



**Data Citation
Working Group Mtg @ P17
April 21 2021, virtually (Edinburgh)**

research data sharing without barriers
rd-alliance.org

Agenda

2

- Introduction, Welcome
- Short description of the WG recommendations
- Paper on adoption stories: lessons learned
- Q&A on recommendations
- On-going adoption activities, other WGs
- Other issues, next steps

Welcome!

to the maintenance meeting
of the
WGDC

Agenda

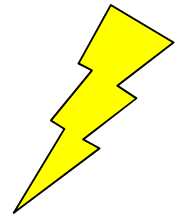
4

- Introduction, Welcome
- Short description of the WG recommendations
- Paper on adoption stories: lessons learned
- Q&A on recommendations
- On-going adoption activities, other WGs
- Other issues, next steps

Identification of Dynamic Data

5

- 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**

Granularity of Subsets

6

- 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)

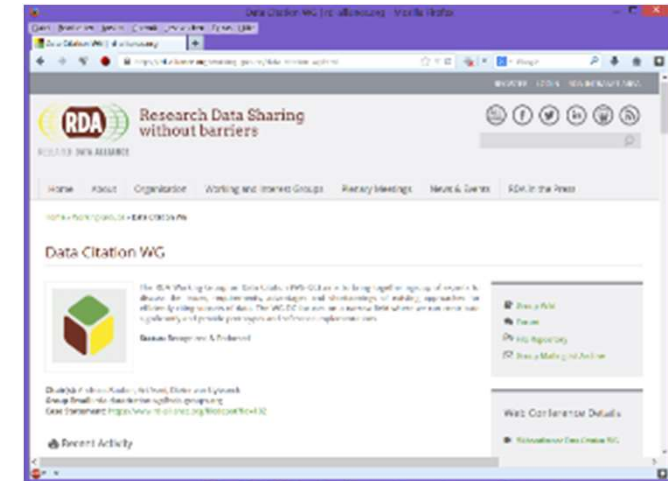


■ Would like to be able to identify precisely the **subset of (dynamic) data used** in a process

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 Citation



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

Dynamic Data Citation



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

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

Dynamic Data 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 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:

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

13

- 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

14

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

15

- 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!!!
 - Data (package)
 - PID (e.g. DOI)
 - Hash value
 - Recommendation
- 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

16

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

- Data (package)
- PID (e.g. DOI)
- Hash value
- Recommendation

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

- PID resolves to landing page

- Provides detailed information
- Option to retrieve data

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

- Upon activating

- Query is re-executed against time-stamped and versioned DB
- Results as above are returned

- Query store aggregates data usage

Data Citation – Recommendations

17

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



- 14 Recommendations grouped into 4 phases:

- 2-page flyer

<https://rd-alliance.org/recommendations-working-group-data-citation-revision-oct-20-2015.html>

- More detailed report: Bulletin of IEEE TCDC 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>

- Paper pre-print:

<http://doi.org/10.5281/zenodo.4571616>



- <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 platform**
 - **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**

■ *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**

Large Number of Adoptions

21

- **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

22

- **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

23

■ 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)

- Paper submitted to Harvard Data Science Review
 - Principles
 - Reference implementations
 - Adoptions as Case Studies
 - **Lessons Learned**
- Paper pre-print:
<http://doi.org/10.5281/zenodo.4571616>

Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data

Andreas Reiter¹, Bernhard Grünewald^{2,3}, Carlo Maria Zwißl⁴, Chris Schmeier⁵, Florina Weisner¹, James Duncan⁶, Katharina Flöker¹, Koji Zetsu⁷, Kristof Meirner¹, Leslie McIntosh-Barnill⁸, Reyna Jenkyn⁹, Stefan Pröll¹⁰, Tomasz Miksa¹¹, Mark Parsons²,

¹ TU Wien, Vienna, Austria

² University of Alabama in Huntsville, AL, USA

³ Earth Observation Data Centre, Vienna, Austria

⁴ IERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne University, UPMC Univ Paris, Mersin, 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, Austin

Abstract

Precisely identifying arbitrary subsets of data so that those can be re-produced is a daunting challenge in data-driven science, the more so if the underlying data source is dynamically evolving. Yet, most settings exhibit exactly those characteristics: increasingly larger amounts of data being continuously ingested from a range of sources, 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 consistently, having a precise identification mechanism to reference, retrieve and work with such data is essential. The RDA Working Group on Dynamic Data Citation has published 11 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. This paper provides an overview of the recommendations, reference implementations and pilot systems deployed and analyses key lessons learned from those. This provides a solid basis for institutions and researchers considering adding this functionality to their data infrastructure.

1 Introduction

Accountability and transparency in automated decision making [1] 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, i.e. the ability to re-compute analyses, arriving 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 ingredient 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 non-trivial 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 for decision making 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 pre-processing routines, are re-collected, and data quality is continually improved by correcting errors. 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 as marking an area on an image) and the ambiguity of natural language (e.g. do the measurements in the time period from Jan 7 to June 12 include or exclude the respective start and end dates?).

What it means to adopt an FAQ (1/3)

- 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. Several pilots allow the user to decide when a query should persist.
- Which PID system should be used? — The one that works best for your situation.

MOU1

Definitely an issue that functionality can be lost in the transition to a new repository. Such has always been the case with all previous reference schemes. Anyone got an alternative?

Microsoft Office User, 18-Apr-21

What it means to adopt an FAQ (2/3)

26

- When multiple distributed repositories are queried, do we need complex time synchronization protocols? — No, not if the local repositories maintain time-stamps.
- 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.

What it means to adopt an FAQ (3/3)

27

- 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 to do.**

- The Recommendations work!
- Non-trivial implementation but all pilots found it worthwhile
 - Saved time and effort for users and repos
 - Better provenance and accuracy
 - Improved processes and documentation
- Technical challenges are solvable and pay for themselves over time
- Policy issues are key
 - Versioning – See RDA group and why we rely on the timestamp
 - Migration and maintenance of functionality is still not fully tested.

- Maintaining precise identification of data can be cumbersome but it is **essential**
- Maintenance of reference schemes is almost as essential as maintaining the data.
- Data are worthless unless you know what they are and where they are.

Adoption Stories

30

- 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

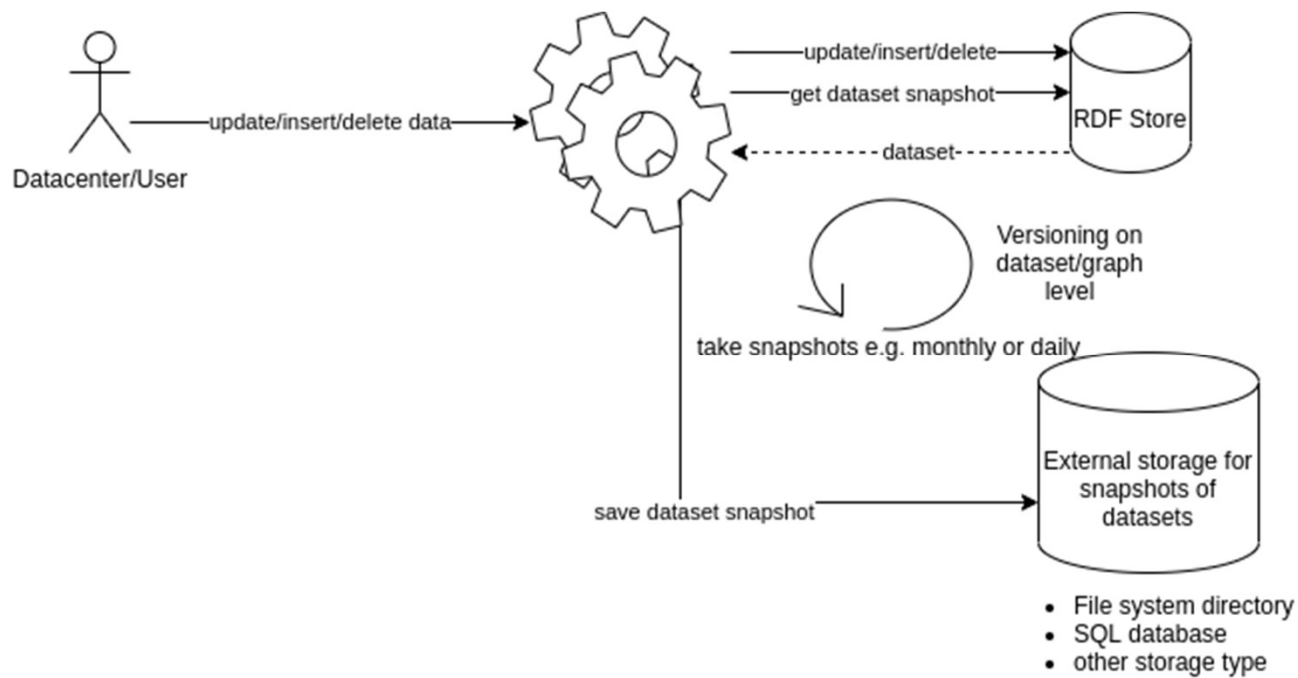
31

- Introduction, Welcome
- Short description of the WG recommendations
- Paper on adoption stories: lessons learned
- Q&A on recommendations
- On-going adoption activities, other WGs
- Other issues, next steps

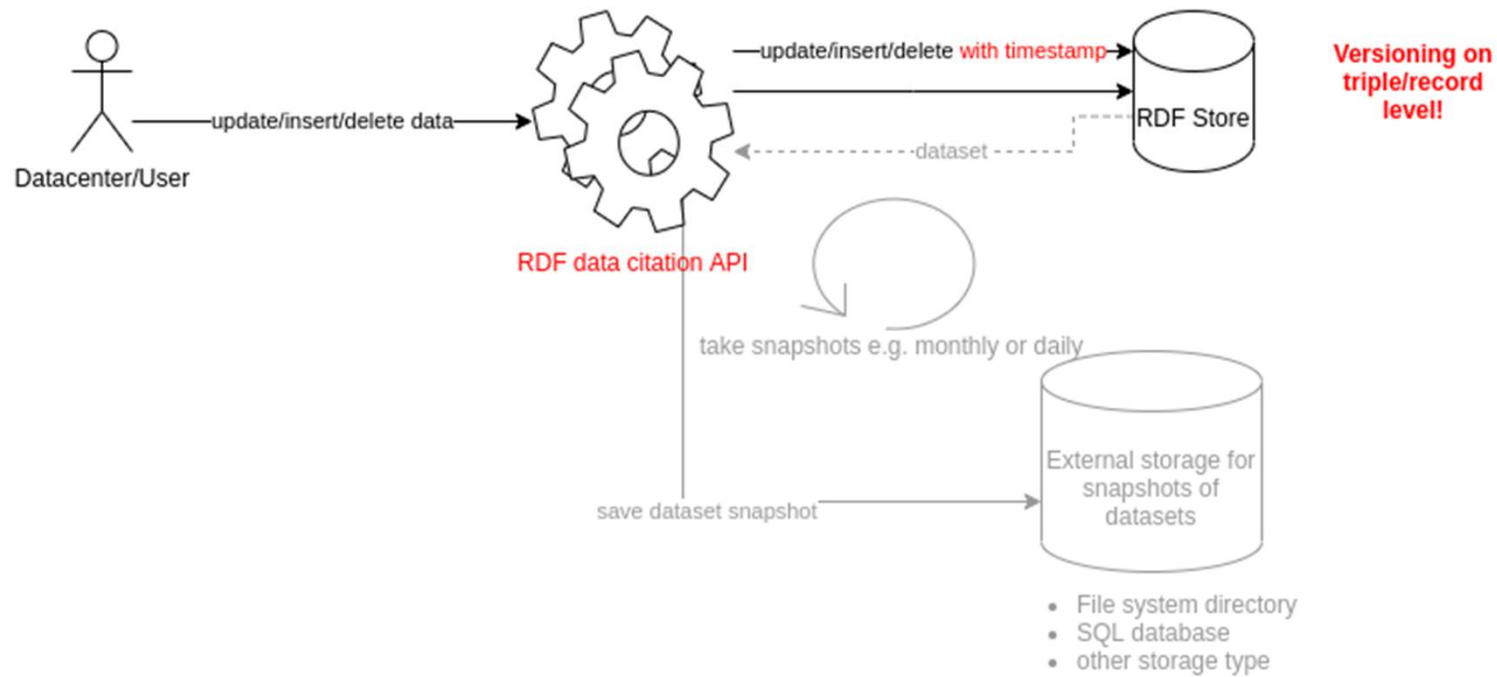
A data citation framework for RDF* stores

by Filip Kovacevic, BSc

Versioning RDF data - Current situation



Versioning RDF data - New world

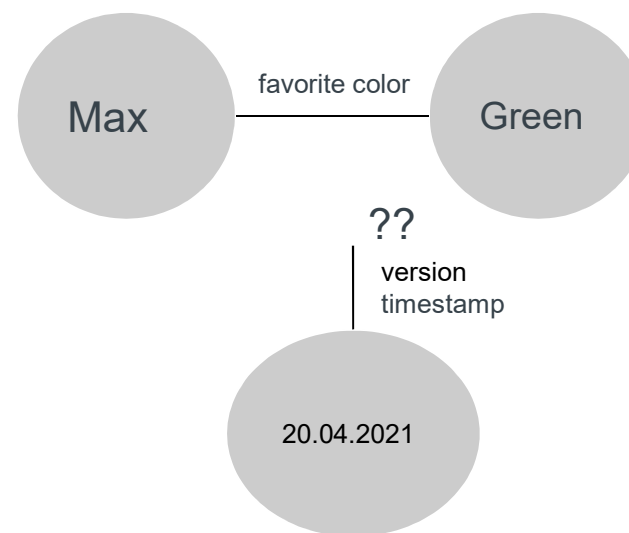


Versioning RDF data on triple level - current situation

Relational databases

Name	Date of birth	favorite color	version timestamp
Max	01.01.2000	green	19.04.2021
Max	01.01.2000	blue	20.04.2021

RDF data/triple stores

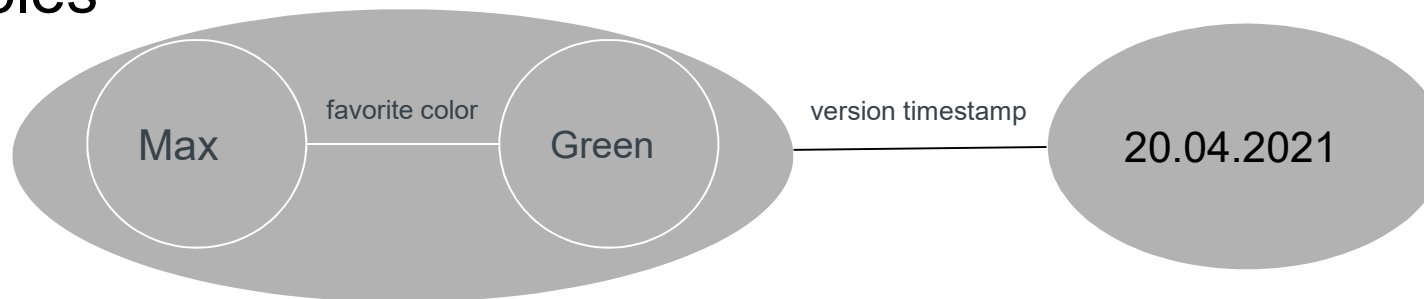


How to put a timestamp on the whole triple?

Versioning RDF data on triple level - solution

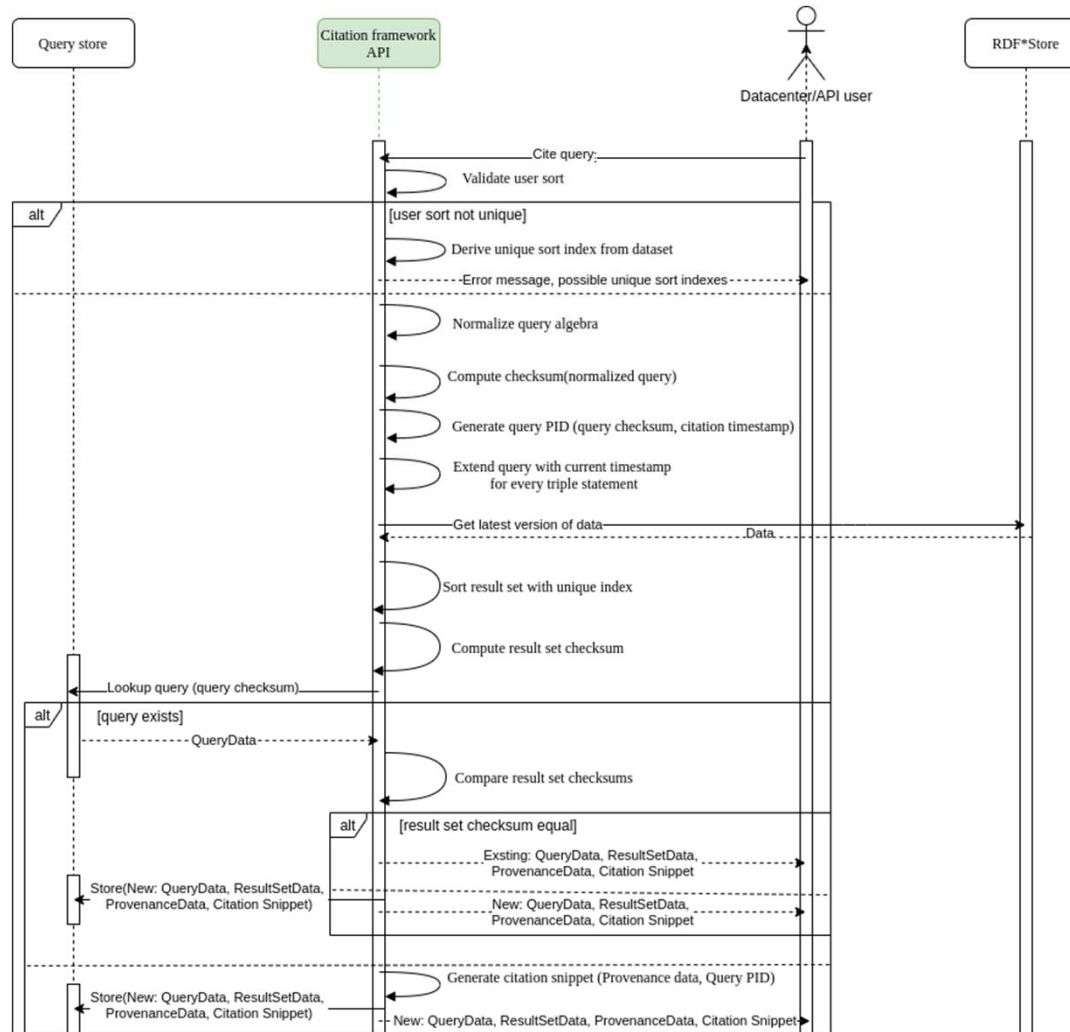
36

RDF* and SPARQL* allow for nested triples



Data citation is more than just versioning

API service - Citation use case



Live demonstration

Live demonstration

40

[Home](#) [News](#) [Contact](#)

Query editor

Add query here...

Execute

Result set

Metadata

Identifier: <https://doi.org/pid/of/query>
Creator: Filip Kovacevic
Title: Judy Chu occurrences
Publisher: Filip Kovacevic
Publication year: 2021
Resource type: Dataset/RDF data
Contributor: Tomasz Miksa

Execution and citation details

Execution timestamp:
Number of rows:
Query already cited?:
Result set changed since last citation?:
Order by attributes yield a unique sort?:

Follow Me

<https://github.com/GreenfishK/DataCitation>

Citation Snippet

Live demonstration

41

Home News Contact

Query editor

```
PREFIX publishing: <http://ontology.ontotext.com/publishing#>

select ?personLabel ?party_label ?document ?mention {
  ?mention publishing:hasInstance ?person .
  ?document publishing:containsMention ?mention .
  ?person pub:memberOfPoliticalParty ?party .
  ?person pub:preferredLabel ?personLabel .
  ?value pub:preferredLabel ?party_label .
  ?party pub:hasValue ?value .
  filter(?personLabel = "Barack Obama"@en)
} order by ?document
```

Execute

Result set

	personLabel	party_label	document	mention
0	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-health-ebola-usa-idUSKCN0HY2A520141010	http://data.ontotext.com/publishing#Mention-705a80ff154a6c9cce8ad6fa1aca37249c12bada619cdaffa5afd55aee07953f
1	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-73bb312bbde27a704a4e4cb4c25942b37c771e6f93cb7d9f163650728a67f75c
2	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-96cd9530c126974107c405f240907337db267d369e851e904b21ad75955473af
3	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-a7a06bb06c91289740a7888691757b1b2ec39a7cf3908aa885dc9cf070852b06
4	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-aea1154aa5adc65e96bf204e790b762f01cf1f87380358222f33ba182278ba82
5	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-d29176c43a5c6f95bb0820be4e215f9f6c6baca52e85a09da5a50dc46a348c2d

Metadata

Identifier: https://doi.org/pid/of/query
Creator: Filip Kovacevic
Title: Judy Chu occurrences
Publisher: Filip Kovacevic
Publication year: 2021
Resource type: Dataset/RDF data
Contributor: Tomasz Miksa

Execution and citation details

Execution timestamp:
Number of rows: 6
Query already cited?:
Result set changed since last citation?:
Order by attributes yield a unique sort?:

Follow Me

<https://github.com/GreenfishK/DataCitation>

Citation Snippet

Live demonstration

42

Home News Contact

```
filter(?personLabel = "Barack Obama"@en)  
} order by ?document
```

Execute

Result set

	personLabel	party_label	document	mention
0	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-health-ebola-usa-idUSKCN0HY2A520141010	http://data.ontotext.com/publishing#Mention-705a80ff154a6c9cce8ad6fa1aca37249c12bada619cdaffa5afd55aee07953f
1	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-73bb312bbde204b21ad75955473af163650728a67f75c
2	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-96cd9530c126904b21ad75955473af
3	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-a7a06bb06c91885dc9cf070852b06
4	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-aea1154aa5acd2f33ba182278ba82
5	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-d29176c43a5c6f95bb0820be4e215f9f6c6baca52e85a09da5a50dc46a348c2d

Cite

Home · Citations · About · Faq · Contact

Example Datacenter © 2021

[f](#) [t](#) [in](#) [g](#)

Execution and citation details

Execution timestamp: ?
Number of rows: 6
Query already cited?: ?
Result set changed since last citation?: ?
Order by attributes yield a unique sort?: **False**

Follow Me

<https://github.com/GreenfishK/DataCitation>

Citation Snippet

?

Error message

The "order by"-clause in your query does not yield a uniquely sorted dataset. Please provide a primary key or another unique sort index

Live demonstration

Home News Contact

PREFIX publishing: <http://ontology.ontotext.com/publishing#>

```

select ?personLabel ?party_label ?document ?mention {
  ?mention publishing:hasInstance ?person .
  ?document publishing:containsMention ?mention .
  ?person pub:memberOfPoliticalParty ?party .
  ?person pub:preferredLabel ?personLabel .
  ?value pub:preferredLabel ?party_label .
  ?party pub:hasValue ?value .
  filter(?personLabel = "Barack Obama"@en)
} order by ?mention
        
```

Execute

Result set

	personLabel	party_label	document	mention
0	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-health-ebola-usa-idUSKCN0HY2A520141010	http://data.ontotext.com/publishing#Mention-705a80ff154a6c9cce8ad6fa1aca37249c12bada619cdaffa5afd55aee07953f
1	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-73bb312bbde27a704a4e4cb4c25942b37c771e6f93cb7d9f163650728a67f75c
2	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-96cd9530c126974107c405f240907337db267d369e851e904b21ad75955473af
3	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-a7a06bb06c91289740a7888691757b1b2ec39a7cf3908aa885dc9cf070852b06
4	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-aea1154aa5adc65e96bf204e790b762f01cf1f87380358222f33ba182278ba82
5	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-d29176c43a5c6f95bb0820be4e215f9f6cbaca52e85a09da5a50dc46a348c2d

Cite

Identifier: https://doi.org/pid/of/query
Creator: Filip Kovacevic
Title: Judy Chu occurrences
Publisher: Filip Kovacevic
Publication year: 2021
Resource type: Dataset/RDF data
Contributor: Tomasz Miksa

Execution and citation details

Execution timestamp: 2021-04-21T13:16:33.619027+01:00
Number of rows: 6
Query already cited?: False
Result set changed since last citation?: False
Order by attributes yield a unique sort?: True

Follow Me

<https://github.com/GreenfishK/DataCitation>

Citation Snippet

DOI_to_landing_page, Filip Kovacevic, Judy Chu occurrences, Filip Kovacevic, 2021, Dataset/RDF data, pid: 676d5f53a7d8bb3cd3387ceda9d44666c714f4c620b615cc6d3

Live demonstration

Home News Contact

PREFIX publishing: <http://ontology.ontotext.com/publishing#>

```
select ?personLabel ?party_label ?document ?mention {
  ?mention publishing:hasInstance ?person .
  ?document publishing:containsMention ?mention .
  ?person pub:memberOfPoliticalParty ?party .
  ?person pub:preferredLabel ?personLabel .
  ?value pub:preferredLabel ?party_label .
  ?party pub:hasValue ?value .
  filter(?personLabel = "Barack Obama"@en)
} order by ?mention
```

Execute

Result set

	personLabel	party_label	document	mention
0	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-health-ebola-usa-idUSKCN0HY2A520141010	http://data.ontotext.com/publishing#Mention-705a80ff154a6c9cce8ad6fa1aca37249c12bada619cdaffa5afd55aee07953f
1	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-73bb312bbde27a704a4e4cb4c25942b37c771e6f93cb7d9f163650728a67f75c
2	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-96cd9530c126974107c405f240907337db267d369e851e904b21ad75955473af
3	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-a7a06bb06c91289740a7888691757b1b2ec39a7cf3908aa885dc9cf070852b06
4	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-aea1154aa5adc65e96bf204e790b762f01cf1f87380358222f33ba182278ba82
5	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-d29176c43a5c6f95bb0820be4e215f9f6c6baca52e85a09da5a50dc46a348c2d

Cite

Identifier: https://doi.org/pid/of/query
Creator: Filip Kovacevic
Title: Judy Chu occurrences
Publisher: Filip Kovacevic
Publication year: 2021
Resource type: Dataset/RDF data
Contributor: Tomasz Miksa

Execution and citation details

Execution timestamp: 2021-04-21T13:17:14.309114+01:00
Number of rows: 6
Query already cited?: True
Result set changed since last citation?: False
Order by attributes yield a unique sort?: True

Follow Me

<https://github.com/GreenfishK/DataCitation>

Citation Snippet

DOI_to_landing_page, Filip Kovacevic, Judy Chu occurrences, Filip Kovacevic, 2021, Dataset/RDF data, pid: 676d5f53a7d8bb3cd3387ceda9d44666c714f4c620b615cc6d3

Live demonstration

45

Home News Contact

PREFIX publishing: <http://ontology.ontotext.com/publishing#>

```
select ?personLabel ?party_Label ?document ?mention {
  ?mention publishing:hasInstance ?person .
  ?document publishing:containsMention ?mention .
  ?person pub:memberOfPoliticalParty ?party .
  ?value pub:preferredLabel ?party_Label .
  ?person pub:preferredLabel ?personLabel .
  ?party pub:hasValue ?value .
  filter(?personLabel = "Barack Obama"@en)
} order by ?mention
```

Execute

Result set

	personLabel	party_Label	document	mention
0	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-health-ebola-usa-idUSKCN0HY2A520141010	http://data.ontotext.com/publishing#Mention-705a80ff154a6c9cce8ad6fa1aca37249c12bada619cdaffa5afd55aee07953f
1	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-73bb312bbde27a704a4e4cb4c25942b37c771e6f93cb7d9f163650728a67f5c
2	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-96cd9530c126974107c405f240907337db267d369e851e904b21ad75955473af
3	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-a7a06bb06c91289740a7888691757b1b2ec39a7cf3908aa885dc9cf070852b06
4	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-aea1154aa5adc65e96bf204e790b762f01cf1f87380358222f33ba182278ba82
5	Barack Obama@en	Democratic Party@en	http://www.reuters.com/article/2014/10/10/us-usa-california-mountains-idUSKCN0HZ0U720141010	http://data.ontotext.com/publishing#Mention-d29176c43a5c6f95bb0820be4e215f9f6c6baca52e85a09da5a50dc46a348c2d

Cite

Identifier: https://doi.org/pid/of/query
Creator: Filip Kovacevic
Title: Judy Chu occurrences
Publisher: Filip Kovacevic
Publication year: 2021
Resource type: Dataset/RDF data
Contributor: Tomasz Miksa

Execution and citation details

Execution timestamp: 2021-04-21T13:17:46.788631+01:00
Number of rows: 6
Query already cited?: True
Result set changed since last citation?: False
Order by attributes yield a unique sort?: True

Follow Me

<https://github.com/GreenfishK/DataCitation>

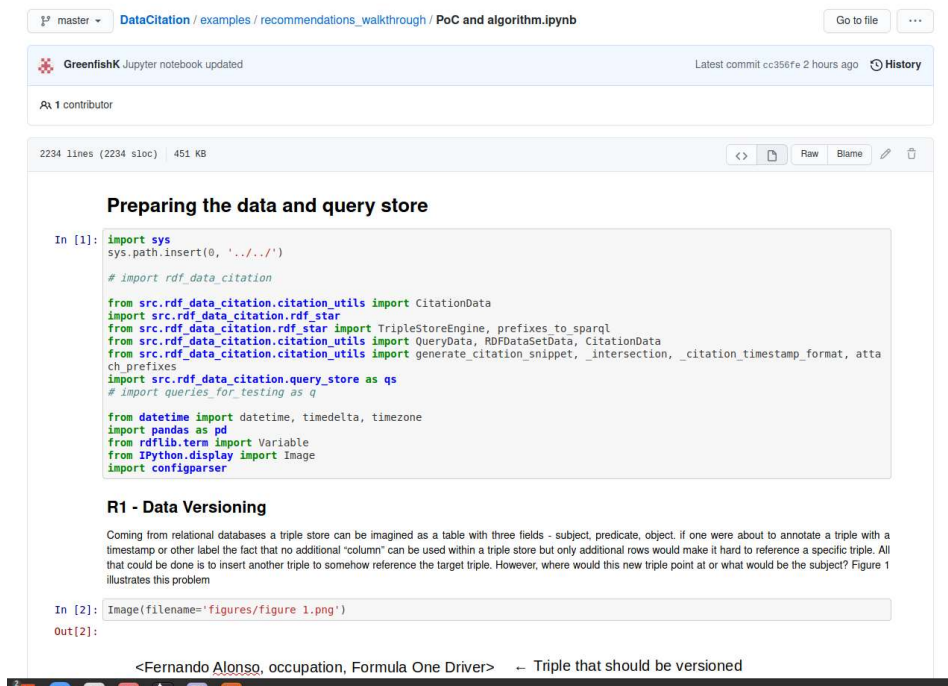
Citation Snippet

DOI_to_landing_page, Filip Kovacevic, Judy Chu occurrences, Filip Kovacevic, 2021, Dataset/RDF data, pid: 676d5f53a7d8bb3cd3387ceda9d44666c714f4c620b615cc6d3

Link to Proof of Concept (Jupyter Notebook)

46

- https://github.com/GreenfishK/DataCitation/blob/master/examples/recommendations_walkthrough/PoC%20and%20algorithm.ipynb



The screenshot shows a GitHub repository page for the file 'PoC and algorithm.ipynb' in the 'DataCitation' repository. The file is 2234 lines long, 451 KB, and was last committed by GreenfishK 2 hours ago. The notebook content is displayed in a Jupyter interface, showing the following code and text:

```
In [1]: import sys
sys.path.insert(0, '../..')

# import rdf_data_citation

from src.rdf_data_citation.citation_utils import CitationData
import src.rdf_data_citation.rdf_star
from src.rdf_data_citation.rdf_star import TripleStoreEngine, prefixes_to_sparql
from src.rdf_data_citation.citation_utils import QueryData, RDFDataSetData, CitationData
from src.rdf_data_citation.citation_utils import generate_citation_snippet, _intersection, _citation_timestamp_format, attach_prefixes
import src.rdf_data_citation.query_store as qs
# import queries_for_testing as q

from datetime import datetime, timedelta, timezone
import pandas as pd
from rdflib.term import Variable
from IPython.display import Image
import configparser
```

R1 - Data Versioning

Coming from relational databases a triple store can be imagined as a table with three fields - subject, predicate, object. If one were about to annotate a triple with a timestamp or other label the fact that no additional "column" can be used within a triple store but only additional rows would make it hard to reference a specific triple. All that could be done is to insert another triple to somehow reference the target triple. However, where would this new triple point at or what would be the subject? Figure 1 illustrates this problem

```
In [2]: Image(filename='figures/figure 1.png')
```

Out[2]:

<Fernando Alonso, occupation, Formula One Driver> -- Triple that should be versioned



Data Versioning WG

Versioning and Data Citation

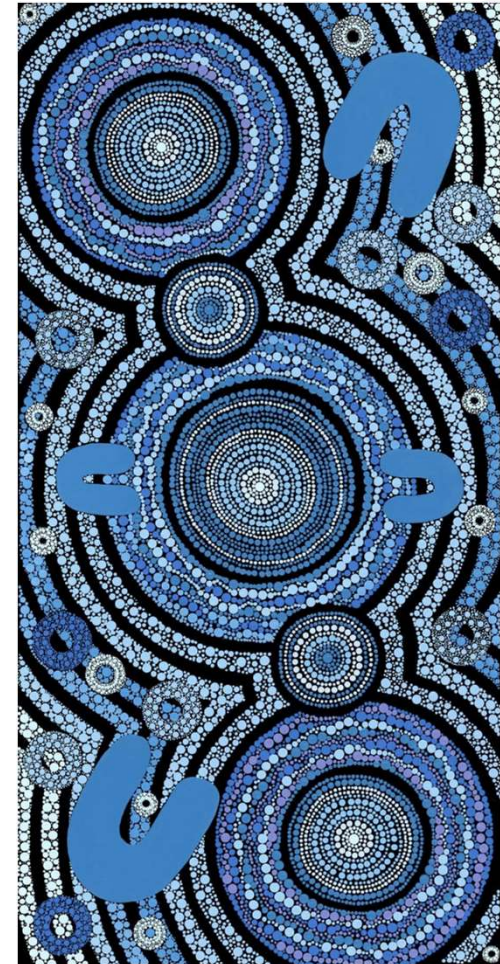
Jens Klump, Lesley Wyborn, Mingfang Wu

RDA Virtual Plenary 17
21 April 2021



ARDC, ANU and CSIRO acknowledge the Traditional Owners of the land, sea and waters, of the area that we live and work on across Australia. We acknowledge their continuing connection to their culture and we pay our respects to their Elders past and present.

Gadi artwork, NCI





What is a version?

- Different content?
- Different presentation?
- Different bit stream?

- Is the significance of a version change linked to the magnitude of the change in the bit stream?
- Can we use hashes to identify versions?





Size doesn't matter

- The first major split in the Christian church happened in the 380 CE over the words

ὁμοούσιον
vs.
ὁμοιούσιον

Levenshtein Distance = 1





How the Working Group Came About

- We tried to implement the Dynamic Data Citation WG recommendations at the National Computational Infrastructure (NCI) in Canberra.
- We realised that we need to collect and analyse use cases to understand versioning.





Versioning IG, WG: a Brief History

The group formed in 2016 as a BOF in Denver at P8 and has met at each plenary ever since as follows:

- P8 Denver (Sept 2016): BoF on Data Versioning
- P9 Barcelona (April 2017): Constituting the Data Versioning IG
- P10 Montreal (Sept 2017): Data Versioning IG, reforming as a WG
- P11 Berlin (March 2018): Data Versioning WG first meeting
- P12 Gaborone (Nov 2018): Data Versioning WG working meeting
- P13 Philadelphia (April 2019): Data Versioning WG draft report and recommendations
- P14 Helsinki (October 2019): Data Versioning WG final report and recommendations, preparation for TAB adoption.
- VP15 Melbourne (March 2020): Report on TAB adoption, discuss future of WG and agreed to go for IG
- VP16 Costa Rica (November 2020): Transition to Data Versioning IG to promote adoption and work on emerging topics in data versioning
- VP17 Edinburgh (April 2021): Advancing Data Versioning: From Principles to Actionable Recommendations





Use Cases

- The RDA Data Versioning Working Group collected 39 data versioning practice use cases from 33 organisations from around the world that cover different research domains, such as social and economic science, earth science, and molecular bioscience, and different data types.
- The use cases describe current practices reported by data providers.
- These use case descriptions are useful in identifying differences in data versioning practices between data providers and highlighting encountered issues.





Designated User Community

- An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time.
- Open Archival Information Systems (OAIS, CCSDS, 2012 page 1–11)





The Principles

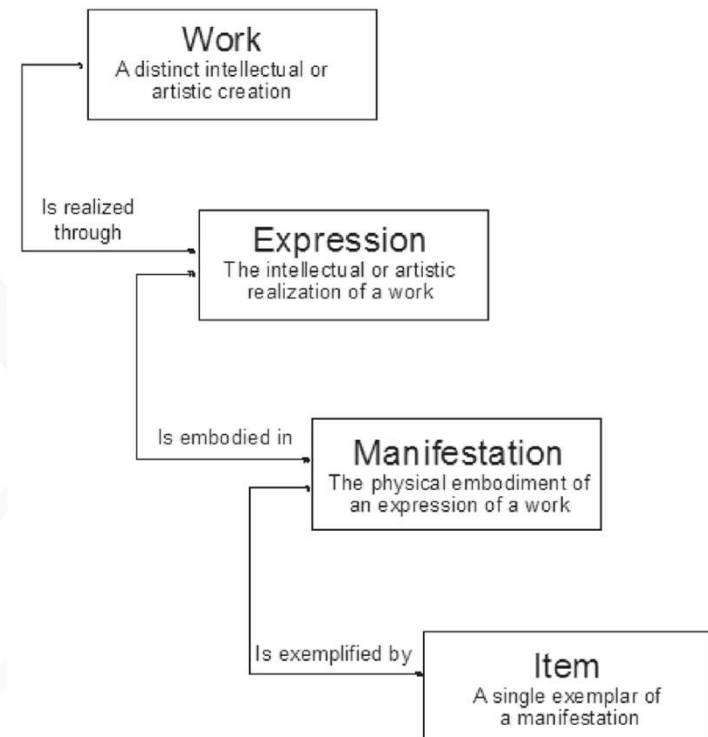
- Version Control (Revision)
 - Identify each change (revision), e.g. dynamic data versioning
- Data Production (Release)
 - Communicate the significance of the change, e.g. Semantic Versioning
- Objects and Collections (Granularity)
 - Identify collections of objects, time series, aggregates
- Formats (Manifestation)
 - Identify different formats of the same work
- Derived Products (Provenance)
 - Information about how this object was derived from a precursor object





Reusing FRBR and Software Versioning

- We used the Functional Requirements for Bibliographic Records (FRBR) to provide a conceptual framework
- The International Federation of Library Associations and Institutions (IFLA) developed FRBR to describe how information resources relate to each other
- We also used fundamentals of software versioning





The importance of mapping the Full-path of data

NASA Processing levels

L0 = Reconstructed, unprocessed instrument data at full resolution

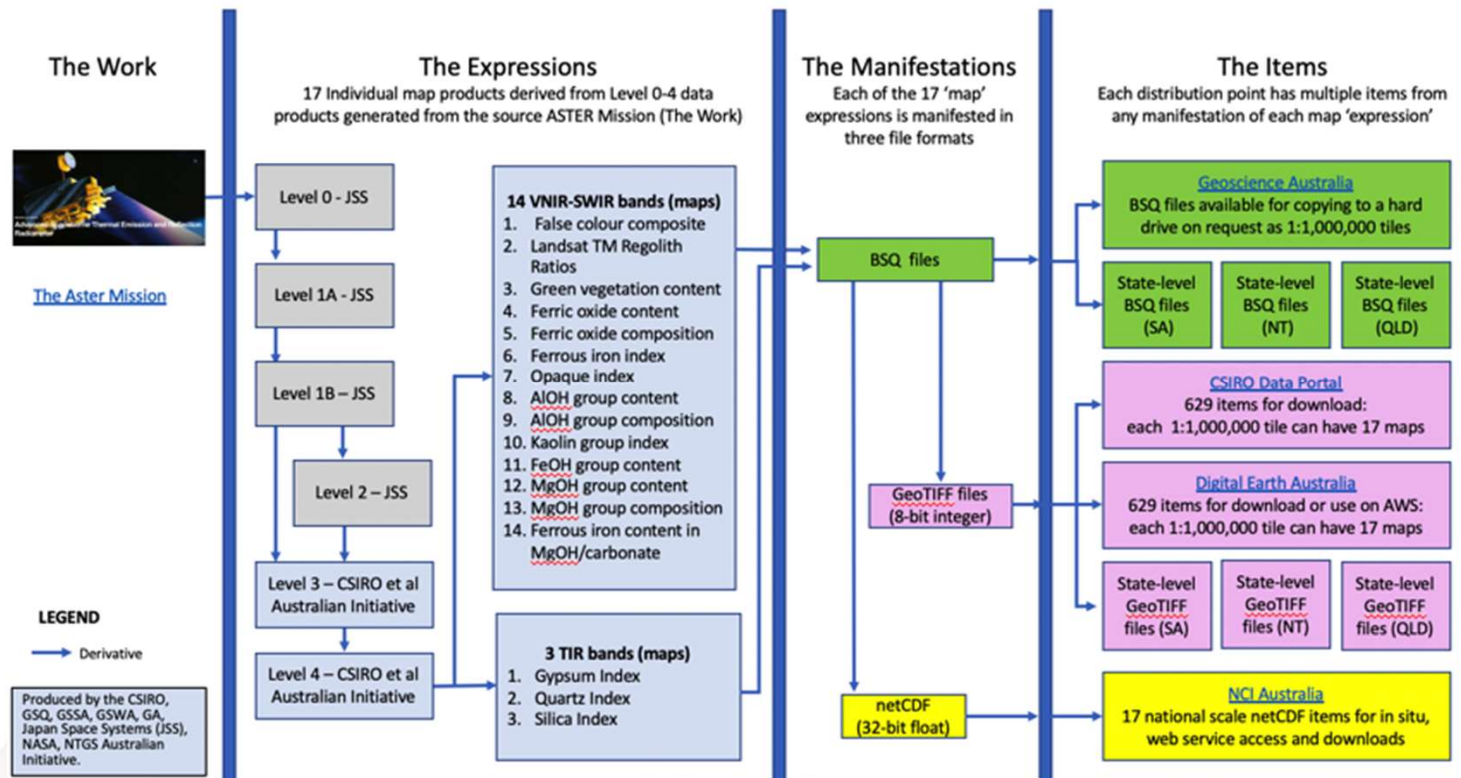
L1 = L0 data time-referenced, annotated & processed to sensor units

L2 = Derived geophysical variables at the same resolution

L3 = Variables mapped onto uniform space-time grid scales

L4 = Model outputs or results from analyses of lower level data

<https://earthdata.nasa.gov/collaborate/object-data-services/earthdata-processing-levels>





Data Science Journal Paper: March 2021

- Used concepts from software versioning as well as the Functional Requirements for Bibliographic Records (FRBR) as a conceptual framework
- Developed 6 foundational principles:
 1. Revision
 2. Release
 3. Granularity
 4. Manifestation
 5. Provenance
 6. Citation.
- Introduced the concept of the 'Full Path of data'
- Note: nowadays it is very rare for research to be based on a single dataset collected in one campaign
- doi:10.5334/dsj-2021-012

Home About Contact Content Research Integrity Search...

DATA SCIENCE JOURNAL

Reading: Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles Share: f t g+ in

Special Collection: [Research Data Alliance Results](#)

Research Papers

Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles

Authors: [Jens Klump](#), [Lesley Wyborn](#), [Mingfang Wu](#), [Julia Martin](#), [Robert R. Downs](#), [Ari Asmi](#)

Abstract

A dataset, small or big, is often changed to correct errors, apply new algorithms, or add new data (e.g., as part of a time series), etc. In addition, datasets might be bundled into collections, distributed in different encodings or mirrored onto different platforms. All these differences between versions of datasets need to be understood by researchers who want to cite the exact version of the dataset that was used to underpin their research. Failing to do so reduces the reproducibility of research results. Ambiguous identification of datasets also impacts researchers and data centres who are unable to gain recognition and credit for their contributions to the collection, creation, curation and publication of individual datasets.

Although the means to identify datasets using persistent identifiers have been in place for more than a decade, systematic data versioning practices are currently not available. In this work, we analysed 39 use cases and current practices of data versioning across 33 organisations. We noticed that the term 'version' was used in a very general sense, extending beyond the more common understanding of 'version' to refer primarily to revisions and replacements. Using concepts developed in software versioning and the Functional Requirements for Bibliographic Records (FRBR) as a conceptual framework, we developed six foundational principles for versioning of datasets: Revision, Release, Granularity, Manifestation, Provenance and Citation. These six principles provide a high-level framework for guiding the consistent practice of data versioning and can also serve as guidance for data centres or data providers when setting up their own data revision and version protocols and procedures.

Keywords: [Data Versioning](#), [File formats](#), [Provenance](#), [Citation](#), [Reproducibility](#), [Attribution](#)





The two critical needs for versioning

Reproducibility

- Reproducibility relies on the precise identification of the actual extract of the data used in a research project
- Failing to do so reduces the reproducibility of research results.

Authority, Identity and Ethics

- Unambiguous identification of datasets enables identification of authority and identity as well as ethical sharing of data
- Ambiguous identification impacts researchers, funders and data centres who are unable to gain recognition and credit for their contributions to the collection, creation, curation and publication of individual datasets.





Thank you!

- RDA Data Versioning WG co-chairs and co-authors:
 - Lesley Wyborn (ANU)
 - Mingfang Wu (ARDC)
 - Julia Martin (ARDC)
 - Ari Asmi (Univ. Helsinki)
 - Robert Downs (Columbia U.)
- Thank you!
 - Contributors of use cases
 - RDA TAB Liaisons
 - ARDC and Gerry Ryder
 - Reviewers



Thanks

61

Thanks!

And hope to see you at the
next meeting

of the

WGDC