MAKING DYNAMIC DATA CITABLE: APPROACHES TO **DATA CITATION** WITHIN AS A **RDA WORKING GROUP**

RDA DATA CITATION WG CHAIRS:

Andreas Rauber, Vienna U. of Technology
Ari Asmi, U. of Helsinki
Dieter Van Uitvanck, CLARIN ERIC

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

RDA
RESEARCH DATA ALLIANCE

OUTLINE

SCIENTIFIC METHOD AND DATA, DATA CITATION PRINCIPLES

RDA WORKING GROUPS

WORKING GROUP ON DATA CITATION OF DYNAMIC DATA

- WHY ARE CURRENT SOLUTIONS INSUFFICIENT?
- HOW TO MAKE DATA CITABLE?
  - SOME INITIAL CASE STUDIES
  - WHAT ARE THE NEXT STEPS?

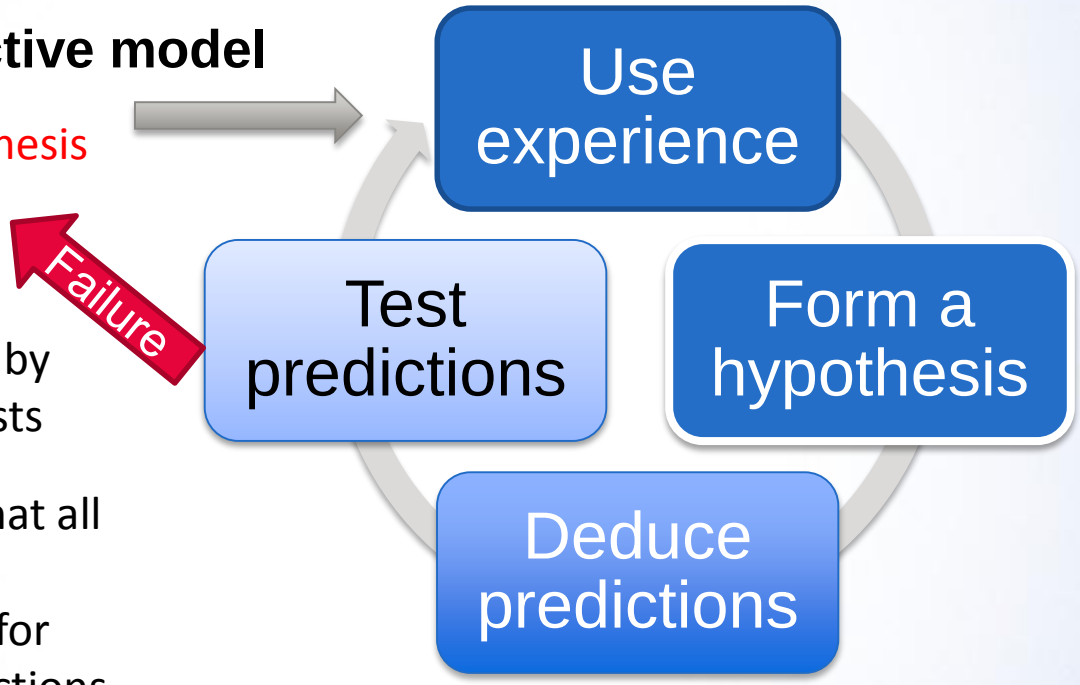
SUMMARY

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

SCIENCE IS BASED ON EVIDENCE

- **Hypothetico-deductive model**



This part is usually done by other scientists

It is critical that all **necessary information** for testing predictions are available!

Thus we need to know what data is used -> Data citation

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

PRINCIPLES OF DATA CITATION

FORCE11
The Future of Research Communications and e-Scholarship

- Data is key enabler
  - Evidence in empirical studies
  - Basis for comparing models, approaches
  - Re-use (investment)
  - Aggregation (meta-studies)
- **Joint Declaration of Data Citation Principles**

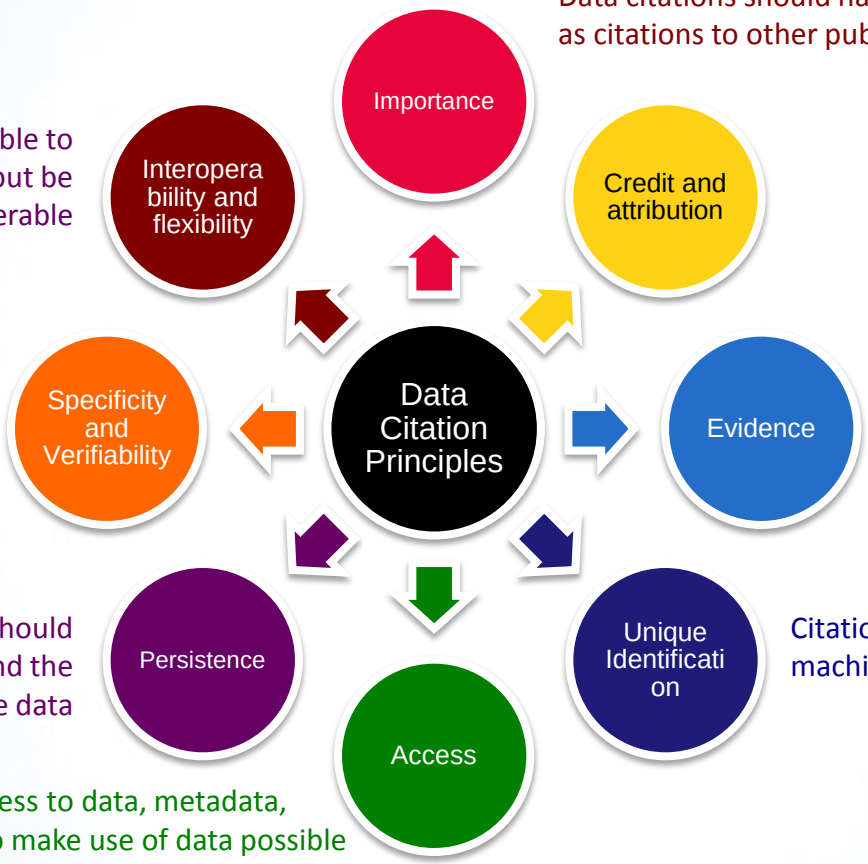
Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

DATA CITATION PRINCIPLES

FORCE11
The Future of Research Communications and e-Scholarship

(Paraphrased)



Data citations should have the **same importance** as citations to other publications

Scholarly credit and normative legal attribution

When claim relies upon data, the data must be cited

Citation must have persistent, machine readable identification

Citations should facilitate access to data, metadata, documentation, code, etc., to make use of data possible

Unique identifiers should persistent – even beyond the lifetime of the data

Citations should facilitate access, identification, and verification of data

Citation must be flexible to different practices, but be interoperable

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

https://www.force11.org/datacitation

-> DATA CITATION IS A COMPLEX ISSUE

SCIENTIFIC METHOD AND DATA, DATA CITATION PRINCIPLES

RDA WORKING GROUPS

WORKING GROUP ON DATA CITATION OF DYNAMIC DATA

- WHY ARE CURRENT SOLUTIONS INSUFFICIENT?
- HOW TO MAKE DATA CITABLE?
  - SOME INITIAL CASE STUDIES
  - WHAT ARE THE NEXT STEPS?

SUMMARY

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

RESEARCH DATA ALLIANCE

RDA Working Groups

- Provide case statement
- **18 months, "picking low-hanging fruit"**
- **Concrete solutions**
  - Data Citation WG
  - Data Description Registry Interoperability
  - Data Foundation and Terminology WG
  - Data Type Registries WG
  - Metadata Standards Directory Working Group
  - PID Information Types WG
  - Practical Policy WG
  - Standardisation of Data Categories and Codes WG
  - Wheat Data Interoperability WG

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

ONLY 18 MONTHS – PRACTICAL RESULT

- Scientific method and data, Data citation principles
- RDA Working Groups
- Working Group on Data Citation of Dynamic Data
  - Why are current solutions insufficient?
  - How to make data citable?
    - Some initial case studies
    - What are the next steps?
- Summary

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

FOCUS: SPECIFICITY AND VERIFIABILITY

FULL TEXT:

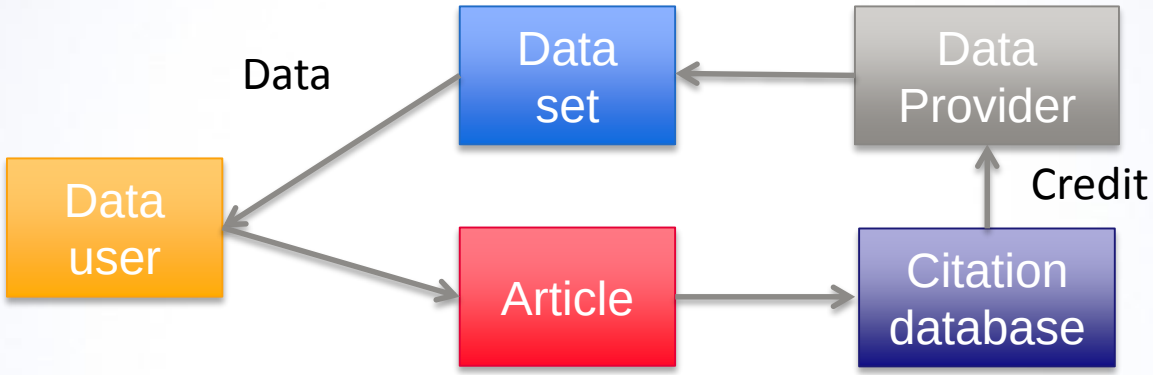
Data citations should facilitate **identification of, access to, and verification of the specific data** that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the **specific time slice, version and/or granular portion of data retrieved** subsequently is the same as was originally cited

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

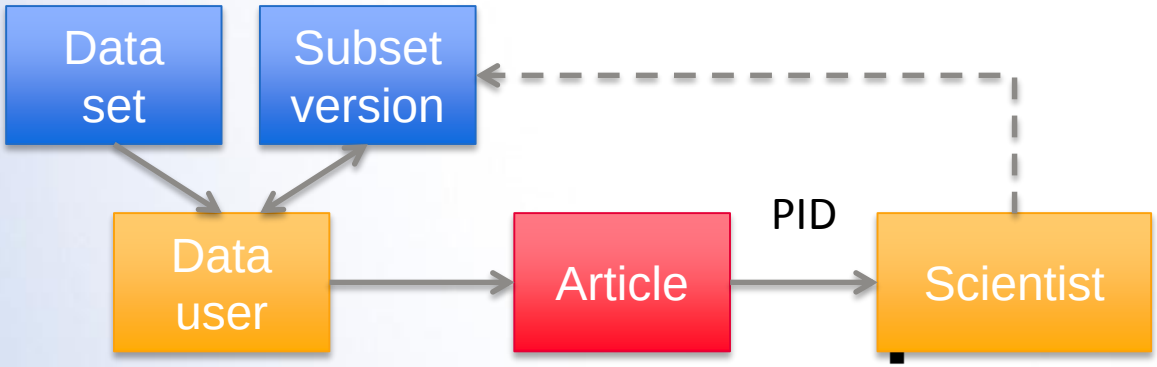
UNIVERSITY OF HELSINKI

DUAL NATURE OF CITATION

- Citations are used to provide credit for the information originator



- They should also provide access to the original object (data in this case)



Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

DATA CITATION – CURRENT APPROACHES

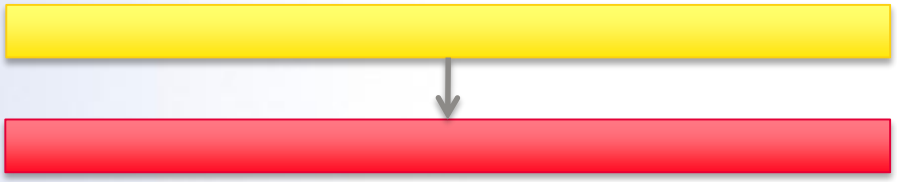
- Persistent Identifier (PID) e.g. DOI, URI, ARK, …
  currently provided for
  - **entire data sets**, copies of subsets
  - **static data**, sometimes releases of versions
  - cited in their entirety with **textual description of subsetting**
- This is insufficient in many settings
  - **not machine-actionable**
  - **not scalable** for large data sets
  - insufficient support for **data that changes**
  - insufficient support for **arbitrary subsets** (rows/columns)

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

.. AS A GRAPH

Arbitrary subsetting



Changing datasets

Growing datasets

These are surprisingly common in e.g. Earth System Sciences and many social fields

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

DATA CITATION - REQUIREMENTS

Need means to support citation of

- arbitrary subsets of data
  - (rows/columns, time sequences, …)
- when data is changing
  - (corrections, additions,…)
- stable across technology changes
  - (e.g. migration to new database)
- machine-actionable
  - (not just machine-readable, definitely not just human-readable and interpretable)
- scalable to very large datasets

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

… HOW TO IMPROVE THE SITUATION ?

SCIENTIFIC METHOD AND DATA, DATA CITATION PRINCIPLES

RDA WORKING GROUPS

WORKING GROUP ON DATA CITATION OF DYNAMIC DATA

WHY ARE CURRENT SOLUTIONS INSUFFICIENT?

HOW TO MAKE DATA CITABLE?

SOME INITIAL CASE STUDIES

WHAT ARE THE NEXT STEPS?

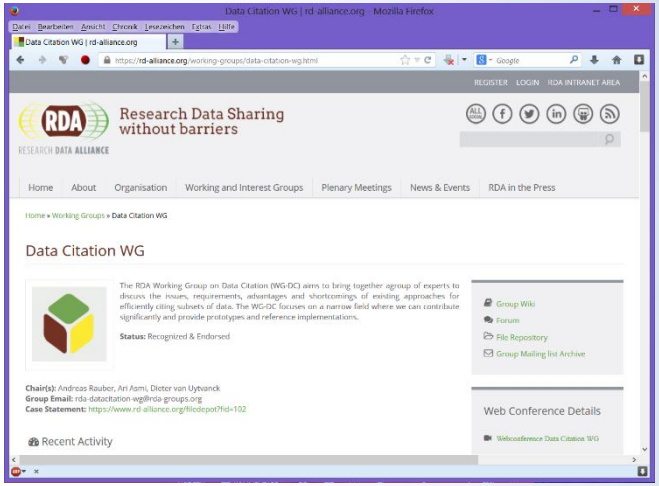
SUMMARY

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

RDA WG DATA CITATION

- WG officially endorsed in March 2014
  - Concentrating on the problems of **dynamic (changing) datasets**
  - Focus!
  - Liaise with other WGs on attribution, metadata, …
  - Liaise with other initiatives on data citation (CODATA, DataCite, Force11, …)
- Cooperation
  - periodic WG teleconferences
  - meetings every 6 months at RDA plenaries
  - special workshops on specific pilots



Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

RDA WG DATA CITATION

Approach

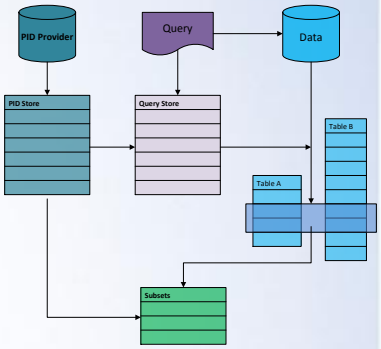
- Conceptual solution devised in initial discussions
- Identifying pilot settings:
  - variety in domains
  - variety in types of data (SQL, XML, CSV, RDF, nosql, …)
  - variety in dynamics, size, usage
- Pilot data centres to test the approach
- Starting with conceptual evaluation
studying fitness, impact, scalability, changes required, …
- Followed by actual pilot implementation

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

DATA CITATION APPROACH

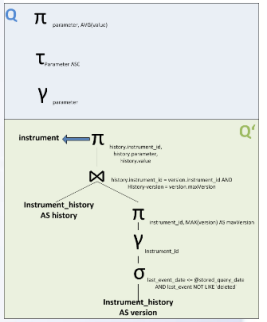
Data Citation: Data + means-of-access



1.Data → time-stamped & versioned

2.Access → **query assigned a PID**, enhanced with

- 2.1 time-stamp: for re-execution against versioned DB
- 2.3 unique-sort: processes are sequence-based, data stores mostly set-based
- 2.3 result-set hash: verifying identity/correctness



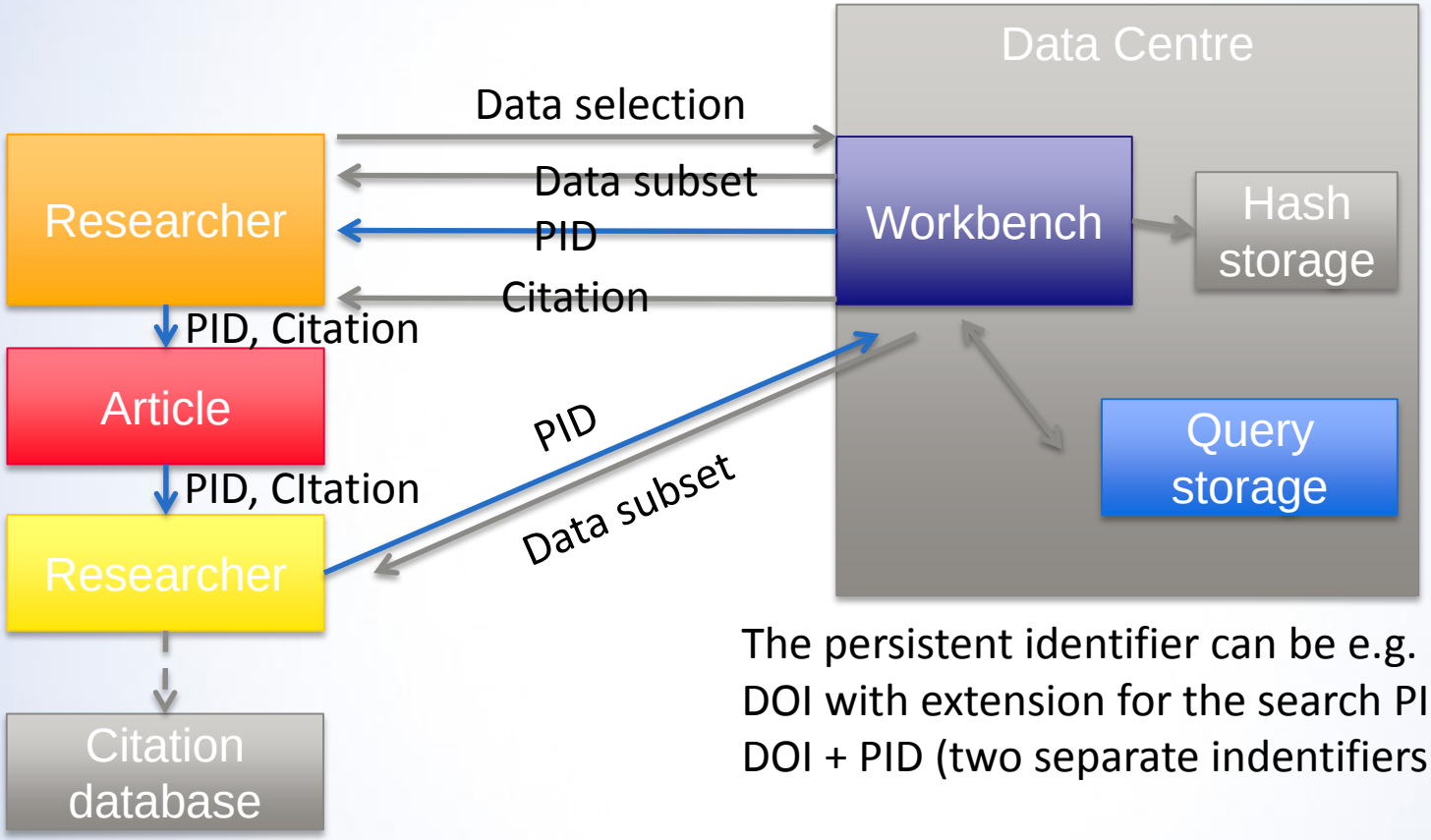
•Stable across data source migrations (e.g. diff. DBMS), scalable, machine-actionable

Stefan Pröll and Andreas Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation.** In IEEE International Conference on Big Data 2013 (IEEE BigData 2013), October 6-9 2013, Santa Clara, CA, USA. 2013. IEEE.
http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

Data Citation – Deployment



The persistent identifier can be e.g.
DOI with extension for the search PID
DOI + PID (two separate indentifiers)

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

… FIRST STEPS TOWARDS IMPLEMENTATION

- Scientific method and data, Data citation principles
- RDA Working Groups
- Working Group on Data Citation of Dynamic Data
  - Why are current solutions insufficient?
  - How to make data citable?
    - Some initial case studies
    - What are the next steps?
- Summary

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

Data Citation – Pilots

- LNEC: Infrastructure Sensor Network (dams, bridges)
- VAMDC - Atomic and Molecular Data
- SCOR/IODE/MBLWHOI Library Data Publication Project: Oceanographic datasets
- Million Song Dataset: Benchmark collection(s) in music IR
- Earth System Grid Federation:
  Earth system modelling data, CMIP experiments, netcdf
- Field Linguistics:
  language archive: transcriptions

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

Data Citation – Pilots

- Global Biodiversity Information Facility:
  species occurrence records
- DataNet Federation Consortium:
  Hydrology, Oceanography, Social Science, plant Biology, Engineering, Cognitive Science
- NASA MODIS Data: remote sensing satellite data
- NERC ECN citing dynamic monitoring data
  Long-term environmental monitoring from automatic and manual recording across the UK
- UK Butterfly Monitoring Scheme annual species metrics
- Additional pilots joining almost on a weekly basis!
  http://rd-alliance.org/groups/data-citation-wg/wiki/collaboration-environments.html

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

- Scientific method and data, Data citation principles
- RDA Working Groups
- Working Group on Data Citation of Dynamic Data
  - Why are current solutions insufficient?
  - How to make data citable?
    - Some initial case studies
    - What are the next steps?
- Summary

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

NEXT STEPS

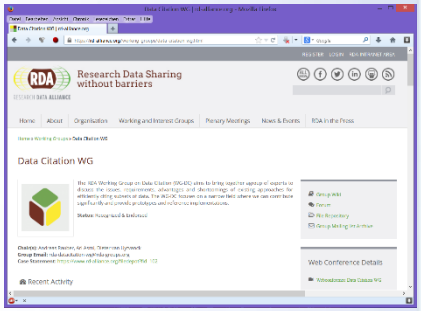
- Solution devised for SQL -> expand to other data types
  - pilot for CSV
  - analyze how to make XML and RDF time-stamped, versioned
- Verify pilots conceptually
  - does it work?
  - impact on data center (size, operations, APIs, …)
  - how to integrate in workbenches?
- Implement several pilots and verify
- Test stability under migrations of data management systems

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

JOIN RDA AND THE WORKING GROUP

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



- 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
- Description of SQL pilot solution:
  - Stefan Pröll and Andreas Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation.** In IEEE International Conference on Big Data 2013 (IEEE BigData 2013),Santa Clara, CA, USA. 2013. http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

SUMMARY

- Data and process re-use as basis for data driven science
  - evidence
  - investment
  - efficiency
- Machine-readable and –actionable data citation in large-scale and dynamic environments
- Database: versioned and time-stamped
- PIDs assigned to time-stamped “queries”
- Need to move beyond concept of data
- Process Management Plans (PMPs)
- Join the RDA Working Group & Discussion!

Dr. Ari Asmi
Research Coordinator
Faculty of Science
Department of Physics

UNIVERSITY OF HELSINKI

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