Enabling Reproducibility for Small and Large Scale Research Data Sets

Stefan Pröll, Andreas Rauber

Secure Business Austria &
Vienna University of Technology

sproell@sba-research.org

Outline

- Research data challenges
- Reproducible data sets and subsets
- Small and large scale data settings
- Summary

Research Data Subsets

- Results are based on data
  - Data analysis
  - Data mining
  - Visualisation

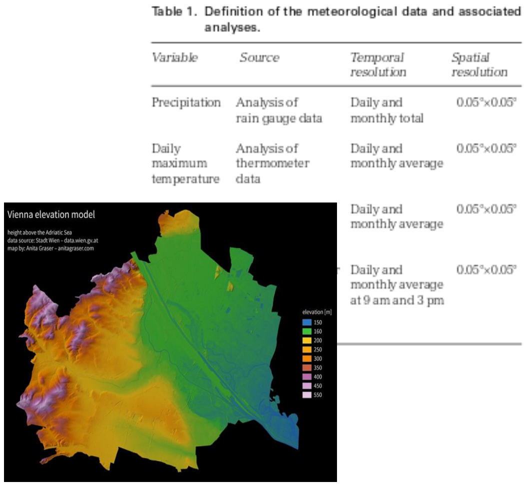
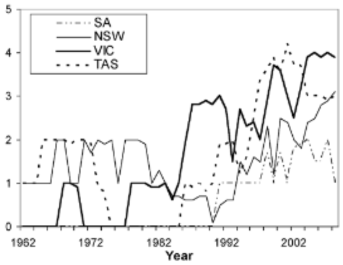

Vienna elevation model
height above the Adriatic Sea
data source: Stadt Wien – data.wien.gv.at
map by: Anita Graser – anitagraser.com
elevation [m]
150
160
200
250
300
350
400
450
550

Table 1. Definition of the meteorological data and associated analyses.

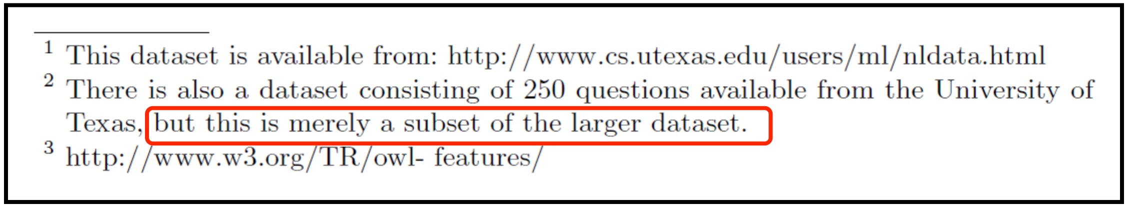
Variable	*Source*	*Temporal resolution*	*Spatial resolution*
Precipitation	Analysis of rain gauge data	Daily and monthly total	0.05°×0.05°
Daily maximum temperature	Analysis of thermometer data	Daily and monthly average	0.05°×0.05°
		Daily and monthly average	0.05°×0.05°
		Daily and monthly average at 9 am and 3 pm	0.05°×0.05°

Source: open.wien.gv.at

Fig. 4 The average number of high-elevation stations operating in January of the listed year. High-elevation stations are defined as those above 1500 metres in NSW and Victoria, above 1000 metres in Tasmania and above 700 metres in South Australia.



- Publications often contain just an aggregated result
  - Images
  - Tables
  - Graphs
  - Visualisations

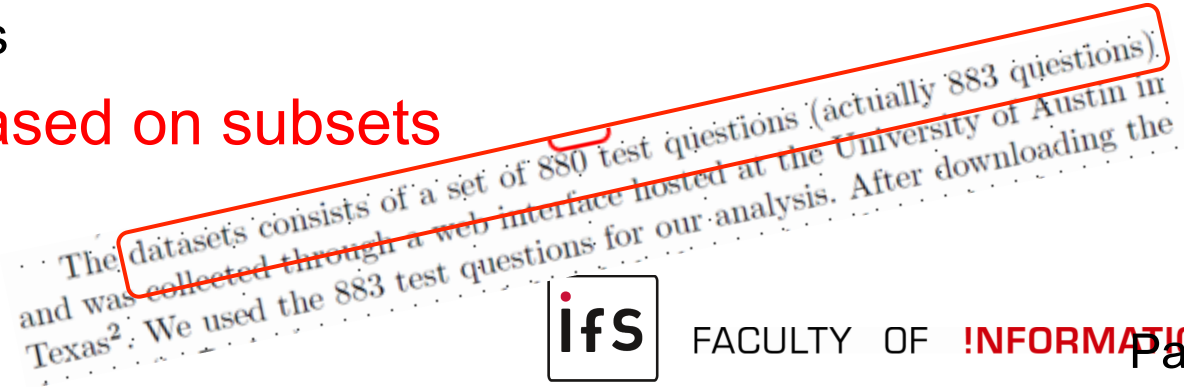


[1] This dataset is available from: http://www.cs.utexas.edu/users/ml/nldata.html

[2] There is also a dataset consisting of 250 questions available from the University of Texas, but this is merely a subset of the larger dataset.

[3] http://www.w3.org/TR/owl- features/

- Results are based on subsets



The datasets consists of a set of 880 test questions (actually 883 questions) and was collected through a web interface hosted at the University of Austin in Texas[2]. We used the 883 test questions for our analysis. After downloading the

Research Data

- Increasingly large amounts of data
  - Sensors
  - Streaming data
  - Time series
  - Satelite images
  - Real time analytics

Source: CartoDB

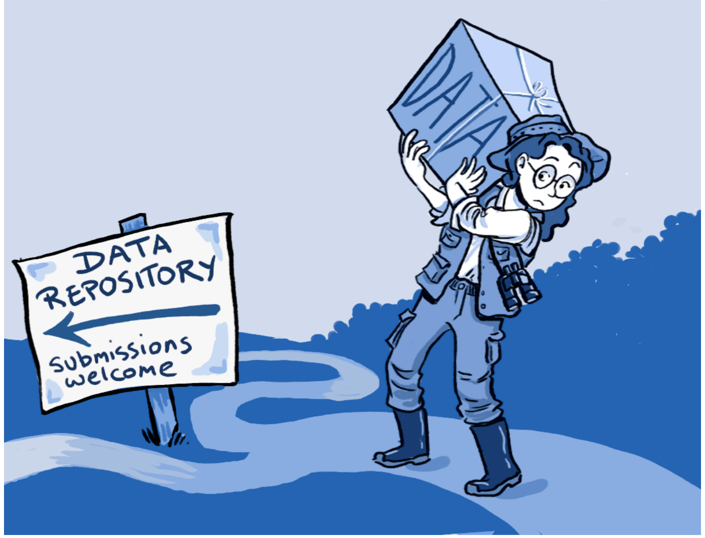
- Data is heterogenous across domains but homogenous within a domain
  - Silos prevent data and methods exchange
- How to improve and maintain accessibility to data?

Data and Data Citation

- Data as a “1st-class citizen” in science
- We need to be able to
  - preserve data and keep it accessible
  - cite data to give credit and show which data was used
  - **identify precisely the subset of data used** in a study/ process for repeatability, verifyability,…
- Why is this difficult?
  (after all, it’s being done…)

Main Challenges

- Scalability
  - More and more data sets
  - Growing amounts of data
  - Granularity
- Infrastructure
  - Sophisticated data management is not always available
  - Processes not defined well
- Dynamics
  - Frequent updates
  - Evolving data
- Precise identification
  - Ambiguity?



Src: CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=30978545

Granularity of Subsets

- What about the **granularity** of data to be identified?
  - Databases collect enormous amounts of data over time
  - 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

Identification of Dynamic Data

- Citable 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**

Goals

- Would like to cite precisely the **data as it existed at certain point in time**, without delaying release of new data
- We want to be able to **access previously existing version of a subset** and track the changes
- We want to **identify precisely the subset of (dynamic) data used** in a study
- Improve scalability

Solution

Idea: Versioned data + timestamped queries

- Data: timestamped and versioned (aka history)
- Query: Timestamped

- Access: Re-execute query on versioned data with the appropriate timestamp.

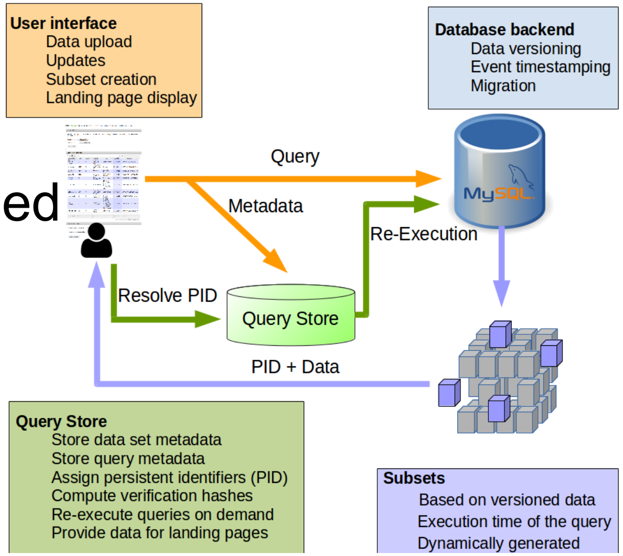
- **Trick: Assign the PID to the query**
- **Store queries enhanced with:**
  - **Time-stamping** for re-execution against versioned DB
  - **Normalize queries** for detecting duplicates
  - **Apply unique sorting**
  - **Compute hash** of the result-set for verifying identity/correctness

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

Query Store

- The Query Store is a central concept
  - Stores queries, parameters and metadata
  - Identifies all queries and data sets with unique PIDs
  - Establishes a link between the timestamped query and the versioned data set
  - Allows to re-execute queries and access the data
  - Provides information for landing pages
  - Allows to verify data sets and subsets
  - Analyse data usage
  - Can enforce policies
  - …



User interface
Data upload
Updates
Subset creation
Landing page display

Database backend
Data versioning
Event timestamping
Migration

Query Store
Store data set metadata
Store query metadata
Assign persistent identifiers (PID)
Compute verification hashes
Re-execute queries on demand
Provide data for landing pages

Subsets
Based on versioned data
Execution time of the query
Dynamically generated

Data Citation – Deployment

- Researcher uses workbench or tool 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!

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

- ubset of data
- ser gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - 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

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

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!

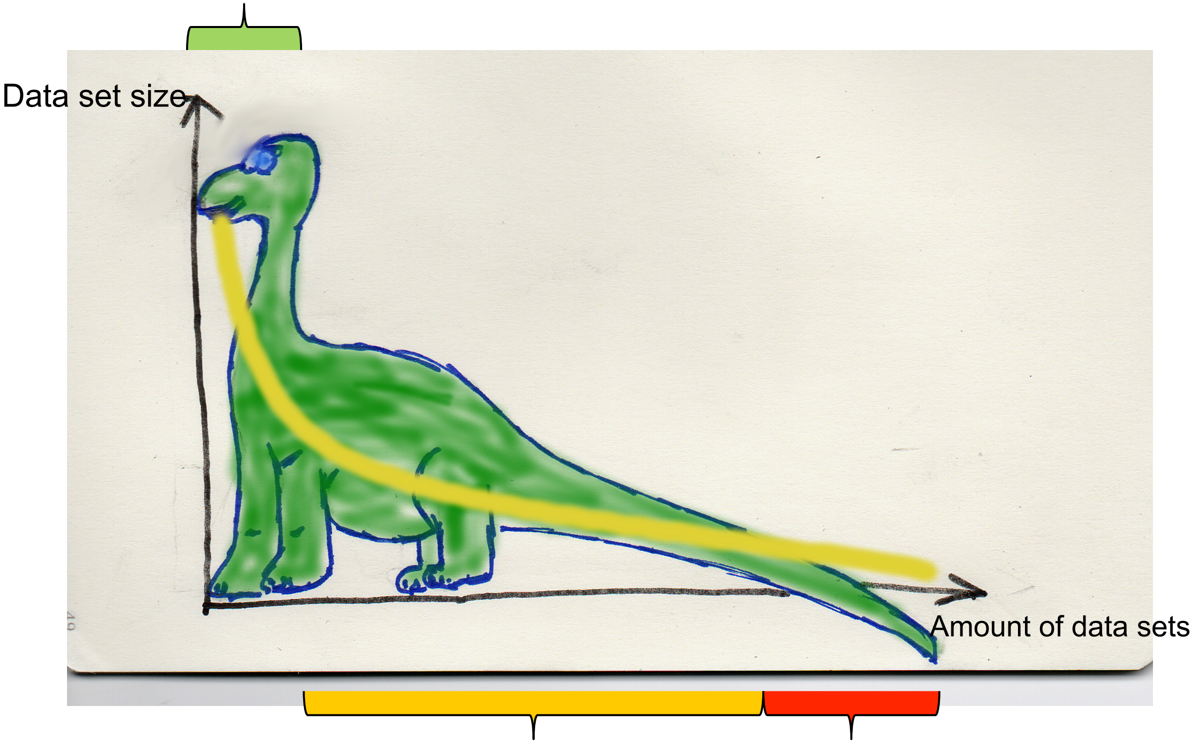
- PID resolves t…
  - Provides deta…
  - Option to retr…

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

- 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

Long Tail Research Data

Big data,
well organized,
often used and cited



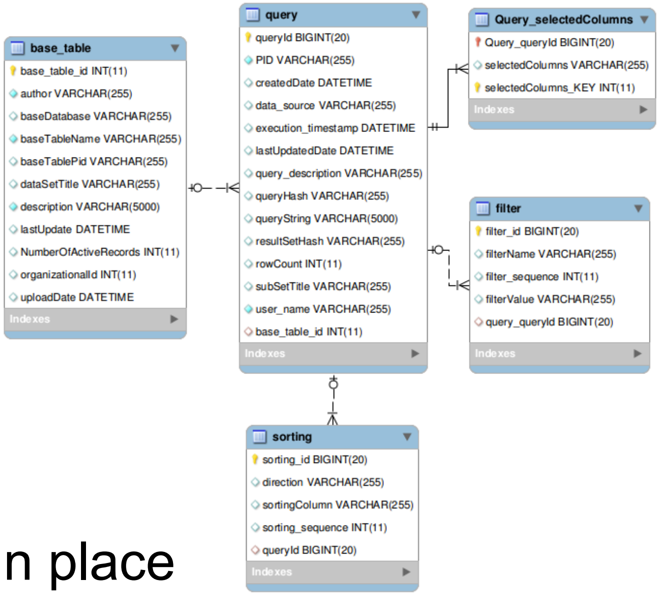
Less well organized,
non-standardised
no dedicated infrastructure

“Dark data”

[1] Heidorn, P. Bryan. "Shedding light on the dark data in the long tail of science." Library Trends 57.2 (2008): 280-299.

Large Scale Research Settings

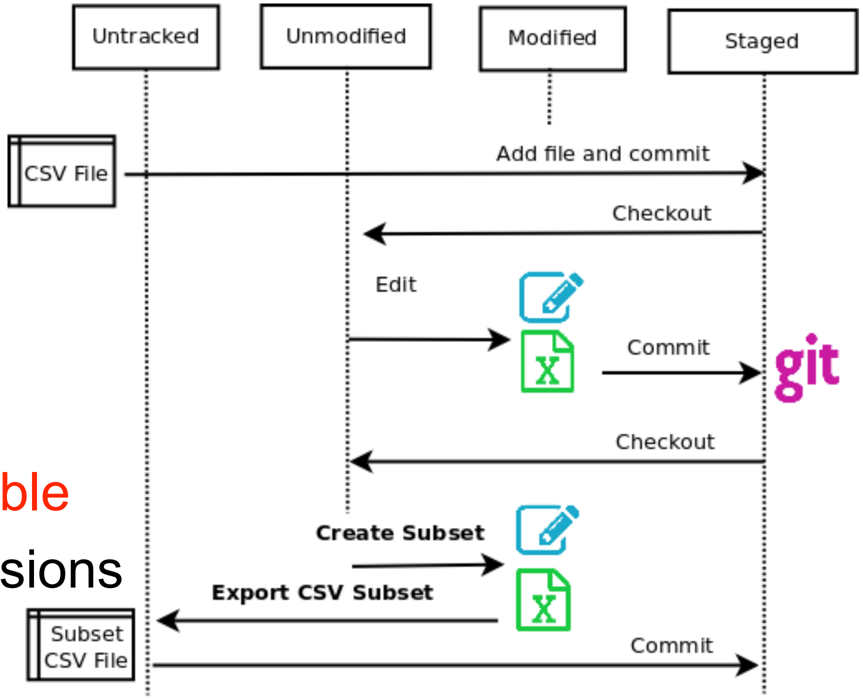
- Advanced data infrastructure
  - Big data
  - Database driven
  - Defined interfaces
  - Trained experts available



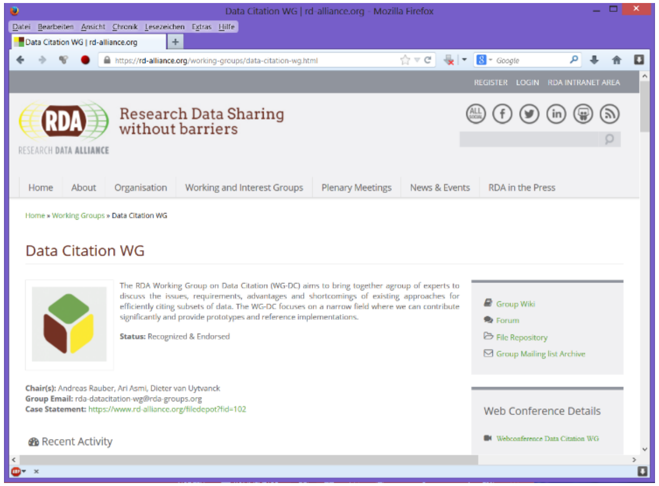
- Required adaptions
  - Introduce versioning, if not already in place
  - Capture subset process
  - Implement dedicated query store

Small Scale Research Settings

- Local workstations
  - Smaller data sets
  - Local storage and tools
  - Scripting languages
- Required adaptions
  - Data versioning, e.g. with Git
  - Store scripts versioned as well
  - Make subset creation reproducible
  - Document software and OS versions
  - Share repositories



RDA WG Data Citation


Data Citation WG | rd-alliance.org - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
Data Citation WG | rd-alliance.org
https://rd-alliance.org/working-groups/data-citation-wg.html
Google
REGISTER LOGIN RDA INTRANET AREA
RDA
RESEARCH DATA ALLIANCE
Research Data Sharing without barriers
Home About Organisation Working and Interest Groups Plenary Meetings News & Events RDA in the Press
Home » Working Groups » Data Citation WG
Group Wiki
Forum
File Repository
Group Mailing list Archive
Web Conference Details
Webconference Data Citation WG
Recent Activity

Data Citation WG

The RDA Working Group on Data Citation (WG-DC) aims to bring together agroup of experts to discuss the issues, requirements, advantages and shortcomings of existing approaches for efficiently citing subsets of data. The WG-DC focuses on a narrow field where we can contribute significantly and provide prototypes and reference implementations.

Status: Recognized & Endorsed

Chair(s): Andreas Rauber, Ari Asmi, Dieter van Uytvanck
Group Email: rda-datacitation-wg@rda-groups.org
Case Statement: https://www.rd-alliance.org/filedepot?fid=102

- Research Data Alliance
- WG on **Data Citation:**
  Making Dynamic Data Citeable
- WG officially endorsed in March 2014
  - Concentrating on the problems of **large, dynamic (changing) datasets**
  - Focus! Identification of data!
    Not: PID systems, metadata, citation string, attribution, …
  - Liaise with other WGs and initiatives on data citation (CODATA, DataCite, Force11, …)

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

Data Citation – Output

- 14 Recommendations
  grouped into 4 phases:
  - Preparing data and query store
  - Persistently identifying specific data sets
  - Resolving PIDs
  - Upon modifications to the data infrastructure
- 2-page flyer
- More detailed Technical Report:
  https://rd-alliance.org/group/data-citation-wg/wiki/wgdc-recommendations.html
- Reference implementations
  (SQL, CSV, XML) and Pilots

Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

Draft - Request for Comments

Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use

ABSTRACT

1. INTRODUCTION

NERC SCIENCE OF THE ENVIRONMENT   Argo   ESIP   VAMDC consortium

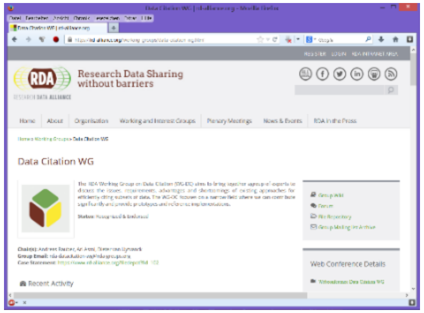
ifs   FACULTY OF !NFORMATICS

RDA Data Citation WG Pilots

Name	Data	Type	Status	Notes
Timbus	RDBMS	research	finished	Sensor data, pilot
XML-Reference	XML	research	finished	eXist-DB
DEXHELPP	CSV/RDBMS	research	running	Social security data
CSV-Reference	CSV/RDBMS	reference	running - β	Reference implem.
GIT-Reference	<ASCII>	reference	running - α	Reference implem.
VAMDC	SQL/NoSQL/ ASCII -> XML	deployment	running	Distributed data center
CBMI@wustl	RDBMS	deployment	starting	integration into i2b2
CCCA	NetCDF	deployment	starting	climate data
ENVRIplus	NetCDF	deployment	starting	ICOS: Carbon Obs.Infr.
ARGO	NetCDF	deployment	starting	ODIP-II, RDA-Europe
BCO-DMO	CSV	deployment	starting	RDA-US
VMC (Vermont)	VMC data cat.	deployment	starting	Forest Research Data
<a few others>	CSV, RDBMS	deployment	planned	Conceptual evaluation, seeking funding

Join RDA and Working Group

If you are interested in joining the discussion, contributing a pilot, wish to establish a data citation solution, …



Research Data Sharing without barriers

RESEARCH DATA ALLIANCE

Data Citation WG

The RDA Working Group on Data Citation (WG-DC) aims to bring together a group of experts to discuss the issues, requirements, advantages and shortcomings of existing approaches for efficiently citing subsets of data. The WG-DC focuses on a narrow field where we can contribute significantly and provide prototypes and reference implementations.

Status: Recognised & Endorsed

- Register for the RDA WG on Data Citation:
  - Website: https://rd-alliance.org/working-groups/data-citation-wg.html
  - Mailinglist: https://rd-alliance.org/node/141/archive-post-mailinglist
  - Web Conferences: https://rd-alliance.org/webconference-data-citation-wg.html
  - List of pilots: https://rd-alliance.org/groups/data-citation-wg/wiki/collaboration-environments.html

Thank you!

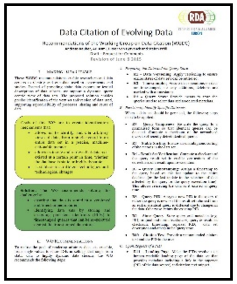
- Questions?
- Comments?

Thank you very much for your attention!

sproell@sba-research.org

@stefanproell

Data Citation – Recommendations



Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

[illegible]

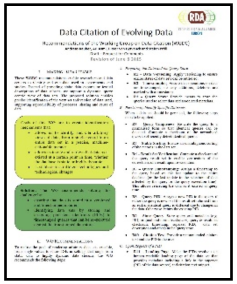
A) Preparing the Data and the Query Store

- **R1 – Data Versioning:** Apply versioning to ensure earlier states of data sets the data can be retrieved

- **R2 – Timestamping:** Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp

- **R3 – Query Store:** Provide means to store the queries used to select data and associated metadata

Data Citation – Recommendations


Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

Draft – Request for Comments

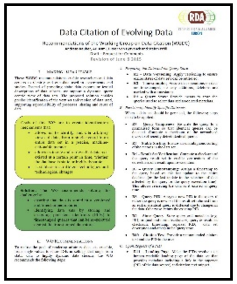
Revision of June 9 2015

B) Persistently Identify Specific Data sets (1/2)

When a data set should be persisted:

- **R4 – Query Uniqueness:** Re-write the query to a normalised form so that identical queries can be detected. Compute a checksum of the normalized query to efficiently detect identical queries
- **R5 – Stable Sorting:** Ensure an unambiguous sorting of the records in the data set
- **R6 – Result Set Verification:** Compute a checksum of the query result set to enable verification of the correctness of a result upon re-execution
- **R7 – Query Timestamping:** Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time). This allows retrieving the data as it existed at query time

Data Citation – Recommendations



Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

B) Persistently Identify Specific Data sets (2/2)

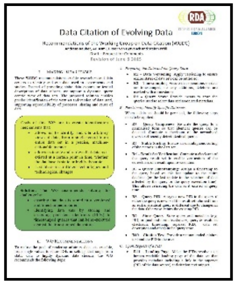
When a data set should be persisted:

- **R8 – Query PID:** Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID

- **R9 – Store Query:** Store query and metadata (e.g. PID, original and normalised query, query & result set checksum, timestamp, superset PID, data set description and other) in the query store

- **R10 – Citation Text:** Provide a recommended citation text and the PID to the user

Data Citation – Recommendations



Data Citation of Evolving Data

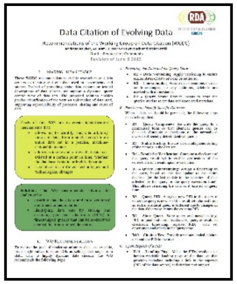
Recommendations of the Working Group on Data Citation (WGDC)

C) Upon Request of a PID

- **R11 – Landing Page:** Make the PIDs resolve to a human readable landing page of the data set that provides metadata including a link to the superset (PID of the data source) and citation text snippet

- **R12 – Machine Actionability:** Make the landing page machine-actionable, allowing to retrieve the data set by re-executing the timestamped query

Data Citation – Recommendations



RDA

Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

D) Upon Modifications to the Data Infrastructure

- **R13 – Technology Migration:** When data is migrated to a new representation (e.g. new database system, a new schema or a completely different technology), migrate also the queries and associated checksums

- **R14 – Migration Verification:** Verify successful query migration should, ensuring that queries can be re-executed correctly