Precisely Identifying Arbitrary Subsets of (Dynamic) Data: Recommendations of the RDA WGDC

Andreas Rauber

Technical University of Vienna
Favoritenstr. 9-11/188
1040 Vienna, Austria
rauber@ifs.tuwien.ac.at
http://ww.ifs.tuwien.ac.at/~andi

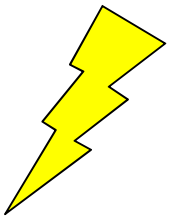
FACULTY OF !NFORMATICS

Outline

- Two challenges in data identification for citation
  - How to identify dynamic data?
  - How to deal with different granularity levels?
- Recommendations of the RDA WGDC
- Deployments

Identification of Dynamic Data

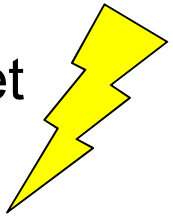
- Usually, datasets have to be static
  - Fixed set of data, no changes:
    no corrections to errors, no new data being added
- But: (research) data is **dynamic**
  - Adding new data, correcting errors, enhancing data quality, …
  - Changes sometimes highly dynamic, at irregular intervals
- Current 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 any(!) specific point in time

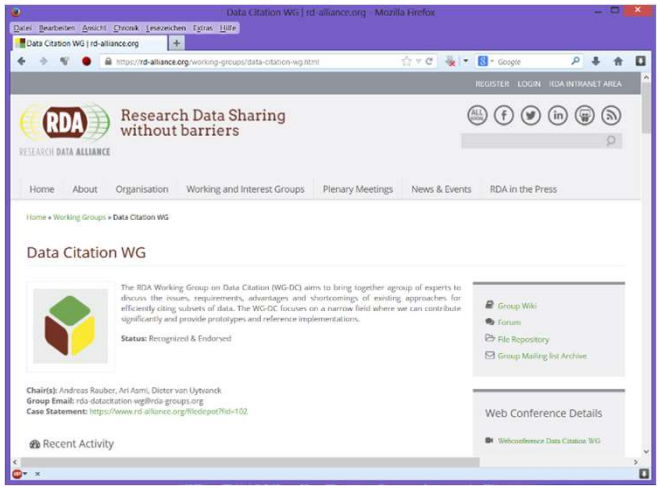
Granularity of Subsets

- What about the **granularity** of data to be identified?
  - Enormous amounts of 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 identify **precisely & machine-actionably any 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 then: supporting adopters

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

RDA WGDC - Solution

- **We have**
  - **Any** kind of data & **some** means of access („query“)

Dynamic Data Citation

We have: Data + Means-of-access

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

Dynamic Data Citation

We have: Data + Means-of-access

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

Steps:

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

Dynamic Data Citation

We have: Data + Means-of-access

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 Citation

We have: Data + Means-of-access

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 data
   - **Re-writing** for normalization, unique-sort, …
   - **Hashing** result-set: verifying identity/correctness

   leading to landing page

S. Pröll, A. Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation.** In IEEE Intl. Conf. on Big Data 2013 (IEEE BigData2013), 2013
http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

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
- ser 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[illegible] 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[illegible]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[illegible] 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
  - Provides det
  - Option to ret
- Identify which parts of the data are used. If data changes, identify which queries (studies) are affected
- Upon activating PID associated w[illegible]n a data citation
  - Query is re-executed against time-st[illegible]mped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

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.

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

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

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, P19)

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, P20, Webinar)

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 (information accessed).
- **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 – “shopping cart”
- **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.

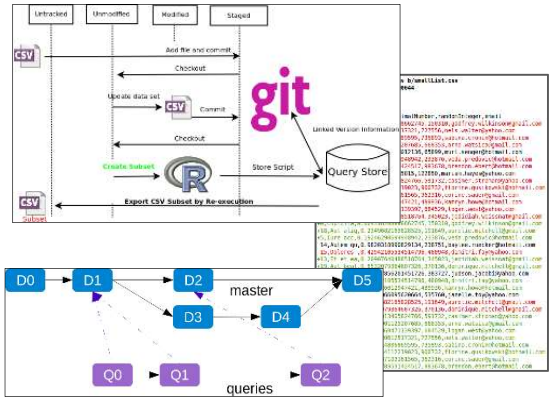
Lessons Learned as an FAQ (2 of 2)

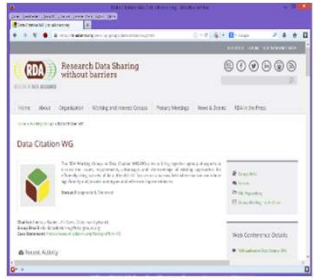
- **How does this support giving credit and attribution?**
  By including a reference to the overall data set as well as the subset.
- **How does this support reproducibility and science?**
  By providing a reference to the exact data used in a study.
- **Does this data citation imply that the underlying 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.’ No minor/major “updates”.
- **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.

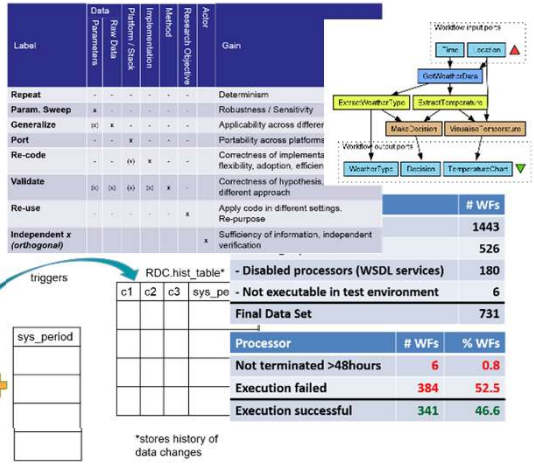
RDA Recommendations - Summary

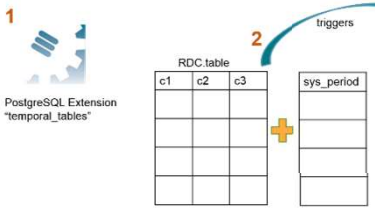
- ***Benefits***
  - Allows **identifying, retrieving and citing the precise data subset** with minimal storage overhead by only storing the versioned data and the queries used for extracting it
  - Allows retrieving the data both **as it existed** at a given point in time as well as the **current view** on it, by re-executing the same query with the stored or current timestamp
  - It allows to 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**

Thank you!



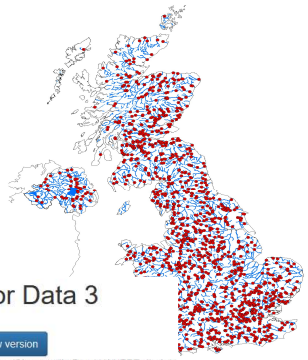




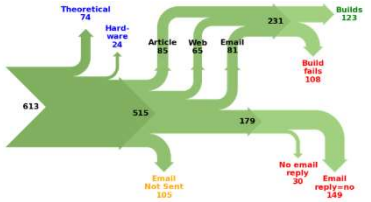


Thanks!

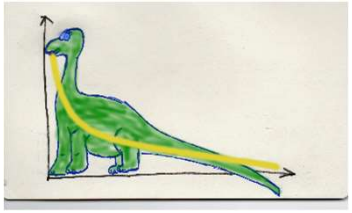
https://rd-alliance.org/working-groups/data-citation-wg.html



DC1
Data Citation Principles



VAMDC consortium



Editing data for Data 3

Back to Dataset Versions

Show changes | Save to a new version

1 UPDATE Z0001_test SET `SiteID` = 'Stevensville Brook' WHERE db_tabl
2 DELETE FROM Z0001_test where db_table_pk=30
3 DELETE FROM Z0001_test where db_table_pk=35

Actions	SiteID	LabID	Date	MeanDensity	Mean
	Stevensville Brook	2000.187	0000-00-00	4644322354	39.0
	Winhall River	2011.081	2011-10-07	201	47.5
	Winhall River	2012.089	2012-09-27	1981	52.0
	Winhall	2013.150	2013-10-15	1092	39.0

2010