# Working Group Meeting:

# WG Data Citation: Making Dynamic Data Citeable

March 9 2015, San Diego

Notes prepared by Peter Kraker
pkraker@know-center.at

## Agenda

- Brief re-cap of the WG goals and recommendation
- Prototype for CSV data
- Report on workshops / Update on pilots
- Future pilots
- Open issues
- Writing up reports/recommendations

## Introduction (Andreas Rauber)

WG officially endorsed in March 2014
2 areas of focus:

1. Citing arbitrary subsets of data
2. Citing data that is dynamic

### Citation of dynamic data

Problem: citable datasets have to be static
But: research data is dynamic!
Current approaches not usable

--> Cite and retrieve precisely the data as it existed at a certain period in time

### Granularity of Data Citation

Current approaches do not take that into account
--> Cite the exact subset of (dynamic) data used in a study

### Principles of Dynamic Data Citation

Cite data dynamically via query

Prerequisites:

1) data is timestamped
2) data is versioned

Access: assign PID to "Query" enhanced with time-stamping, re-writing, hashing

Data Citation - Deployment:

- Researcher uses workbench to identifiy subset of data
- Upon executing selection ("download") user gets
- PID resolves landing page
- Note: query string provides valuable provenance information on the data set
- Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation, query is re-exectued against time-stamped and versioed DB

## Questions and Comments

Q: Query re-execution could be a time and computationally expensive task
A: Depends on setting, control by storing original execution time; depends also on versioning; one option is to outsource historic data to a second database

Comment: Like approach that you can see changes -> you can notify researchers that the data has changed
A: You can build meta-studies with that; show how inferences have changed over time

Q (NIH): Is there any constraint on the size of the dataset, or the size of the slices?
A: No constraints, but there will be problems if there are too many versions (but the only limitation is cost of storage)

Comment: Great stuff, we have enormous datasets and we had no way of identifying a subset
Historical data: if you have no more storage space

Q: Can you comment on citation metrics and how they would work?
A: The person in charge may be interested in what part of the data is more interesting
Creator of the subset as author
Identifiies subset and superset; when is query unique enough to warrant a citation?

Q: Same query at a different point in time is different?
A: Yes, you could view that as a different query. But technically, there is no difference if underlying data has not changed in the meantime.

Q: Versioning of software of queries?
A: Yes, we have considered that. You can migrate queries over different software, but the order may be a problem. Unique default sorts before user-requested sorts

Q: When the hash of the query and the hash of the results are the same, it should be a different PID because the question is asked on a different day.

A: We are not talking about research questions; you could assign a different PID, but we don't advise that.

Q: Vulnerable to change in DB structure?
A: Whenever you have a schema migration, you also need to change the resulting query translations. Should not happen too frequently as schema migraton is usually a major change for the entire RI (adapting APIs, …)

Q: Is NoSQL considered?
A: 1st pilot: SQL, 2nd pilot: CSV (most widespread data format), pilots with XML data. Versioning and timestamps for linked data.
GOAL: Running as many pilots as possible in the last half year. Sign up to mailing list! Focused workshops --> implementation. Offer to come for a workshop or meeting. Lots of many smaller meetings. Collect feedback -> write up recommendations. Don't start with the most difficult database. Looking for a pilot on No-SQL.

Q: Subset not created by query but by files downloaded?
A: Fileset is a kind of query: select those files WHERE ... discussed already in several mtgs., should be doable. Easiest solution via repository infrastructure supporting selection (querying) of file sets.

First pilot on distributed data

## Prototype on CSV

- Upload CSV files
- Migrate CSV file into RDBMS:
    - Generate table structure
    - Add metadata columns for versioning
    - Add indices

- Dynamic data
- Access interface

Protoype: Little security, basic interface, but fully functional
DEMO

**Upload**

- Upload CSV files -> Upload new data
- When you upload, you can give it a title; change table name, specify author, add description, upload CSV file
    - WHO CSV file: country data
- Process file -> identifies columns; set unique id (or system sets it)
- Migrate into database

**Create a subset**

- Select columns
- Apply filter

- Provide title for subset
- Provide description

-> Creates link to a landing page:

- Subset PID
- Author
- Authors of original dataset

- Suggested citation text
- Download dataset, subset and latest subset or diff

-> Upload new CSV file:

- Only new data (append-only), or check for updates / deletes (requires primary key to identify identity of lines/records)

Prototype will be polished and published on github

## Pilot presentation: John Watkins

Progress on Data Citation within UK NERC Data Centres

### Recap

- Joined WG in Plenary 3.
- British Library Workshop - July 2014
- Pragmatic approach: needs-driven and adoption in real world
- Various different angles: view points and how we would address them and take them forward
- Reported to Plenary 4

### Argo global array

- 3000 free-drifting profiling floats
- Real need (data collected right now)
- DOI-based
- Snapshot method
- The RDA conceptual model is being used to guide how
- As of this morning there is a DOI landing page -> long way from facilities presented in the earlier prototype
- Snapshot DOIs
- Move to dice and slice setup- > happens already, automate that with the RDA model. timeslice
- Include timestamps in DOI
- Incentive: track where ARGO data is being used
- How to integrate? Short DOIs?
- Thanks to Justin Buck

### Questions and Comments

Comment: Does not make sense to put intelligence into identifier - you can either manage it at the registry or at the de-referencing point.
A: People like to have some kind of readability in a DOI.

Q: There are lots of discussions why information should be put into the ID; shortening inhibits tracking
Andi: Time and location semantics problematic in IDs

Q (DRYAD): We have some trouble resolving query strings. Do you have that?
A: This syntax is not live; but I can confirm that there have been issues

Q: Not all browser handle query strings the same way
A: Part of the implementation

## Dynamic Data Citation Wokshop at ESIP (Ruth Duerr)

### Presentation

Use cases:

- MODIS data: level 2500 m snow product
- BCO-DMO ship and aerosol data
- LASP Interactive Solar Irradiance Datacenter

ESIP has had guidelines for citing dynamic data for many years -> machine-readable solution for reproducibility

LISIRD system only needs minor tweaks
BCO-DMO needs to investigate costs
MODIS problematic: different access services and federated nature of them

Identified a simple tool that would be helpful:

- Researcher would point to directory tree
- The tool would record file names and checksums
- ...

### Questions and Comments

Q: Multiple agencies involved. Publisher allows only one agency, but all want recognition. Is that a problem?
A: Not really. Two ID mode: regular dataset ID & subset ID. Not the problem that two groups created the same subset
A: Nothing prohibits you to list two editors of a book
A: The landing page can provide any further credits; does not rely on citation blurb

## Solar physics data (J.A.Hourcle)

### Presentation

Remote observations of the sun: images

- Every 12 seconds on a wavelength: millions of records
- Different processing applied; different groups may distribute multiple variants

- Space weather data is processed for speed, not accuracy
- Problem: artefacts in images (UFO hunters beware!)

Useful in astronomy, earth science and planetary science

- Sequences of images
- Aggregation issues
- Subsetting issue: sampled vs. binned (reductions) vs. cut-outs vs. time-ranges vs. observing modes

Use case: AIA level 1:

- 56.7k images per day ~ 1 TB compressed
- Daily batches
- Data: FITS files (archived to tape), PostgreSQL catalog (journaled)
- Access: it's complicated
- Current citation approach: acknowledgement string in paper, cite the 'First results' and 'Instrument' papers
- Ideal citation: something concrete and reproducible, specific files used + timestamps
- Not universally useful: browse products are just files in a directory rsynced around

Solution: A program to generate an inventory of files....

http://dx.doi.org/10.5281/zenodo.13802

### Comments and Questions

Q: Why not do the second easiest?
A: Because they all kind of suck

Andi: Another use case: distributed dataset of VAMDC

If anybody is interested in running a workshop, please let us know.

Next steps: Report, Compact handout

Thank you for joining!

---

# Summary of WG presentations at other sessions

## Plenary (Andreas)

### Presentation

Results:

- Recommendation how to support citation of arbitrary subsets of data, even if that data is highly dynamic

- Principle can be applied across different types of storage systems

Expected impact:

- Data centers/data providers: support dynamic citation mechanisms
- Researchers: identify subset
- Also works for static data

**Pilots!**

### Questions and Comments

Q: Works only in an ideal world. What if the researcher did a number of queries and does not remember which one he used; or if he uses only a subset of the queried data in his work?
A: Want to provide a mechanism, but we can't make up for errors on the researcher's side. Make it as easy as possible; does not replace good academic practice.

## IG Data Sharing (Ari Asmi)

### Questions and Comments

Comment: Data centers are not equipped for that kind of system at the moment. But it's a great thing for large datasets (related to citing parts of an article). There is a need for finer granularity - but it is still a long way.

Q: Citation as a pointer vs. credit -> two roles of citations that have different requirements regarding granularity. Do you recognize both roles?
A: Two numbers: DOI for whole dataset / data center, and PID for dataset to get the reproducible data

Q: With some of the data centers we have been trying to do that -> none of the datasets stay the same. But the query does not replicate what is used in the publication.
A: Step by step implementation.