

# Working Group on Data Citation: Making Dynamic Data Citeable Andreas Rauber, Ari Asmi, Dieter van Uytvanck, Stefan Pröl

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**



- 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



Research Data Sharing

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

Recommensations of the Working Group on Data Charles 6665C Andreas Rauber, Art North Dieter von Descare cond Stafen Fish Draft - Dansact for Commands

#### 1. Macore Data Crosper

There WOD'S minigrated traces make assumption and date realist. Sectoral of providing rates days organic or topics. Securificate of this extension we expect a figurest, participate there of that new The provided mixings, carbles species identification of the very swinger treasure of time evel.

#### the of the William to said distributes

- affices to an absorbt out title account tienes of data finance single count time, write form for its at machine, modified-
- about as to the had report that the re-y conducts sent in past or time, or ense the Bridges is experiented by Contract 6 Add com dalmas adaption on

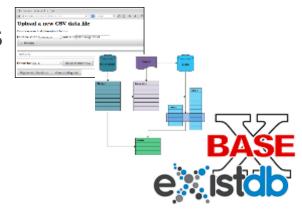
#### white it we present your is

- and the stand of married shapping our on to stone an enteres morning theretal to the terms of t patential principle in the sealers speaked
- is Withoutper-species. To regime the send of managing artistacy data sets desable. from legal value to control 20, in lessing the raise from the first operation of the same in the excellent the first operation.

- A Departy or Benevator Gare Star.
- 11. But through approximation in more contract. In the characteristics.
   32 Teachington Learn the spiritual in fail.
- no knowlegit, i.e. air iddress, cidolius no iaideorrika knowleg
- Ist i type y more into the same to more the species and transport describes a variety or and the
- constructive and specification of

When all or dealers product to fill any upgreat to be necked.

- RE Come Langueuro de vetir la gera a secrativa de forma en der relevant operar ou bestelle de la companya del companya del companya de la companya del companya del companya de la companya del companya del companya de la companya del companya del companya de la companya de la companya de la companya de la companya del companya del compan
- of the second series have been been series
- to the distribution to appear a reliant of the court could not be suble touchable of the communical country appearance and
- \$7 Chery Theoreuping steep a country in the carry based on the fell update in the other declaration for the first quittee to the consecut of Just
- so your ten and a recommend of the part of other the congruency or other medium manned from to Jun Charuse from the energy PD.
- 85 Nor Gang Man per an remain y AD edges to complete our grant a rest or declare, reasons, report AD, and in declarity of the professors are
- . Will Finale that from the management has miretie Hits force
- di una vanco de est
- \* All . Suprier Play State the PEDs proces to a hause metrics bodies note of the dra net that provides metalom actuates a late to the reporter PID of the data senses and interse met sugger





## **WG Pilots**

- Pilots and implementations by
  - LNEC: Critical Infrastructure Monitoring System
  - Virtual Atomic and Molecular Data Centre



- ata Centres) NERC SCIENCE OF THE
- 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:
     <a href="http://europe.rd-alliance.org/rda-europe-call-collaboration-projects">http://europe.rd-alliance.org/rda-europe-call-collaboration-projects</a>
  - Revise / enhance recommendations
  - Tackle some open issues and more challenging settings (LOD, distribution, generalized views, permissions, ...)

