Working Group on Data Citation: Making Dynamic Data Citeable

Andreas Rauber, Ari Asmi,
Dieter van Uytvanck, Stefan Pröll

research data sharing without barriers
rd-alliance.org

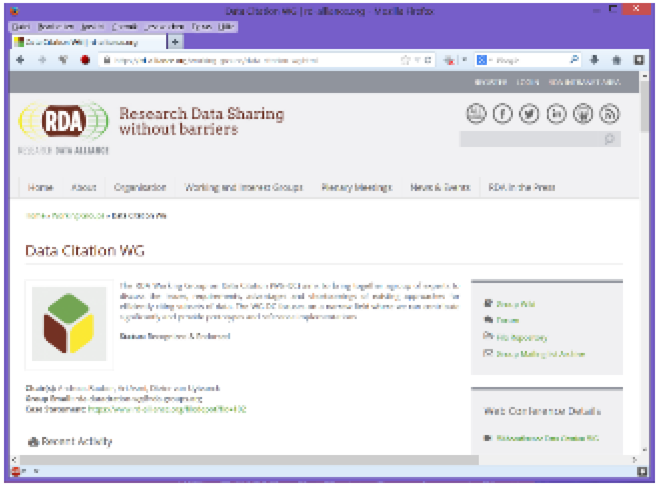
What is the Problem?

- 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, …
- Would like to cite precisely the **data as it existed at certain point in time**, without delaying release of new data

- What about the **granularity** of data to be identified/cited?
  - Databases collect enormous amounts of data over time
  - Researchers use specific subsets of data
  - Need to identify precisely the subset used
- Would like to be able to indentify & cite precisely the **subset of (dynamic) data used** in a study

RDA WG Data Citation



Data Citation WG | rd-alliance.org - Mozilla Firefox

Research Data Sharing without barriers

Home | About | Organization | Working and Interest Groups | Plenary Meetings | News & Events | RDA in the Press

Data Citation WG

[illegible]

Web Conference Details

Recent Activity

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

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

Solution

Data Citation: Data + Means-of-access

- Data → time-stamped & versioned (aka history)

Researcher creates working-set via some interface:

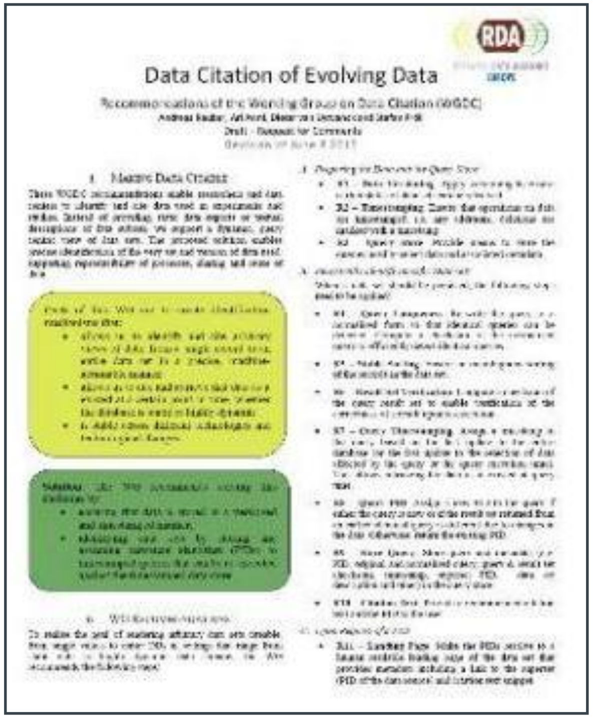
- Access → **assign PID to QUERY**, enhanced with
  - **Time-stamping** for re-execution against versioned DB
  - **Re-writing** for normalization, unique-sort, mapping to history
  - **Hashing** result-set: verifying identity/correctness

  leading to landing page

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

Output / Results

- 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
- Technical Report to follow
- Reference implementations (SQL, CSV, XML)
- Pilots



Data Citation of Evolving Data

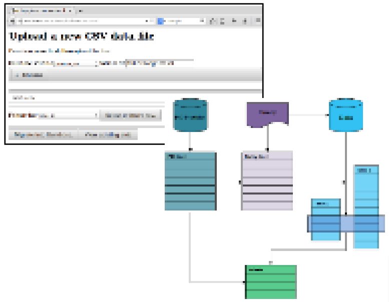
Recommendations of the Working Group on Data Citation (WGDC)

Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll

Draft - Request for Comments

[illegible]

[illegible]



Upload a new CSV data file

WG Pilots

- Pilots and implementations by
  - LNEC: Critical Infrastructure Monitoring System
  - Virtual Atomic and Molecular Data Centre
  - NERC (UK Natural Environment Research Council Data Centres)
    - ARGO Buoy Network
    - River Flow Dataset
  - ESIP (Earth Science Information Partners)
    - BCO-DMO
  - DEXHELPP – Social Security Data
  - ENVRIplus: Carbon Observation System
  - Million Song Database, IR Benchmark DBs
  - Several others under discussion…

Next steps

- Wrapping up this WG:
  - Finalize detailed report
  - Wrap up reference implementations
  - Publish results
  - Get them adopted as RDA outputs

- Follow-up activities
  - Help with adoption!
    Support for implementations – RDA Collaboration Projects:
    http://europe.rd-alliance.org/rda-europe-call-collaboration-projects
  - Revise / enhance recommendations
  - Tackle some open issues and more challenging settings (LOD, distribution, generalized views, permissions, …)