary Subsets

- Each query issued to the database is saved in the Query Store
- Attaching metadata to a query statement, following the DataCite schema
- Mirror the query metadata to DBRepo's central database, ensure that the metadata is always available even when the database is not.

Query Store   Metadata Database

DBRepo: Evaluation

Open Data Catalogue

- Hourly mean measurements for 3 locations within Zürich, CH
- 1 table, 210.192 tuples
- Ozone (O_3), Nitrogen Oxides (NO_x), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO_2), Particulate Matter (PM_{10} and $PM_{2,5}$), Carbon Monoxide (CO), Sulfur Dioxide (SO_2)
- Public Test Instance: https://dbrepo.ossdip.at

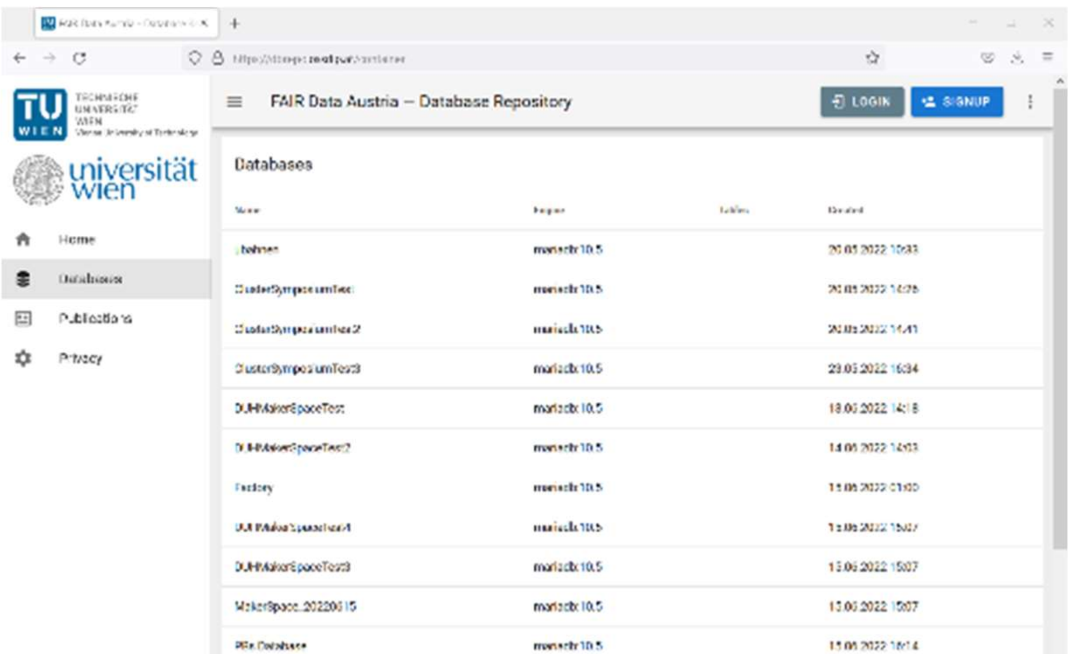
_id	Datum	Standort	Parameter	Intervall	Einheit	Wert	Status
1	2021-01-...	Zch_Sta...	CO	h1	mg/m3	0.44	provisori...
2	2021-01-...	Zch_Sta...	SO2	h1	µg/m3	4.88	provisori...
3	2021-01-...	Zch_Sta...	NOx	h1	ppb	29.46	provisori...
4	2021-01-...	Zch_Sta...	NO	h1	µg/m3	9.85	provisori...
5	2021-01-...	Zch_Sta...	NO2	h1	µg/m3	41.24	provisori...
6	2021-01-...	Zch_Sta...	O3	h1	µg/m3	8.51	provisori...
7	2021-01-...	Zch_Sta...	PM10	h1	µg/m3	88.34	provisori...
8	2021-01-...	Zch_Sta...	PM2.5	h1	µg/m3	75.72	provisori...
9	2021-01-...	Zch_Sch...	NOx	h1	ppb	41.66	provisori...
10	2021-01-...	Zch_Sch...	NO	h1	µg/m3	21.64	provisori...

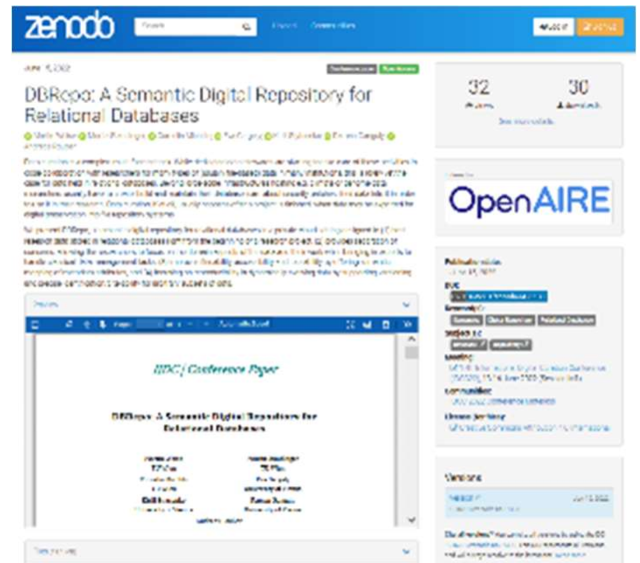
Public dataset with sensor measurements to showcase AMQP API

https://data.stadt-zuerich.ch/dataset/ugz_luftschadstoffmessung_stundenwerte/resource/4466ec4a-b215-4134-8973-2f360e53c33d

DBRepo: Further Material

- https://doi.org/10.5281/zenodo.6637333 (manuscript)
- https://dbrepo.ossdip.at (Public Test Instance)
- https://dbrepo-docs.ossdip.at (Documentation)
- https://gitlab.phaidra.org/fair-data-austria-db-repository/fda-services (Source)





Agenda

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Harvard Data Science Review Paper
- New Ref Implementations:
  - RDF
- New Pilots:
  - DBRepo: Open source database repository system
  - OSSDIP: Secure data visiting platform for sensitive data
- “New" directions:
  - Information Retrieval Systems
  - AI online learning systems
- Other issues, next steps

RESEARCH DATA ALLIANCE

OSSDIP – Reference Implementation of a Secure Data Visiting Infrastructure

Martin Weise, TU Wien

research data sharing without barriers

rd-alliance.org

OSSDIP: Secured Data Visiting

- Sensitive Data (privacy issues, commercial interests, …)
- Provide access for analysis,
  but ensure data is not leaked / misused
- Standard approach: pseudonymization / anonymization
  - k-anonymity, l-diversity, t-closeness
- **Data Visiting** instead of Data Sharing!
- **Data owner maintains full control over data and use:**
  - **Who** to allow access,
  - over **which period of time**,
  - for **which subset of data,**
  - to answer **which research question / analysis goals,**
  - while **monitoring what they are doing**

OSSDIP: Core Concepts

Secure data infrastructure, controlled access

- Physical protection:
  - specific server rooms, locked server racks
  - 4-eye principles
- Encrypted storage
- VPN
- Gateway Firewall allowing access
  - Incoming: only to a specific VM per user
  - Outgoing: read access to package servers for SW updates plus manually configured license servers
- 2-factor authentication
- Transfer of credentials via separate channels

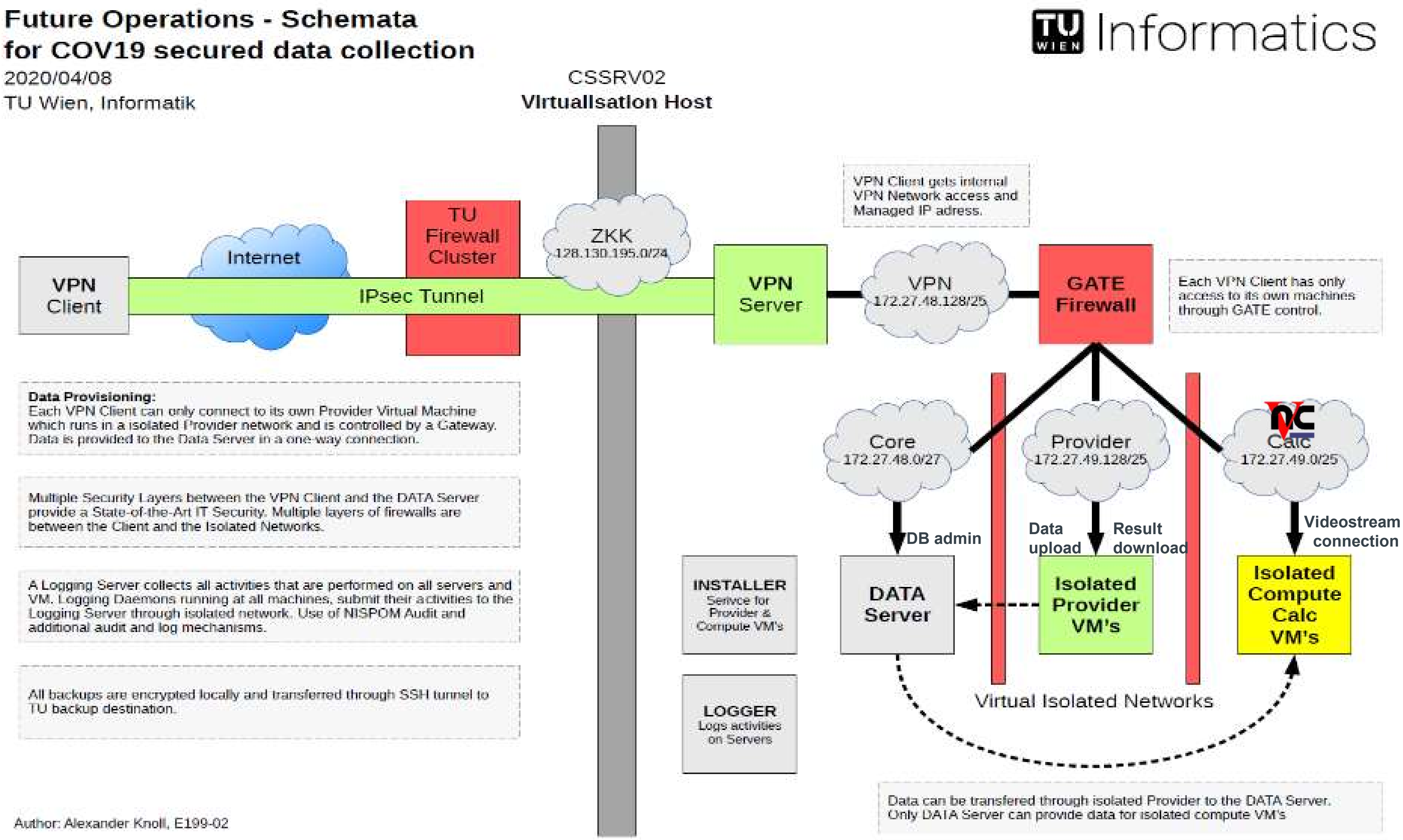
„Standard“

OSSDIP: Core Concepts

Provisioning of data subsets on isolated machines

- Dedicated VMs for each task and individual user
- Subsets of data extracted from central repository
- Metadata on subsets may be shared (FAIRness for closed data!)
- Customized data provisioning per VM
  - Individual subsets (+ data citation, + metadata -> FAIR)
  - Individual k-anonymity, l-diversity, t-closeness
  - Individual fingerprints
  - (Homomorphic encryption, Data Shielding)

OSSDIP: Technical Architecture Set-up

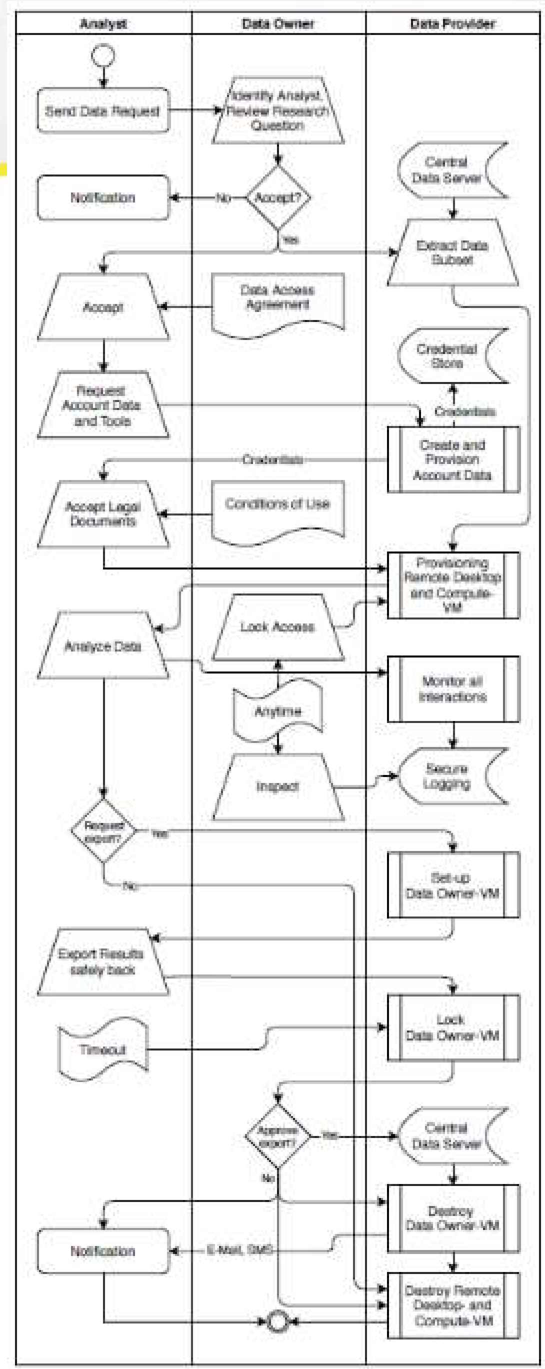


http://www.ifs.tuwien.ac.at/~andi/secure_data_infrastructure.html

OSSDIP Processes: Data Access

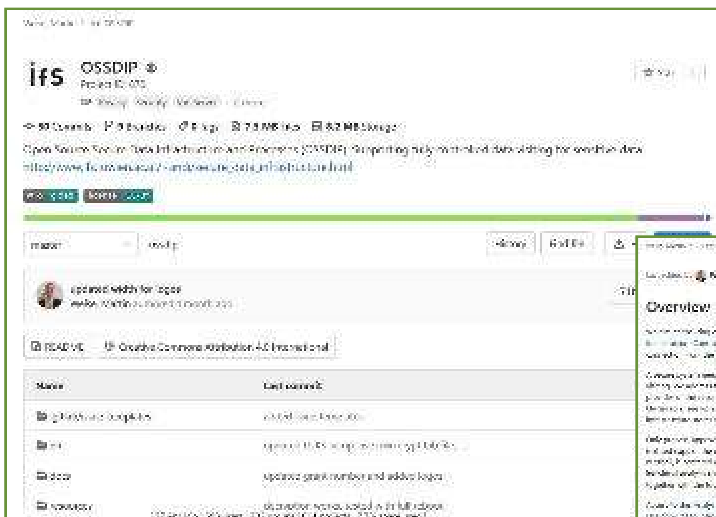
- (selected subset of steps)

1. Researcher sends **request** to data owner (*Person, question, required data*)
2. In case of **permission** being **granted: subset of data**, at specific **aggregation level**, potentially with **fingerprint** is extracted onto a VM for a dedicated **researcher** for a dedicated **time period** to address the **question** posed
3. Expose metadata of data subsets (**F**AIRness)
4. (…)
5. Provisioning of VNC and Compute VMs with dedicated SW and data
6. Monitoring of all interactions on machine on secured logging server
7. Transfer of results via dedicated Provider-VM
8. Destruction of VNC and Compute VMs



OSSDIP: Sources and Further Reading

- Reference-Implementation for Data Visiting System:
  - **Paper**: Weise, M., Kovacevic, F., Popper, N., & Rauber, A. (2022). OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting. Data Science Journal, 21(1), 4. DOI: http://doi.org/10.5334/dsj-2022-004
  - **Source**: https://gitlab.tuwien.ac.at/martin.weise/ossdip

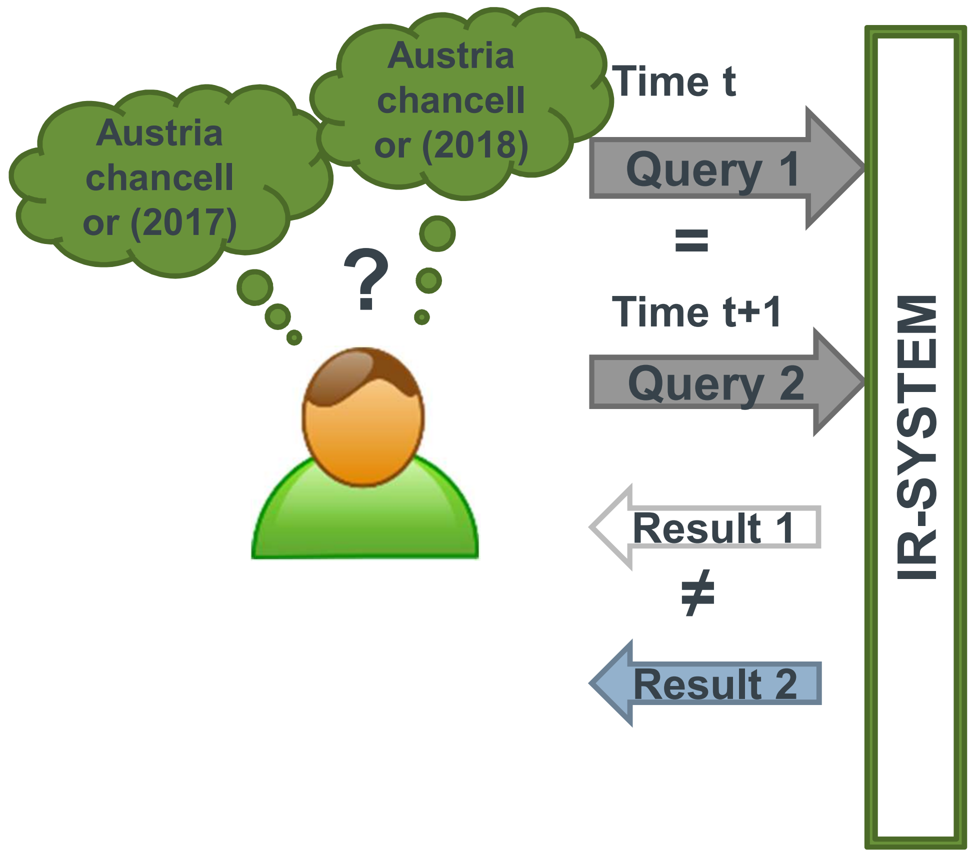


Agenda

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Harvard Data Science Review Paper
- New Ref Implementations:
  - RDF
- New Pilots:
  - DBRepo: Open source database repository system
  - OSSDIP: Secure data visiting platform for sensitive data
- “New" directions:
  - Information Retrieval Systems
  - AI online learning systems
- Other issues, next steps

Reproducibility of IR-Rankings

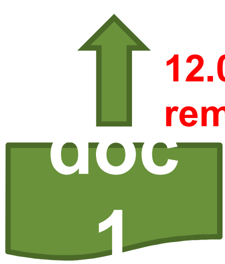
- Typically, retrieval result rankings are not reproducible
- If documents are added, changed or deleted, the resulting rankings differ (even if changed documents do not appear in ranked list!)



Reproducibility of IR Rankings

Term(s)	Document(s)	Term-freq. (tf)	Document freq. (fq)
dog	{~~doc1~~, doc5, doc7,…}	{~~5~~,3,2,…}	~~5895~~5894
house	{doc2, doc112, doc7,...}	{4,8,1,…}	12897
interest	{doc9, doc11, doc12,…}	{1,2,1,…}	5485
right	{doc2, doc18, doc4,…}	{10,2,1,…}	63201

- Document frequency changes when collection is updated
- No tracking of changes
- No straightforward solution to reconstruct values



12.04.2017 remove

Reproducibility of IR-Rankings

Why / where do we need reproducible rankings?

Retrieval results often the basis for

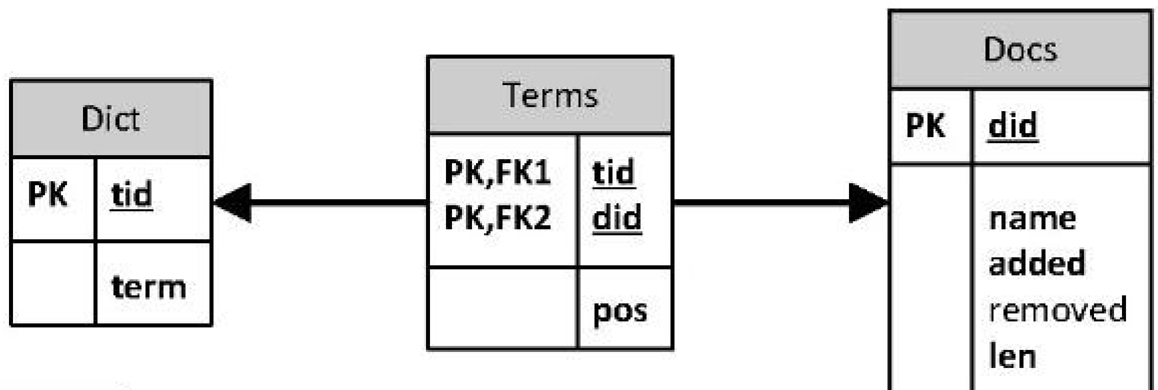
- Scientific experiments, should be reproducible
  - Publication databases, Medline, Google Books, …
  - Social Media / Twitter Feeds / Wikipedia
- Business intelligence reports
  - Press monitoring
  - Social Media surveillance
- On-line learning systems
  - Classifiers (spam filters, document routing)
  - Chatbots
- Decisions that are required to be auditable
  - Patent retrieval, due diligence evaluations, …

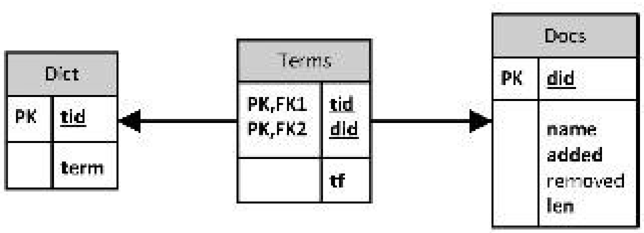
IR Using Column Store Database

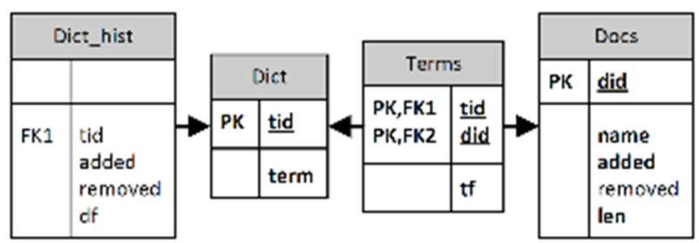
- Mühleisen et al.: retrieval based on a column-store DB
- Retrieval algorithm is translated to SQL
- Promising benchmark results
- Hannes Mühleisen,Thaer Samar, Jimmy Lin, Arjen de Vries. *Old Dogs Are Great at New Tricks: Column Stores for IR Prototyping*. SIGIR2014, ACM
  https://hannes.muehleisen.org/SIGIR2014-column-stores-ir-prototyping.pdf

IR Using Versioned Column Store Database

- Different data models
- Each document conforms to one record in the docs-table
- Each term in every document corresponds to a single record in the ter
- (100) term







OKAPI BM25 translated to SQL

```
WITH
/* filter valid documents */
qdocs AS (SELECT * FROM docs WHERE added <= $timestamp AND (removed IS NULL OR removed > $timestamp)),
/* valid terms containing one of the search strings */
qterms AS (SELECT terms.tid, terms.did, tdic.term FROM
(SELECT tid, term FROM dict WHERE term IN ($term1, $term2, $term3, ...., $termx) AS tdic
JOIN terms ON terms.tid = tdic.tid
JOIN qdocs ON qdocs.did = terms.did ),
/* average document length (avg(len)) and number of documents (N) */
stats AS (SELECT avg(len) AS anr, count(*) AS tnr FROM qdocs),
/* frequency of terms in documents (tf) = term frequency */
term_tf AS (SELECT tid, did, COUNT(*) AS tf FROM qterms GROUP BY tid, did),
/* compute number of documents containing search term (df) */
term_df AS (SELECT tid, count(tid) AS df from term_tf GROUP BY tid),
/* compute document term scores */
subscores AS (SELECT qdocs.did, qdocs, qdocs."len", term_tf.tid, term_df.df, term_tf.tf, (SELECT tnr FROM stats) AS n, (SELECT anr FROM stats) as av,(log(((SELECT tnr FROM stats) - term_df.df + 0.5)/(term_df.df + 0.5)) * term_tf.tf * (1.2 + 1) / (term_tf.tf + 1.2 * (1 - 0.75 + 0.75 * ((qdocs."len")/((SELECT anr FROM stats)))))) AS subscore
FROM term_tf
JOIN qdocs ON term_tf.did=qdocs.did
JOIN term_df ON term_df.tid = term_tf.tid)
/* summing up document scores and order by score descending */
SELECT subscores.did, sum(subscores.subscore) AS rnk
FROM subscores GROUP BY subscores.did ORDER BY rnk desc LIMIT 1000;
```

Document filtering

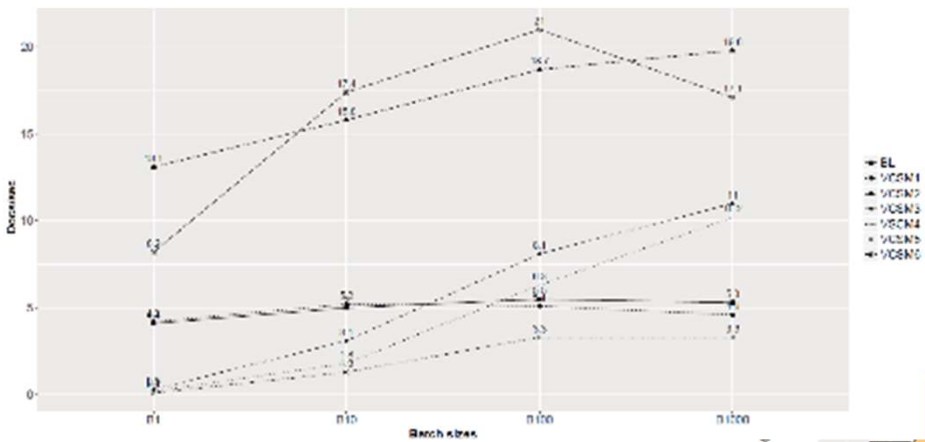
Search terms

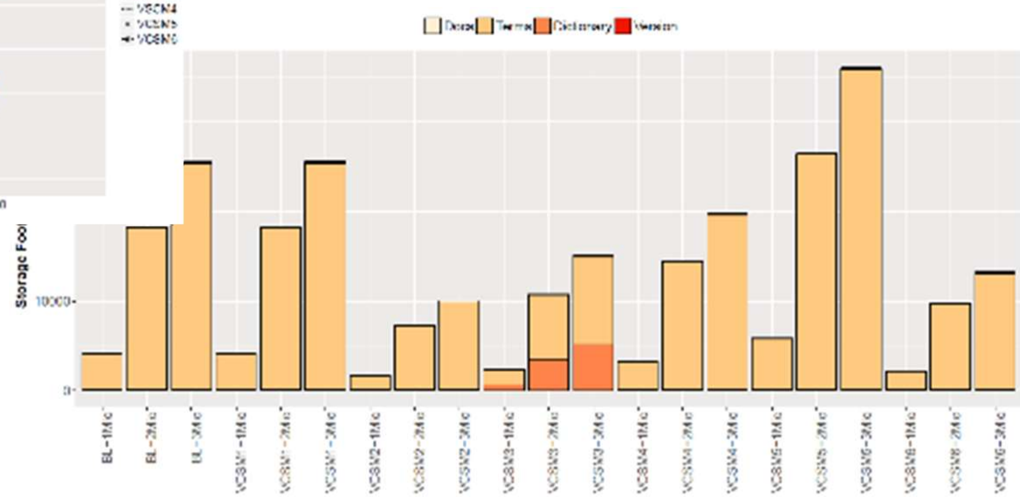
tf and df values

OKAPI BM25, Documen

IR Using Versioned Column Store Database

- Evaluation: Slower but acceptable
- Combined with standard search engine (Lucene) for “live” searches



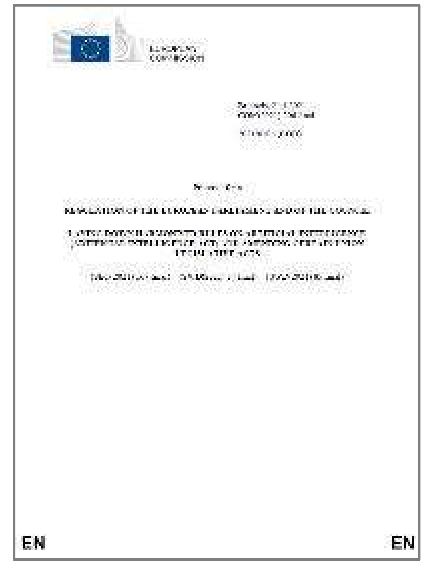


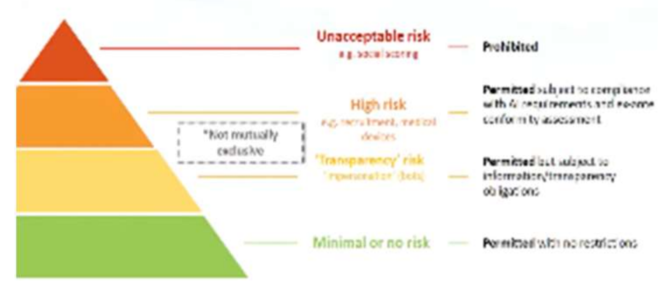
Agenda

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Harvard Data Science Review Paper
- New Ref Implementations:
  - RDF
- New Pilots:
  - DBRepo: Open source database repository system
  - OSSDIP: Secure data visiting platform for sensitive data
- “New" directions:
  - Information Retrieval Systems
  - AI online learning systems
- Other issues, next steps

EU Regulation on AI

- Proposal for a regulation of the European parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.
- Risik Classification of AI-Systems
- Obligations on their creation, operations and monitoring
- Strong impact on data analysis and data management

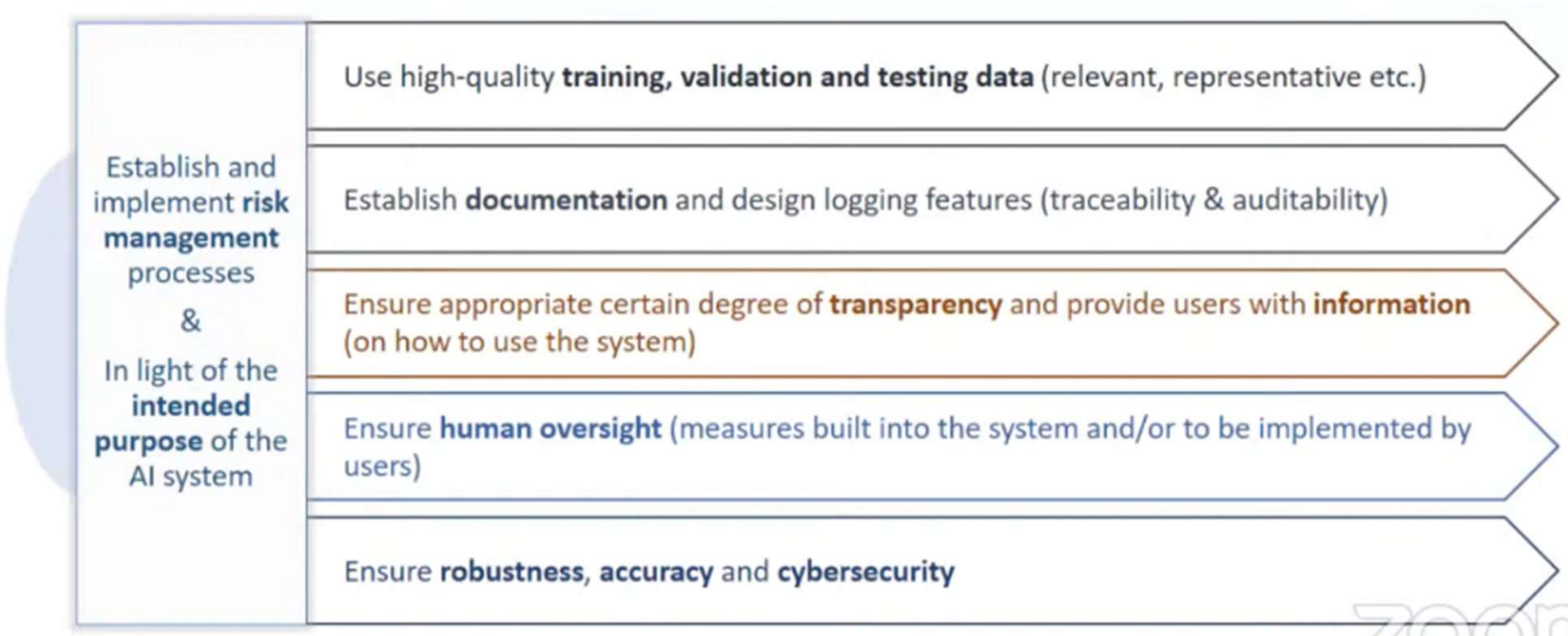




https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

EU Regulation on AI

- Requirements towards high-risk AI systems
- → How to fulfill these?
- → How to demonstrate fullfillment?



Irina Orssich, "The New EU Proposal for AI Regulation", Digital Humanism Lecture, 8.6.2021
https://www.youtube.com/watch?v=9rkH1C1n9sQ

EU Regulation on AI

- Specific challenge: on-line learning systems
- Regular (mini-batch) updates based on data received
- Evolving ML model
  - Which model state used at specific point in time?
  - How to re-activate a specific model state to verify processing?
- Applying WGDC principles to evolving ML model
  - Different approaches to versioning
  - Impact on training speed (less an issue with online learning – small batch updates – few iterations)
  - Evaluation prototype using Tensorflow
- In progress: stand by for further updates…

Agenda

- Introduction, Welcome
- Short description of the WG recommendations
- Q&A on recommendations
- Harvard Data Science Review Paper
- New Ref Implementations:
  - RDF
- New Pilots:
  - DBRepo: Open source database repository system
  - OSSDIP: Secure data visiting platform for sensitive data
- “New" directions:
  - Information Retrieval Systems
  - AI online learning systems
- Other issues, next steps

Thanks

Thanks!

And hope to see you at the next meeting

of the

WGDC