

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
 - Al online learning systems
- Other issues, next steps

research data sharing without barriers rd-alliance.org



Welcome and Intro

Welcome! to the maintenance meeting of the WGDC



3

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
 - Al online learning systems
- Other issues, next steps

research data sharing without barriers rd-alliance.org



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

sets

Research Data Sharing without barriers

TO NOT ALL ADD

Data Citation WG

A Recent Activity

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

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





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



research data sharing without barriers rd-alliance.org



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

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



research data sharing without barriers rd-alliance.org

Dynamic Data Identification and Citation



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

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

Steps:

1. Data \rightarrow 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 \rightarrow 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 \rightarrow 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



- 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



- Note: query string provides excellent provenance information on the data set! er gets
 - Data (package, acce API, ...)
 - PID (e.g. DOI) (Que is time-stamped and stored)
 - Hash value compute over the data for local storage
 - Recommended citation 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



- Note: query string provides excellent provenance information on the data set! er gets
 - Data (pad This is an important advantage over
 - PID (e.g. traditional approaches relying on, e.g.
 - Hash value storing a list of identifiers/DB dump!!!
 - Recommended citating text (e.g. pibrex)
- 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



- Note: query string provides excellent provenance information on the data set! er gets
 - Data (pad This is an important advantage over
 - PID (e.g. traditional approaches relying on, e.g.
 - Hash valu storing a list of identifiers/DB dump!!!
 - Recommended citati riext (e.g. pibrex)
- PID resolves Identify which parts of the data are used.
 - Provides det If data changes, identify which queries
 - Option to ret (studies) are affected
- Upon activating PID associated win a data citation
 - Query is re-executed against time-st nped 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 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

When Resolving a PID

- R11 Landing Page
- R12 Machine Actionability

Upon Modifications to the Data Infrastructure

- R13 Technology Migration
- R14 Migration Verification



research data sharing without barriers rd-alliance.org

RESEARCH DATA ALLIANCE

Data Citation – Output

- I4 Recommendations grouped into 4 phases:
- 2-page flyer
 <u>https://rd-alliance.org/recommendations-working-group-data-citation-revision-oct-20-2015.html</u>
- Detailed report: Bulletin of IEEE TCDL 2016 <u>http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-</u> <u>TCDL-DC-2016_paper_1.pdf</u>
- Adopter's reports, webinars
 <u>https://www.rd-alliance.org/group/data-citation-</u>
 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.

10000	Data Citation modes of the We remotion is set and the	of Evolving Data		
1 Vietness Theory Viet Concernment New York Concernment New Yor	Deta (most in white concerns of a data that a concerns and practice that in appendix because the second states of the second states of the second states of the termine the second states	Angle of the constraints of the constraints A Constraints		
		mand the it for proving the	**	
antes con en antes con en al biologia e real de contes e an al contes e al contes e an al contes e al contes e al contes e al contes e al contes e al contes	Autor and the second se	klentification of Reproduct Sharing	ble Subsets for De and Re-Use	na Citution.
Alarma de ser		Access faits was a control of the control of the control of the control of the state of the control of the control of the control of the state of the control of the control of the control of the state of the control of the control of the control of the state of the control of the control of the control of the state of the control of the control of the control of the state of the control of the control of the control of the state of the control of the control of the control of the control of the state of the control of the control of the control of the control of the state of the control of the control of the control of the control of the state of the control of the contro	Endows The second states and a second state Second states Second states Seco	i total I total I total
STREET, STREET		And the other is a second seco	 Antonio ed Richerto, N. Antonio ed Richerto, M. antonio ed Richerto, M. Antonio	
Ference Data	Citation W	G	the second state	the local day
 Marca and a second secon	Million y searches ne constantes endores endores de character sersoa aporte de centre plande constantes con	n Al serie in digt ARM The Britships Cost of the will a lease of the second second second at match Alman second second	And a second sec	
The second s	Project and P	waladu waning goon was uwa a	e Sino i	and the second
Ministry (Ministry) (M	Dranke Dra. Animaliethy Be Nordel Estaman Reach Train	daed Giller, "A Cole State Series Cole State of a General Cole Series Series American Soliton and Series The Planet Series States of Series Sciences Series Sciences	Coles Alexiet Cole General Reported you	-
Bit-Dissection of end of experiments of the end of the endly of the endly of the endly of the official and an endly	Description of the second seco	alad Caro, A. Californi, M. Salar, C. S. Sharay, Salar Californi, Carolina Californi, Salar Sa	Color Nameri - Lon 197 Bigo, Lobyrt Marcan I. Alf, London	
De Deserrando el la Carteria de La companya de la contecta de la companya en estado en el companya de la contecta de la companya de la contecta de la contecta de la contecta de la contecta del la contecta de la conte	Departs Dec. Automatication for the hear of the automatication of of the Automatication of the Automatication of the Hear of the Automatication of the Automatication of the Automatication of the Hear of the Automatication of the Automatication of the Automatication of the Automatication of the Hear of the Automatication of the Hear of the Automatication of the Automa	shared Games, A. Chair Hank, Shirin Chair Hank, and Share Hank, Share Share Share Share Share Share and the state of the s	Geber Marant Josef Landi Barr, Salayan Marana (ABP Constance 275 B) Ma	
Der Deservice direct of Statistics of the second second second Theorem with a second	Descara de Desc la casa de destructura terrar en la companya de la companya de la companya de la companya de la companya de la companya de la companya de la companya de la companya de la companya de la companya de la companya de la	shared below in a constraint where the states of a strain "A support of the states of a states of a strain "A support of the states of a states of a strain a straint of the states of a states of a straint of the states of the	Dans Manada Lann Berlin Berg, Salayar Manada (JAK Karatana JAK B ¹ M	
An end of the second se	De santé Direc- locaras destrés las locaras destrés de la locara de la locaras en la companya de la esta de la locaras la calité des astrés la calité des astrés la calité des astrés la calité des la la calité de la calité de la calité de la la calité de la la la la la la la la la la	shared Galaxies of the Argon Cold and an of the structure of the cold and argon Cold and argon of structure of the cold and the cold and argon of the cold and the cold and the cold argon of the cold and argon of the cold argon of the cold and the cold argon of the cold argon of the cold and the cold argon of the cold argon of the cold and the cold argon of the cold argon of the cold and the cold argon of the cold argon of the cold and the cold argon of the cold argon of the cold and the cold argon of the cold argon of	Canal Palanasi Junia Gana Palanasi Mangalan (PAP University Mangalan Mangal	
(a) Show is not device a first one of the second of the ment of the second of the second of the second of the second of	De park Elocation Science Science Sciences Science Sciences H. Weiss Sciences H. Weiss Sciences H. Weiss Sciences Sciences Sciences Control Sciences	shared Calibration of the Arrival Constraints of the Array of the Arra	The Property of the second sec	
4. Street and order the street of the str	The start E Deci- closure Education for the start of the start of the transformation of the start of the star	share the second	The Branch of the Second secon	
4.5 Shows in and order to be a stress of the second of the intervence of the second of the physical second of the second of the intervence of the second of	De para 2 Deci. de la constante de la constan	share the second	Carlo Table Control on the second sec	

HDSR Paper: From Principles to Adoption¹⁹

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** <u>10.1162/99608f92.be565013</u>

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

research data sharing without barriers rd-alliance.org



Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data Andreas Rauber¹, Bernhard Gößwein^{1,3}, Carlo Maria Zwölf⁴, Chris Schubert⁵, Florina Wörister¹, James Duncan⁶, Katharina Flicker¹, Koji Zettsu⁷, Kristof Meixner¹, Leslie D. McIntosh⁸, Reyna Jenkyns⁹, Stefan Pröll¹⁰, Tomasz Miksa^{1,11} Mark A. Parsons²

 TU Wien, Vienna, Austria
 University of Alakama in Huntsville, AL, USA
 Earth Observation Data Centre, Vienna, Austria
 LERMA, Observatoire de Paris, PSI. Research University, CNRS, Sorbonne University, UPMC Univ Paris Meudon, France
 Climate Change Centre Austria, Vienna, Austria
 Forest Ecosystem Monitoring Cooperative, University of Vermont, Burlington, VT, USA
 National Institute of Information and Communications Technology, Tokyo, Japan
 Ripeta, Saint Louis, MO, USA
 Ocean Networks Canada, University of Victoria, Victoria, BC, Canada
 Cropster, Innsbruck, Austria
 SIA Research, Austria

Abstract

Precisely identifying arbitrary subsets of data so that these can be reproduced is a damnting 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 implementations. This article provides an overview of the recommendations, reference implementations, and ploit 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 processes 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 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 reckilitated, data enditis 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



Large Number of Adoptions

Standards / Reference Guidelines / Specifications:

- Joint Declaration of Data Citation Principles: Principle 7: Specificity and Verifiability (<u>https://www.force11.org/datacitation</u>)
- 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)



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

- <u>https://www.rd-alliance.org/group/data-citation-wg/</u> 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

research data sharing without barriers rd-alliance.org



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
 - Al online learning systems
- Other issues, next steps research data sharing without barriers rd-alliance.org





Any questions? Any issues identified? Anybody in the progress of (planning to) implement the recommendations?



research data sharing without barriers rd-alliance.org

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/recommendationsoutputs/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
 - Al specifically online learning systems
- Other issues, next steps

research data sharing without barriers rd-alliance.org





Implementing the RDA WGDC Recommendations to Make RDF Data Citeable

Filip Kovacevic, TU Wien

research data sharing without barriers rd-alliance.org

WGDC Implementation for RDF Data

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



... 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.



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



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

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

research data sharing without barriers rd-alliance.org

<<

<c: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.

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)
}</pre>

E.g. GraphDB, Jena TDB



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

35

RESEARCH DATA ALLIANCE

Further Reading

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

Que a management out			×				The second second second	• 1	-
 3 U. S. Carly Information Million and 		W.	3 ± =				C	$\overset{\circ}{\otimes}$ wave carry r and accorring in the conductor are converged where according out	E G
) Parket - The Diky of Solar - Sheredari Pang -	Seat	·	54.14 (<u>2.8.4)</u>				GA/1	Served Cheere and period so there is, and the second strain to the term	····
trade Constituio (14)	0.1	toth or Y M X	06.8 -1					w.n. Linde before derived Recardies the indensis deservations that	
the Grand Character S and Clother in the	Cont Ender		Gan arrente	+			togin	Star/ers - Venioring and Timestamping RDF data by means of FBF* - An Approach based on Annotated Triplet	and seed a ticks
" man	Geralda (1996a)	Aboxs	a 26	durightening alle have		ŵ		CONTRACTOR CONTRACTOR CONTRACTOR	 All many oil in Allow All model of the Allow
the second start of the reserves the second others, for the second		N 911 8 9 10 - 7 1	0	-		с б .	2000	Transfer & State and a state street	many effectively enter
sa santas tes	Mary Januar	C 21- 000						artur	· Sector of reasons with
Any server a service share sha	week and the second	the Case	· Green (Stationed)	toy res	C 1	B. Star V. Lawrence	Consumption of	sub-states	- united sames
en en alter soulledays a start star	and an Inches	A learning	State Glass 3 feb		Prins a		Los 1	where the second s	dens - den
acid orsicator to control or control of the control	1946-24122							Parent de effert	40220
343 Adde RURATORE Second edu 5,4	for spool investigation	E-bear	Fra - Pabase &		with code-	(ka)	ED WHO BOATS	5 m	antistuents
in regione in the second degree in the second degree.	and the second	to mile a stable .	the lotter of th	- maa	w fishion .	Apropulside voor faar de	Sec. Cates	a starse	A COMPLET ARTIST
Therease dependent and any increase international	200-00-00					ET for a serar sed sex	Liby Million J	dense is her remained of a suble exactly site of a suble site of a suble site of the	Autor
PTST menanan Cause is she train salitines in	the Canada and	trainin		and the second sec	2 Section 2	49.001	Land, Const.	For the second	
Water a store that make a set by the app	fore the ferrer is	In solution with data	0.0	enforce data and leave the could be stated.	2 months ago	1 An mithen	True Brok h	HE Share the start of the first and and the start start and the start of the Start Start Start Start Start Start	tro
			11	money and developer inserved follows motion	Insubsequ	T 1. Pro	in an invite		1 . (a) 4.00 feet
			11 Sector	and as the office prevent a provide	Zr. ob.gr.	Y C db			
			D +++++++	wild creat	Privals car			AT A	H.
			0	star Gages (FSG) is dealer and experimental	2 Zevelsop	P desert		(#Ph)	
research data sharing	without ba	rriers	O view second	strates of the address only and uniper de-	at vide sa	to managers of			
			D married	enforces there said have the solel to state the	2 months age				

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
 - Al online learning systems
- Other issues, next steps research data sharing without barriers rd-alliance.org





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





RESEARCH DATA ALLIANCE

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 ⁴⁰

- 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 (O₃), Nitrogen Oxides (NO_x), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO₂), Particulate Matter (PM₁₀ and PM_{2,5}), Carbon Monoxide (CO), Sulfur Dioxide (SO₂)
- Public Test Instance: <u>https://dbrepo.ossdip.at</u>

_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	03	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

 AMOP API

 research data sharing without barriers
 https://data.stadt-zuerich.ch/dataset/ugz_luftschadstoffmessung_stundenwerte/

 rd-alliance.org
 resource/4466ec4a-b215-4134-8973-2f360e53c33d



DBRepo: Further Material

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

1	🖉 FOR form Family - Departure (C.K.	+					4	×
÷ -	o o	Billips//doops.conf.put/scolator			17	8	ŝ.	=
Ţ	TROMMECHE UNATERITZ WEN Versi Driveskyst Tertraloge	FAIR Data Austria – Database F	lepository		E LOGIK	SIGNUP	1	Ĵ
(A)	universität wien	Databases						
		No	bagar	Labims	Denied			
ń	Home	, bohnen	menecbr10.5		20.05 2022 10:33			
\$	Outschooless	Cander Sympton am Test	menisch 10.5		20.05 2022 14:25			
	Publications	(SustarD)mpeatantises2	mariaclentes		20.05.2022.14.41			
\$	Privacy	Cluster9ymposiumTest3	mariacto: 10.5		28.05 2022 16:34			
		DUHMakerEpaceTest	mariactic 10.5		18.05 2022 14:18			
		0.14MakerSpaceTest2	moniectr 10.5		14 05 2022 14:03			
		Factory	marjach 10.5		15 05 2022 01:00			
		DUI IMaka Space (en/A	majachettes		15.052022.15027			
		DUHMakerEpaceTest3	mariactic 10.5		15.05 2022 15:07			
		Maker8pace_20220615	mariactic 10.5		15.05.2022 15:07			
		PEr Detabase	manach 10.5		15.05 2022 16:14			





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
 - Al online learning systems
- Other issues, next steps research data sharing without barriers rd-alliance.org





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, I-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 research data sharing without barriers rd-alliance.org



*



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







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, I-diversity, t-closeness
 - Individual fingerprints
 - (Homomorphic encryption, Data Shielding)



OSSDIP: Technical Architecture Set-up



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 (FAIRness)
- **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

research data sharing without barriers rd-alliance.org



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: <u>http://doi.org/10.5334/dsj-2022-004</u>

50

• **Source**: <u>https://gitlab.tuwien.ac.at/martin.weise/ossdip</u>

ifs OSSDIP &		(#**# (). ()	VIEW IX. DO	Cres Control	
The straig county water	and Cheve		Open Seusce Secure	Pata lefractivicture and	79 72
≪sociants Pathadic Ot	xg: B75NBino B&2MB10regs		Processes for Data V	isiting	anto Arokan'
Open Source Secure Oata Infraction	tion and Processes (CASIDIE). Surporting thely root-blood data visiting for sens	Her Are	Val (Provid Follow Theory		
nilos/www.louwiet.ac.a? - andro	ecrected unbalant and and		Challing Restanding responsed and every set of the re- tain well predicables as to restart shall in the Second Bill	n ing sailes, ever an all residence in alle insister as many lesses a antimation sine access adapter along one has the selected and	
WIN GODD BORTH LAND			Arristin in the statement solid of the statement of a most solid	profiles, and the provider prevention result a contribution of	
10			Here's internets of take inter-or control	NOT STREET, ST	OpenAIRE
marat oxst p	-firmy God Ba	A - missioner met in a recome	 Contains to obtain a contained method in the and individual to 12 or 10 an execution operation 	targets and a submark logan and markets to choose. Supremptions of a superiment of a subgroup of a summer	
		San adam to 🏨 New Made 1 and says	Administration of the effort of the second state of the second sta	ale e a color de la color Color con el color de la col	National date
apparent width for logar	and and the	70	ADDDY C COMPT MIDDLE HOME		6.6.170,702 760
		Overview	11 0.4 1 Mar 1.4.4	+ erester 20 in m av	State Backward and State
B READ VE - U Orathe Common	statisation 4.0 (variational	(a) the entropy dependence of the debug decogle of the administration of the fact of the state of the stat		6	Contraction of Contraction
		sants wange and roman and any dealers may be any second and any			Sati Base Contact
Natio	Cast connects	der sy water side a state man a second a prime to deal a state systematical states and they are the side of the	2		 PSM contract - PMSS method on S0 mag
In plantary popular	ented on elaster and	On an experiment, and Andreas K. Maddels (2008) The Physics of a state of the balance of the bal	Open Kolery Karan Bona Tahu Tar	na turnare and Praestan far Data Kishing Taloni largear	Actual Modifiers Registeration and an
		integrate approximately a stretch to the instant active stretch at the stretch and the stretch active stretch and	-	a service and	N/S/SPOK/SPOK/SPOK
	diameters and assume thread as a	In the datagets the state, where where the interpretation is the state of a set of the data of the state o	in a Tanker of Backson's system	Ng wang Pilinia Bauhag U Satalah Ja	Assessments and a second secon
E den	operate grant number and addee leave	builded and your out to there a write the write out of 50. Also write provide the sound working to provide the sound work out out out out.	And a second second second second	a salasati ana ana ana ana ana ana ana ana ana an	 Comparison (Section 2) - Section (C) - Section (Section 2)
Distances -	standor sono, solo in later antenna	Conclude indjectifie party all is applicable as toget celeric under sever on allerity lation concernates de-	6 Alexandra and a second second	As the set of the set	
3025 53	AN ARE NOT SHOULD TRANSFORMED AND AND	sandri avas atri Wilandaha biladi salapishi, desati sishe diftaarda. Wilati simoada		- International data and the	And And
osoarch data sha	ring without harriers	and the second system with the second state of the second se	1 (100 (127.4)	+	ANTI-
escar cir uata sila	ing without barriers		Televenter and	Tao	Construction of the second
d-alliance.org			and an a fair of	and and and	And enterly were as fronted over a 22-

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
 - Al online learning systems
- Other issues, next steps research data sharing without barriers rd-alliance.org



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	{ <mark>doc1</mark> , doc5, doc7,…}	{ <mark>5</mark> ,3,2,}	5895 5894
house	{doc2, doc112, doc7,}	{4,8,1,}	12897
interest	nterest {doc9, doc11, doc12,}		5485
right	{doc2, doc18, doc4,…}	{10,2,1,}	63201



53

- No tracking of changes
- No straightforward solution to reconstruct values



(RDA)

research data sharing without barriers rd-alliance.org

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://https//https://https://https://https//htttps//https//https//https//https//https//https//https//https/



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



OKAPI BM25 translated to SQL

WITH	Document
/* filter valid documents */	Document
qdocs AS (SELECT * FROM docs WHERE added <= \$timestamp AND (removed IS \$timestamp)),	filteringed >
/* 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) */	Search
stats AS (SELECT avg(len) AS anr, count(*) AS tnr FROM qdocs),	Ocarcii
/* frequency of terms in documents (tf) = term frequency */	terme df
term_tf AS (SELECT tid, did, COUNT(*) AS tf CROW qterms GROUP By tid, did),	
/* compute number of documents containing search term (df) */	values
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_AS n, (SELECT anr FROM stats) as av,(log(((SELECT tnr FROM stats) - term_df.df term_tf.tf * (1.2 + 1) / (term_tf.tf + 1.2 * (1 - 0.75 + 0.75 * ((qdocs."len")/((SELECT an	tf.tf, (SELECT tnr FROM stats) + 0.5)/(term_df.df + 0.5)) * nr 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)	UKAPI
/* summing up document scores and order by score descending */	BM25m
SELEGesenbearstand without largescores.subscore) AS rnk	
FROM subscores GROUP BY subscores.did ORDER BY rnk desc LIMIT 1000;	Documen

IR Using Versioned Column Store Database

58

RESEARCH DATA ALLIANCE

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



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
 - Al online learning systems
- Other issues, next steps research data sharing without barriers rd-alliance.org



EU Regulation on Al

- 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



60



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



EU Regulation on Al

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

		Use high-quality training, validation and testing data (relevant, representative etc.)
1	Establish and implement risk management processes & In light of the intended purpose of the Al system	Establish documentation and design logging features (traceability & auditability)
		Ensure appropriate certain degree of transparency and provide users with information (on how to use the system)
		Ensure human oversight (measures built into the system and/or to be implemented by users)
		Ensure robustness, accuracy and cybersecurity

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

research data sharing without barriers rd-alliance.org



EU Regulation on Al

- 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
 - Al online learning systems
- Other issues, next steps research data sharing without barriers rd-alliance.org





Thanks! And hope to see you at the next meeting of the WGDC



64

research data sharing without barriers rd-alliance.org