RDA

RESEARCH DATA ALLIANCE

Dynamic Data Citation RDA Outputs Plenary @ P5 9th March 2015, San Diego

research data sharing without barriers
rd-alliance.org

RDA WG Data Citation - Agenda

- Brief re-cap of the WG goals and recommendation
- Prototype for CSV data
- Report on workshops / Update on pilots
  - SeaDataNet, NERC
  - ESIP Winter Meeting in Washington
  - CLARIN
  - NASA
  - VAMDC
- Future pilots: workshops planned, support for take-up
- Open issues concerning recommendations (timestamps, hashes,...)
- Writing up reports / recommendations
  - Finalizing recommendation
  - Summary report on pilots
  - Material for adoption support

RDA WG Data Citation

- Research Data Alliance
- WG on **Data Citation:**
  Making Dynamic Data Citeable
- WG officially endorsed in March 2014
- 2 areas of **focus**
  - Citing arbitrary **subsets** of data
  - Citing data that is **dynamic**
  - Stable across technology changes, scalable, implementable
- **Not** in focus
  - Metadata for citing data, landing page design
  - Which PID solution to adopt (DOI, URI, ARK, …)
- Liaise with other initiatives on data citation
  (CODATA, DataCite, Force11, …)

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

Citation 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 (how useful is this on its own?)
  - “Artificial” versioning by identifying batches of data (e.g. annual), aggregating changes into releases
- Would like to cite (and retrieve) precisely the **data as it existed at certain point in time**

Granularity of Data Citation

- What about **granularity** of data to be cited?
  - Databases collect enormous amounts of data over time
  - Researchers use specific subsets (“rows/columns”) of data
  - Need to identify precisely the subset used
- Current approaches
  - Storing pre-defined sub-sets -> not flexible enough
  - 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 (“rows”) -> scalability, not for arbitrary subsets (e.g. when not entire record selected)
- Would like to be able to cite precisely the **subset of (dynamic) data used** in a study

Principles of Dynamic Data Citation

We have: Data + Means-of-access

Dynamic Data Citation: Cite data dynamically via query!

Steps:

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

Researcher creates working-set via some interface:

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

   leading to landing page

Data Citation – Deployment

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

Data Citation – Deployment

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

- … set of data
- … 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

Data Citation – Deployment

- set of data
- r gets

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

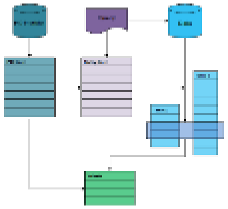
  - Data (pac kage, acc s API, …)
  - PID (e.g.
  - Hash val
  - Recomm

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

- PID resolves to lan ing page
  - Provides detailed metadata, lin 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

WG Pilots

- Pilot workshops and implementations by
  - Various EU projects (TIMBUS, SCAPE,…)
  - NERC (UK Natural Environment Research Council Data Centres)
  - ESIP (Earth Science Information Partners)
  - CLARIN (Field Linguistics Transcriptions)
  - Virtual Atomic and Molecular Data Centre (Worldwide e-infrastructure federating 41 heterogeneous and interoperable Atomic and Molecular databases)
- Prototype solutions for
  - SQL, XML, CSV
  - LOD/RDF, triple-store DBs in the queue
  - Distributed data

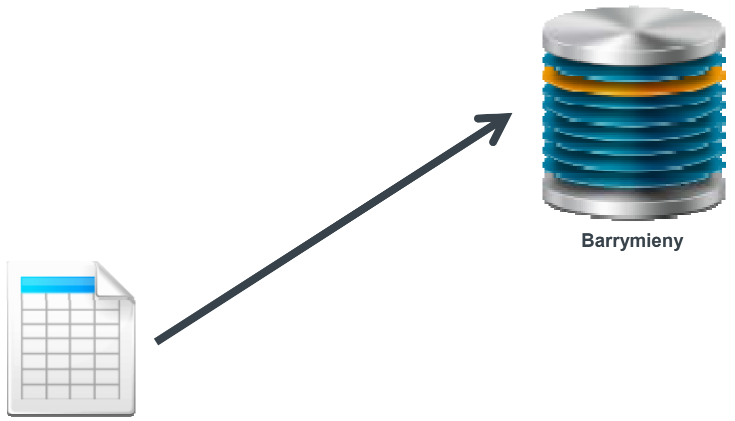


RDA WG Data Citation - Agenda

- Brief re-cap of the WG goals and recommendation
- Prototype for CSV data
- Report on workshops / Update on pilots
  - SeaDataNet, NERC
  - ESIP Winter Meeting in Washington
  - CLARIN
  - NASA
  - VAMDC
- Future pilots: workshops planned, support for take-up
- Open issues concerning recommendations (timestamps, hashes,...)
- Writing up reports / recommendations
  - Finalizing recommendation
  - Summary report on pilots
  - Material for adoption support

CSV Data Prototype

- Upload interface
  - Upload CSV files
- Migrate CSV file into RDBMS
  - Generate table structure
  - Add metadata columns for versioning
  - Add indices
- Dynamic data
  - Update existing records
  - Append new data
- Access interface
  - Track subset creation
  - Store queries

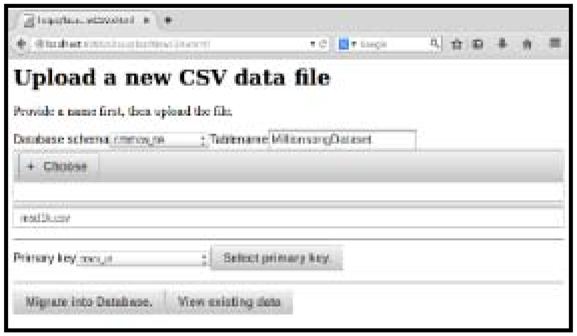


Prototype Implementation: CSV

Remember

- It’s a prototype implementation: proof of concept!
- No sophisticated authentication, security mechanisms, privacy protection, …
- No decorated UI, color schemes, logos, …
- No extensive list of metadata elements describing the dataset

- But, basically, a fully functional prototype
- Relying on a SQL database as a back-end
- Performing versioning / timestamping
- Storing queries persistently
- Assigning PIDs
- That can be resolved, leading to a landing page, leading to data



Upload a new CSV data file

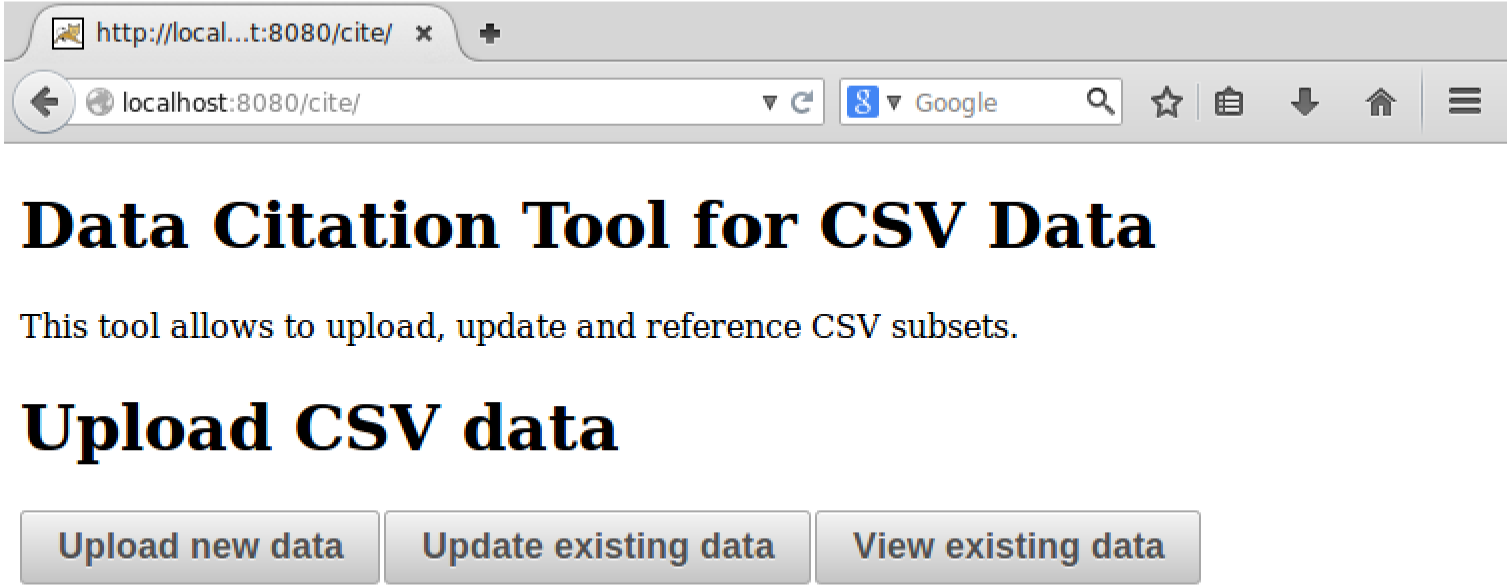
Provide a name first, then upload the file.

Database schema [illegible] Tablename MillionSongDataset

[illegible].csv

Primary key [illegible]

CSV Data Prototype



Data Citation Tool for CSV Data

This tool allows to upload, update and reference CSV subsets.

Upload CSV data

http://ate.ifs.tuwien.ac.at/cite/

RDA WG Data Citation - Agenda

- Brief re-cap of the WG goals and recommendation
- Prototype for CSV data
- Report on workshops / Update on pilots
  - SeaDataNet, NERC
  - ESIP Winter Meeting in Washington
  - CLARIN
  - NASA
  - VAMDC
- Future pilots: workshops planned, support for take-up
- Open issues concerning recommendations (timestamps, hashes,...)
- Writing up reports / recommendations
  - Finalizing recommendation
  - Summary report on pilots
  - Material for adoption support

Dynamic Data Citation - Pilots

Progress update from VAMDC

Carlo Maria Zwölf

Virtual Atomic and Molecular Data Centre
carlo-maria.zwolf@obspm.fr

VAMDC

- Virtual Atomic and Molecular Data Centre
- Worldwide e-infrastructure federating 41 heterogeneous and interoperable Atomic and Molecular databases
- Nodes decide independently about growing rate, ingest system, corrections to apply to already stored data
- Data-node may use different technology for storing data (SQL, No-sql, ASCII files),
- All implement VAMDC access/query protocols
- Return results in standardized XML format (XSAMS)
- Access directly node-by-node or via VAMDC portal, which relays the user request to each node

VAMDC

Workshop prior to P4

Issues identified

- Each data node could modify/delete/add data without tracing
- No support for reproducibility of past data extraction

Proposed Data Citation WG Solution:

- Considering the distributed architecture of the federated VAMDC infrastructure, it seemed very complex to apply the "Query Store" strategy
  - Should we need a QS on each node?
  - Should we need an additional QS on the central portal?
  - Since the portal acts as a relay between the user and the existing nodes, how can we coordinate the generation of PID for queries in this distributed context?

VAMDC

Status / Progress since P4

- Versioning adopted prior to P4
- Central service registering user interactions with data
- At each client SW notifies tracing service that a given **user** is using, at a given **time**, that specific **software** for submitting a given **query**
- Will assign single identifier for each unique query centrally
- Query store initially private (confidentiality issues)

Dynamic Data Citation - Pilots

Results from NERC Workshop

June 1-2 2014, London

John Watkins

Centre for Ecology and Hydrology

jww@ceh.ac.uk

Dynamic Data Citation - Pilots

Pilots discussed at ESIP Winter Mtg

Ruth Duerr

(separate slide set)

Dynamic Data Citation - Pilots

Dynamic Data: Solar Physics

J.A. Hourclé

(separate slide set)

RDA WG Data Citation - Agenda

- Brief re-cap of the WG goals and recommendation
- Prototype for CSV data
- Report on workshops / Update on pilots
  - SeaDataNet, NERC
  - ESIP Winter Meeting in Washington
  - CLARIN
  - NASA
  - VAMDC
- Future pilots: workshops planned, support for take-up
- Open issues concerning recommendations (timestamps, hashes,...)
- Writing up reports / recommendations
  - Finalizing recommendation
  - Summary report on pilots
  - Material for adoption support

Further Pilots, Workshops

Upcoming workshops, presentations

- Workshop on Dynamic Data Citation, Riva del Garda, April 20/21
- Research Infrastructures meet RDA, SARA, Amsterdam, May 26/27
- RDA/DINT Workshop: RDA Results Workshop, KIT; Karlsruhe, May 28/29

- Webinar AIS&T, DCMI: Approaches to Making Dynamic Data Citable: Recommendations of the RDA Working Group, April 8, 10:00 EDT
- RDA Webinar: Doodle Poll

- Further individual pilots workshops encouraged!

RDA WG Data Citation - Agenda

- Brief re-cap of the WG goals and recommendation
- Prototype for CSV data
- Report on workshops / Update on pilots
  - SeaDataNet, NERC
  - ESIP Winter Meeting in Washington
  - CLARIN
  - NASA
  - VAMDC
- Future pilots: workshops planned, support for take-up
- Open issues concerning recommendations (timestamps, hashes,...)
- Writing up reports / recommendations
  - Finalizing recommendation
  - Summary report on pilots
  - Material for adoption support