



RESEARCH DATA ALLIANCE

**Data Citation
Working Group Mtg @ P7
March 2nd 2016, Tokyo**

research data sharing without barriers
rd-alliance.org

Agenda

2

- 11:00 - Welcome and Intro
- 11:10 – Recommendations: Re-cap, Flyer, Report
- 11:30 - Recommendations Q&A
- 11:45 - Adoption activities
- 12:15 - Future plans & new roles

Welcome!

to the adoption meeting
of the
WGDC

4

- [illegible]

Data Citation – Next steps

5

- From now on: supporting adoption!
 - Finalize detailed report
 - Support reference implementations
 - Collect and publish feedback, lessons learned, ...
 - Address any upcoming issues

- Follow-up activities
 - Details to be discussed today
 - Help with adoption !!
(what do you need?)
 - Ambassadors / Adoption roles: who wants to join?

Agenda

6

- 11:00 - Welcome and Intro
- 11:10 – Recommendations: Re-cap, Flyer, Report
- 11:30 - Recommendations Q&A
- 11:45 - Adoption activities
- 12:15 - Future plans & new roles

Granularity of Data Identification

7

- What about the **granularity** of data to be identified?
 - Databases collect enormous amounts of data over time
 - Researchers use specific subsets of data
 - Need to identify precisely the subset used
- Current approaches
 - 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 -> scalability, not for arbitrary subsets (e.g. when not entire record selected)
- Would like to be able to identify precisely the **subset of (dynamic) data used** in a study

Identification of Dynamic Data

8

- 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
 - “Artificial” versioning by identifying batches of data (e.g. annual), aggregating changes into releases (time-delayed!)
- Would like to identify precisely the **data as it existed at any point in time**

RDA WG Data Citation



- Research Data Alliance
- WG on **Data Citation: Making Dynamic Data Citeable**
- Ran March 2014 – Sep 2015
 - Concentrating on the problems of **large, dynamic (changing) datasets**
 - Focus! Identification of data!
Not: PID systems, metadata, citation string, attribution, ...
 - Liaise with other WGs and initiatives on data citation (CODATA, DataCite, Force11, ...)
 - Continuing support for adoption



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

Making Dynamic Data Citeable



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

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
- Query store aggregates data usage

Data Citation – Deployment



- Note: query string provides excellent provenance information on the data set!
- 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
- Query store aggregates data usage

Data Citation – Deployment



- Note: query string provides excellent provenance information on the data set!
- subset of data
- user gets

- Data (package)
- PID (e.g. doi)
- Hash value
- Recommended citation text (e.g. doi:10.1111/1365-3113.12111)

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

- 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
- Query store aggregates data usage

Data Citation – Deployment



- Note: query string provides excellent subset of data
- provenance information on the data set! per gets

- Data (package)
- PID (e.g. DOI)
- Hash value
- Recommended citation text (e.g. PID text)

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

- PID resolves
 - Provides details
 - Option to retrieve
- Identify which parts of the data are used. If data changes, identify which queries (studies) are affected
- Upon activating PID associated with a data citation
 - Query is re-executed against time-stamped and versioned DB
 - Results as above are returned
- Query store aggregates data usage

Data Citation – Recommendations

15

Preparing Data & Query Store

- R1 – Data Versioning
- R2 – Timestamping
- R3 – Query Store

When Resolving a PID

- R11 – Landing Page
- R12 – Machine Actionability

When Data should be persisted

- R4 – Query Uniqueness
- R5 – Stable Sorting
- R6 – Result Set Verification
- R7 – Query Timestamping
- R8 – Query PID
- R9 – Store Query
- R10 – Citation Text

Upon Modifications to the Data Infrastructure

- R13 – Technology Migration
- R14 – Migration Verification



Agenda

16

- 11:00 - Welcome and Intro
- 11:10 – Recommendations: Re-cap, Flyer, Report
- 11:30 - Recommendations Q&A
- 11:45 - Adoption activities
- 12:15 - Future plans

WG Pilots

17

Name	Data	Type	Status	Notes
Timbus	RDBMS	research	finished	Sensor data, pilot
XML-Reference	XML	research	finished	eXist-DB
DEXHELPP	CSV/RDBMS	research	running	Social security data
CSV-Reference	CSV/RDBMS	reference	running - β	Reference implem.
GIT-Reference	<ASCII>	reference	running - α	Reference implem.
VAMDC	SQL/NoSQL/ ASCII -> XML	deployment	running	Distributed data center
CBMI@wustl	RDBMS	deployment	starting	integration into i2b2
CCCA	NetCDF	deployment	starting	climate data
ACDH	RDBMS, LOD	deployment	starting	thesaurus
ARGO	NetCDF	deployment	planned	ODIP-II
BCO-DMO	CSV	deployment	planned	
ENVRlplus		deployment	running	
<a few others>	CSV, RDBMS	deployment	planned	Conceptual evaluation, seeking funding

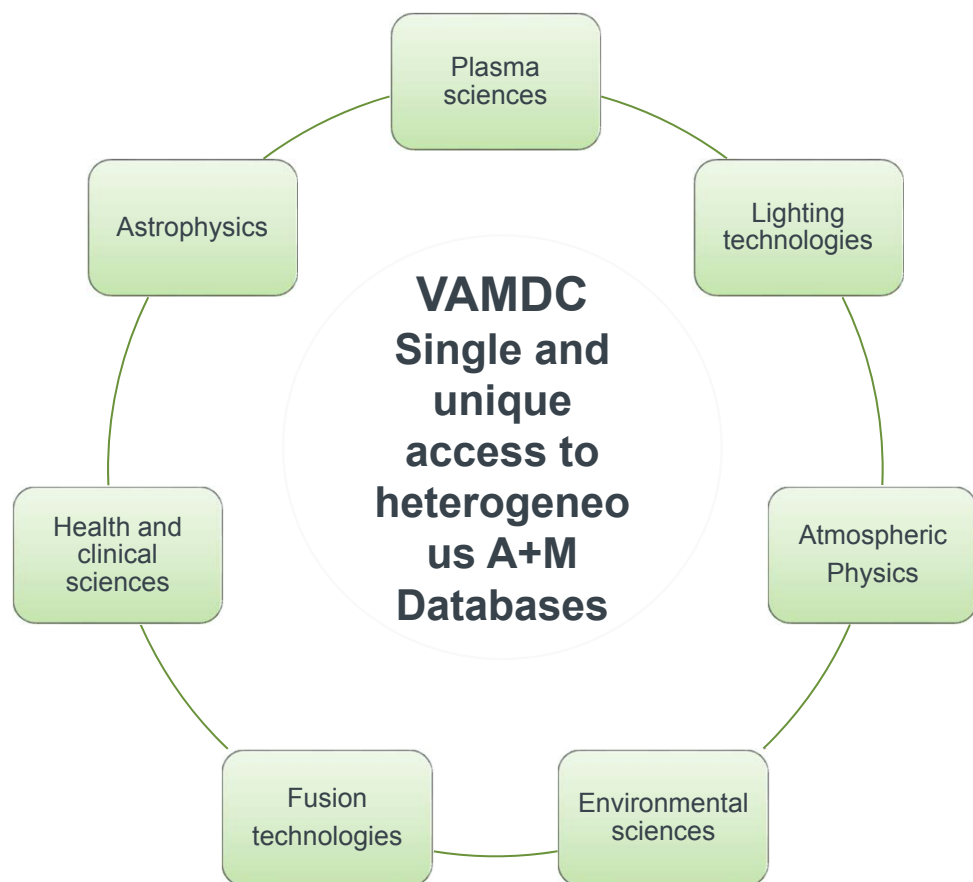


**Progress on Data Citation within
VAMDC**
C.M. Zwölf and VAMDC Consortium
carlo-maria.zwolf@obspm.fr

research data sharing without barriers
rd-alliance.org

The Virtual Atomic and Molecular Data Centre

19



- Federates 28 heterogeneous databases
<http://portal.vamdc.org/>
- Distributed infrastructure with no central management system
- The “V” of VAMDC stands for Virtual in the sense that the e-infrastructure does not contain data. The infrastructure is a wrapping for exposing in a unified way a set of heterogeneous databases.
- Relies on a strong and sustainable technical and political organisation.

We have started the implementation of the recommendations

Query Store

Tagging Datasets with Ids
(Relational Database case)

- On January 2016 we submitted a paper for a special issue of the *Journal of Molecular Spectroscopy* dedicated to *New visions in Spectroscopic Databases*
 - ***New paradigm for datasets citation and extraction reproducibility in VAMDC***, C.M. Zwölf, N. Moreau and M.-L. Dubernet.
- We describe in this paper
 - How the internal VAMDC standards evolve for achieving data-tagging and versioning
 - How a query store will be integrated within the existing VAMDC infrastructure.

We have started the implementation of the recommendations

Query Store

Tagging Datasets with Ids
(Relational Database case)

- The design of the query store is almost finished.
- The development will start on spring 2016, in joint collaboration with the RDA Europe project.
- We have started collaborations with editors for the integration of the query store into their digital platforms and journal repositories.

Concluding remarks / open questions about data citation activity

22

- How to deal with confidentiality of the information?
 - Should we have an authentication/authorization policy on the query store?
 - Is the designed log service compliant with the EU law about confidentiality?
- We are providing to users the tools for efficiently cite our dynamic data, but
 - How can we be sure that they will use it for citing our data?
 - In other words, how to enforce the 'citation instincts' in our final users?
 - *Partial answer from our collaboration with editors*
- We propose a 'reverse approach':
 - We may cite the users accessing our data.
 - They will accept these terms, that will be explained in the condition of usage of the VAMDC services.
- How to prevent plagiarism?:
 - A user might extract data, modify and cite them as the original extracted ones.
 - Do we have tools for preventing such behaviors? MD5 of extracted data on query-store?



RESEARCH DATA ALLIANCE

WG Data Citation Pilot CBMI @ WUSTL

**Leslie McIntosh, Cynthia Hudson Vitale,
Snehil Gupta**

Washington University in St.Luis

research data sharing without barriers
rd-alliance.org

- Center for Biomedical Informatics, Washington University in St. Louis
- Electronic medical health record aggregator **i2b2** (Informatics for Integrating Biology and the Bedside) NIH-funded Health Care System, OS SW
- Electronic patient medical records (EMR)
- i2b2 instance with de-identified data from local hospitals and outpatient clinics
 - Overall approx. 2 billion records
 - 4 mio patients, 48 mio encounters, 82 mio medications, 674 mio lab results, 385 mio vital sign data, ...
- Obtained funding to implement WGDC recommendations
- Timeframe: 9 months

The screenshot displays the i2b2 Query & Analysis Tool interface. The top navigation bar includes the URL <https://www.i2b2.org/webclient/>, project name 'i2b2 Demo', user 'i2b2 User', and links for 'Find Patients', 'Analysis Tools', 'Message Log', 'Help', 'Change Password', and 'Logout'. The left sidebar contains 'Navigate Terms' and 'Previous Queries'. The 'Navigate Terms' section shows a tree view of data categories, with 'Diagnoses (ICD10)' highlighted by a red line. The 'Previous Queries' section lists several queries, with 'Oasis-Express@08:19:26 [9-11-2015] [demo]' circled in red. The main 'Query Tool' section shows a query being built with three groups: Group 1 (Cardiovascular agents), Group 2 (Biologicals), and Group 3 (empty). Logical connectors 'AND' are placed between the groups. Below the query tool are buttons for 'Run Query', 'Clear', and 'Print Query'. At the bottom, there is a 'Show Query Status' and 'Graph Results' section. The 'Graph Results' section shows a bar chart titled 'Number of Patients' with 'Race' on the x-axis. The y-axis ranges from 0 to 10. The x-axis categories include Aleutian, American Indian, Asian, Asian Pacific Isld, Black, Eskimo, Hispanic, Indian, Middle Eastern, Multiracial, Native American, Navajo, Not recorded, Oriental, Other, and White. The 'Graph Results' section also shows a legend for 'Race'.

Goals of this project are to:

1. **Integrate the RDA recommendations for Data Citation of Evolving Data into the WU CBMI instance of i2b2** by modifying the open source code to account for the recommendations at the query or project level.
2. **Contribute all source code back to the i2b2 open source community via GitHub** for wider adoption and implementation among all i2b2 users (currently 50 translational science centers, 34 academic health centers, 2 HMO's, 20 international organizations, 4 companies).
3. **Gather feedback about RDA WGDC-compliant i2b2 code from established i2b2 installations**; obtain projected timelines of compliant code integration, barriers to implementation



RESEARCH DATA ALLIANCE

WG Data Citation Pilot Climate Control Centre Austria

research data sharing without barriers
rd-alliance.org

- <https://www.ccca.ac.at/de/home/>
- Newly founded data centre
- 4TB of NetCDF files
- Kick-off Workshop March/April for detailed planning
- Implementation pending approval by board, budget earmarked





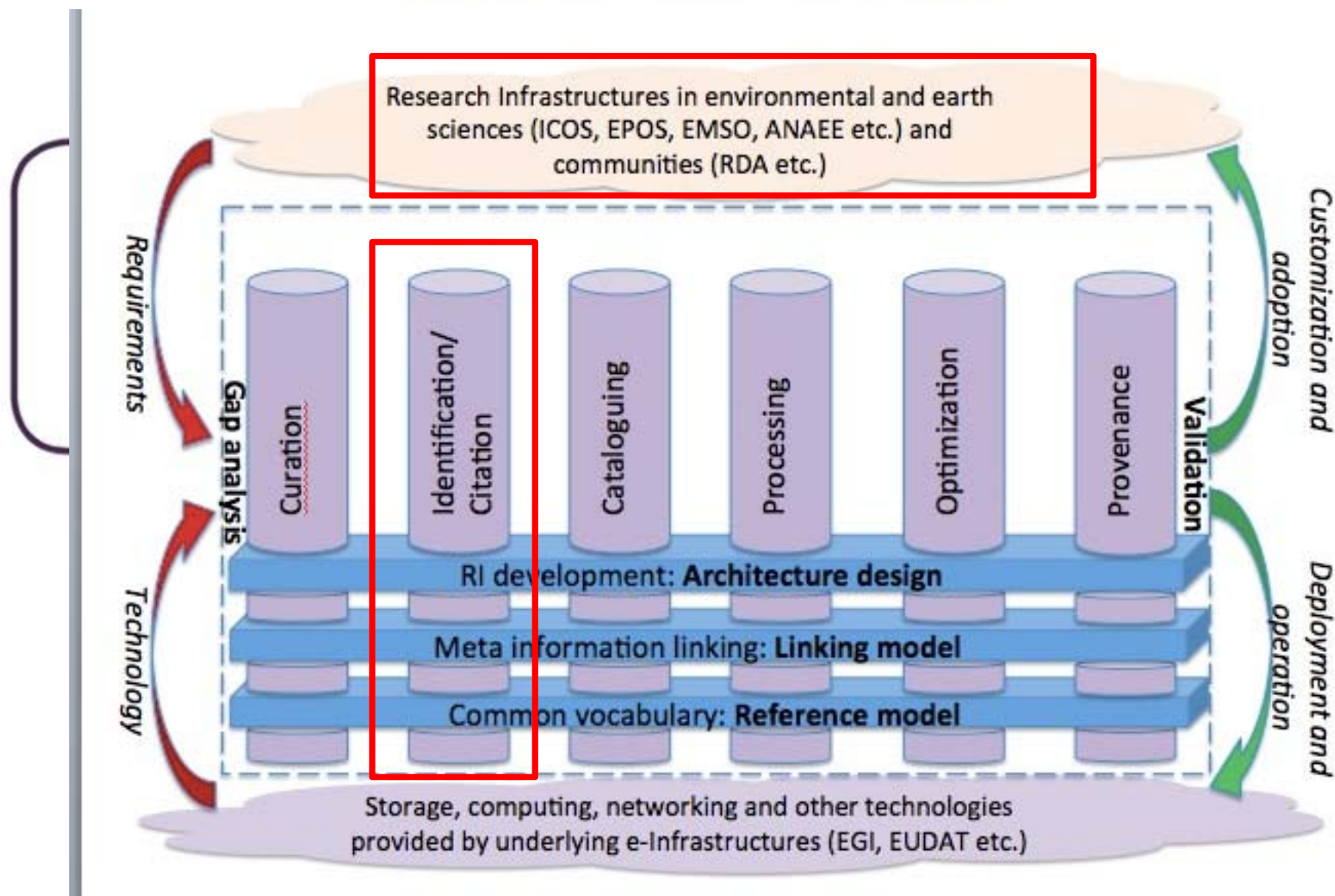
Data Citation for ENVRplus

Ari Asmi
ari.asmi@helsinki.fi

research data sharing without barriers
rd-alliance.org

ENVRI PLUS – Cluster of Environmental Research Infrastructures

30





ENVRI PLUS – ICOS DATA CITATION

32

- Part of ENVRI PLUS data citation Workpackage
- ICOS – Integrated Carbon Observation System (infrastructure)

- Distributed data production
- Distributed data storage
 - Centralized “high level” data sets
- Updated daily
- Usage wide in Carbon observation science
 - Some NRT
 - Some “high level” data storage

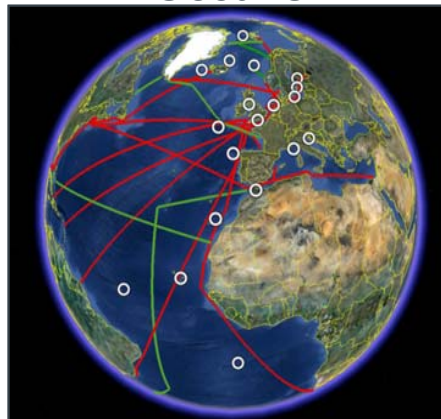
Atmosphere



Ecosystems



Oceans

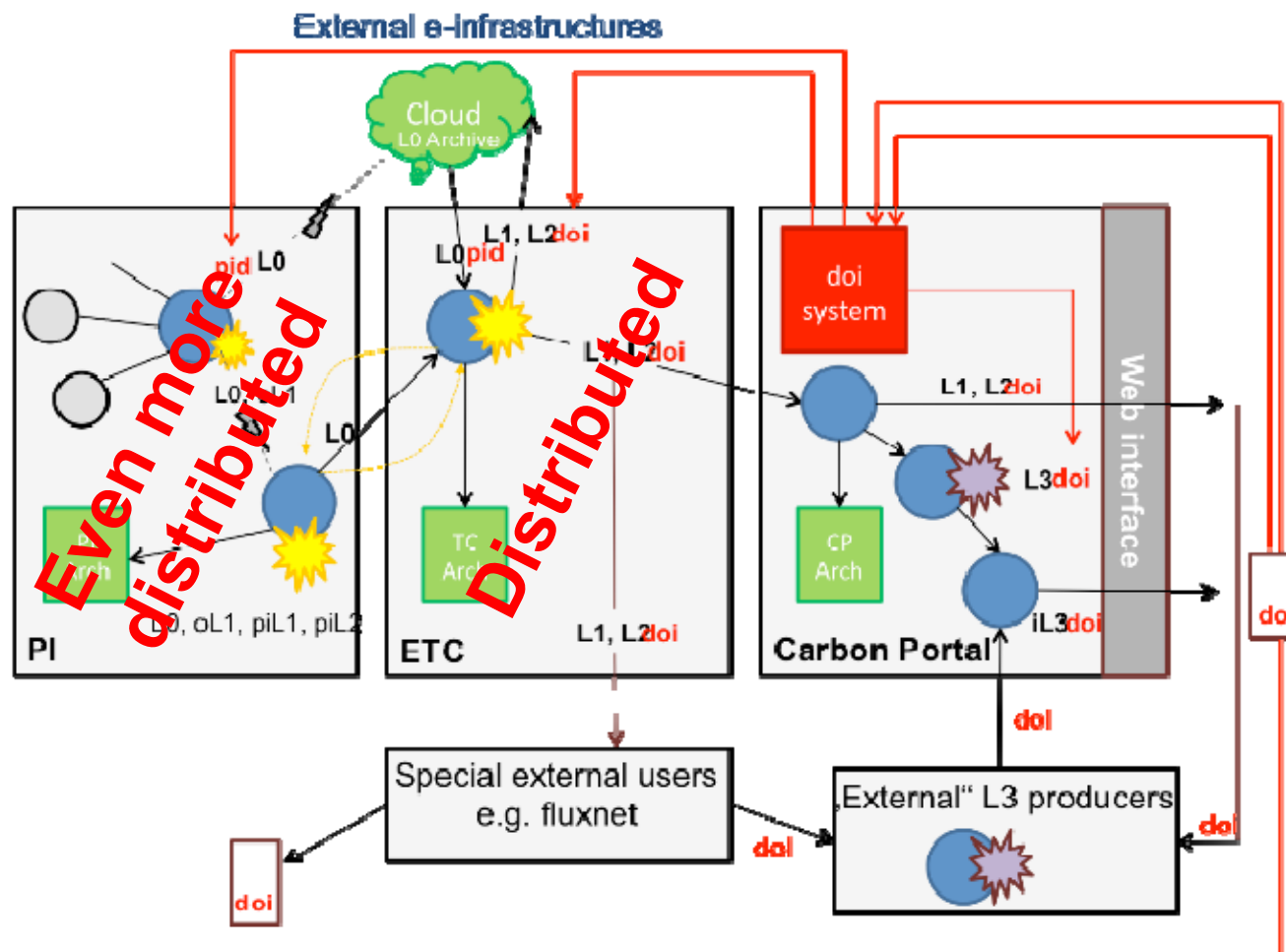


ICOS IMPLEMENTATION (IN PROGRESS)

33

Versioning DB –
OK!

Distributed
Data delivery –
Potential issue
(if users bypass
web interface)



- ICOS development has issues, but they are manageable
- However, in many other ENVRI PLUS Research Infrastructures have issues even with the versioning DB overhead
- ENVRI PLUS WP on Data Citation will be running for the next 3 years, so we will know more in the future
 - RDA recommendations will be closely investigated in the participating Research Infrastructures
 - Each case study will make the issue easier to sell to rest of the RIs...



RESEARCH DATA ALLIANCE

Data Citation for ARGO (ODIP II Project)

Helen Graves

hmg@bgs.ac.uk

research data sharing without barriers
rd-alliance.org

Adoption of DCWG outputs

- RDA Collaboration meeting, Karlsruhe: joint workshop at 6th Plenary
- Data Citation WG webinar: preparation
- RDA 6th plenary: joint workshop
- Co-chaired by:
 - Andi Rauber
 - Ari Asmi
 - Helen Graves

research data sharing without barriers
rd-alliance.org



6th Plenary (Paris) Session Objectives

- Develop a plan for the evaluation and potential adoption of the outcomes of the Data Citation WG for the citation of dynamic data by the marine community
- Identify a small number of suitable use cases
- Evaluate potential proposal for the RDA Collaboration call

- Citation of static datasets well defined
- Marine data capture and delivery is becoming increasingly dynamic
- Increasing need to be able to unambiguously cite these evolving datasets
- Repositories responsible for archiving the data must be able to identify and provide access to a specific instance of the dataset at a given point in time

RDA collaboration project proposal: Argo data use case

39

- Aims & objectives
 - Resolve the ambiguity in the syntax for citation of dynamic data
 - Agree and ratify a common syntax for dynamic data citation
 - Publish results in authoritative documentation e.g. DataCite metadata schema
 - Implement dynamic data citation for Argo data



Use case: Argo Data

- Apply the 14 DCWG recommendations to legacy Argo data architecture
- Validate proposed methodology for real-world exemplar
- Provide template for adoption of the proposed solution by other disciplines

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

Revision of September 24th 2015

I. MAKING DATA CITABLE

These WGDC recommendations enable researchers and data centers to identify and cite data used in experiments and studies. Instead of providing static data exports or textual descriptions of data subsets, we support a dynamic, query-centric view of data sets. The proposed solution enables precise identification of the very subset and version of data used, supporting reproducibility of processes, sharing and reuse of data.

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends solving this challenge by:

- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

II. WG RECOMMENDATIONS

To realise the goal of rendering arbitrary data sets citable, from single values to entire DBs in settings that range from static data to highly dynamic data streams, the WG recommends the following steps:

A. Preparing the Data and the Query Store

Prepare existing data sources and provide the required infrastructure, which is needed for implementing the query-based approach.

- **R1 – Data Versioning:** Apply versioning to ensure earlier states of data sets can be retrieved.
- **R2 – Timestamping:** Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- **R3 – Query Store Facilities:** Provide means for storing queries and the associated metadata in order to re-execute them in the future.

B. Persistently Identify Specific Data Sets

When a data set should be persisted, the following steps need to be applied:

- **R4 – Query Uniqueness:** Re-write the query to a normalised form so that identical queries can be detected. Compute a checksum of the normalized query to efficiently detect identical queries.
- **R5 – Stable Sorting:** Ensure that the sorting of the records in the data set is unambiguous and reproducible
- **R6 – Result Set Verification:** Compute fixity information (checksum) of the query result set to enable verification of the correctness of a result upon re-execution.
- **R7 – Query Timestamping:** Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time). This allows retrieving the data as it existed at the time a user issued a query.
- **R8 – Query PID:** Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID.
- **R9 – Store Query:** Store query and metadata (e.g. PID, original and normalized query, query & result set checksum, timestamp, superset PID, data set description, and other) in the query store.
- **R10 – Automated Citation Texts:** Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing the data. Include the PID into the citation text snippet.

- Argo data
 - Held by several international data centres
 - IFREMER
 - NCEI (formerly NOAA National Climatic Data Center, the National Geophysical Data Center, and the National Oceanographic Data Center)
 - BODC

- Validation of method using a real world exemplar
- Results reported to RDA via DCWG and MDH IG
- Feed into related activities in ODIP, ENVRIplus, EUDAT etc.



Adoption of Data Citation Outcomes by BCO-DMO

Cynthia Chandler, Adam Shepherd

research data sharing without barriers
rd-alliance.org

US Ocean Science Domain Repositories

44

- BCO-DMO
 - Biological and Chemical Oceanography Data Management Office (WHOI)
 - Curation of marine ecosystem system data contributed by NSF funded investigators
- R2R
 - Rolling Deck to Repository
 - Curation of routine, underway data from US academic fleet, and authoritative expedition catalog
- Members of Marine Data Harmonization IG

BCO-DMO Adoption of Data Citation Outputs

45

■ Evaluation

- Evaluate recommendations
- Try implementation in existing systems

■ Trial

- BCO-DMO: R1-11 fit well with current architecture; R12 doable; test as part of DataONE node membership
- R2R: curation of original field data and selected subset of post-field products (ship track); so no evolving data

RDA Recommendations for Data Citation of Evolving Data

R1 Data Versioning
R2 Time Stamping
R3 Query Store
R4 Query Verification
R5 Stable Sorting
R6 Result Set Verification
R7 Query Time Stamping
R8 Query PID
R9 Citation Text
R10 Landing Page
R11 Machine Actionability
R12 Technology Migration
R13 Migration Verification

BCO-DMO Dataset Landing Page

46

Dataset: zooplankton biomass

[Get Data](#)

[Map It](#)

Project: [CARIACO Ocean Time-Series Program](#) (CARIACO)

Principal Investigator: [Jaimie Rojas](#) (Estacion de Investigaciones Marinas de Margarita, EDIMAR-FLASA)

Contact: [Laura Lorenzoni](#) (University of South Florida, USF)

BCO-DMO Data Manager: [Theresa McKee](#) (Woods Hole Oceanographic Institution, WHOI BCO-DMO)

Validated: Yes

Data version: 8 July 2009

Version Date: 07/08/2009

Data URL: <http://www.bco-dmo.org/dataset/3149/data> [↗](#)

Current State: Final with updates expected

[Expand/Collapse All](#)

▼ Description

Brief Description: Zooplankton biomass data from CARIACO time series stations

Zooplankton sampling is done during each CARIACO time-series cruise, and BONGO net tow samples are analyzed to determine dry weight (biomass), and ash (inorganic remains).

CARIACO zooplankton data subset, since 2000

47

/OCB/CARIACO/Zooplankton.html?Date>20000101,
Cruise_ID,lon,lat,Date,zoop_DW_200,zoop_ash_200,
zoop_DW_500,zoop_ash_500

Directory

Documentation

Download & Other Operations

Level 0

Next Level

Flat Listing

```
# version 08 July 2009
#
# PI: Jaimie Rojas (EDIMAR/FLASA)
#
# CARIACO Zooplankton biomass data
#
```

```
=====
Cruise_ID  lon      lat      Date      zoop_DW_200  zoop_ash_200  zoop_DW_500  zoop_ash_500
-----
93HG_071   -64.67  10.5    20011009   26.83        6.84          15.95         3.31
93HG_072   -64.67  10.5    20011106   7.43         0.64          2.34          0.07
93HG_073   -64.67  10.5    20011211   5.60         0.14          0.71          0.18
93HG_074   -64.67  10.5    20020110   10.24        0.51          4.93          0.18
93HG_075   -64.67  10.5    20020214   12.92        2.56          nd            nd
93HG_076   -64.67  10.5    20020312   10.55        1.78          5.47          0.77
93HG_077   -64.67  10.5    20020402   8.71         1.82          8.31          2.15
93HG_078   -64.67  10.5    20020507   4.50         0.43          1.31          0.09
93HG_079   -64.67  10.5    20020612   10.63        1.07          4.54          0.35
```

New capability

48

- Preserve the data subset
- Request a DOI
- Store data subset, query, and create new landing page for data subset DOI

Agenda

49

- 11:00 - Welcome and Intro
- 11:10 – Recommendations: Re-cap, Flyer, Report
- 11:30 - Recommendations Q&A
- 11:45 - Adoption activities
- 12:15 - Future plans

- Support in adoption: what kind of support is needed?
(in the end it all boils down to money, but apart from this...)
 - Webinars: generic
 - Focused workshops for individual pilots
 - Joint projects: proposals, ...
- How could we organize this?
 - New roles to join the WG coordinators
 - Ambassadors
 - Domain-specific contact points
 - Others? Links to supporting infrastructure (PIDs, ...)
 - Collection of on-going pilots
 - Structuring: contact, descriptions, results, lessons learned
 - Outcomes: reports, slides, publications, code, discussions

- Joint projects:
 - Individual funding per data center
 - RDA call for Collaboration Projects (EU, US)
 - H2020 call: *EINFRA-21-2017: Platform driven e-infrastructure innovation*” (and particularly the item *1-Universal discoverability of data objects and provenance*)
 - Others?
- Other needs & means of support?
 - No new WG at the moment, adoption phase, lessons to learn
 - Any crucial open topics?
 - Citation texts, attribution, ...
 - New WG if new issues arise leading to revision of recommendations or addressing new topics

- Any other business? Issues? Whishes?

Thanks!
And hope to see you at the
next meeting
of the
WGDC