



RESEARCH DATA ALLIANCE

**Data Citation
Working Group Mtg @ P8
Sep 16th 2016, Denver**

research data sharing without barriers
rd-alliance.org

Agenda

2

- 11:00 - Welcome and Intro, Recap
- **11:10 – Report from Pilots: Adoption activities**
- 12:20 – Q&A, Future plans

Welcome!

to the adoption meeting
of the

WGDC

(in Maintenance Mode: “MGDC”)

WGDC Goals and Solution

4

- Would like to be able to identify precisely the **subset of (dynamic) data used** in a study
- Would like to identify precisely the **data as it existed at any point in time**

Data Citation: Data + Means-of-access

- Data → time-stamped & versioned
- Access → **assign PID to QUERY + timestamp**
- **Additional benefits:**
 - Query string provides precise provenance info on subset
 - Allows obtaining original data or current view of data

Data Citation – Recommendations

6

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



- 11:00 - Welcome and Intro, Recap
- **11:10 – Report from Pilots: Adoption activities**
 - Stefan Pröll: CSV reference implementation (presented by AR)
 - Leslie McIntosh (WUSTL / i2b2)
 - Cynthia Chandler (WHOI / BCO-DMO)
 - James Duncan (UVM / VMC)
 - Justin Buck (BODC / ARGO)
 - Carlo Maria Zwölf (OBSPM / VAMDC)
 - Matthew Fry (CEH / Riverflow Archive)
 - Nicholas Car (Australian Dynamic Data Citation WG)
 - Martin Fenner (DataCite)
- 12:20 – Q&A, Future plans



RESEARCH DATA ALLIANCE

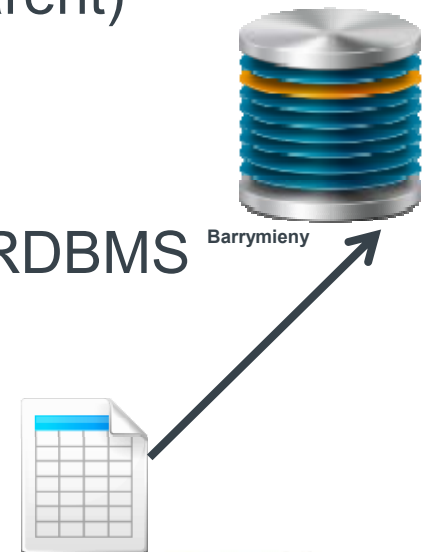
Reference Implementation for CSV Data (and SQL) Stefan Pröll, SBA

research data sharing without barriers
rd-alliance.org

CSV/SQL Reference Implementation 1

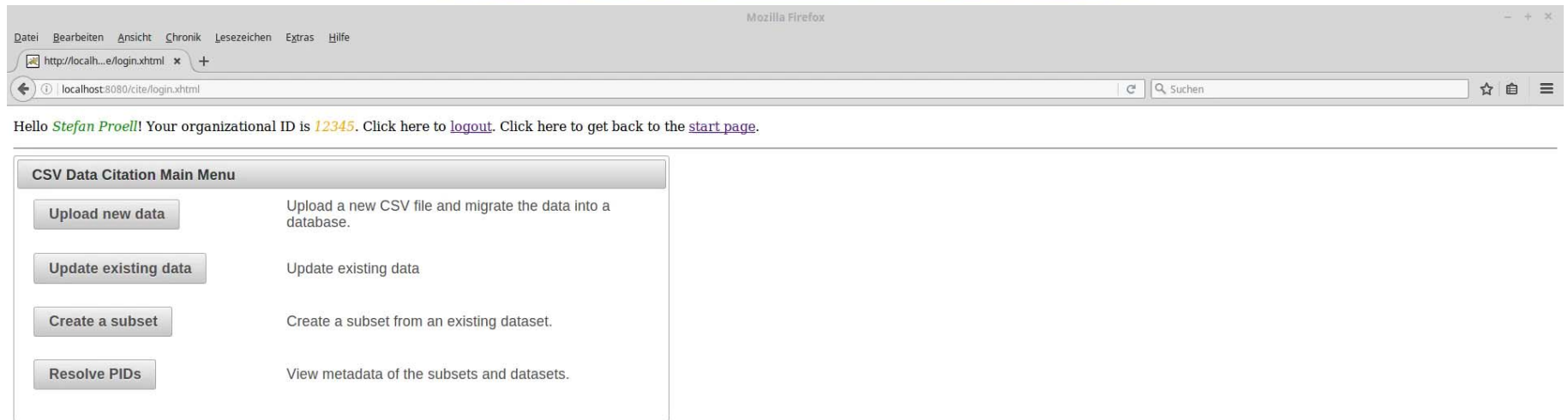
9

- Reference Implementation available on Github
<https://github.com/datascience/RDA-WGDC-CSV-Data-Citation-Prototype>
- Upload interface --> Upload CSV files
- Migrate CSV file into RDBMS
 - Generate table structure, identify primary key
 - Add metadata columns for versioning (transparent)
 - Add indices (transparent)
- Dynamic data: upload new version of file
 - Versioned update / delete existing records in RDBMS
- Access interface
 - Track subset creation
 - Store queries -> PID + Landing Page



CSV Data Prototype

10



CSV Data Prototype

11

Mozilla Firefox

http://localhost:8080/cite/menu.xhtml

Suchen

Hello **Stefan Proell**! Your organizational ID is **12345**. Click here to [logout](#). Click here to get back to the [start page](#).

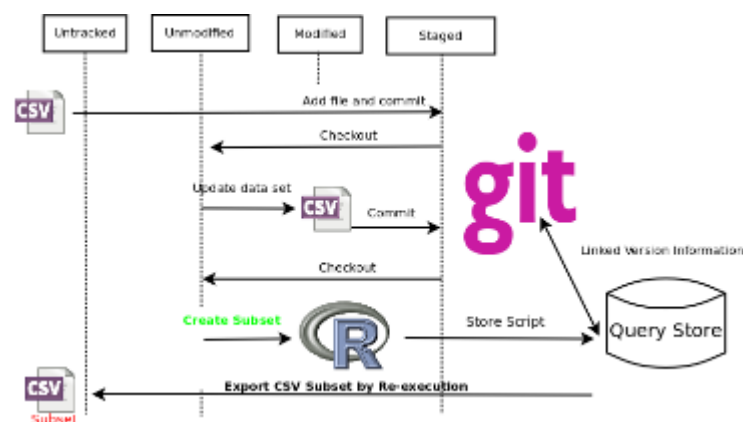
CSV Data Citation Main Menu

Upload new data	Upload a new CSV file and migrate the data into a database.
Update existing data	Update existing data
Create a subset	Create a subset from an existing dataset.
Resolve PIDs	View metadata of the subsets and datasets.

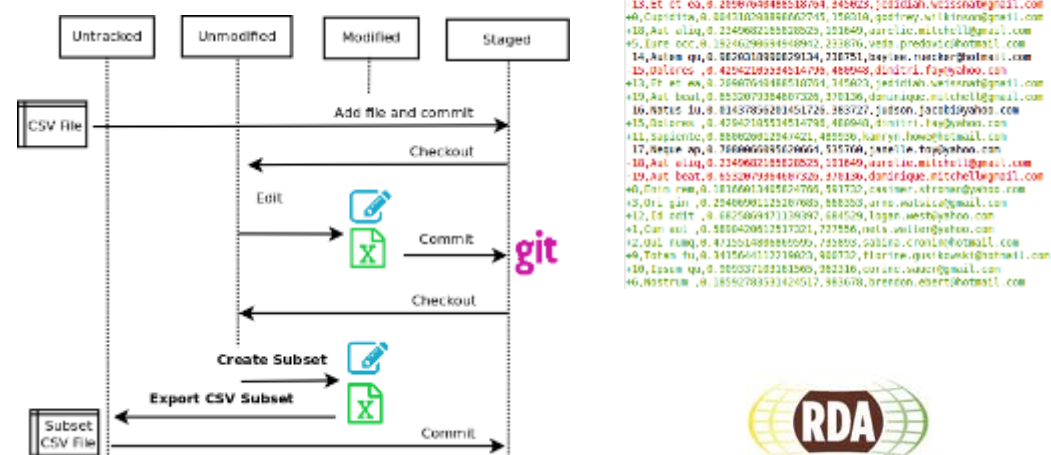
CSV Reference Implementation 2

12

- Based on Git only (no SQL database)
- Upload CSV files to Git repository
- SQL-style queries operate on CSV file via API
- Data versioning with Git
- Store scripts versioned as well
- Make subset creation reproducible



research data sharing without barriers
rd-alliance.org



```
diff --git a/smallList.csv b/smallList.csv
index f6299f...f6b025 100644
--- a/smallList.csv
+++ b/smallList.csv
@@ -1,21 +1,21 @@
sequenceNumber,firstName,lastName,email
0,Capitola,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
-1,Car,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
2,Car,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
3,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
4,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
5,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
6,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
7,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
8,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
9,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
10,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
11,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
12,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
13,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
14,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
15,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
16,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
17,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
18,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
19,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
20,ori,0.00418209888662742,126310,gadfrey.willkinson@gmail.com
```



CSV Reference Implementation 2

13

- Stefan Pröll, Christoph Meixner, Andreas Rauber
Precise Data Identification Services for Long Tail Research Data.
Proceedings of the intl. Conference on Preservation of Digital Objects (iPRES2016), Oct. 3-6 2016, Bern, Switzerland.
- Source at Github:
<https://github.com/datascience/RDA-WGDC-CSV-Data-Citation-Prototype>
- Videos:
 - Login: <https://youtu.be/EnralwbQfM0>
 - Upload: <https://youtu.be/xJruifX9E2U>
 - Subset: <https://www.youtube.com/watch?v=it4sC5vYiZQ>
 - Resolver: <https://youtu.be/FHsvjsUMiiY>
 - Update: <https://youtu.be/cMZ0xoZHUyl>





RESEARCH DATA ALLIANCE

WG Data Citation Pilot CBMI @ WUSTL

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

Washington University in St. Louis

research data sharing without barriers
rd-alliance.org

Moving Biomedical Big Data Sharing Forward

An adoption of the RDA Data Citation of Evolving Data
Recommendation to Electronic Health Records

Leslie McIntosh, PHD, MPH
Cynthia Hudson Vitale, MA
2016

RDA P8
Denver, USA September

@mcintold
@cynhudson



Background

Director, Center for
Biomedical
Informatics

Leslie McIntosh, PHD, MPH

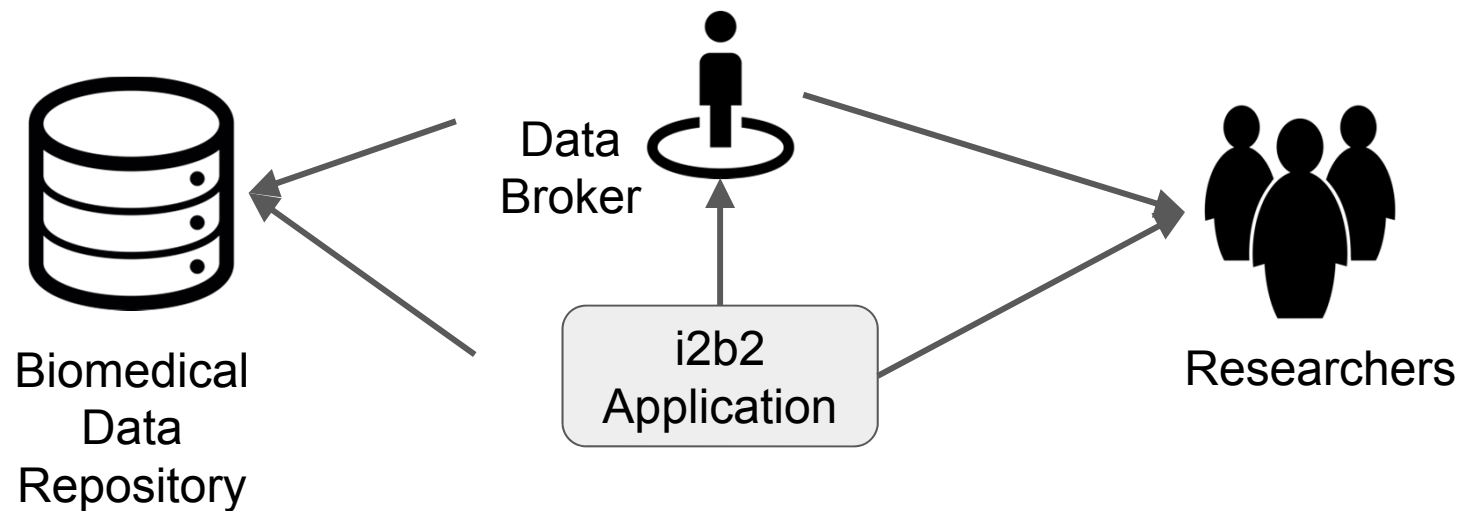
Data Services
Coordinator

Cynthia Hudson Vitale, MA



BDaaS

Biomedical Data as a Service



Move some of the responsibility of
reproducibility

Biomedical
Researcher



Biomedical
Pipeline

RDA/MacArthur Grant

Biomedical Adoption Project Goals

- ▶ Implement RDA Data Citation WG recommendation to local Washington U i2b2
- ▶ Engage other i2b2 community adoptees
- ▶ Contribute source code back to i2b2 community

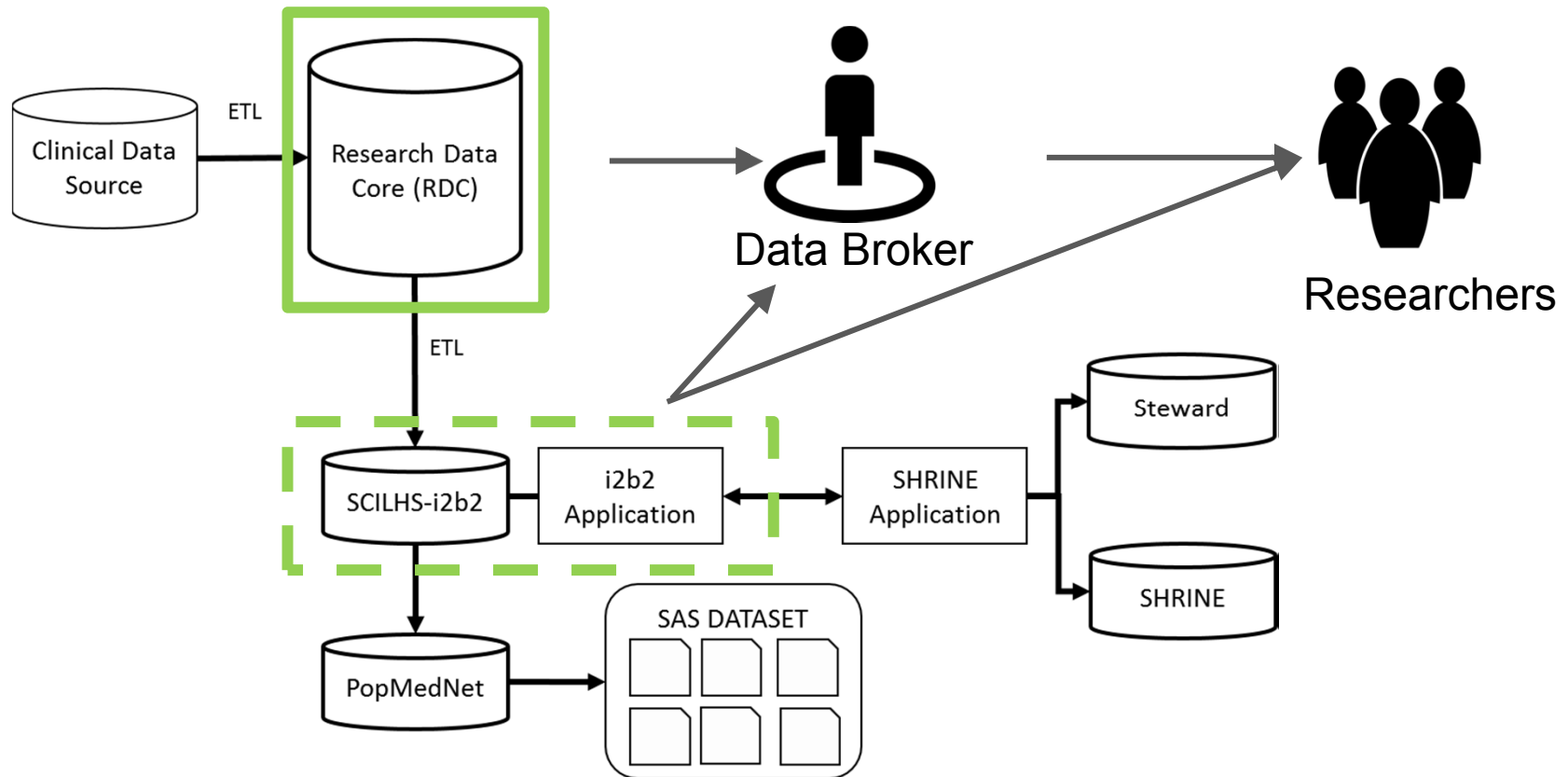
RDA Data Citation WG Recommendations

- ▶ R1: Data Versioning
- ▶ R2: Data Timestamping
- ▶ R3, R9: Query Store
- ▶ R7: Query Timestamping
- ▶ R8: Query PID
- ▶ R10: Query Citation

Internal Implementation Requirements

- ▶ Scalable
- ▶ Available for PostgreSQL
- ▶ Actively supported
- ▶ Easy to maintain
- ▶ Easy for data brokers to use

RDA-MacArthur Grant Focus



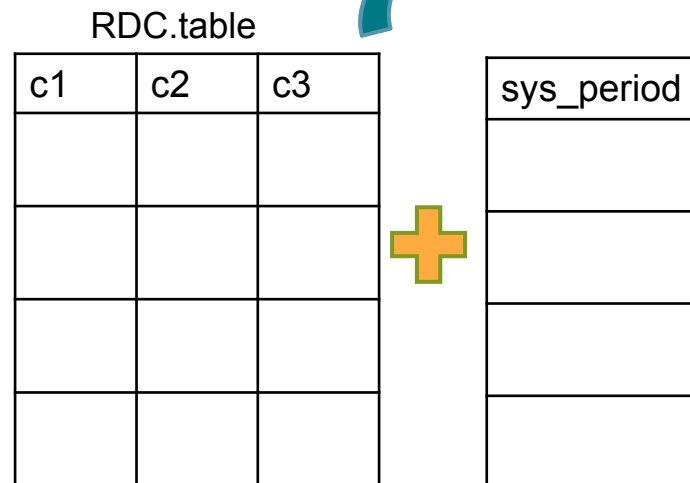
R1 and R2 Implementation

1



PostgreSQL Extension
"temporal_tables"

2



triggers

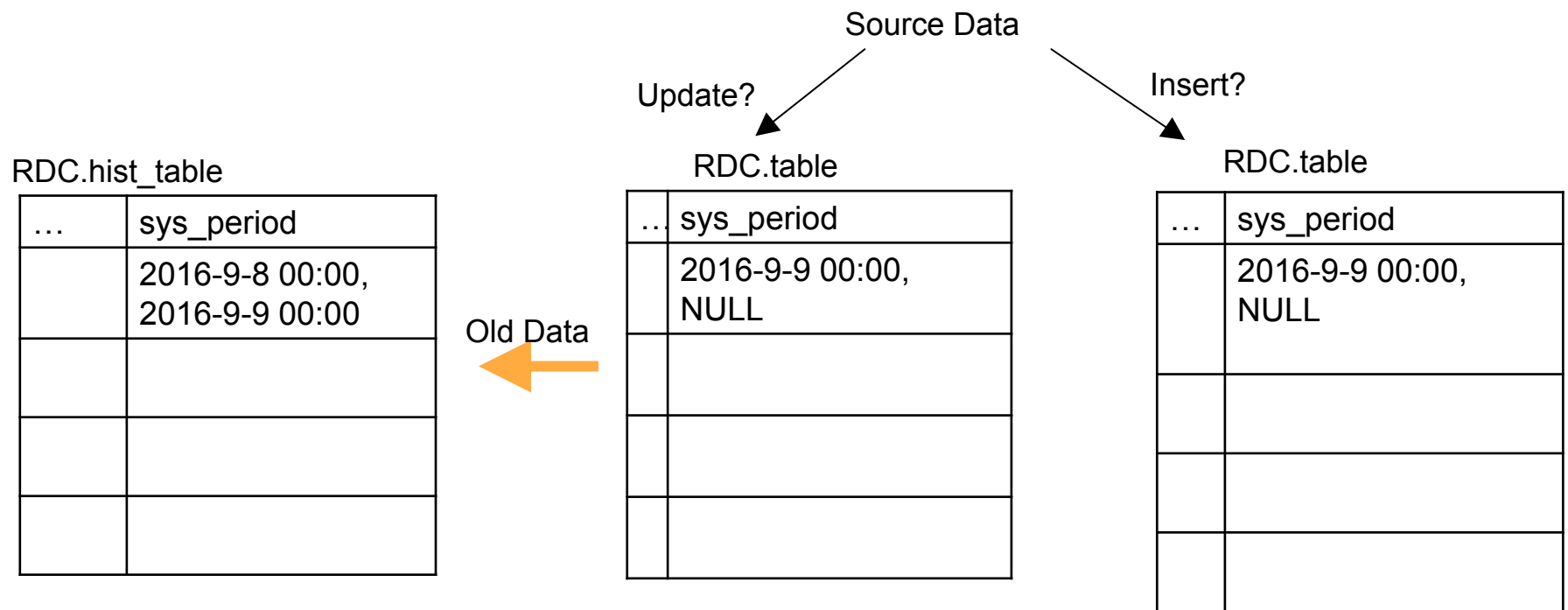
3

RDC.hist_table*

c1	c2	c3	sys_period

*stores history of
data changes

ETL Incrementals



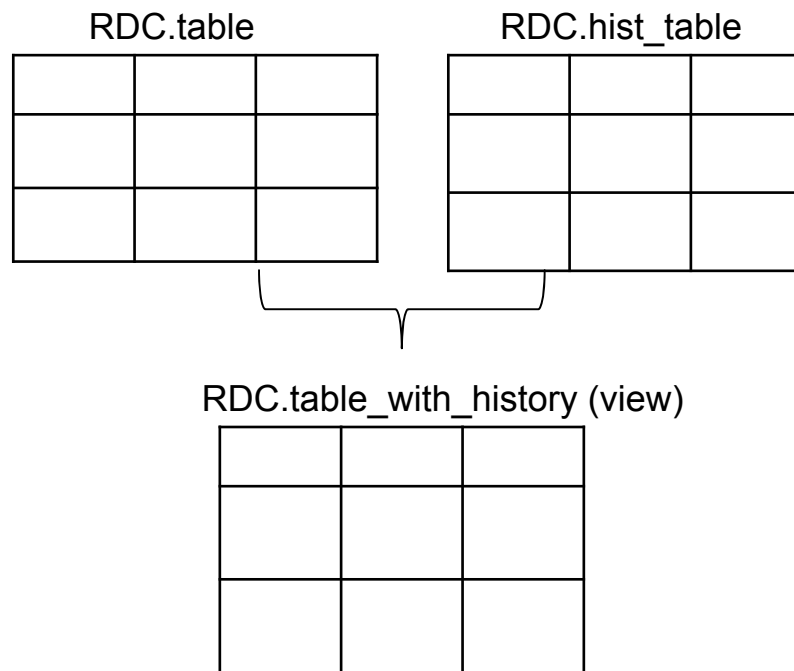
R3, R7, R8, R9, and R10 Implementation

1



PostgreSQL Extension
"temporal_tables"

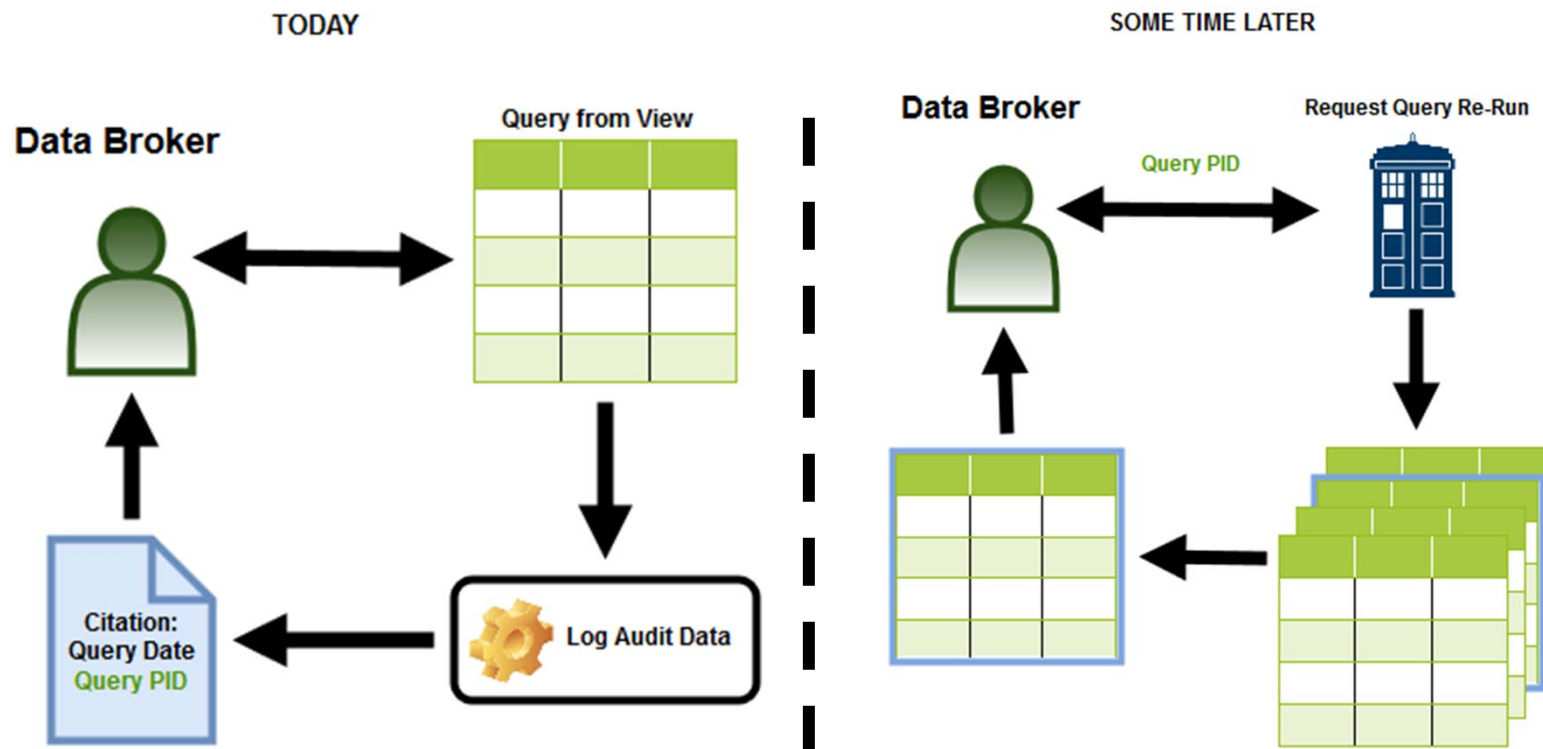
2



3

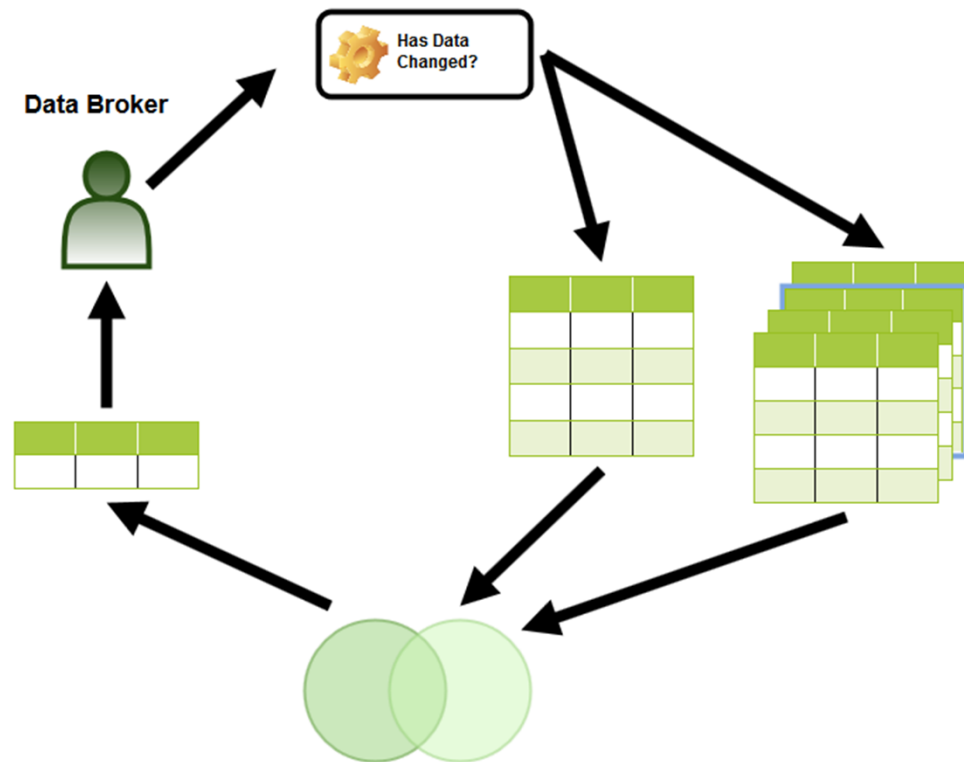
- functions
- triggers
- query audit tables

Data Reproducibility Workflow



*tardis: <http://pc012.deviantart.com/art/TARDIS-Simple-Vector-481284558>

Bonus Feature: Determine if Change Occurred



Future Developments

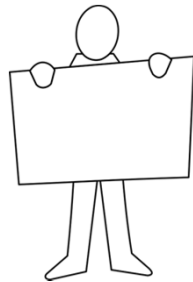
- ▶ Develop a process for sharing Query PID with researchers in an automated way
- ▶ Resolve Query PIDs to a landing page with Query metadata
- ▶ Implement research reproducibility requirements in other systems as possible

Outcomes and Support

Obtained Outcomes

- ▶ Implemented WG recommendations
- ▶ Engaged with other i2b2 adoptees
(Harvard, Nationwide Children's Hospital)

Dissemination



- ▶ Poster presentation (Harvard U, July 2016)
- ▶ Scientific manuscript based on our proof of concept to AMIA TBI/CRI 2017 conference
- ▶ Sharing the code with the community

Return on Investment (ROI) - Estimated

- ▶ 20 hours to complete 1 study
- ▶ \$150/hr (unsubsidized)
- ▶ \$3000 per study
- ▶ 115 research studies per year
- ▶ **14 replication studies**

Funding Support

**MacArthur Foundation 2016 Adoption Seeds program Foundation
through a sub-contract with Research Data Alliance**

Washington University Institute of Clinical and Translational Sciences
NIH CTSA Grant Number UL1TR000448 and UL1TR000448-09S1

Siteman Cancer Center at Washington University
NIH/NCI Grant P30 CA091842-14

Center for Biomedical Informatics @WUSTL

Teams for Reproducible Research

NIH-NLM Supplement

Leslie McIntosh
Cynthia Hudson-Vitale
Anthony Juehne
Rosalia Alcoser
Xiaoyan 'Sean' Liu
Brad Evanoff

RDA Collaborators

Andreas Rauber
Stefan Pröll

Research Data Alliance

Leslie McIntosh
Cynthia Hudson-Vitale
Anthony Juehne
Snehil Gupta
Connie Zabarovskaya
Brian Romine
Dan Vianello

WashU CBMI Research Reproducibility Resources

Repository

https://github.com/CBMIWU/Research_Reproducibility

Slides

<http://bit.ly/2cnWorU>

Bibliography

https://www.zotero.org/groups/biomedical_informatics_respro



Adoption of Data Citation Outcomes by BCO-DMO

Cynthia Chandler, Adam Shepherd

research data sharing without barriers
rd-alliance.org

A story of success enabled by RDA

39

- An existing repository (<http://bco-dmo.org/>)



- Marine research data curation since 2006
- Faced with new challenges, but no new funding
- e.g. data publication practices to support citation
- Used the outcomes from the RDA Data Citation Working Group to improve data publication and citation services

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



BCO-DMO Curated Data

40


- BCO-DMO is a thematic, domain-specific repository funded by NSF Ocean Sciences and Polar Programs
- BCO-DMO curated data are
 - Served: <http://bco-dmo.org> (URLs, URIs)
 - Published: at an Institutional Repository (CrossRef DOI) <http://dx.doi.org/10.1575/1912/4847>
 - Archived: at NCEI, a US National Data Center <http://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:0078575>

▼ Archival Copy

for Linked Data URI: <http://lod.bco-dmo.org/id/dataset/3046>

Version Date	Archive	Persistent Identifier	Date Assigned
2007-05-24	National Oceanographic Data Center (NODC)	0078575 	2011-10-26
2007-05-24	Marine Biological Laboratory/Woods Hole Oceanographic Institution Library (MBLWHOI DLA)	10.1575/1912/4847 	2011-10-12

BCO-DMO Dataset Landing Page (Mar '16) ⁴¹



Biological & Chemical Oceanography Data Management Office

[HOME](#)[DATA](#)[RESOURCES](#)[ABOUT US](#)

DATABASE


Programs	36
Projects	649
Deployments	2300
Datasets	7892
Instruments	395
Parameters	1366
People	2026
Affiliations	472
Funding	79
Awards	1304

Dataset: larval krill pigments

[Get Data](#)[Map It](#)

Project: U.S. GLOBEC Southern Ocean (SOGLOBEC)
Principal Investigator: Dr Robin Ross (University of California-Santa Barbara, UCSB) Dr Langdon Quetin (University of California-Santa Barbara, UCSB)
Contact: Dr Robin Ross (University of California-Santa Barbara, UCSB)
BCO-DMO Data Manager: Nancy Copley (Woods Hole Oceanographic Institution, WHOI BCO-DMO)
Validated: Yes
Data version: 2010-02-03
Version Date: 02/03/2010
Data URL: http://www.bco-dmo.org/dataset/3300/data ↗
Current State: Final no updates expected

GEOSPATIAL ACCESS

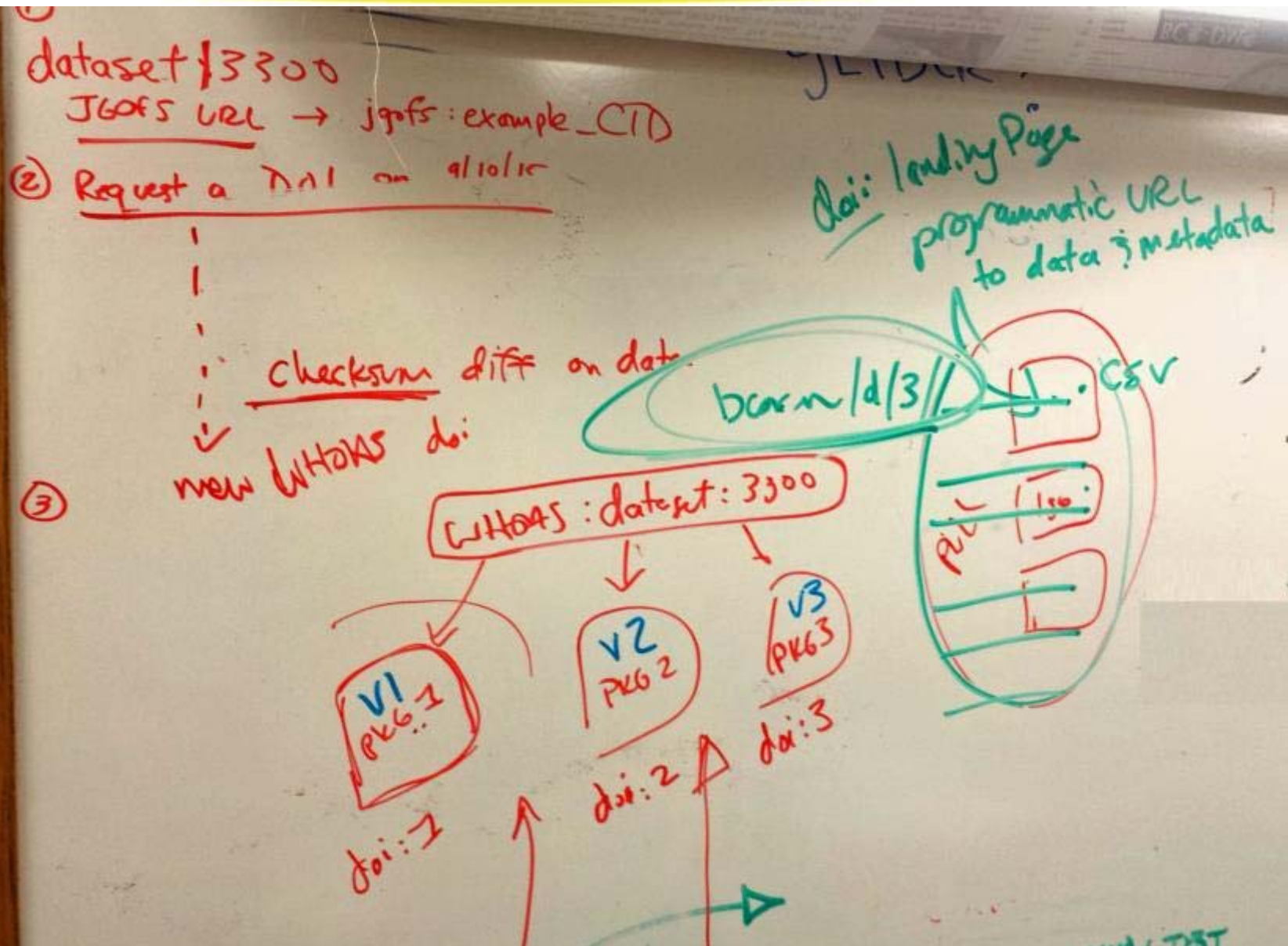


[Expand/Collapse All](#)

[Archival Copy](#)

Archive Location	Accession ID	Archive Copy URL
National Oceanographic Data Center (NODC)	0112635	http://accession.nodc.noaa.gov/0112635

42



Modified Architecture (March 2016)

43

BCO-DMO DATA PUBLICATION & CITATION

28 MAR 2016

Archive Record: W10015 [mid = 567890]

DFD 90001 [ISO] downloaded: DOI
doi: 10.175/1915/bco-dmo.567890

DFD 90002 [DATA 1]

DFD 90003 [DATA 2]

W10015

ISO

DATA 1

DATA 2

QUESTION:

supporting docs
are recorded as DFD
at W10015 for
reuse w/ NCET

DATA FILE DESCRIPTOR [DFD]

OSPREY CONTENT TYPE

checksum: text
checksum algorithm: controlled vocabulary
filesize_in_bytes: integer
resolvable_url: URL
dataset: Dataset or Dataset Deployment
type: controlled vocabulary [Data File Type]

DATA File Type

name: text (ex: Tab Separated Values, ISO 19115-2)
data_one_identifier: text
data_one_type: controlled vocabulary
mime_type: text (text/csv, application/xml)

Archive Record: NCET [mid = 67890]

DFD 90001

DFD 90002

DFD 90003

Serialized from W10015,
Stored in WAF (https)
SHA 384 MANIFEST & Summary

Archive Record: DATAONE [mid = 789012]

DFD 90001 [ISO] downloaded: DOI
Accession # at NCET
DOI: 10.175/1915/bco-dmo.567890

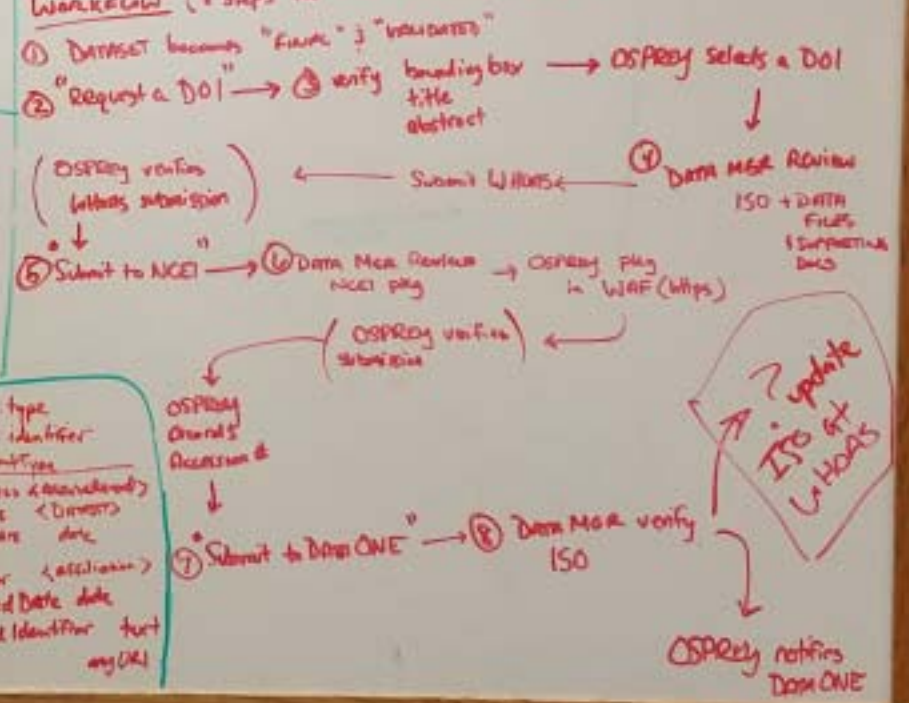
DFD 90002

DFD 90003

(stored locally
in OSPREY) or update
versions?

(Serialized from
W10015)

WORKFLOW (8 steps for new DATASET version)



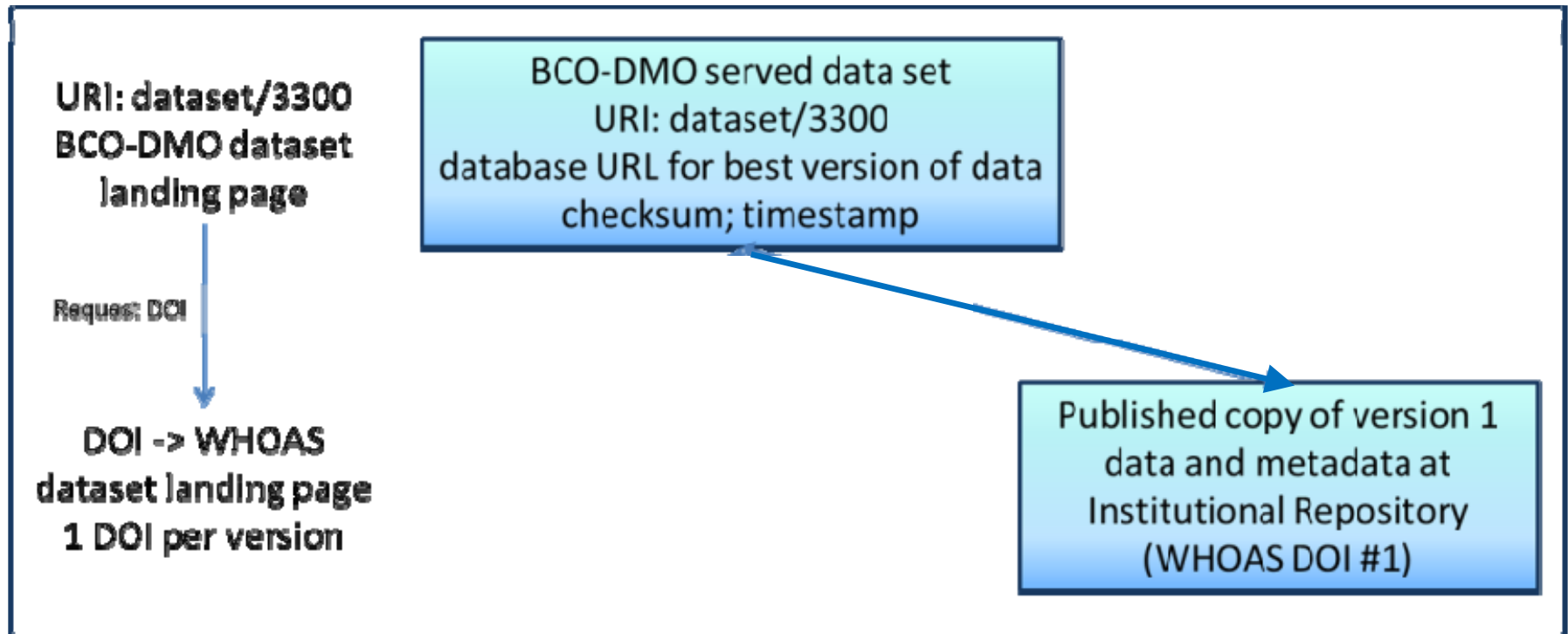
<DATASET>	archived By	<Archive Record>
Archive Record	his Bitstream	
Data File Descriptor		
checksum	string	
checksum algorithm	string	
filesize_in_bytes	integer	
resolvable_url	URL	
dataset	Dataset	
type	controlled vocabulary	
mime_type	text	

<Data File Descriptor>	<Data File Type>
checksum	string
checksum algorithm	string
filesize_in_bytes	integer
resolvable_url	URL
dataset	Dataset
type	controlled vocabulary
mime_type	text

<Archive Record>	<Data File Descriptor>
checksum	string
checksum algorithm	string
filesize_in_bytes	integer
resolvable_url	URL
dataset	Dataset
type	controlled vocabulary
mime_type	text

BCO-DMO Data Publication System Components

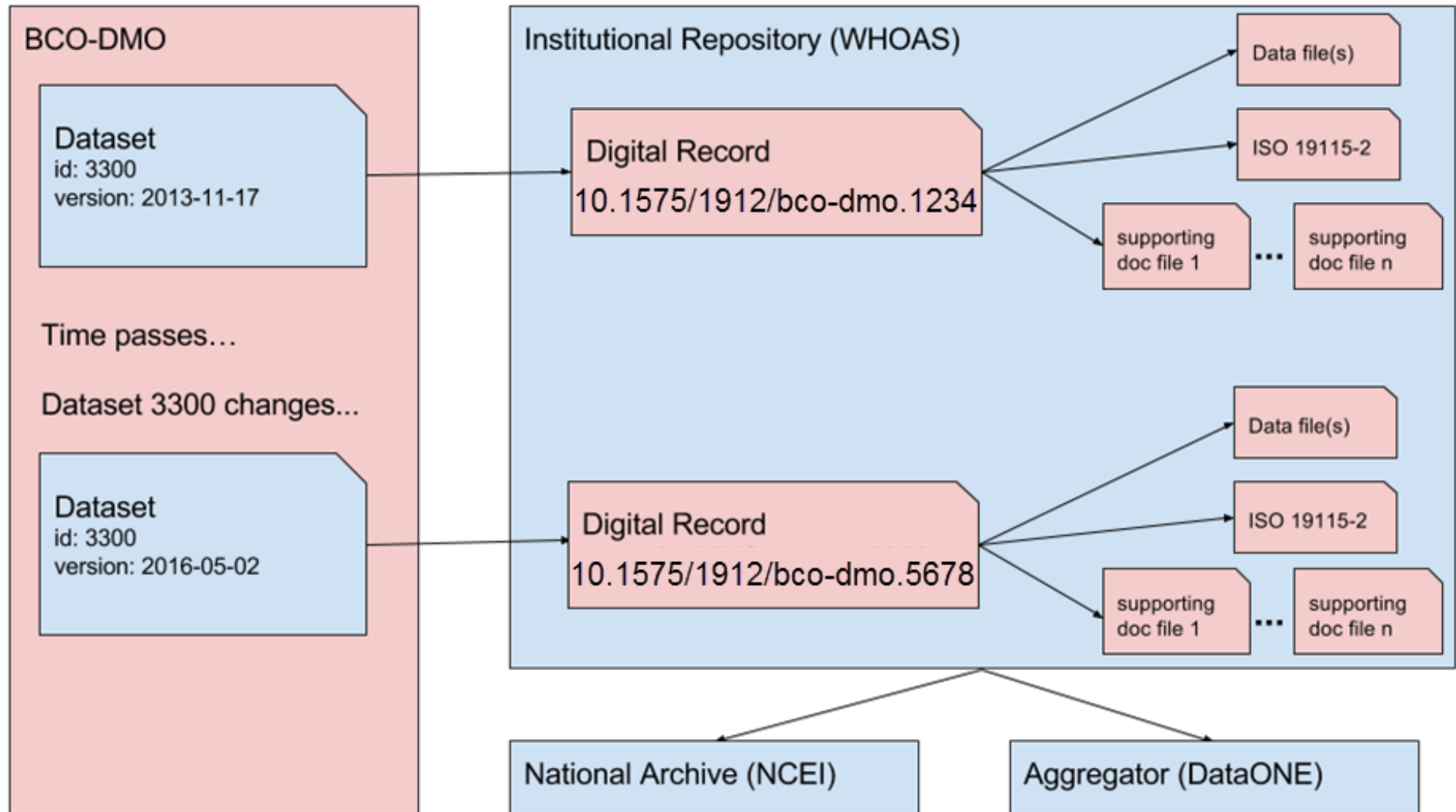
44



BCO-DMO publishes data to WHOAS and a DOI is assigned.
The BCO-DMO architecture now supports data versioning.


BCO-DMO Data Citation System Components

45



BCO-DMO Data Set Landing Page

46



Biological & Chemical Oceanography Data Management Office

[DATA](#) [RESOURCES](#) [ABOUT US](#)

DATABASE

- New Entry
- Programs 36
- Projects 696
- Deployments 2360
- Platforms 486
- Datasets 8079
- Instruments 408

Dataset: Cellular element quotas: Si in Synechococcus cells

[Get Data](#) [Map It](#)

Project: [Understanding the Role of Picocyanobacteria in the Marine Silicate Cycle \(Si_in_Syn\)](#)

Principal Investigator: [Dr Benjamin Twining](#) (Bigelow Laboratory for Ocean Sciences, Bigelow)

Contact: [Daniel Ohnemius](#) (Bigelow Laboratory for Ocean Sciences, Bigelow)

Data version: Final

Version Date: 05/06/2016

[Archival Copy](#)

Archival Copy

Version Date	Archive	Persistent Identifier	Date Assigned
2016-05-06	Marine Biological Laboratory/Woods Hole Oceanographic Institution Library (MBLWHOI DLA)	10.1575/1912/bco-dmo.651474 	2016-07-08

GEOSPATIAL ACCESS



References:

Twining, B.S., et al. Metal contents of phytoplankton and labile particulate material in the North Atlantic Ocean. *Prog. Oceanogr.* (2015)
<http://dx.doi.org/10.1016/j.pocean.2015.07.001>
https://www.researchgate.net/publication/282626294_Metal_contents_of_phytoplankton_and_labile_particulate_material_in_the_North_Atlantic_Ocean

[More Information about this dataset](#)

Search



☒ Search WHOAS

☐ This Collection

BROWSE

All of WHOAS

Communities & Collections

Element quotas of individual *Synechococcus* cells collected during Bermuda Atlantic Time-series Study (BATS) cruises aboard the R/V Atlantic Explorer between dates 2012-07-11 and 2013-10-13

dc.description.sponsorship NSF Ocean Sciences (NSF OCE) OCE-1131139, NSF Ocean Sciences (NSF OCE) OCE-1335012, NSF Ocean Sciences (NSF OCE) OCE-1131046

dc.publisher Biological and Chemical Oceanography Data Management Office (BCO-DMO). Contact: bco-dmo-data@whoi.edu


dc.relation <http://lod.bco-dmo.org/id/dataset/644840>

By Issue Date

Authors

Titles

View Item

 cellular-element-quotas-si-synechococcus-cells.tsv (10.09Kb)

 115-2.xml (88.65Kb)

 Names.pdf (12.10Kb)

31.6691

DOI


[10.1575/1912/bco-dmo.651474](https://doi.org/10.1575/1912/bco-dmo.651474)

Keyword

SXRF; Synchrotron radiation X-Ray Fluorescence; *Synechococcus*; sil

BCO-DMO Data Set Landing Page

48



Biological & Chemical Oceanography Data Management Office

[DATA](#) [RESOURCES](#) [ABOUT US](#)

DATABASE

- New Entry
- Programs 36
- Projects 696
- Deployments 2360
- Platforms 486
- Datasets 8079
- Instruments 408
- Parameters 1373

Dataset: Cellular element quotas: Si in Synechococcus cells

[Get Data](#) [Map It](#)

Project: Understanding the Role of Picocyanobacteria in the Marine Silicate Cycle (Si_in_Syn)

Principal Investigator: Dr Benjamin Twining (Bigelow Laboratory for Ocean Sciences, Bigelow)

Contact: Daniel Ohnemus (Bigelow Laboratory for Ocean Sciences, Bigelow)

Data version: Final

Version Date: 05/06/2016

▼ **Archival Copy**

Version	Archive	Persistent Identifier	Date
---------	---------	-----------------------	------

References:

Twining, B.S., et al. Metal contents of phytoplankton and labile particulate material in the North Atlantic Ocean. Prog. Oceanogr. (2015)

<http://dx.doi.org/10.1016/j.pocean.2015.07.001>

<https://www.researchgate.net/publication>

[/282626294_Metal_contents_of_phytoplankton_and_labile_particulate_material_in_the_North_Atlantic_Ocean](https://www.researchgate.net/publication/282626294_Metal_contents_of_phytoplankton_and_labile_particulate_material_in_the_North_Atlantic_Ocean)

are also provided.

GEOSPATIAL ACCESS



References:

Twining, B.S., et al. Metal contents of phytoplankton and labile particulate material in the North Atlantic Ocean. Prog. Oceanogr. (2015)

<http://dx.doi.org/10.1016/j.pocean.2015.07.001>

<https://www.researchgate.net/publication>

[/282626294_Metal_contents_of_phytoplankton_and_labile_particulate_material_in_the_North_Atlantic_Ocean](https://www.researchgate.net/publication/282626294_Metal_contents_of_phytoplankton_and_labile_particulate_material_in_the_North_Atlantic_Ocean)





▼ **More information about this dataset**


Linked to Publication via DOI

49


ScienceDirect

JournalsBooksSign in ?



 Download PDF  Add to online library  Export  Advanced search



Progress in Oceanography
Volume 137, Part A, September 2015, Pages 261–283



Metal contents of phytoplankton and labile particulate material in the North Atlantic Ocean

Benjamin S. Twining^a,  , Sara Rauschenberg^a, Peter L. Morton^{b, 1}, Stefan Vogt^c

[Show more](#)

doi:10.1016/j.pocean.2015.07.001

Get rights and content

Highlights

- First basin-wide measurements of plankton metal quotas in the N. Atlantic Ocean.
- Fe and Mn quotas significantly higher on western side of section.
- Cu and Ni quotas significantly elevated on eastern side of section.
- Evidence for Al scavenging by biogenic silica.
- Dissolved ratios not an accurate measure of cellular Fe quotas.

Recommended articles

Evaluation of approaches to estimate biogenic ...
2015, Marine Chemistry [more](#)

High-frequency sea level oscillations in the Me...
2015, Progress in Oceanography [more](#)

Composition of metals in suspended particulate...
2016, Regional Studies in Marine Science [more](#)

[View more articles »](#)

Citing articles (2)

Related book content

New Capabilities ... BCO-DMO becoming a DataONE Member Node

50

<https://search.dataone.org/>

The screenshot displays the DataONE search interface. At the top, the DataONE logo is on the left, and navigation links (About, News, Participate, Resources, Education, Data) are in the center. A search bar with the text "DATAONE SEARCH:" is on the right, followed by "Search", "Summary", and "Jump to: DOI or ID Go" buttons. On the far right, "Sign in" and "Sign up" buttons are visible.

Below the navigation bar, a "Clear all filters" link is present. The left sidebar contains a "Search" section with a search phrase input field and a "My Search" section with a "BCO-DMO" filter button. Below these are "Filter by:" options: Data attribute, Data files, Member Node, Creator, Year, Identifier, and Taxon.

The main content area shows "Datasets 1 to 25 of 149". It includes a pagination bar with numbers 1, 2, 3, ..., 6, and a "Next" button. A "Sort by" dropdown menu is set to "Most recent". Three dataset entries are listed:


- Dr David Siegel and Dr Kenneth O. Buesseler. 2009. urn:node:mnTestBCODMO. <http://lod.bco-dmo.org/id/dataset-file/555907>.
- Dr Carl Lamborg and Dr Kenneth O. Buesseler. 2009. urn:node:mnTestBCODMO. <http://lod.bco-dmo.org/id/dataset-file/555902>.
- Dr Uta Passow. 2014. urn:node:mnTestBCODMO. <http://lod.bco-dmo.org/id/dataset-file/555897>.

Each entry has icons for information, document, and location. At the bottom of the list, a logo for "DOC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, NESDIS, NOAA, U.S." is shown.

On the right, a map of the United States is displayed with a "Hide Map" button. The map shows several numbered regions: 2, 4, 19, 1, 25, and 1. A "Satellite" and "Terrain" toggle is at the bottom left of the map. The Google logo and "Map data ©2016 Google, INEGI 1000 km Terms of Use" are at the bottom right.

New Capabilities ... BCO-DMO Data Set Citation

51



Biological & Chemical Oceanography Data Management Office

DATA RESOURCES ABOUT US

Enter search terms

DATABASE

Programs 37

Projects 724

Get Data Map It

Dataset: Prochlorococcus_narB

Cite This Dataset

Data Citation:

Chisholm, Sallie (2016) The abundance of Prochlorococcus cells containing the nitrate reductase gene (narB) at the HOT and BATS sites in the Pacific and Atlantic Oceans between October 2005 and January 2008. Biological and Chemical Oceanography Data Management Office (BCO-DMO) Dataset version 2016-03-31 [if applicable, indicate subset used]. doi:10.1575/1912/bco-dmo.641735 [access date]

Terms of Use


All data sets are licensed under a [Creative Commons Attribution 4.0 International License](#) (CC BY 4). Per the CC BY 4 license it is understood that any use of the data set will properly acknowledge the individual(s) listed above using the suggested data citation. If you wish to use this data set, it is highly recommended that you contact the original principal investigator(s) (PI). Should the relevant PI be unavailable, please contact BCO-DMO (info@bco-dmo.org) for additional guidance. For general guidance please see the BCO-DMO [Terms of Use](#) document.

Awards 1412

Current State: Final no updates expected

Archival Copy

GEOSPATIAL ACCESS



Version Date	Archive	Persistent Identifier	Date Assigned
2016-03-31	Marine Biological Laboratory/Woods Hole Oceanographic Institution Library (MBLWHOI DLA)	10.1575/1912/bco-dmo.641735	2016-04-04

Thank you ...

52

- To the Data Citation Working Group for their efforts
<https://www.rd-alliance.org/groups/data-citation-wg.html>
- RDA US for funding this adoption project
- TIMELINE:
 - Redesign/protoype completed by 1 June 2016
 - New citation recommendation by 1 Sep 2016
 - Report out at RDA P8 (Denver, CO) September 2016
 - Final report by 1 December 2016

Cyndy Chandler @cynDC42
ORCID: 0000-0003-2129-1647

@bcodmo
cchandler@whoi.edu

- Removed these to reduce talk to 10-15 minutes

■ Evaluation

- Evaluate recommendations (done December 2015)
- Try implementation in existing BCO-DMO architecture (work began 4 April 2016)

■ Trial

- BCO-DMO: R1-11 fit well with current architecture; R12 doable; test as part of DataONE node membership; R13-14 are consistent with Linked Data approach to data publication and sharing

NOTE: adoption grant received from RDA US (April 2016)

- **DC goals:** to create identification mechanisms that:
 - allow us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
 - allow us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- **DC outcomes:** 14 recommendations and associated documentation
 - ensuring that data are 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

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

- Preparing the Data and the Query Store
 - R1 - Data Versioning
 - R2 - Timestamping
 - R3 - Query Store Facilities
- Persistently Identifying Specific Data Sets
 - R4 - Query Uniqueness
 - R5 - Stable Sorting
 - R6 - Result Set Verification
 - R7 - Query Timestamping
 - R8 - Query PID
 - R9 - Store the Query
 - R10 - Automated Citation Texts
- Resolving PIDs and Retrieving the Data
 - R11 - Landing Page
 - R12 - Machine Actionability
- Upon modifications to the Data Infrastructure
 - R13 - Technology Migration
 - R14 - Migration Verification

»» Data Versioning: For retrieving earlier states of datasets the data need to be versioned. Markers shall indicate inserts, updates and deletes of data in the database.

»» Data Timestamping: Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.

»» Data Identification: The data used shall be identified via a PID pointing to a time-stamped query, resolving to a landing page.

Oct 2015 version w/ 14 recommendations

DC WG chairs: Andreas Rauber, Ari Asmi, Dieter van Uytvanck

New capability (implemented)

57

procedure: when a BCO-DMO data set is updated ...

- A copy of the previous version is preserved
- Request a DOI for the new version of data
- Publish data, and create new landing page for new version of data, with new DOI assigned
- BCO-DMO database has links to all versions of the data (archived and published)
- Both archive and published dataset landing pages have links back to best version of full dataset at BCO-DMO
- BCO-DMO data set landing page displays links to all archived and published versions

- Extended description of recommendations
**Identification of Reproducible Subsets for Data Citation,
Sharing and Re-Use**

Andreas Rauber
Vienna University of Technology, Austria
rauber@ifs.tuwien.ac.at

Dieter van Uytvanck
CLARIN ERIC, Utrecht, Netherlands
dieter@clarin.eu

Ari Asmi
University of Helsinki, Finland
ari.asmi@helsinki.fi

Stefan Pröll
SBA Research, Vienna, Austria
sproell@sba-research.org

- Altman and Crosas. 2013. “Evolution of Data Citation ...”
- CODATA-ICSTI 2013. “Out of cite, out of mind”
- FORCE11 <https://www.force11.org/about/mission-and-guiding-principles>
- R. E. Duerr, et al. “On the utility of identification schemes for digital earth science data”, ESI, 2011.



RESEARCH DATA ALLIANCE

Vermont Monitoring Cooperative

James Duncan, Jennifer Pontius
VMC

research data sharing without barriers
rd-alliance.org



IMPLEMENTATION OF DYNAMIC DATA CITATION

James Duncan and Jennifer Pontius

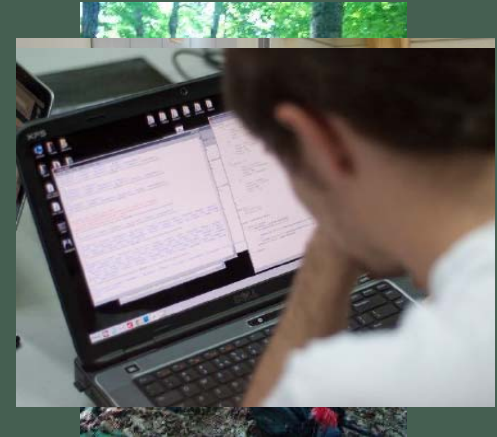
9/16/2016

james.duncan@uvm.edu, www.uvm.edu/vmc



Vermont Monitoring Cooperative

Ecosystem Monitoring
Collaborator Network
Data Archive, Access and
Integration



Soil



Water



Forest



Air



Wildlife

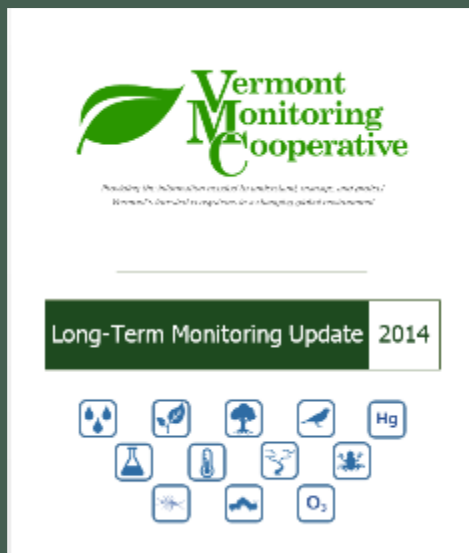
MANY DISCIPLINES, MANY CONTRIBUTORS

VMC houses any data related to forest ecosystem condition, regardless of affiliation or discipline



WHY WE NEED IT

- Continually evolving datasets
- Some errors not caught till next field season
- Frequent reporting and publishing



DYNAMIC DATA CITATION – FEATURES NEEDED

- Light footprint on database resources
- Works on top of existing catalog and metadata
- Works in an institutionally managed PHP/MySQL environment
- User-driven control of what quantity of change constitutes a version
- Integration with management portal
- Track granular changes in data

USER WORKFLOW TO DATE

- Modify a dataset
 - changes tracked
 - original data table unchanged
- Commit to version, assign name
 - computes result hash (table pkid, col names, first col data) and query hash
 - updates data table to new state
 - formalizes version
- Restore previous version
 - creates new version table from current data table state,
 - walks it back using stored SQL.
 - Garbage collected after a period of time

Manage Dataset Versions

Original Version (created 2016-09-09 00:00:00, creator ID = 2360) [Get Version](#)

V1.3 (created 2016-09-14 21:07:05, creator ID = 2360) [Get Version](#)

v1.4.1 (created 2016-09-14 21:10:09, creator ID = 2360) [Get Version](#)

v1.5 (created 2016-09-15 00:31:12, creator ID = 2360) [Get Version](#)

[Create New Version](#)

Editing data for Data 3

[Back to Dataset Versions](#)

[Show changes](#)

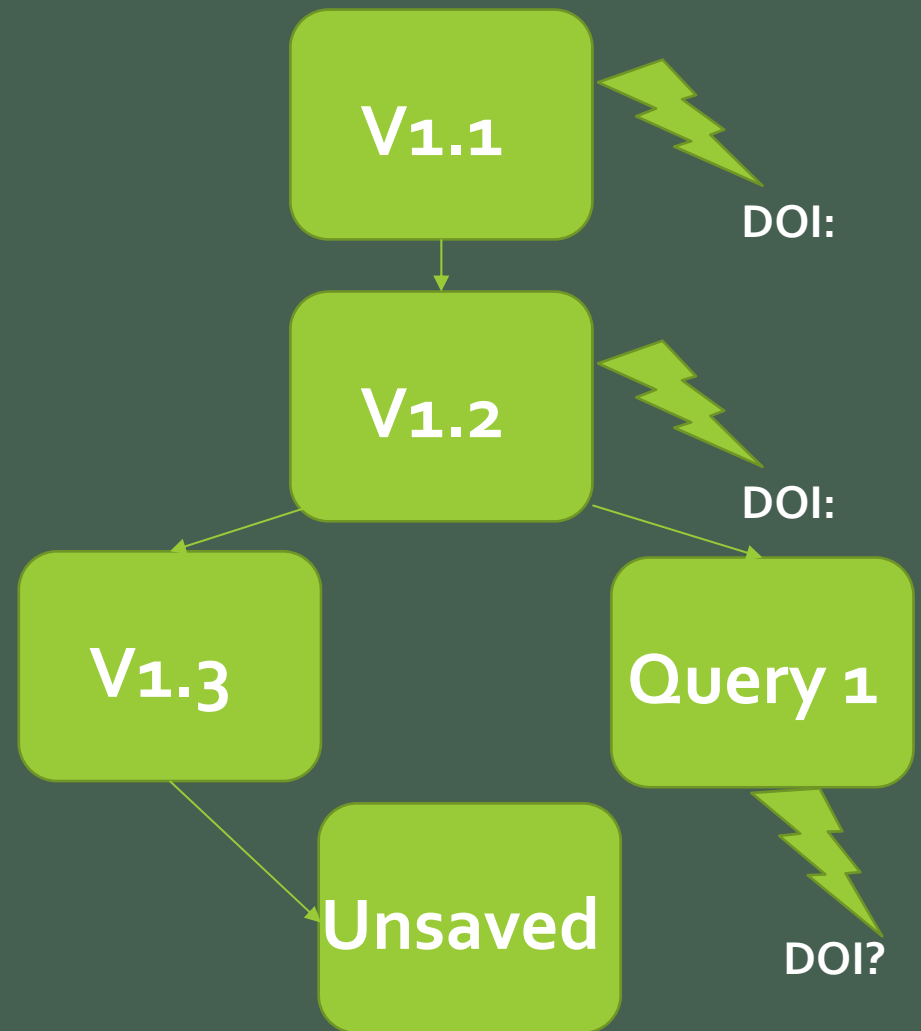
[Save to a new version](#)

```
1 UPDATE Z0001_test SET 'SiteID' = 'Stevensville Brook' WHERE db_table_pk=30
2 DELETE FROM Z0001_test where db_table_pk=30
3 DELETE FROM Z0001_test where db_table_pk=35
```

Actions	SiteID	LabID	Date	MeanDensity	Mean
 	Stevensville Brook	2000.187	0000-00-00	4644322354	39.0
 	Winhall River	2011.081	2011-10-07	201	47.5
 	Winhall River	2012.089	2012-09-27	1981	52.0
 	Winhall	2013.150	2013-10-15	1092	39.0

USER WORKFLOW, STILL TO COME

- Web-based data editing validation
- DOI minting integration
- Public display
- Subsetting workflow
- *Other methods of data modification?*
- Upgrade to rest of system



TECHNICAL DETAILS

Version Info Table

Version PID	Datase t ID	Version Name	Version ID	Person ID	Query Hash	Result Hash	Timesta mp
23456	3525	Version 1.5	3	
23574	3525	Unsaved	-1		NULL

Step Tracking Table (Child of Version Info)

Step ID	Version PID	Forward	Backward	Order
983245	23574	DELETE FROM...	INSERT INTO...	1
983245	23574	UPDATE SET site="Winhall" ...	UPDATE SET site="Lye Brook"...	2

IMPLEMENTATION CHALLENGES AND QUESTIONS

Challenges

- Large updates
- Re-creation of past versions, in terms of garbage collection and storage

Questions

- Query uniqueness checking and query normalization
- Efficient but effective results hashing strategies
- Linear progression of data, versus branching network

ACKNOWLEDGMENTS

- Adoption seed funding - MacArthur Foundation and the Research Data Alliance



MacArthur Foundation

- The US Forest Service State and Private Forestry program for core operational funding of the VMC



- Fran Berman, Yolanda Meleco and the other adopters who have been sharing their experiences.
- All the VMC cooperators that contribute





THANK YOU!





ARGO

Justin Buck, Helen Glaves
BODC

research data sharing without barriers
rd-alliance.org



RDA P8 2016, Denver, 16th September 2016

WG Data Citation: Adoption meeting

Argo DOI pilot

Justin Buck, National Oceanography Centre (UK), juck@bodc.ac.uk

Thierry Carval, Ifremer (France), thierry.carval@ifremer.fr

Thomas Loubrieu, Ifremer (France), thomas.loubrieu@ifremer.fr

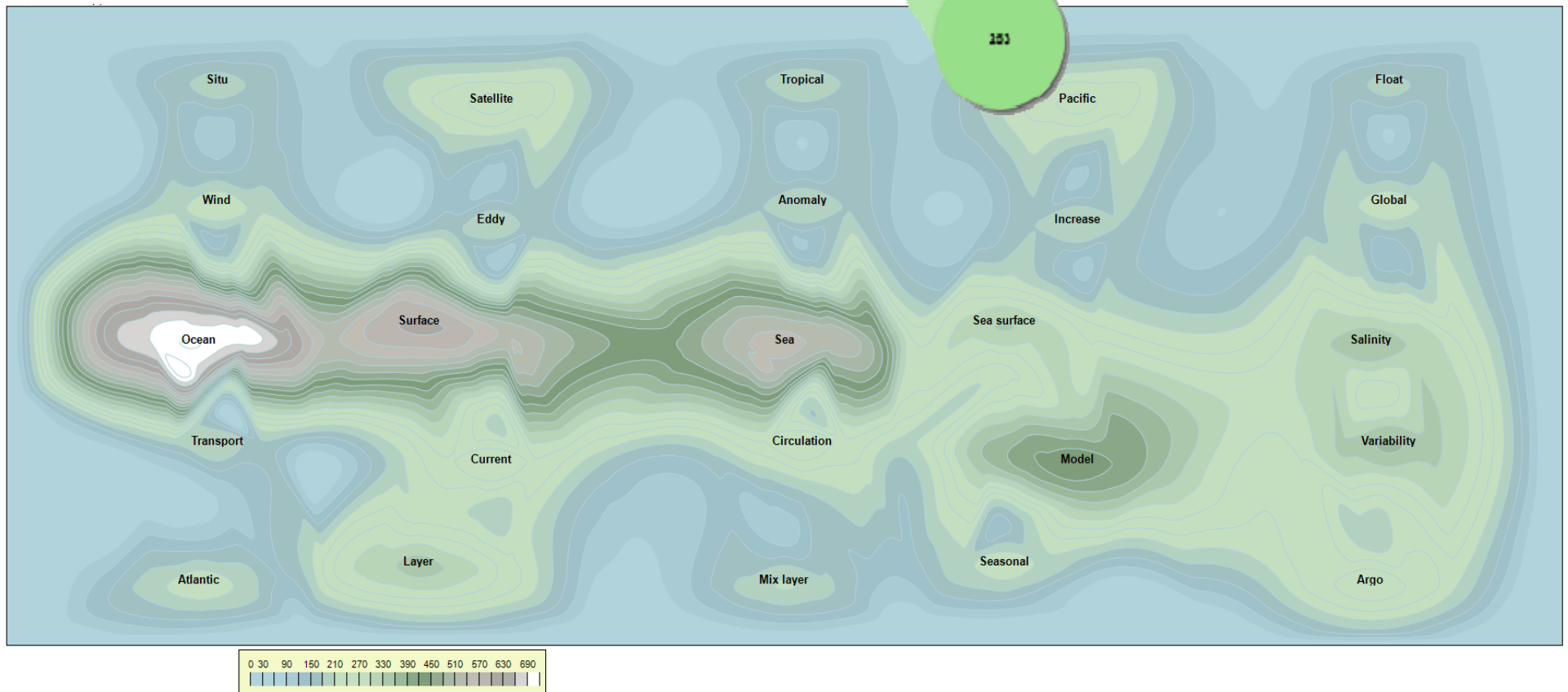
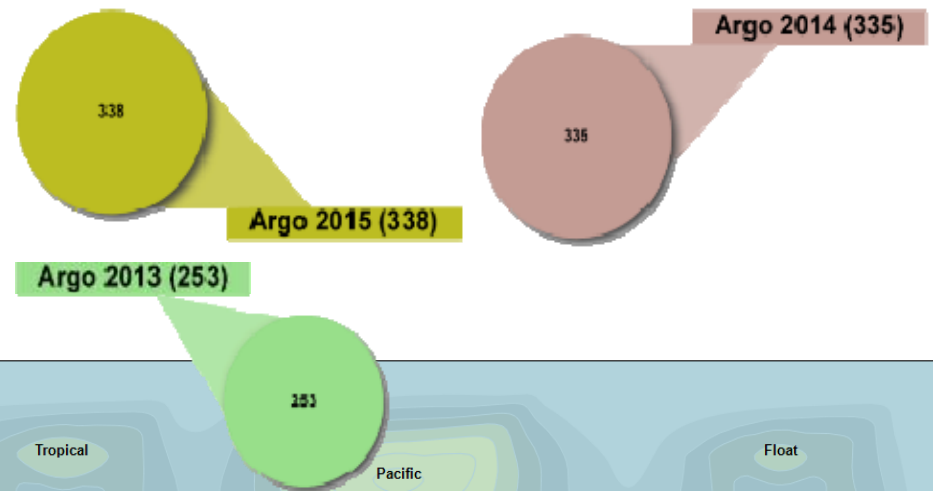
Frederic Merceur, Ifremer (France), frederic.merceur@ifremer.fr



**National
Oceanography Centre**
NATURAL ENVIRONMENT RESEARCH COUNCIL

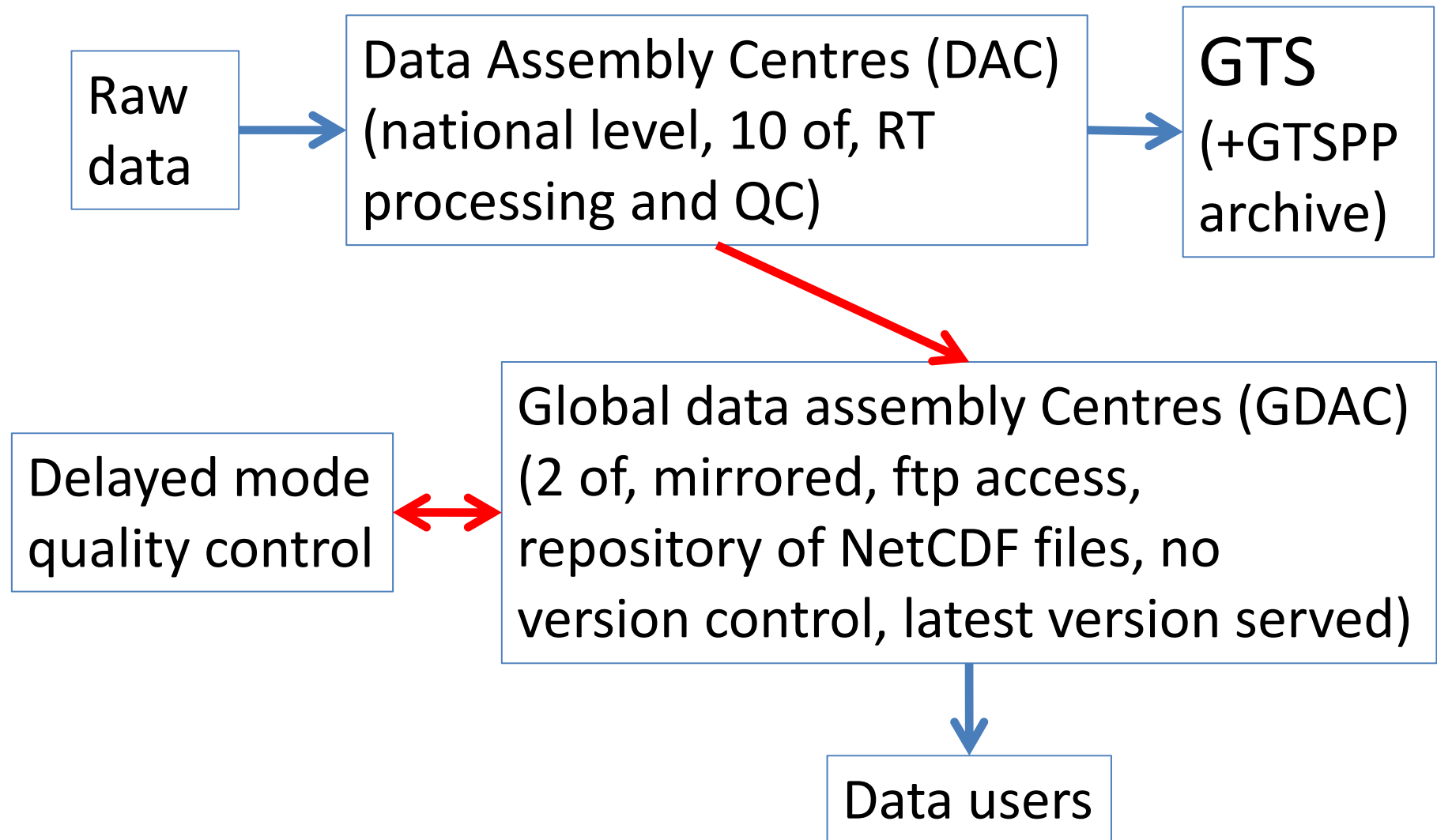


300+ citations
per year



How to cite Argo data at a given point in time?
Possible with a single DOI?

Argo data system (simplified)

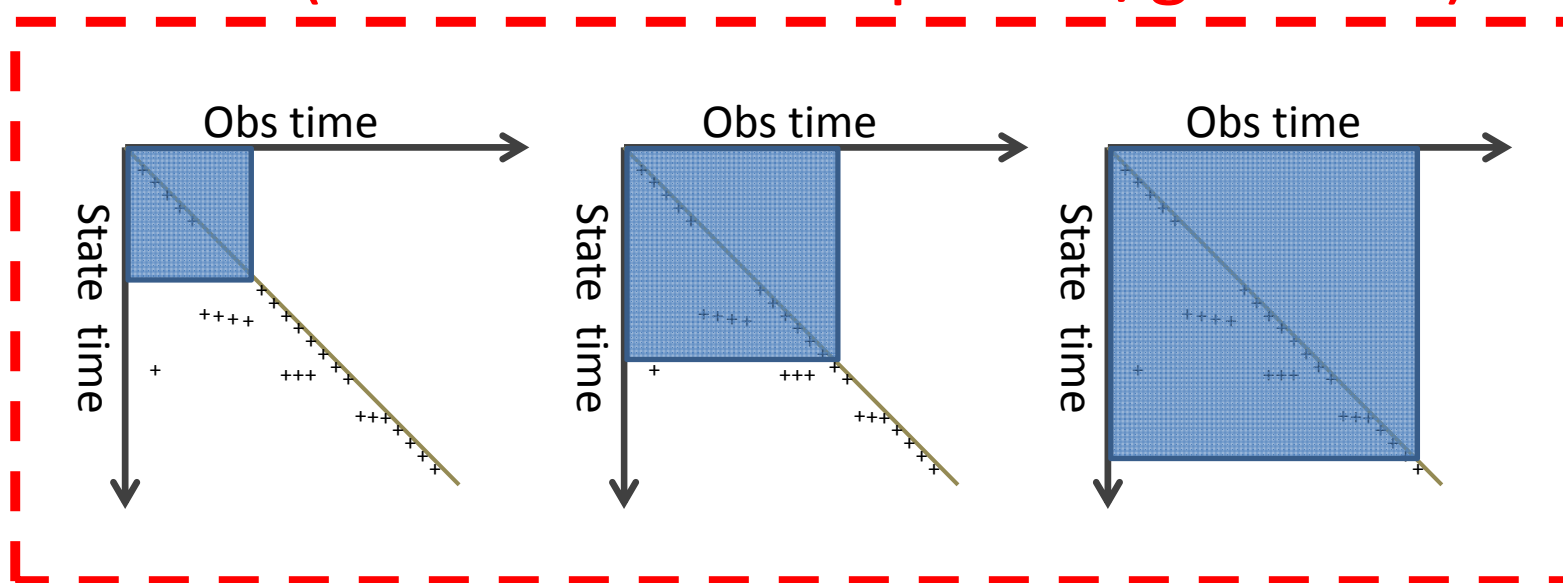


Key associated milestones

- 2014 – Introduction of dynamic data into the DataCite metadata schema
 - <https://schema.datacite.org/>
 - Snapshots were used as an interim solution for Argo
- 2015 – RDA recommendations on evolving data:
 - https://rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf
 - Legacy GDAC architecture does not permit full implementation

How to apply a DOI for Argo

Archive (collection of snapshots/granules)



To cite a particular snapshot one can potentially cite a time slice of an archive i.e. the snapshot at a given point in time.

New single DOI

Argo (2000). **Argo float data and metadata from Global Data Assembly Centre (Argo GDAC)**. SEANOE. <http://doi.org/10.17882/42182>

16 notes - J... x


Argo float data and metac... x

Justin

www.seanoe.org/data/00311/42182/

Argo

part of the integrated global observation strategy



Argo float data and metadata from Global Data Assembly Centre (Argo GDAC)

Publication date

2000-09-12

Author(s)

Argo

DOI

[10.17882/42182](http://doi.org/10.17882/42182)

Publisher

[SEANOE](http://seanoe.org)


Keyword(s)


float, Argo, global ocean observing system, ocean circulation, in-situ, ocean pressure, sea water salinity, sea water temperature, multi-year, weather climate and seasonal observation, global-ocean, Installations de suivi environnemental

Abstract

Argo is a global array of 3,000 free-drifting profiling floats that measures the temperature and salinity of the upper 2000 m of the ocean. This allows, for the first time, continuous monitoring of the temperature, salinity, and velocity of the upper ocean, with all data being relayed and made publicly available within hours after collection. The array provides 100,000 temperature/salinity profiles and velocity measurements per year distributed over the global oceans at an average of 3-

Click to download the data

 DATA



key used to identify snapshot

File	Size	Format	Processing	Access	Key
2016-09-08	10 GB	NC, NetCDF	Quality controlled data	Open access	45724
2016-08-08	10 GB	NC, NetCDF	Quality controlled data	Open access	45420
2016-07-08	10 GB	NC, NetCDF	Quality controlled data	Open access	45418

key used to identify snapshot

File	Size	Format	Processing	Access	Key
2016-09-08	10 GB	NC, NetCDF	Quality controlled data	Open access	45724
2016-08-08	10 GB	NC, NetCDF	Quality controlled data	Open access	45420
2016-07-08	10 GB	NC, NetCDF	Quality	Open access	45418

<http://www.seanoe.org/data/00311/42182/#45420>

Argo part of the integrated global observation strategy

Argo float data and metadata from Global Data Assembly Centre (Argo GDAC) - Snapshot of Argo GDAC of August, 8th 2016

Step towards RDA recommendation

Archive snapshots enables R1 and R2 at monthly granularity

R1 – Data Versioning

R2 – Timestamping

Argo Pilot effectively uses predetermined referencing of snapshots removing the need for requirements R3 to R7. # keys are PIDs for the snapshots and have associated citation texts.

R3 – Query Store Facilities

R4 – Query Uniqueness

R5 – Stable Sorting

R6 – Result Set Verification

R7 – Query Timestamping

R8 – Query PID

R9 – Store Query

R10 – Automated Citation Texts

SEANOE landing page architecture means R11 and R12 effectively met

R11 – Landing Page

R12 – Machine Actionability

Final two requirements untested at this stage

R13 – Technology Migration

R14 – Migration Verification

Summary

- There is now a single DOI for Argo
 - Takes account of legacy GDAC architecture
 - Monthly temporal granularity
 - Enables both reproducible research and simplifies the tracking of citations
 - ‘#’ rather than ‘?’ in identifier takes account of current DOI resolution architecture
- Extensible to other observing systems such as OceanSITES and EGO
- The concept allows for different subsets of Argo data e.g. ocean basins, Bio-Argo data



**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

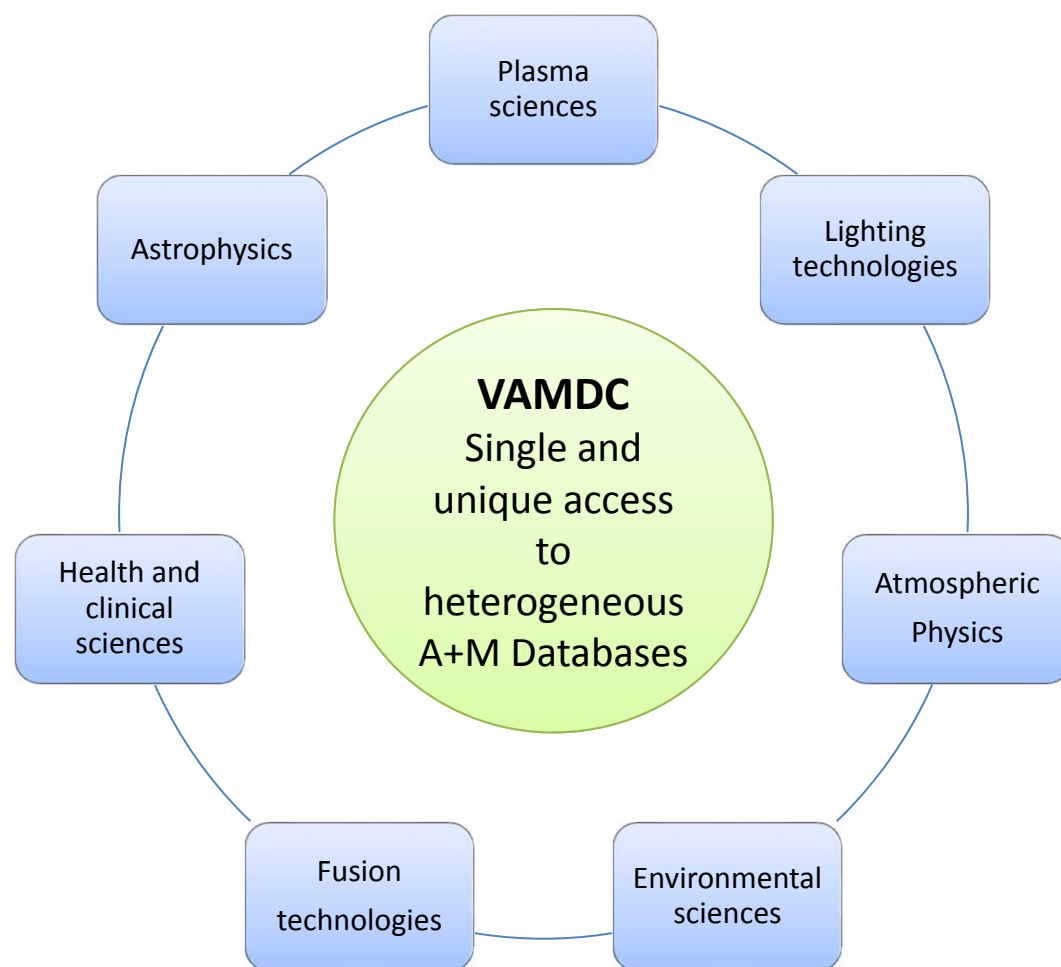
From RDA Data Citation Recommendations to new paradigms for citing data from VAMDC

C.M. Zwölf and VAMDC consortium

RDA 8th Plenary - Denver



The Virtual Atomic and Molecular Data Centre



➤ Federates 29 heterogeneous databases

<http://portal.vamdc.org/>

➤ 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.

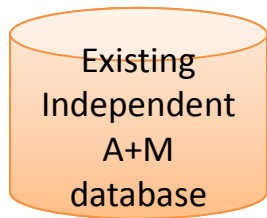
➤ The consortium is politically organized around a Memorandum of understanding (15 international members have signed the MoU, 1 November 2014)

➤ High quality scientific data come from different Physical/Chemical Communities

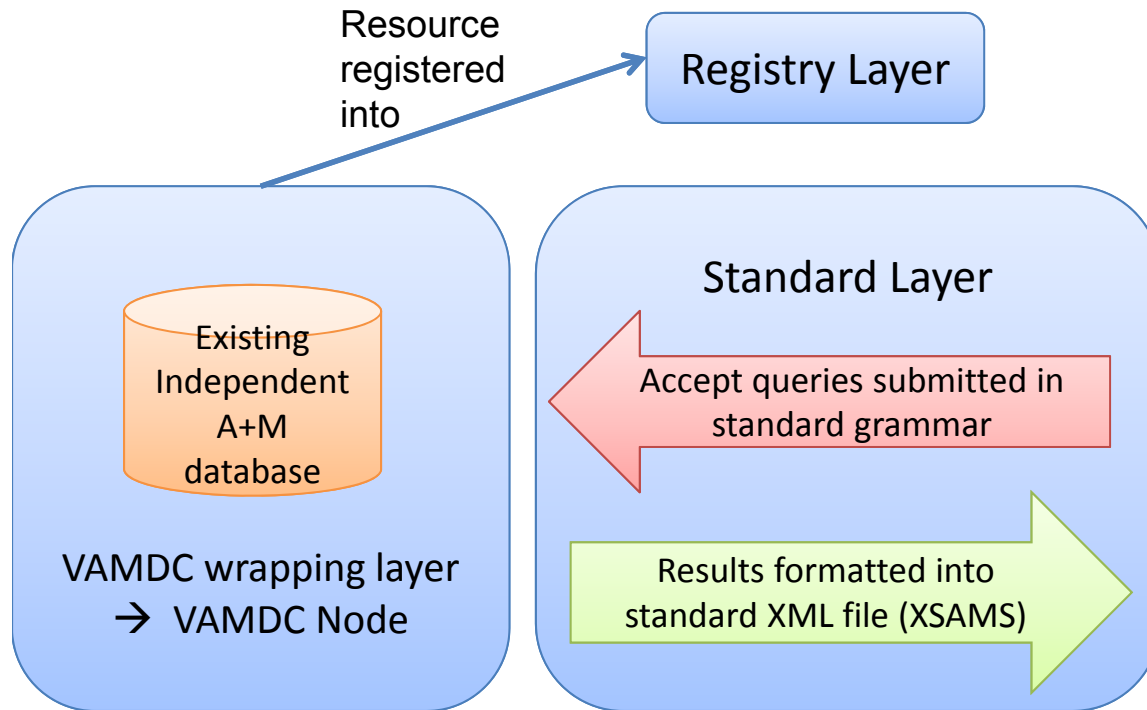
➤ Provides data producers with a large dissemination platform

➤ Remove bottleneck between data-producers and wide body of users

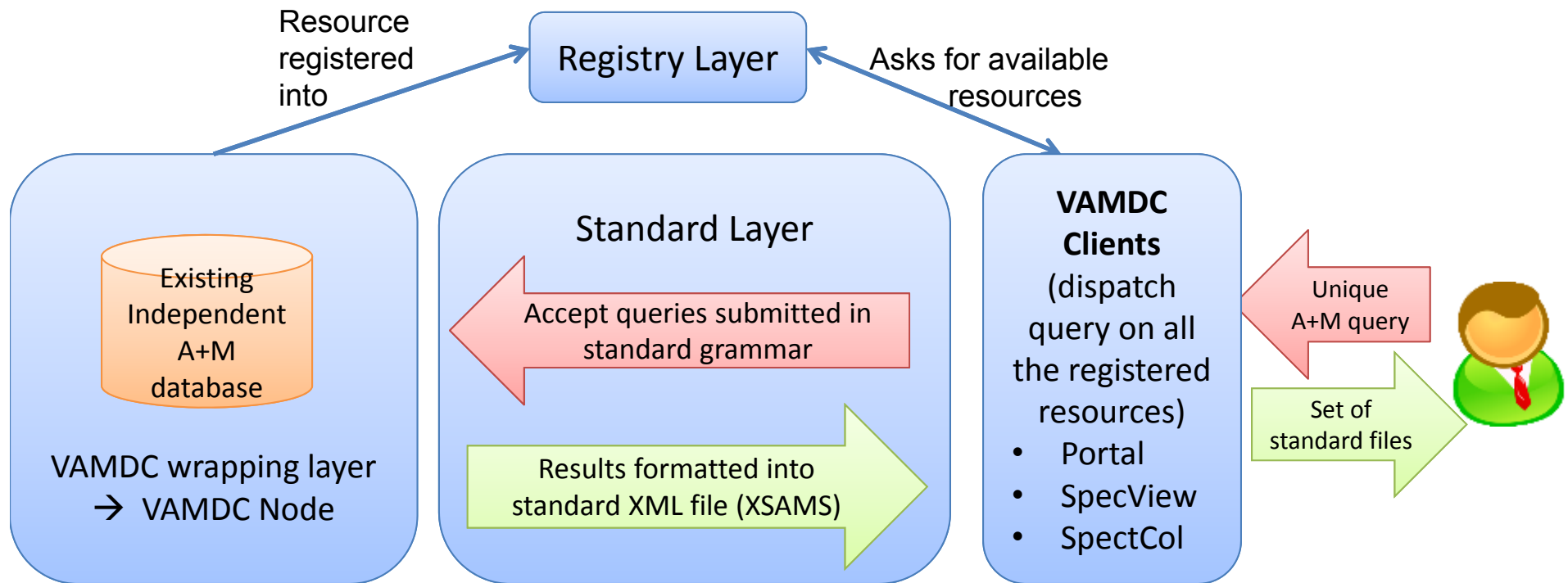
The VAMDC infrastructure technical architecture



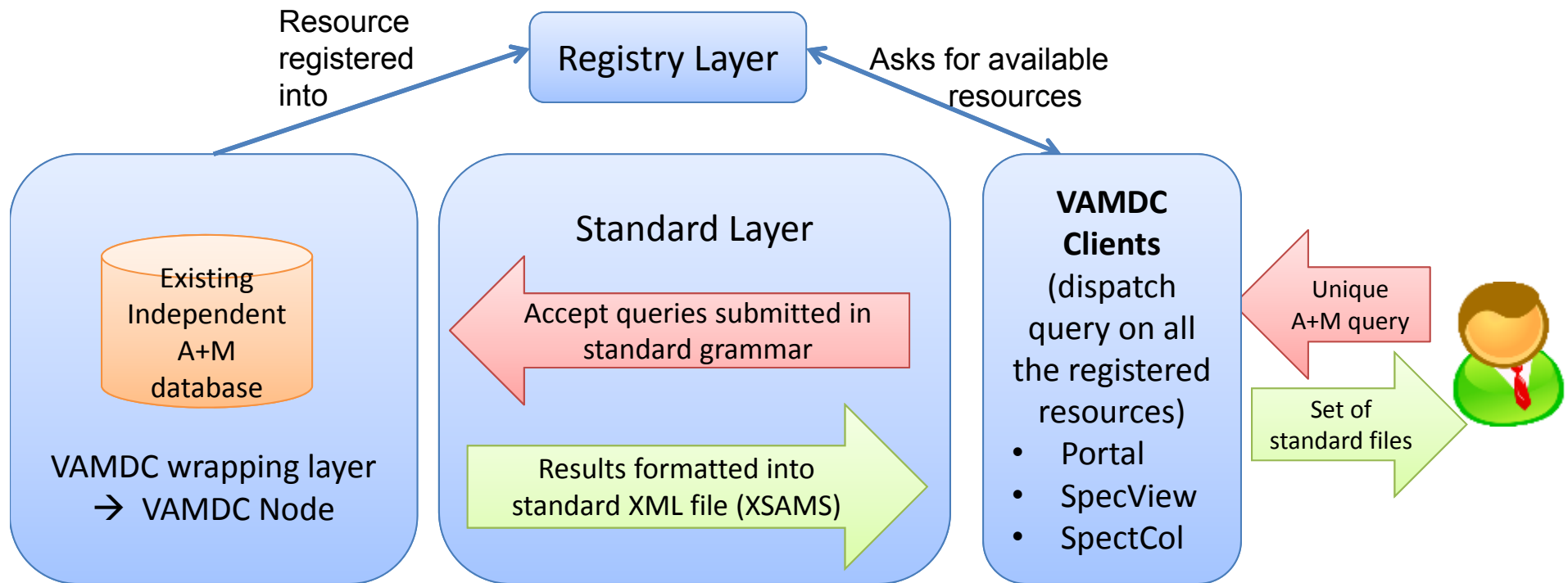
The VAMDC infrastructure technical architecture



The VAMDC infrastructure technical architecture



The VAMDC infrastructure technical architecture



- VAMDC is agnostic about the local data storage strategy on each node.
- Each node implements the access/query/result protocols.
- There is no central management system.
- Decisions about technical evolutions are made by consensus in Consortium.

➤ It is both technical and political challenging to implement the WG recommendations.

Let us implement the recommendation!!

The problem is more **anthropological** than technical...

Tagging and versioning data

What does it really mean *data citation*?

Let us implement the recommendation!!

The problem is more **anthropological** than technical...

Tagging and versioning data

What does it really mean *data citation*?

Let us implement the recommendation!!

The problem is more **anthropological** than technical...

Tagging and versioning data

We see technically how to do that

Ok, but What is the data granularity for tagging?

Naturally it is the dataset (A+M data have no meaning outside this given context)

But each data provider defines differently what a dataset is.

What does it really mean *data citation*?

Let us implement the recommendation!!

The problem is more **anthropological** than technical...

Tagging and versioning data

We see technically how to do that

Ok, but What is the data granularity for tagging?

Naturally it is the dataset (A+M data have no meaning outside this given context)

But each data provider defines differently what a dataset is.

What does it really mean *data citation*?

Everyone knows what it is!

Yes, but everyone has its own definition

RDA → cite databases record or output files.
(an extracted data file may have an H-factor)

VAMDC → cite all the papers used for compiling the content of a given output file.

Let us implement the recommendation!!

Implementation will be an overlay to the standard / output layer, thus independent from any specific data-node

Tagging versions of data

Two layers
mechanisms

1 → Fine grained granularity:

Evolution of XSAMS output
standard for tracking data
modifications

2 → Coarse grained granularity:

At each data modification to a
given data node, the version of
the Data-Node changes

With the **second mechanism** we know that something changed : in other words, we know that the result of an identical query may be different from one version to the other. The detail of which data changed is accessible using the **first mechanisms**.

Let us implement the recommendation!!

Implementation will be an overlay to the standard / output layer, thus independent from any specific data-node

Tagging versions of data

Query Store

Two layers
mechanisms

1 → Fine grained granularity:

Evolution of XSAMS output standard for tracking data modifications

2 → Coarse grained granularity:

At each data modification to a given data node, the version of the Data-Node changes

Is built over the versioning of Data

Is plugged over the existing VAMDC data-extraction mechanisms.

Due to the distributed VAMDC architecture, the Query Store architecture is similar to a log-service.

With the **second mechanism** we know that something changed : in other words, we know that the result of an identical query may be different from one version to the other. The detail of which data changed is accessible using the **first mechanisms**.

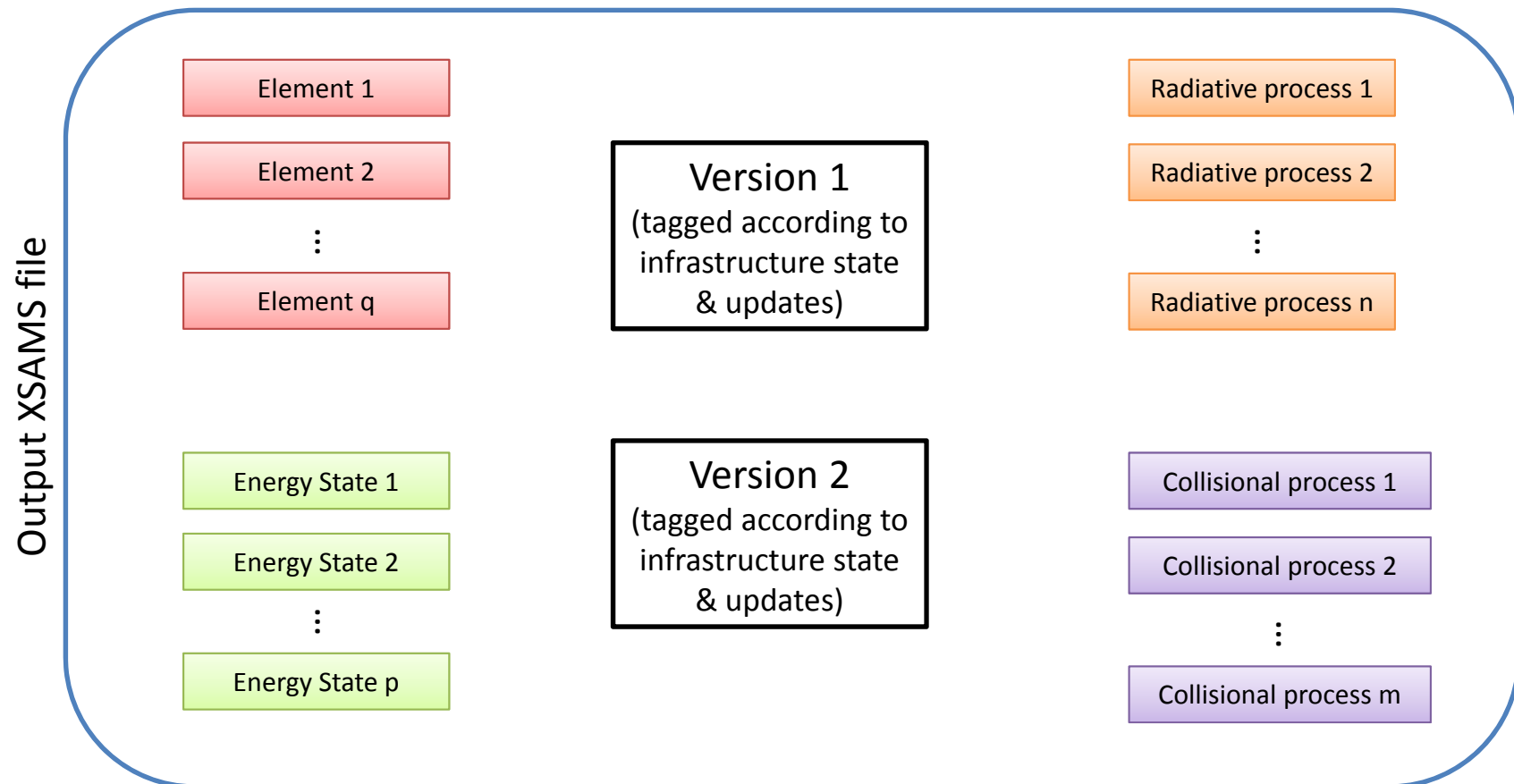
Data-Versioning: overview of the fine grained mechanisms

We adopted a change of paradigm (weak structuration):



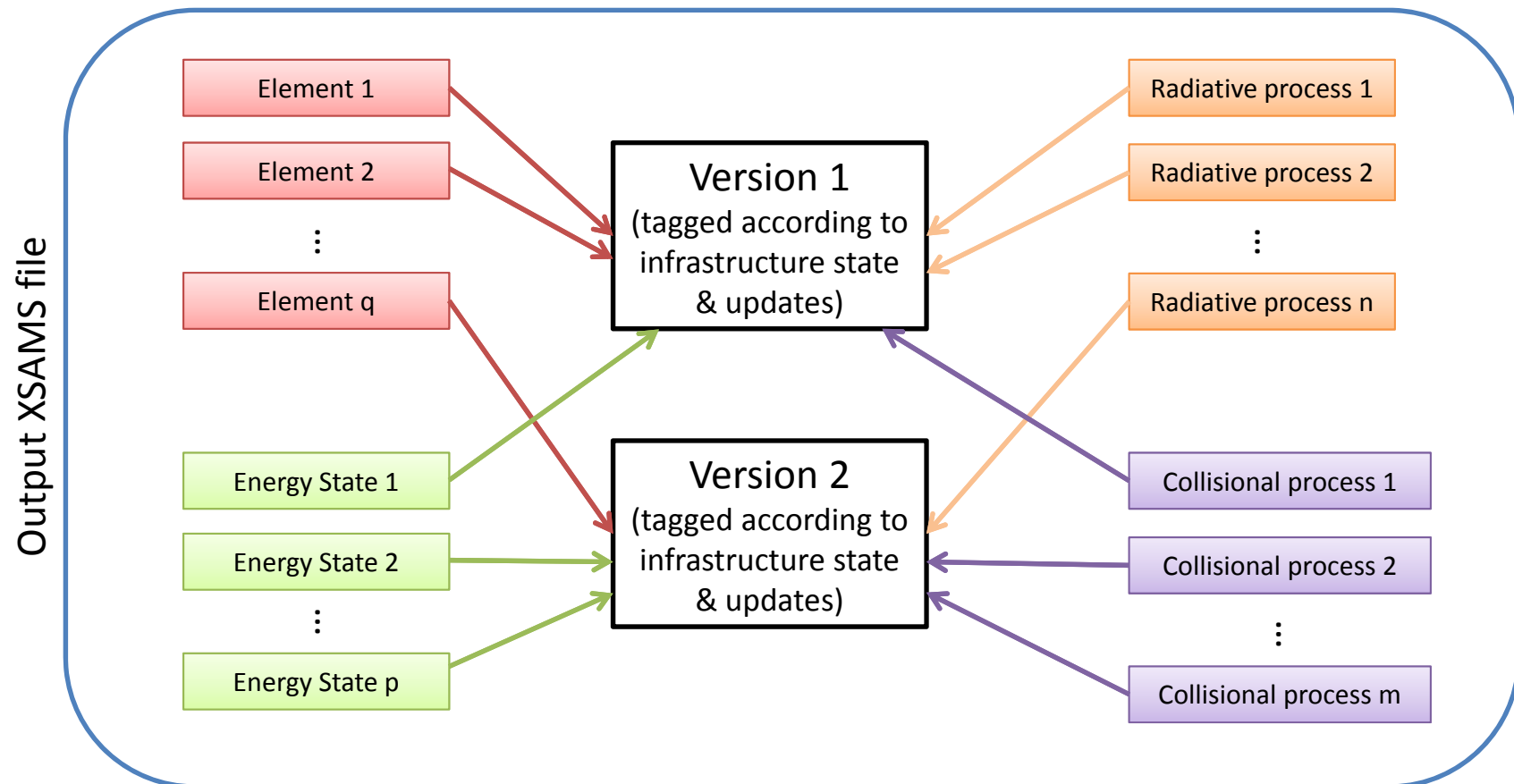
Data-Versioning: overview of the fine grained mechanisms

We adopted a change of paradigm (weak structuration):



Data-Versioning: overview of the fine grained mechanisms

We adopted a change of paradigm (weak structuration):



Data-Versioning: overview of the fine grained mechanisms

We adopted a change of paradigm (weak structuration):

This approach has several advantages:

- It solves the data tagging granularity problem
- It is independent from what is considered a dataset
- The new files are compliant with old libraries & processing programs
 - We add a new feature, an overlay to the existing structure
 - We induce a structuration, without changing the structure (weak structuration)

Data-Versioning: overview of the fine grained mechanisms

We adopted a change of paradigms:

This approach has several advantages:

- It solves the data tagging granularity problem
- It is independent from what is considered a dataset
- The new files are compliant with old libraries & processing programs
 - We add a new feature, an overlay to the existing structure
 - We induce a structuration, without changing the structure (weak structuration)

Technical details described in

New model for datasets citation and extraction reproducibility in VAMDC,

C.M. Zwölf, N. Moreau, M.-L. Dubernet,

In press *J. Mol. Spectrosc.* (2016), <http://dx.doi.org/10.1016/j.jms.2016.04.009>

Arxiv version: <https://arxiv.org/abs/1606.00405>

Let us focus on the query store:

The difficulty we have to cope with:

- Handle a query store in a distributed environment (RDA did not design it for these configurations).
- Integrate the query store with the existing VAMDC infrastructure.

Let us focus on the query store:

The difficulty we have to cope with:

- Handle a query store in a distributed environment (RDA did not design it for these configurations).
- Integrate the query store with the existing VAMDC infrastructure.

The implementation of the query store is the goal of a joint collaboration between VAMDC and RDA-Europe.

- Development started during spring 2016.
- Final product released during 2017.

Let us focus on the query store:

The difficulty we have to cope with:

- Handle a query store in a distributed environment (RDA did not design it for these configurations).
- Integrate the query store with the existing VAMDC infrastructure.

The implementation of the query store is the goal of a jointly collaboration between VAMDC and RDA-Europe.

- Development started during spring 2016.
- Final product released during 2017.

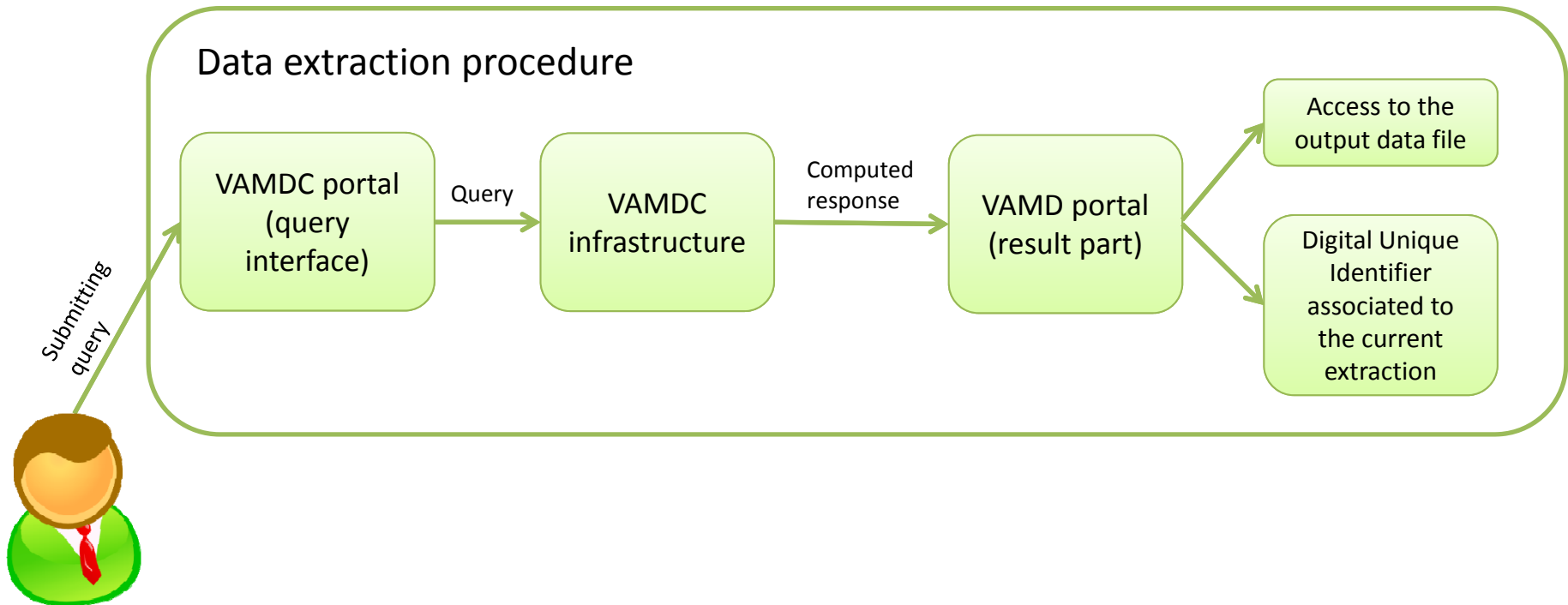
Collaboration with Elsevier for embedding the VAMDC query store into the pages displaying the digital version of papers.

Designing technical solution for

- Paper / data linking at the paper submission (for authors)
- Paper / data linking at the paper display (for readers)

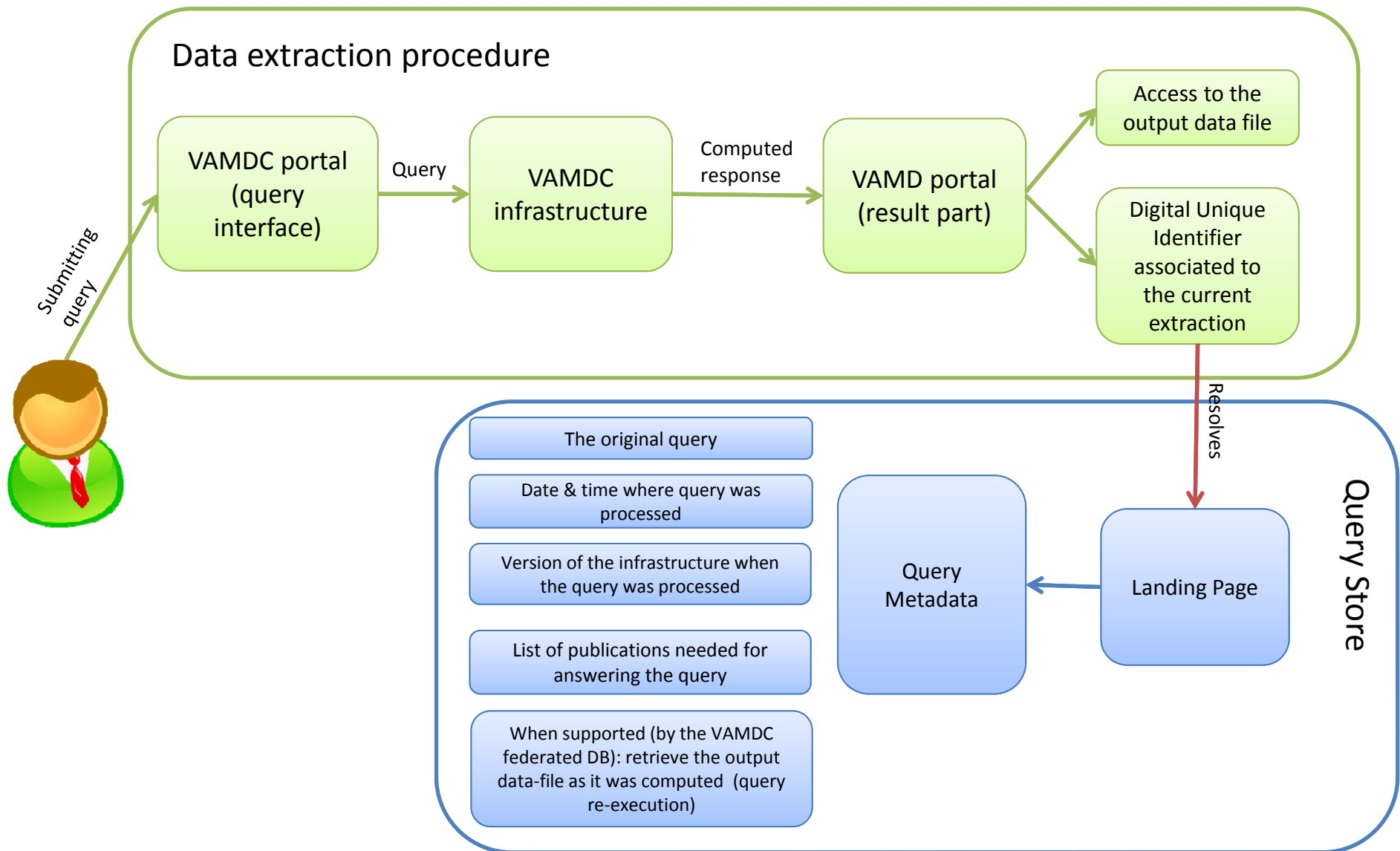
Let us focus on the query store:

Sketching the functioning – From the final-user point of view:



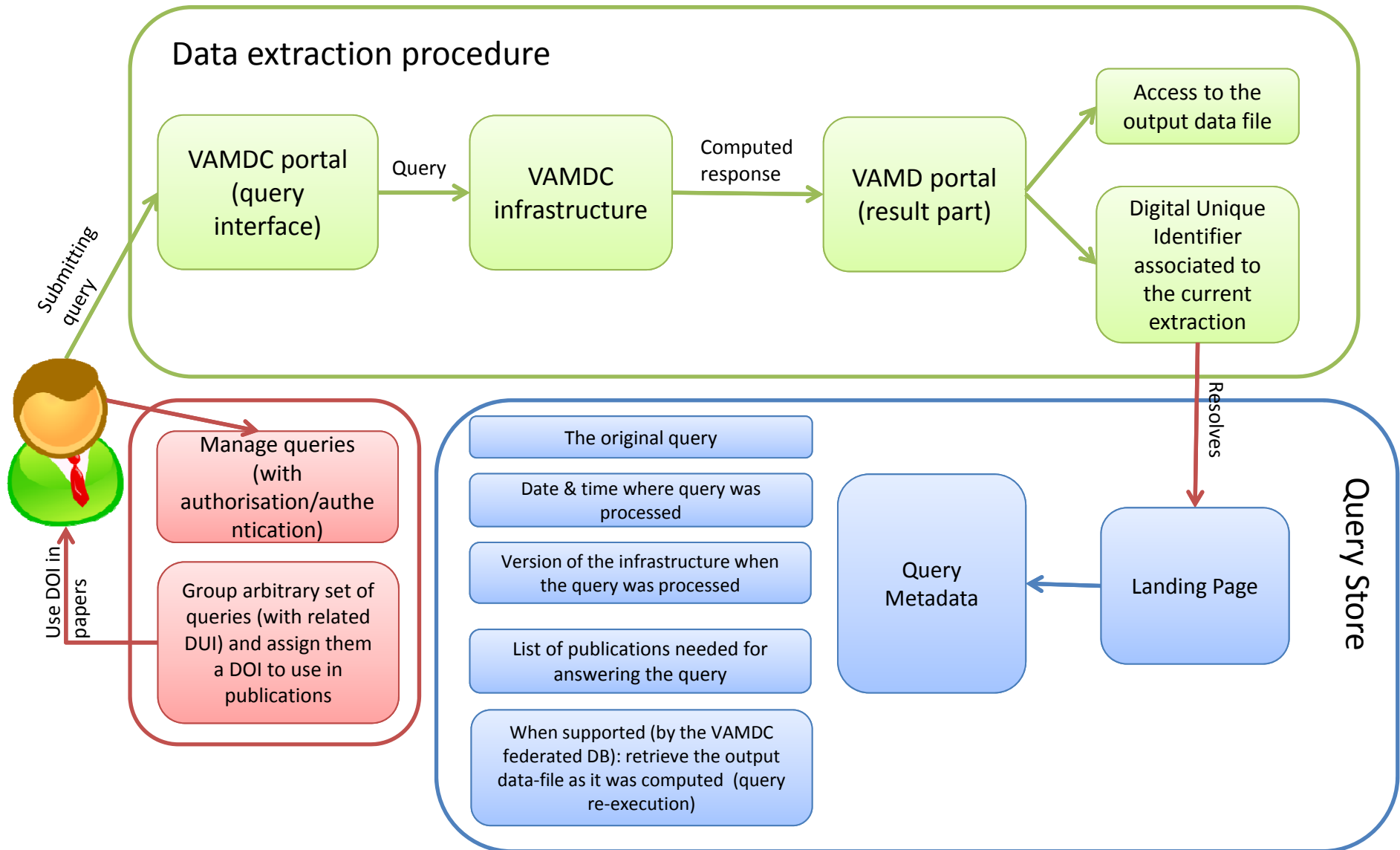
Let us focus on the query store:

Sketching the functioning – From the final-user point of view:



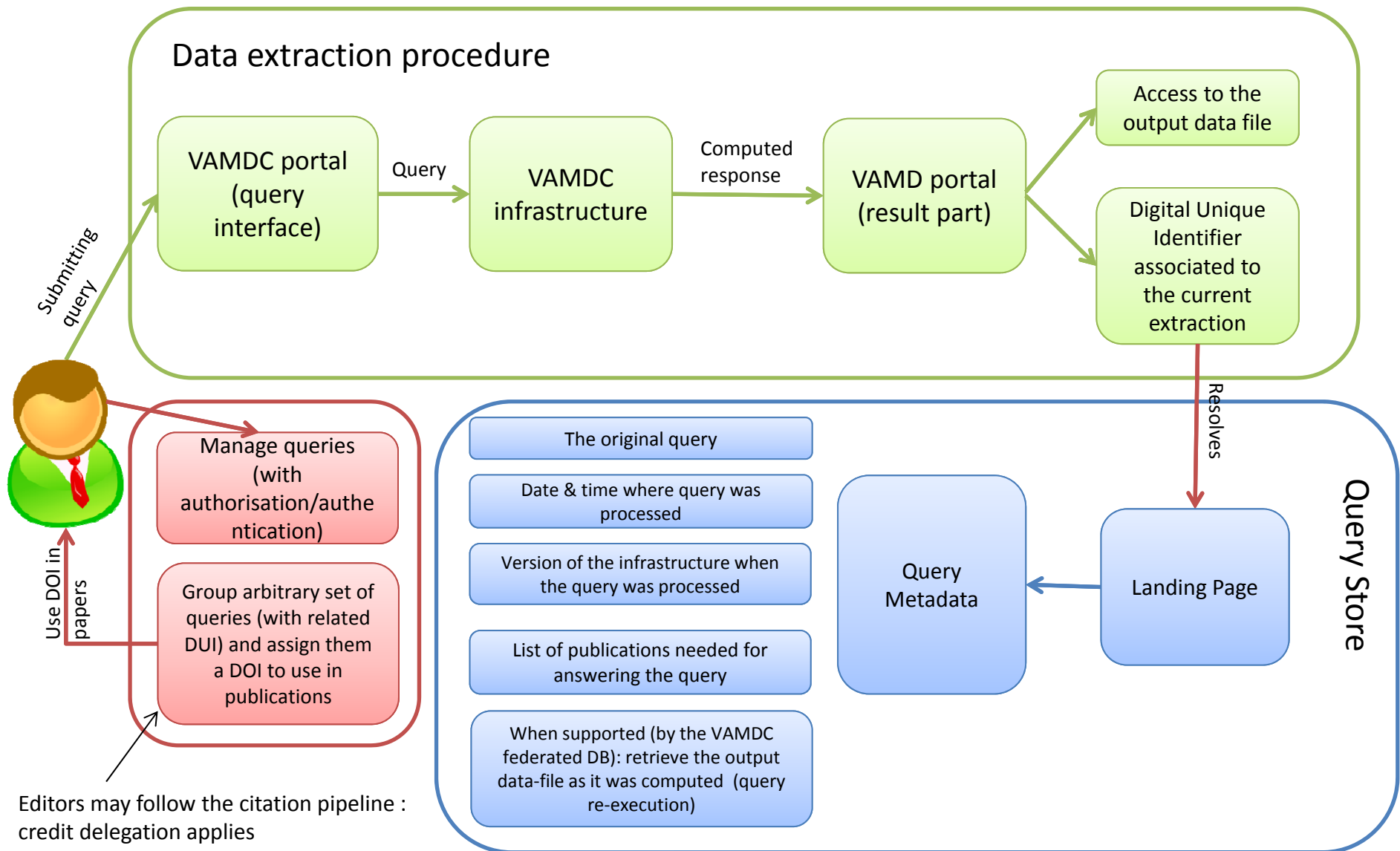
Let us focus on the query store:

Sketching the functioning – From the final-user point of view:



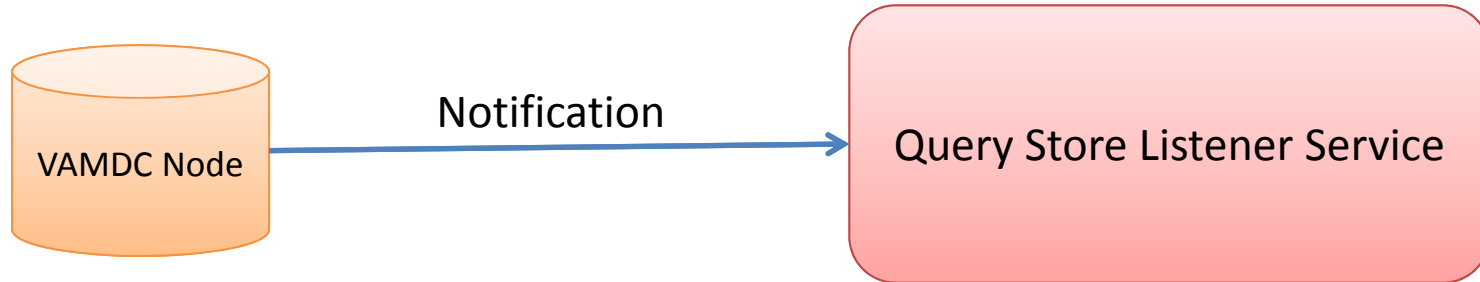
Let us focus on the query store:

Sketching the functioning – From the final-user point of view:



Let us focus on the query store:

Sketching the functioning – Technical internal point of view:

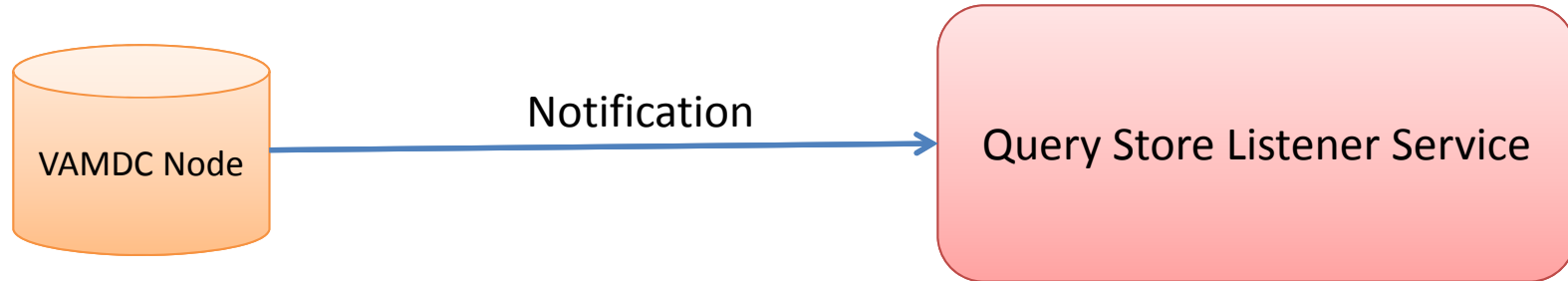


1 → When a node receives a user query, it notifies to the Listener Service the following information:

- The identity of the user (optional)
- The used client software
- The identifier of the node receiving the query
- The version (with related timestamp) of the node receiving the query
- The version of the output standard used by the node for replying the results
- The query submitted by the user
- The link to the result data.

Let us focus on the query store:

Sketching the functioning – Technical internal point of view:

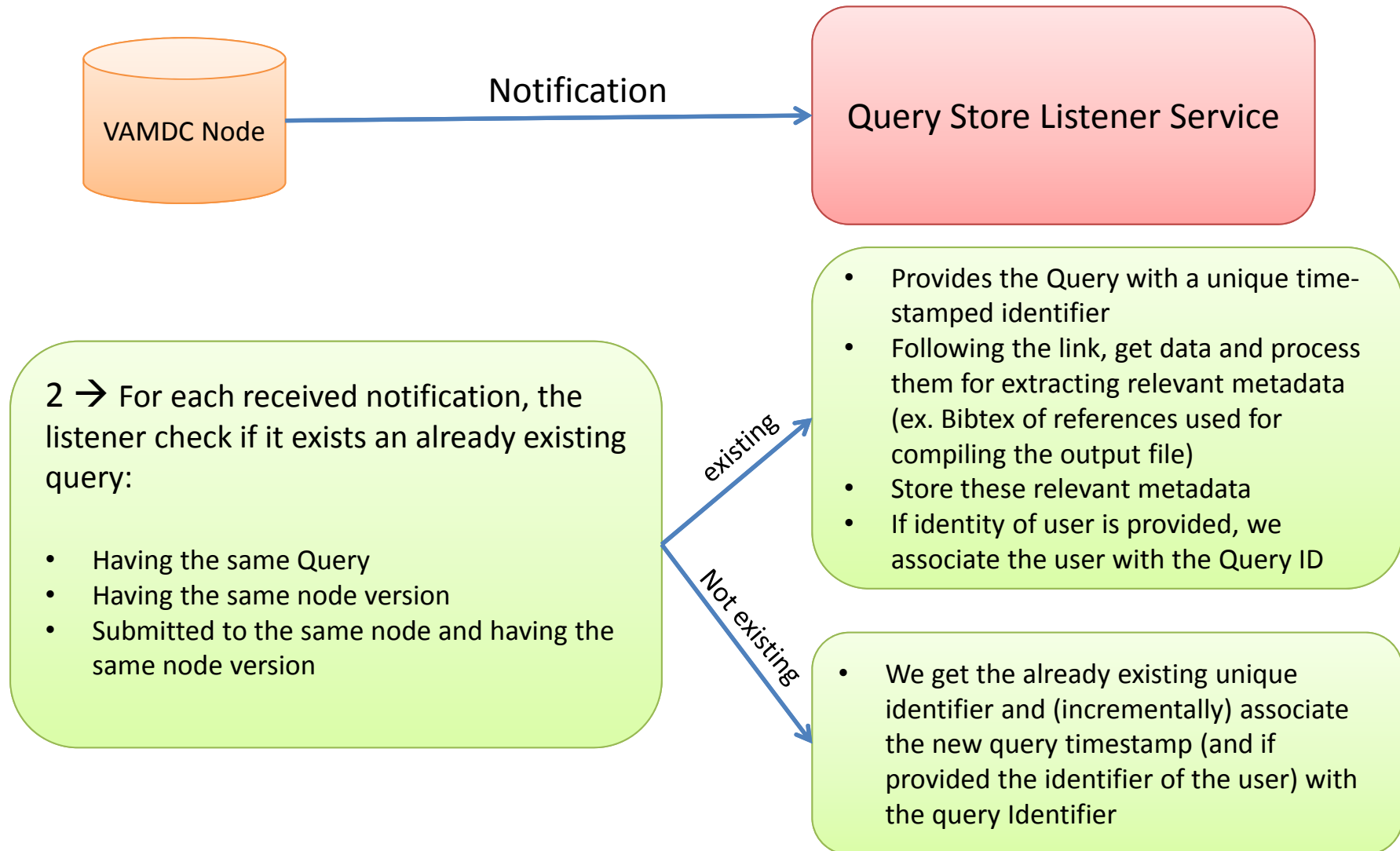


2 → For each received notification, the listener check if it exists an already existing query:

- Having the same Query
- Having the same node version
- Submitted to the same node and having the same node version

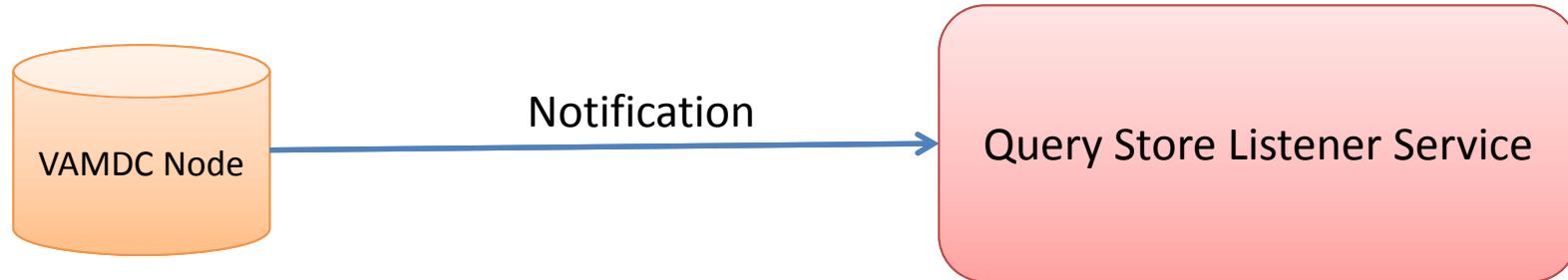
Let us focus on the query store:

Sketching the functioning – Technical internal point of view:



Let us focus on the query store:

Sketching the functioning – Technical internal point of view:



Remark on query uniqueness:

- The query language supported by the VAMDC infrastructure is VSS2 (VAMDC SQL Subset 2, <http://vamdc.eu/documents/standards/queryLanguage/vss2.html>).
- We are working on a specific VSS2 parser (based on Antlr) which should identify, from queries expressed in different ways, the ones that are semantically identical
- We are designing this analyzer as an independent module, hoping to extend it to all SQL.

Final remarks:

- Our aims:
 - Provide the VAMDC infrastructure with an operational query store
 - Share our experience with other data-providers
 - Provide data-providers with a set of *libraries/tools/methods* for an easy implementation of a query store.
 - We will try to build a generic query store (i.e. using generic software blocks)



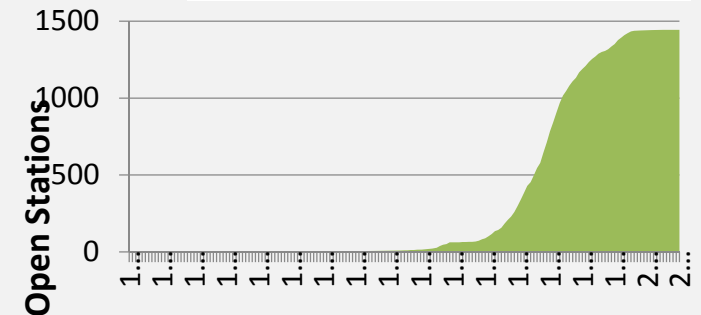
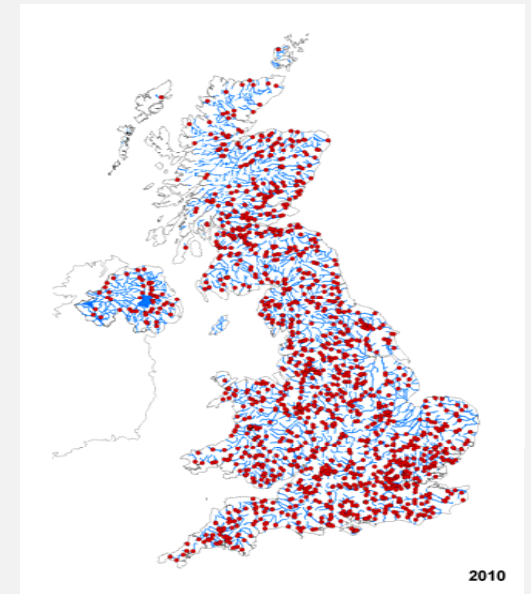
RESEARCH DATA ALLIANCE

UK Riverflow Archive
Matthew Fry
mfry@ceh.ac.uk

research data sharing without barriers
rd-alliance.org

UK National River Flow Archive

- Curation and dissemination of regulatory river flow data for research and other access
- Data used for significant research outputs, and a large number of citations annually
- Updated annually but also regular revision of entire flow series through time (e.g. stations resurveyed)
- Internal auditing, but history is not exposed to users

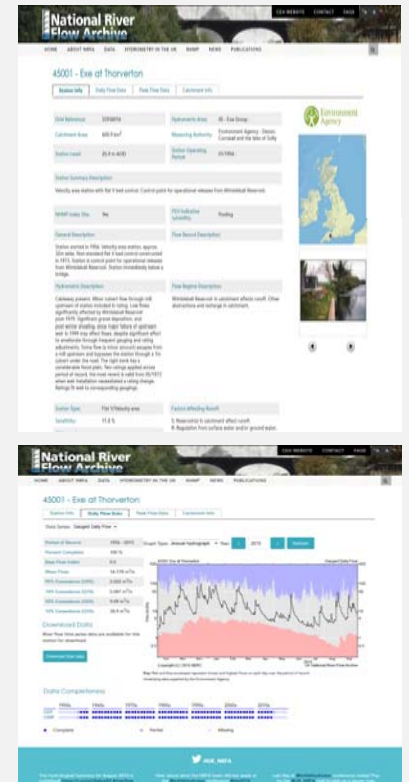


Research Data Sharing
without barriers

Global Water Information Interest Group meeting
RDA 8th Plenary, 15th September 2016, Denver

UK National River Flow Archive

- ORACLE Relational database
- Time series and metadata tables
- ~20M daily flow records, + monthly / daily catchment rainfall series
- Metadata (station history, owners, catchment soils / geology, etc.)
- Total size of ~5GB
- Time series tables automatically audited,
 - But reconstruction is complex
- Users generally download simple files
- But public API is in development / R-NRFA package is out there
- Fortunately all access is via single codeset



Research Data Sharing
without barriers

Global Water Information Interest Group meeting
RDA 8th Plenary, 15th September 2016, Denver

Our data citation requirements

- Cannot currently cite whole dataset
- Allow citation a subset of the data, as it was at the time
- Fit with current workflow / update schedule, and requirements for reproducibility
- Fit with current (file download) and future (API) user practices
- Resilient to gradual or fundamental changes in technologies used
- Allow tracking of citations in publications



Research Data Sharing
without barriers

Global Water Information Interest Group meeting
RDA 8th Plenary, 15th September 2016, Denver

Options for versioning / citation

- “Regulate” queries:
 - limitations on service provided
- Enable any query to be timestamped / cited / reproducible:
 - does not readily allow verification (e.g. checksum) of queries (R7), or storage of queries (R9)
- Manage which queries can be citable:
 - limitation on publishing workflow?



Research Data Sharing
without barriers

Global Water Information Interest Group meeting
RDA 8th Plenary, 15th September 2016, Denver

Versioning / citation solution

- Automated archiving of entire database – version controlled scripts defining tables, creating / populating archived tables (largely complete)
- Fits in with data workflow – public / dev versions – this only works because we have irregular / occasional updates
- Simplification of the data model (complete)
- API development (being undertaken independently of dynamic citation requirements):
 - allows subsetting of dataset in a number of ways – initially simply
 - need to implement versioning (started) to ensure will cope with changes to data structures
- Fit to dynamic data citation recommendations?
 - Largely
 - Need to address mechanism for users to request / create citable version of a query
- Resource required: estimated ~2 person months



Research Data Sharing
without barriers

Global Water Information Interest Group meeting
RDA 8th Plenary, 15th September 2016, Denver



RESEARCH DATA ALLIANCE

Dynamic Data Citation Interest Group Australia

Nicholas Car

GA and Dynamic Data Citation IG Australia

research data sharing without barriers
rd-alliance.org

Report on Adoption: AusAstro, TERN, NCI, GA

Nichlolas Car, Geoscience Australia

Use Case 1: TERN's Australian Ecological Knowledge and Observation System

- Ecological, ontological data management
- 3 forms of dynamic: (i) an extending data model (ii) data accretion (iii) subsetting (both in system and outside)
- How do researchers accurately describe metadata for 3 forms of dynamic data management for PIDs?
- R1, R2, – works for (ii) but not (i)
- R3, R7 – works for iii but misleading external subsetting not correctly linked to parent data

Use Case 1: TERN's Australian Ecological Knowledge and Observation System

- Ecological, ontological data management
 - 3 for (iii)
sub
 - Ho
dynamic data management for (iii):
- Narrow: dynamic scenarios not covered**
- R1, R2, – works for (ii) but not (i)
 - R3, R7 – works for iii but misleading external subsetting not correctly linked to parent data

Use Case 1: TERN's Australian Ecological Knowledge and Observation System

- Ecological, ontological data management
- 3 for (iii) sub
- Ho dyn
- R1
- R3 link

Narrow: dynamic scenarios not covered

Incomplete: recs do not cover all objects required for citation

Use Case 2: Astronomy data (ASKAP)

- CSIRO ASKAP archive data grows by ~5PB per year
- Need the ability to accurately cite static reproducible blocks of data while the collection is still growing
- “progressive” DOI allocation used
- Data scale prohibits snapshots (R1?)
- R3 – R9 irrelevant

Use Case 2: Astronomy data (ASKAP)

- CSIRO ASKAP archive data grows by ~5PB per year

- Too generic: not enough info in R1 & R2

f data

- Data scale prohibits snapshots (R1?)
- R3 – R9 irrelevant

Use Case 2: Astronomy data (ASKAP)

- CSIRO ASKAP archive data grows by ~5PB per year

Too generic: not enough info in R1 & R2

Narrow: dynamic scenarios not covered

Use Case 3: Australian National Environmental Research Data Interoperability Platform

- Large datasets TB to ~3 PB in size
- Data is accessible for in situ access, in particular subsets, via services or as file downloads if required
- Data can be reprocessed quickly: new data sets are easy to generate
- Snap-shots of data sets is not feasible (scale)
- Reproduction possible if enough processing info stored

Use Case 3: Australian National Environmental Research Data Interoperability Platform

- Latency
- Data sets, via
se
- Data can be reprocessed quickly: new data sets are easy to generate
- Snap-shots of data sets is not feasible (scale)
- Reproduction possible if enough processing info stored

Reprocessing: not handled

Use Case 3: Australian National Environmental Research Data Interoperability Platform

- Data sets, via
 - se
 - Data sets, via
 - ge
 - Sr
 - Re

Reprocessing: not handled

Incomplete: recs do not cover all objects required for citation

Use Case 4: GA's web services

- Base data subsetting via standardised, OGC, Web Services (simple dynamic)
- Base data sometimes reprocessed
- Web Service agents change

Use Case 4: GA's web services

- Base data subsetting via standardised, OGC, Web Services (simple dynamic)
- Bas
- We

Reprocessing: not handled

Use Case 4: GA's web services

- Base data subsetting via standardised, OGC, Web Services (simple dynamic)
- Bas
- We

Reprocessing: not handled

Incomplete: recs do not cover all objects
required for citation



DataCite

Martin Fenner
DataCite

research data sharing without barriers
rd-alliance.org

Dynamic Data Citation

Martin Fenner

DataCite Technical Director

<http://orcid.org/0000-0003-1419-2405>



No Discrepancies

No discrepancies to how DataCite recommends to cite data (no fragment identifiers or query strings in PID)

No adjustments to the DataCite Metadata Schema needed

PID at moment of citation

Concept of generating PID with metadata at moment of citation central to RDA recommendations, fully supported by DataCite.

Challenge Query Store

Query Store Facilities (R3):

Provide means for storing queries and the associated metadata in order to re-execute them in the future.

Focus on Principles

Data Versioning (R1): Apply versioning to ensure earlier states of data sets can be retrieved.

PID for cited subset (R8): Assign a new PID to the cited subset of a data set if either the subset is new. Otherwise, return the existing PID.

Store Metadata for cited subset (R9): Store metadata (e.g. PID, superset PID, data set description, and other).

Automated Citation Texts (R10): 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.

Landing Page (R11): Make the PIDs resolve to a human readable landing page that provides the data and metadata, including a link to the superset (PID of the data source) and citation text snippet.

Machine Actionability (R12): Provide an API / machine actionable landing page to access metadata and data.

Add recommendation

PID for whole dataset (R15)

Assign a PID to the whole dataset. Link to this PID in the metadata of every PID for cited subset.

Support multiple implementations

Consistent with ANDS recommendations

Snapshots

Relation database

File-based object store

Query store preferred implementation, but snapshots are common practice

Agenda

140

- 11:00 - Welcome and Intro, Recap
- **11:10 – Report from Pilots: Adoption activities**
- 12:20 – Q&A, 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, ...
- Dissemination of information from on-going pilots
 - Structuring: contact, descriptions, results, lessons learned
 - Outcomes: reports, slides, publications, code, discussions
 - Summary paper on pilots
- Extracting a “principles” layer from the recommendations
 - Principles plus different levels of implementing them
- Anything else? AOB? Wishes?

Thanks!

And hope to see you at the
next meeting
of the

~~WGDC~~

IGDC