



**Data Citation  
Working Group Mtg @ P12  
Nov. 5 2018, Gabarone**

**research data sharing without barriers**  
**[rd-alliance.org](http://rd-alliance.org)**

# Agenda

2

- 14:00 Introduction, Welcome
- 14:10 Short description of the WG recommendations
- 14:30 Report on new issues discussed / lessons learned
- 14:45 Reports on use cases
- 15:20 Other issues, next steps

# Welcome!

to the maintenance meeting  
of the  
WGDC

# Agenda

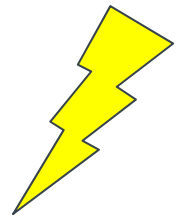
4

- 14:00 Introduction, Welcome
- 14:10 Short description of the WG recommendations
- 14:30 Report on new issues discussed / lessons learned
- 14:45 Reports on use cases
- 15:20 Other issues, next steps

# Identification of Dynamic Data

5

- Usually, 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 a specific point in time**



# Granularity of Subsets

- What about the **granularity** of data to be identified?
    - Enormous amounts of CSV data
    - 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 process

# RDA WG Data Citation



- Research Data Alliance
- WG on **Data Citation: Making Dynamic Data Citeable**
- March 2014 – September 2015
  - Concentrating on the problems of **large, dynamic (changing) datasets**
- Final version presented Sep 2015 at P7 in Paris, France
- Endorsed September 2016 at P8 in Denver, CO
- Since: support for take-up/adoption, lessons-learned  
<https://www.rd-alliance.org/groups/data-citation-wg.html>



# Dynamic Data Citation



**We have: Data + Means-of-access (“query”)**



# Dynamic Data Citation



We have: Data + Means-of-access (“query”)

**Dynamic Data Citation:  
Cite (dynamic) data dynamically via query!**

# Dynamic Data Citation



**We have:** Data + Means-of-access (“query”)

**Dynamic Data Citation:  
Cite (dynamic) data dynamically via query!**

**Steps:**

1. Data → versioned (history, with time-stamps)

# Dynamic Data Citation



**We have:** Data + Means-of-access (“query”)

**Dynamic Data Citation:  
Cite (dynamic) data dynamically via query!**

**Steps:**

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

**We have:** Data + Means-of-access (“query”)

**Dynamic Data Citation:  
Cite (dynamic) data dynamically via query!**

**Steps:**

1. Data → versioned (history, with time-stamps)

Researcher creates working-set via some interface:

2. Access → **store & 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

# Data Citation – Deployment

13

- 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

14

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

15

- Note: query string provides excellent provenance information on the data set!
- This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!
  - Data (package)
  - PID (e.g. DOI)
  - Hash value
  - 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

16

- Note: query string provides excellent provenance information on the data set!
- This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!
  - Data (package)
  - PID (e.g. DOI)
  - Hash value
  - Recommended citation text (e.g. PID/EX)
- PID resolves
  - Provides details
  - Option to return
- 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 – Output

- 14 Recommendations grouped into 4 phases:

- 2-page flyer

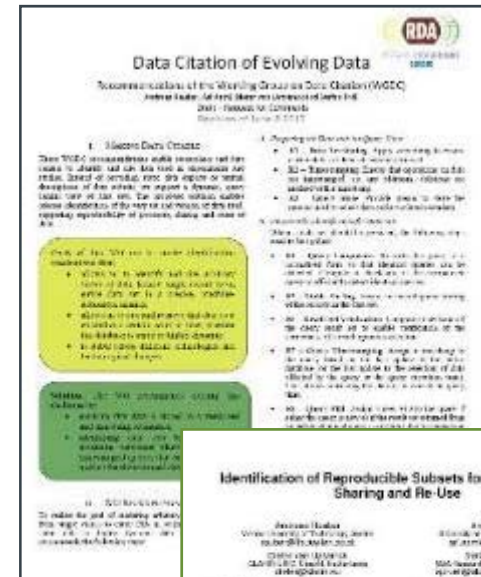
<https://rd-alliance.org/recommendations-working-group-data-citation-revision-oct-20-2015.html>

- More detailed report: Bulletin of IEEE TCDL 2016

[http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016\\_paper\\_1.pdf](http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf)

- Adopter's presentations, webinars and reports

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



- Series of Webinars presenting implementations
  - Recordings, slides, supporting papers
  - <https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html>
  - **Automatically generating citation text from queries (Recommendation 10) for RDBMS and XML data sources**
  - Implementing of the RDA Data Citation Recommendations by the **Climate Change Centre Austria (CCCA) for a repository of NetCDF files**
  - Implementing the RDA Data Citation Recommendations for **Long-Tail Research Data / CSV files**
  - Implementing the RDA Data Citation Recommendations in the **Distributed Infrastructure of the Virtual and Atomic Molecular Data Center (VAMDC)**
  - Implementation of Dynamic Data Citation at the **Vermont Monitoring Cooperative**
  - Adoption of the RDA Data Citation of Evolving Data Recommendation to **Electronic Health Records**

# Data Citation – Recommendations

19

## 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



## ■ *Benefits*

- Allows **identifying, retrieving and citing the precise data subset** with minimal storage overhead by only storing the versioned data and the queries used for extracting it
- Allows retrieving the data both **as it existed** at a given point in time as well as the **current view** on it, by re-executing the same query with the stored or current timestamp
- It allows to cite even an **empty set!**
- The query stored for identifying data subsets provides valuable **provenance data**
- Query store collects **information on data usage**, offering a basis for data management decisions
- **Metadata** such as checksums support the verification of the correctness and **authenticity** of data sets retrieved
- The same principles work for **all types of data**

# Agenda

21

- 14:00 Introduction, Welcome
- 14:10 Short description of the WG recommendations
- 14:30 Report on new issues discussed / lessons learned
- 14:45 Reports on use cases
- 15:20 Other issues, next steps

# Standardization

22

*No news so far*

- RDA applied for WGDC recommendations to become **ICT Technical Specification: TS5 RDA Data Citation of Evolving Data**
- European Multi Stakeholder Platform (MSP) has positively assessed the compliance of these RDA technical specifications in Dec. 2017
- It recommended that these would be officially acknowledged as ICT Technical Specifications and listed for referencing in public procurement
- Official approval pending, keep a watch on: [https://ec.europa.eu/growth/industry/policy/ict-standardisation/ict-technical-specifications\\_en](https://ec.europa.eu/growth/industry/policy/ict-standardisation/ict-technical-specifications_en)

# New contacts

23

- OpenEO is implementing the recommendations for earth observation data -> update during this meeting
- H2020 project discussing adoption of the recommendations for medical data sharing -> planning implementation
- Meeting with Ocean Networks Canada to discuss options for implementing the recommendations in Jan 2018 -> update during this meeting
- Europeana considering a pilot to implement this functionality

## Q&A: R7: Query Timestamping – Distributed Settings <sup>24</sup>

### Distributed Setting

- No need for synchronized timestamps across nodes
- Each node keeps local time
- Solution with one central query store (master node):
  - Master node distributes queries
  - Distributed nodes return query result with local execution timestamp
  - Master stores timestamps per node where response received
- Solution with individual query stores
  - Distributed nodes store own query and timestamps, return their PIDs
  - Central/original query processing node stores query ids of distributed nodes
  - Central node only aggregator



# Q&A: R7: Query Timestamping – Semantic Versioning<sup>25</sup>

## Why timestamps, why not semantic versioning

- Some prefer to use semantic c versioning (minor/major updates that do not / do change behaviour/interface)
  - Advantage: version number indicates relationship btw. versions
  - Disadvantage:
    - Something that was expected to be a not-changing update may turn out to induce changes / side-effects later-on
    - With data, “minor” updates are hard to think of: changing a typo may result in a record being found / not found by a query, encoding changes may break subsequent processing pipelines
    - Different semantics / types of use across different communities
- Recommendation
  - No semantic in identifier (mantra!)
  - Keep identification (version timestamp) and semantics separate
  - Semantic version number in addition to timestamp

- **Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing and sharing the data.**  
**Include the PID in the citation text snippet.**
- **2 PIDs!**
  - **Superset:** the “database” and it’s holder (repository, data center)
    - Changing / evolving
  - **Subset:** based on the query
    - Static / fixed (but: may be retrievable at state of later point in time)
  - Accumulate credits for / trace usage of subset and (dynamic) data collection/holder
  - Similar to article in journal/proceeding series

Suggested citation  
text:

Stefan Proell (2015) "Austria Facts" created at 2015-10-07 10:51:55.0, PID [ark:12345/qmZi2wO2vv]. Subset of CIA: "The CIA WorldFactbook", PID [ark:12345/cLfH9FjxnA]

# Agenda

27

- 14:00 Introduction, Welcome
- 14:10 Short description of the WG recommendations
- 14:30 Report on new issues discussed / lessons learned
- 14:45 Reports on use cases
  - Ocean Networks Canada; Reyna Jenkyns
  - Deep Carbon Observatory: Mark Parson
  - OpenEO: Tomasz Miksa
  - River Flow Archive: Matthew Fry
  - VAMDC: (Carlo Maria Zwölf)
  - Climate Change Centre Austria: (Chris Schubert)
- 15:20 Other issues, next steps



# Ocean Network Canada

Reyna Jenkins

research data sharing without barriers  
[rd-alliance.org](http://rd-alliance.org)

WORLD-LEADING DISCOVERIES AT A CRITICAL TIME

OCEAN  
NETWORKS  
CANADA

## MINTED Data Citations Project Kickoff and Overview

Reyna Jenkyns ([reyna@uvic.ca](mailto:reyna@uvic.ca))

RDA Data Citations Session, International Data Week

Gaborone, Botswana, 2018-11-05

# BRITISH COLUMBIA INFRASTRUCTURE



PACIFIC OCEAN

## CANARIE RDM Funding

- Launch of new Research Data Management funding initiative
- Community consultation occurred in February 2018
- Aims to support FAIR Data Principles and contribute towards an emerging National Data Services Framework
- Defined priority areas:
  - Enriching (Meta)data and Discovery
  - Federated Repositories / Interoperability
  - Domain-Specific Repositories
  - Data Deposit and Curation
  - Preservation
  - Persistent IDs / Citability
  - Data Access and Analytics
  - Data Privacy and Security
- Up to \$2.7 million (CAD) for selected projects
- proposals submitted in June, awarded in August and started in October 2018, complete in March 2020

canarie



# MINTED: Making Identifiers Necessary to Track Evolving Data

- Implement dynamic data citations, applying 14 recommendations from this RDA WG output to the greatest extent possible (site visit from Andreas Rauber in 2018-01 information exchange as precursor to project)
- Improve provenance, versioning, and ISO 19115 metadata records as they relate to data citation framework
- Utilize DataCite Canada membership for access to services for registering datasets
- Introduce ORCIDs for dataset contributors and user accounts, leveraging ORCID-CA frameworks and advice
- Deliver citation text provision service and a citation resolver services to National Data Services Framework
- As a member of the World Data System (WDS), adhere to the new CoreTrustSeal data repository certification requirement for Data Discovery and Identification (R13), such that users can discover and refer to data in a persistent way through proper citation
- Participate in FREYA ambassadorship program
- Consult with RDA Data Versioning, Provenance Patterns and new Data Granularity Working Groups for relevant expertise



# MINTED: Making Identifiers Necessary to Track Evolving Data

- ONC data are very dynamic due to continually accumulating data streams, data reprocessing and data product code versioning
- Highly heterogeneous data – fixed and mobile platforms, instrument types, data formats and processing levels, real-time vs autonomous
- many building blocks already exist (but more to go):
  - local identifiers and metadata for individual data queries,
  - software versioning,
  - metadata history tables,
  - reprocessing records,
  - archived file metadata (timestamping, history of changes –due to manual fixes or re-generation of derived data products, etc)
  - parser modification history
  - data agreement attributions (using ISO 19115:2014 terms) and restriction framework for third party data partners

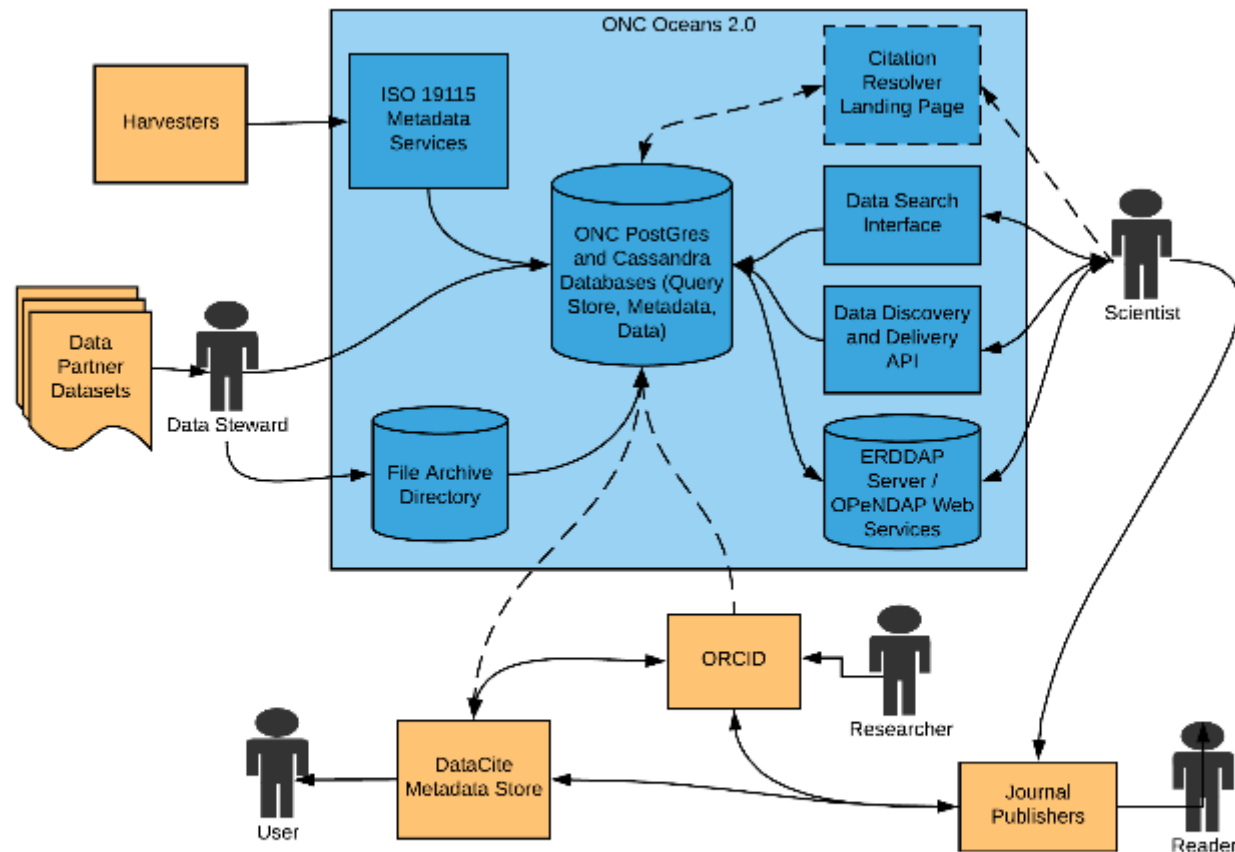
# MINTED: Making Identifiers Necessary to Track Evolving Data

- Challenges in dataset granularity decisions for attributing DOIs - leaning towards
  - a per instrument deployment on a fixed or mobile platform
  - using PROV to isolate processing level (e.g., different data products from raw to derived) that warrant distinct identifier
  - aggregating deployments over time of same device category on same platform and/or all devices on a platform over same time frame (supporting metadata already exists in Oceans 2.0) ...or even user-selected dataset “collection” to streamline citation
- But how to handle device systems?
  - Where data from multiple instruments is combined in its rawest form (e.g., Axys Watchman 500 buoy data)
  - Multiple instruments operate together as a system, but have separate data products (e.g., camera, lights, pan/tilt)
  - Data is fused in derived data products

# MINTED: Making Identifiers Necessary to Track Evolving Data

- Many elements of data framework contribute to versioning ...requires
  - an aggregated versioning solution (e.g., data product code, derivation formula changes, metadata updates, data corrections, data quality flag evolutions),
  - excluding changes that do not (noteably) impact actual dataset content (e.g., technology migrations, minor metadata changes like spelling errors, etc).
- Appropriate citation text and resolver landing page features
- Linkages to same or related data in some cases at other institutions (R2R, IRIS, OTN)?

# MINTED: Making Identifiers Necessary to Track Evolving Data



**System architecture description:** The ONC Oceans 2.0 system (in blue), and third party sources and applications (in orange). Dotted lines indicate aspects that need to be added, while all ONC components would be modified. The ONC components can be directly controlled via the project, with expected modifications to include a new data model and tables within the database, additional web services, integration of third party APIs, and data citation features.

WORLD-LEADING DISCOVERIES AT A CRITICAL TIME

**OCEAN  
NETWORKS  
CANADA**

## THANK YOU!

Ocean Networks Canada is funded by the Canada Foundation for Innovation, Government of Canada, University of Victoria, Government of British Columbia, CANARIE, and IBM Canada.

 @ocean\_networks  OceanNetworksCanada visit: [oceannetworks.ca](http://oceannetworks.ca)



**Deep Carbon Observatory Adoption of  
RDA Recommendations  
Ahmed Eleish, Brenda Thomson, Mark  
Parsons, Peter Fox**

**research data sharing without barriers**  
**[rd-alliance.org](http://rd-alliance.org)**



Rensselaer

# DeepCarbon Observatory adoption of Dynamic Data Citation

Ahmed Eleish, Winona Schroeer-Smith, Brenda Thomson, Shweta Narkar, Mark Parsons, Kathy Fontaine, John Erickson, Peter Fox

International Data Week  
Gaborone, Botswana  
5 November 2018



# Roles

---



- Ahmed - Graduate Student, solution design and architecture
- Brenda - Graduate Student, Analyst
- Shweta - Graduate Student, Backend hardware and software
- Winona - Undergraduate Student, Full-stack software developer
- Mark - Research scientist, project lead, decoder of RDA terminology, scope of recommendations, process and greater landscape of adoption
- Kathy - co-PI, strategic, governance, technical, and schedule guidance
- John - co-PI, knowledge graph consultant
- Peter - PI, DCO portal architect, hires smart and talented people



DCO

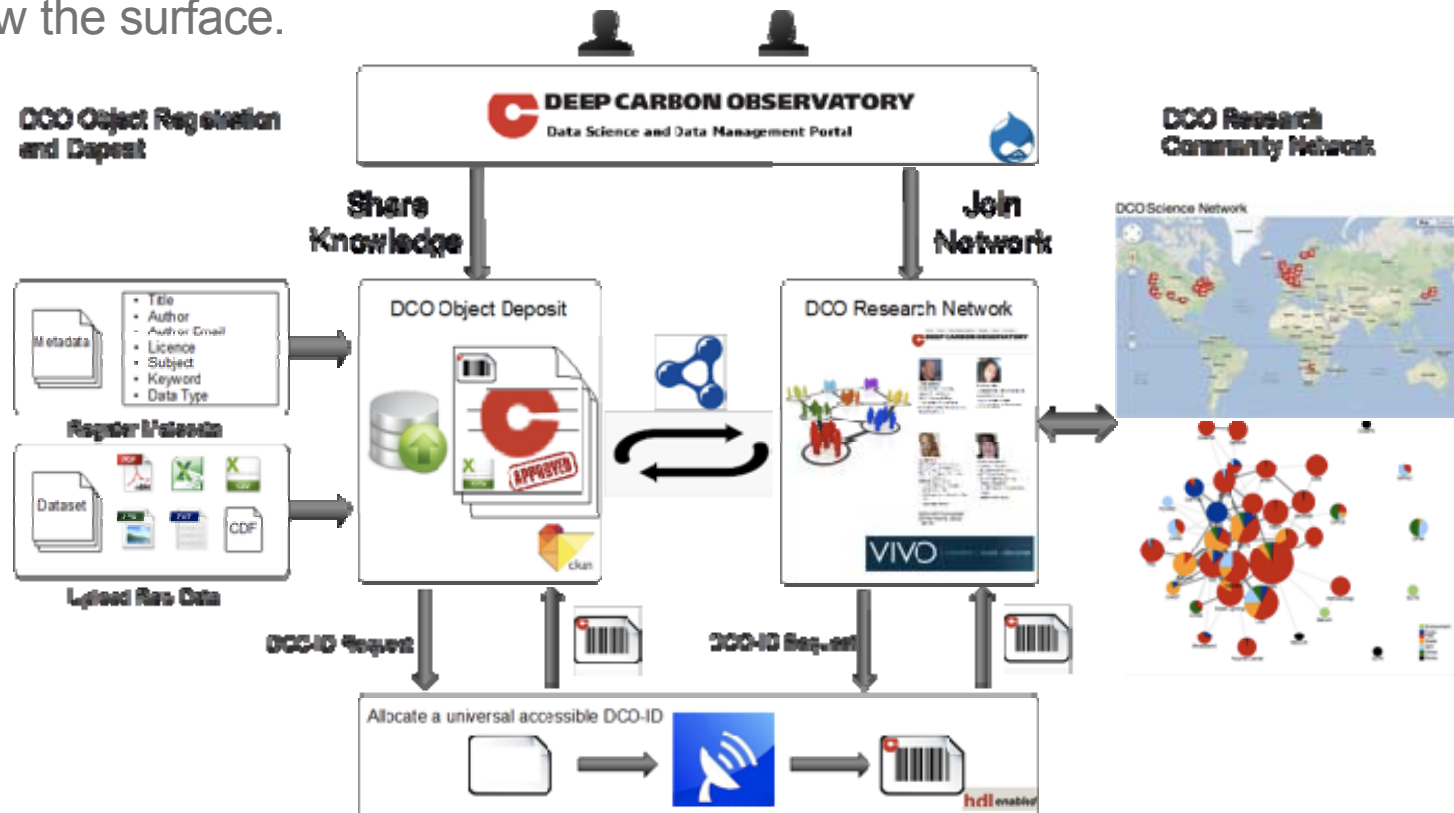
---



# Environment

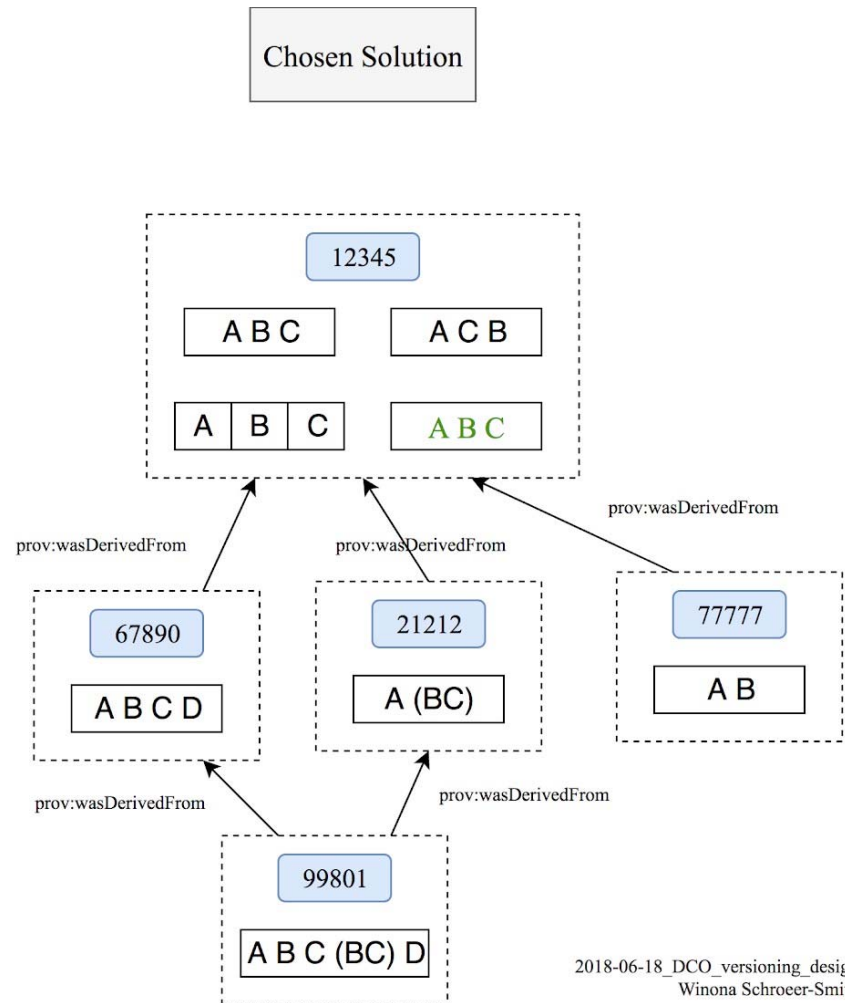


- Resource access portal for more than 1,000 diverse researchers from >40 countries studying carbon reservoirs and fluxes, extreme physics, energy, and life below the surface.



# Implementation of Recommendations 1/5

- R1 - Data Versioning
  - use prov:wasDerivedFrom to link derived datasets in a version-system-agnostic way. Saved as RDF into our knowledge database.



2018-06-18\_DCO\_versioning\_design  
Winona Schroeder-Smith

# Implementation of Recommendations

## 2/5

---



- R2 - Time Stamp: already in place
- R3 - Query Store
  - Added mechanism for research to store query which is a URI
  - Instances of dco:DCOID in VIVO updated with a prov:generatedAtTime relationship from their handle records.
  - When a user selects to store a query, the particulars of the current query, i.e. search keywords, values of filter facets, ordering, etc. are stored along with a standard representation of the current date/time for future recall.
  - When a stored query is re-run, the query is executed along with the recorded original date/time such that only records those DCO IDs minted prior to the query date are returned.
- R4 - Query uniqueness: It's a URI
- R5 - Query sorting: already in place
- R6 - Result Set Verification: already in place

# Implementation of Recommendations

## 3/5

---



- R7 - Query Timestamping
  - Implemented under R2, as we are at the collection level not the subset level.
- R8 - Query PID: in progress
- R9 - Store Query: addressed by R3
- R10 - Automated Citation Texts: in progress

# Implementation of Recommendations

## 4/5



- R11 - Landing Pag

**DEEP CARBON OBSERVATORY**      FAQs    DIRECTORY    LOGIN

DEEP CARBON SCIENCE    NEWS    COMMUNITY PORTAL    ABOUT    Search

2014

[← BACK TO DATASETS](#)

**Dataset Title**  
Noble gas isotope abundances in terrestrial fluids

**DCO ID**  
<http://dx.deepcarbon.net/11121/4317-8058-4791-8747-CC>

**Description**  
Global Data Base on isotopic composition of helium and other data for terrestrial fluids was compiled earlier by the Russian research group (B.Polyak, E.Prasolov, I.Tolstikhin, L.Yakovlev) and then was modified within the frame of this project supported by the Sloan Foundation. The DB is compiled as the Microsoft Excel Table. There are a few tables for regional records and a table for compiled total records of helium. In the dataset there is also an atlas document and a description document. Each MS-Excel file contains two spreadsheets: (1) Master sheet – contains all the compiled data (DATA) (2) References sheet – is the list of the data sources (REFS)

**Suggested citation:** Kikvadze, Olga; Prasolov, Edward; Vereina, Olga; Tolstikhin, Igor; Vetrina, Margarite; Loffe A; Yakovlev L; Polyak B. 2013. Updated 2014-01-13. Noble gas isotope abundances in terrestrial fluids. Deep Carbon Observatory Data Portal. Data set accessed on (date) at <http://dx.deepcarbon.net/11121/4317-8058-4791-8747-CC>

**DCO Community**  
Deep Energy Community

DCO Contributor	DCO ID
Kikvadze, Olga,	<a href="http://dx.deepcarbon.net/11121/2397-9343-7847-5186-CC">http://dx.deepcarbon.net/11121/2397-9343-7847-5186-CC</a>
Prasolov, Edward,	<a href="http://dx.deepcarbon.net/11121/9712-2348-4704-4389-CC">http://dx.deepcarbon.net/11121/9712-2348-4704-4389-CC</a>
Vereina, Olga,	<a href="http://dx.deepcarbon.net/11121/2486-8278-3805-1061-CC">http://dx.deepcarbon.net/11121/2486-8278-3805-1061-CC</a>
Tolstikhin, Igor,	<a href="http://dx.deepcarbon.net/11121/2199-3542-5961-4099-CC">http://dx.deepcarbon.net/11121/2199-3542-5961-4099-CC</a>
Vetrina, Margarite,	<a href="http://dx.deepcarbon.net/11121/3174-4746-5341-8386-CC">http://dx.deepcarbon.net/11121/3174-4746-5341-8386-CC</a>

**Other Contributors**  
Loffe A, Yakovlev L, Polyak B, ,

**Geographic Focus**  
World, Australia, Antarctica, Russian Federation, Africa, Americas, Former Soviet Union, Europe,

**Distribution**      **Distribution URI**  
2013-10-21T11:44      <http://info.deepcarbon.net/individual/n397>

File Name	downloadURL
Introduction_DB_2013_06_11	<a href="http://udco.tw.rpi.edu/ckan/storage/f/2013-11-01-052919/04_Introduction_DB_2013_06_11.doc">http://udco.tw.rpi.edu/ckan/storage/f/2013-11-01-052919/04_Introduction_DB_2013_06_11.doc</a>
Russia_1246	<a href="http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035811/RUSSIA_1246.xlsx">http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035811/RUSSIA_1246.xlsx</a>
Atlas_English_2013_06_10	<a href="http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035847/Atlas_English_2013_06_10.pdf">http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035847/Atlas_English_2013_06_10.pdf</a>
America_906	<a href="http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035634/America_906.xlsx">http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035634/America_906.xlsx</a>
Europe_2462	<a href="http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035730/Europe_2462.xlsx">http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035730/Europe_2462.xlsx</a>
Antarctica_018	<a href="http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035648/Antarctica_018.xlsx">http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035648/Antarctica_018.xlsx</a>
Africa_062	<a href="http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035603/Africa_062.xlsx">http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035603/Africa_062.xlsx</a>
Australia_249	<a href="http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035703/Australia_249.xlsx">http://udco.tw.rpi.edu/ckan/storage/f/2013-10-21-035703/Australia_249.xlsx</a>

# Implementation of Recommendations

## 5/5

---



- R12 - Machine Actionability — in place but we plan to extend for truly dynamic citation
- R13 - Technology Migration adopted migration policy
- R14 - Migration Verification —need to test policy and address issues exposed in R12.

# Next steps

---



- Implement final pieces
- Outreach to sister repositories
- Implementation of same infrastructure for a complex minerals network data set.





# OpenEO

**Tomasz Miksa, Bernhard Gößwein**

research data sharing without barriers  
[rd-alliance.org](http://rd-alliance.org)

# Data Citation @ OpenEO



Tomasz Miksa, Bernhard Gößwein & Andreas Rauber,  
TU Wien



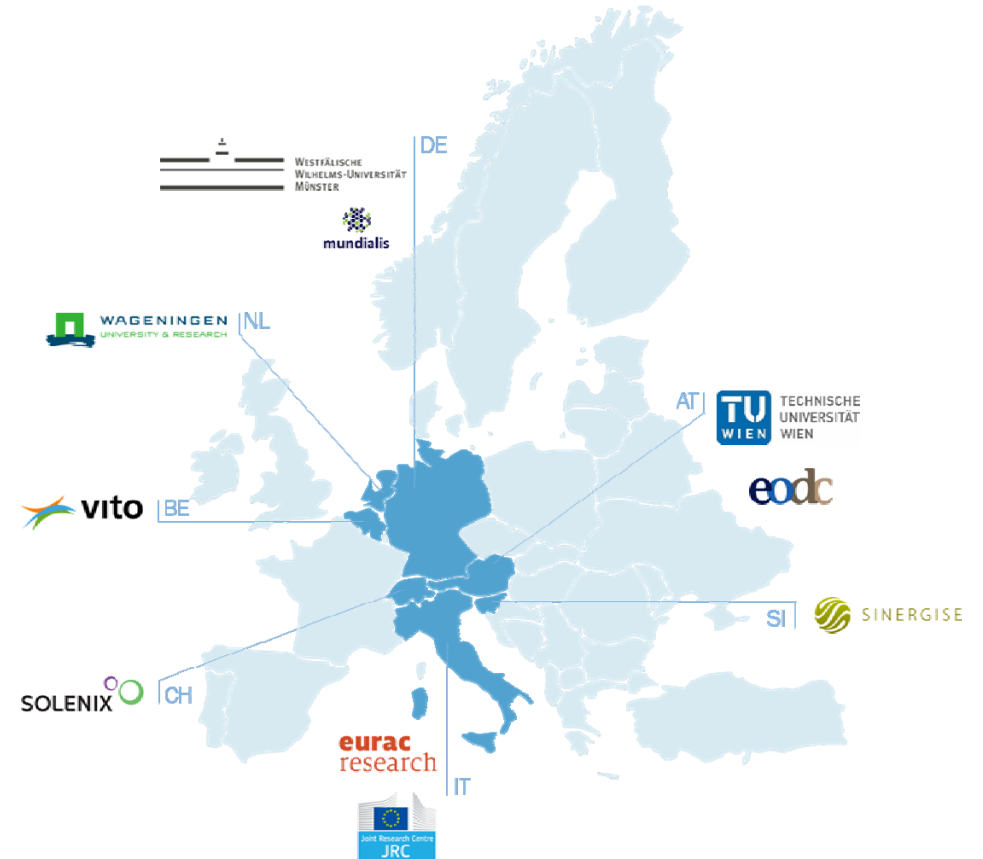
Grant agreement No 776242 | 11/5/2018



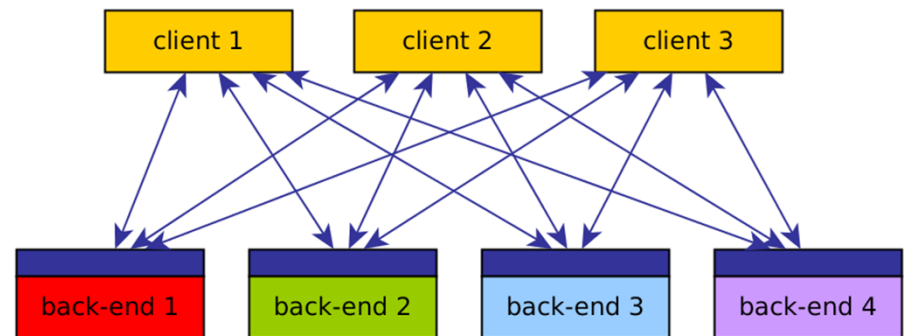
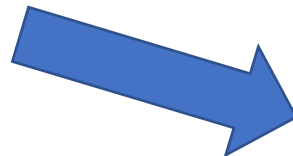
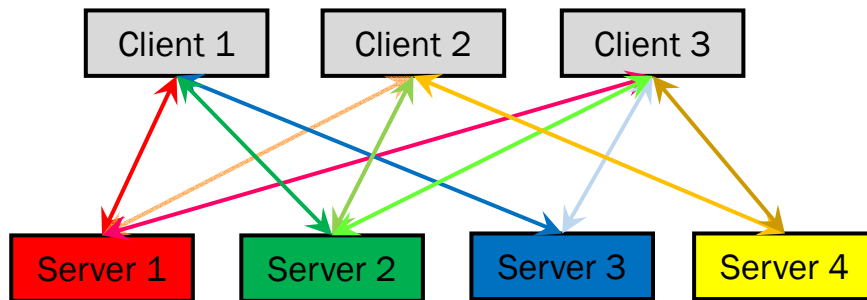
openEO | Grant agreement No 776242

# OpenEO

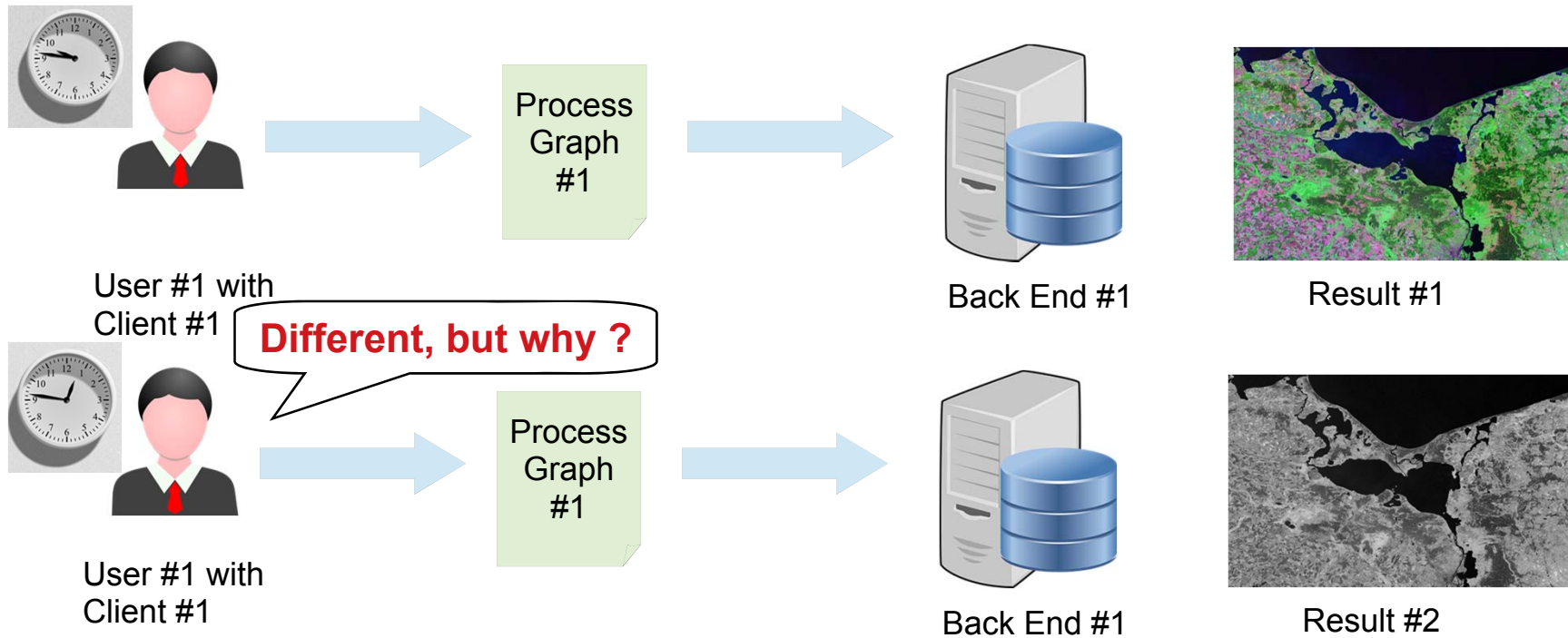
- Earth Observation
- Data
  - Too big for local processing
  - Code visits data
- Back-end operators
- Goal
  - Develop common API



# OpenEO Overview



# Data Citation – why?



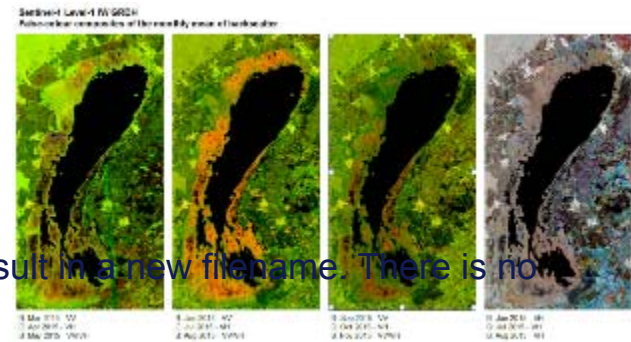
# Current state at back-ends

- European Space Agency (ESA)
  - Provides data to back-ends
  - Pre-processes data (instrument calibration)
  - Updates data

- Current situation at back-ends
  - File based data management

- Every source data has a unique path. Updates on the data result in a new filename. There is no guaranteed persistence of deprecated data objects.

- Querying the data happens through a Web API using the OGC standard CSW (see <https://csw.eodc.eu>)
  - Queries are not persisted



# Ongoing developments for data citation

- OpenEO API
  - Methods in API to provide information on data used in computation
- Back-end implementations
  - Query store
    - Modify existing CSW API used by back-ends
    - Store queries and timestamps
- Old file versions
  - May be deleted – index data still retained, query can be processed
  - API will indicate missing files, still exist at ESA – can be imported
- Data citation combined with execution context capturing improves reproducibility

<http://openeo.org/>

55



# Dynamic data citation: Implementation in the UK National River Flow Archive

Matthew Fry  
*[mfry@ceh.ac.uk](mailto:mfry@ceh.ac.uk)*

research data sharing without barriers  
[rd-alliance.org](http://rd-alliance.org)



# UK National River Flow Archive

~1500 river flow and rainfall time series + spatial data + metadata

All openly downloadable over the web (csv files, but also API)

The screenshot shows the website interface for station 7001. The header includes the site logo and navigation links. The main content area is titled '7001 - Findhorn at Shenachie' and features several tabs: Station Info, Daily Flow Data, Peak Flow Data, Trends, and Catchment Info. The 'Station Info' tab is active, displaying a table of metadata:

Grid Reference:	NH825335	Hydrometric Area:	7 - Findhorn Group
Catchment Area:	415.6 km <sup>2</sup>	Measuring Authority (local station number):	Scottish Environment Protection Agency - North (234306)
Station Level:	252.4 m AOD	Station Operating Period:	01/1960 - N/A

Below the table is a map of Scotland with a red dot indicating the station location. Further down, there are sections for 'Station Summary Description', 'NHMP Index Site', 'FEH Indicative suitability', 'General Description', 'Flow Record Description', and 'Hydrometric Description'. A small video player is visible at the bottom right of the page.

The screenshot shows the website interface for station 28052. The header is identical to the previous screenshot. The main content area is titled '28052 - Sow at Great Bridgford' and features tabs for Station Info, Daily Flow Data, Peak Flow Data, and Catchment Info. The 'Catchment Info' tab is active, displaying a 'Catchment Description' and a table of 'Catchment statistics'.

**Catchment Description:**  
Low relief agricultural catchment, primarily on Mercia Mudstone, with some Sherwood Sandstone in headwaters. Glacial gravel in valleys maintain baseflows.

Select spatial data type to view:

Statistic	Value	Unit
Minimum Altitude:	78.7	mAOD
10 Percentile:	94.3	mAOD
50 Percentile:	119.8	mAOD
90 Percentile:	162.7	mAOD
Maximum Altitude:	234.6	mAOD

Below the table is a map of the catchment area with a legend and a 'Download catchment boundary' button. The map shows the river network and the catchment boundary. A 'Transparency' slider and checkboxes for 'Show OS background?' and 'Show CEH Rivers?' are also visible.

## The dataset

- RDBMS of time series + metadata (~20M daily flow records), only 5-10GB
- Updated on an annual basis, with occasional additional interim updates
- Currently most edits are audited, but reconstruction is complex
- Many users downloading small subsets via an API / website – too many queries to log / checksum them all individually
- We would like to allow citation of a subset, but principally citation of a version
- We would also like users to be able to query older versions via API

## Solution defined

- Entire database archived on a semi-regular basis (~twice yearly)
- Copy / backup of tables is automated, basically adding suffixes to table names and moving to archive schema
- Version numbering is explicit part of this process
- Workflow defined for creating new versions at appropriate intervals, expect ~2 per year (based on our data update schedule)

## Current state of implementation

- System currently being used but versions not exposed
- We have yet to implement versioning throughout data access code, including API (but plan to)
- It has helped us to simplify our database structure
- But it has meant we have to be more careful to ensure updates are complete before release
- What will be the mechanism for citing these versions?

## Use of recommendations

Recommendations were hugely useful in doing this work!

But some not necessary, largely due to our way of working (occasional updates + mainly API access)

- R1 – Data versioning: Yes
- R2 – Timestamping: Already done, but not directly tied in to this work
- R3 – Query store facilities: No
- R4 – Query uniqueness: Not relevant (R2/3)
- R5 – Stable sorting: Not relevant (R2/3)
- R6 – Result set verification (checksums) – possibly.

## Use of recommendations

- R7 – Query timestamping: No (apart from API logs)
- R8 – Query PID: No (but we should have a version PID)
- R9 – Query store: No
- R10 – Automated citation texts: we should do!
- R11 – Landing page: we should do!
- R12 – Machine Actionability: Sure
- R13 – Technology Migration: Essential (probably more so)
- R14 – Migration verification: Yes, but we need to think about this!

## Summary

- The need to provide users with API access highlighted some issues with direct uptake of all recommendations (we can log but not checksum and archive all API calls)
- Small size of database meant version / archive of entire database was preferable option
- Because of this not all recommendations relevant
- However, could consider version of the entire database as a pre-defined “query”; we are just allowing individual sub-queries
- Recommendations were very useful



## VAMDC Query Store implementation

C.M. Zwölf, N. Moreau,

VAMDC Consortium

[carlo-maria.zwolf@obspm.fr](mailto:carlo-maria.zwolf@obspm.fr)

research data sharing without barriers  
[rd-alliance.org](http://rd-alliance.org)



### The Virtual Atomic and Molecular Data Centre

- Is a digital infrastructure federating ~30 heterogeneous autonomous databases in an interoperable way
- Is a distributed set with no central management system
- Was a pilot for the Data Citation WG since 2014 (Plenary 3): does the recommendation contain blocking point for VAMDC distributed architecture → NO blocking points!

### The recommendation was implemented from 2016 to 2017:

- Technical details for data versioning: “New model for datasets citation and extraction reproducibility in VAMDC”, C.M. Zwölf, N. Moreau, M.-L. Dubernet, <http://dx.doi.org/10.1016/j.jms.2016.04.009>
- The Query Store was implemented in collaboration with the RDA-EU3 project, during 2017.

# The VAMDC Query Store

66

- We successfully implemented a query store for the distributed asynchronous VAMDC architecture: source code at <https://github.com/VAMDC/QueryStore>
- The data-citation capabilities are deployed over 1/3 of the VAMDC databases, the other are joining → <https://cite.vamdc.eu> (~200 queries stored)
- Experience shows it is quite difficult to inculcate the “Query Store reflex” in our final users’ minds. This does not fit directly with their usual workflow.
  - We are working for a better adoption from our user community.

# The VAMDC Query Store

67

- <https://cite.vamdc.eu> (~200 queries stored)

The screenshot displays the VAMDC Query Store interface. On the left, there are navigation links for Home, Queries, and Credits. Below these, there are input fields for 'Query executed between' (08/05/2018 12:00 AM and 11/01/2018 12:00 AM) and a 'Submit' button. A section titled 'Accessed resources' lists various databases and interfaces, including SHECaSDa, MeCaSDa, RuCaSDa, Stark b, TIPMeCaSDa, TIPbase, and VALD (atoms).

Request	Accessed resource	Last execution	UUID
select species	Stark-b	2018-11-4 0:24:39	8f387172-afa0-16c8-ba0f-c885e2e950de
select species	VALD (atoms)	2018-11-3 17:42:35	1ac3bd31-9182-44d2-84bf-2212b88ab7d5
select species	TOPbase : VAMDC TAP interface	2018-11-2 16:44:42	34875a6c-4bc2-4497-9c25-44ecd39cbb0f
select species	TIPbase : VAMDC TAP interface	2018-11-2 16:44:42	f29e648a-7c00-4f7-afe5-5c1f36891a3d
select species	MeCaSDa - Methane Calculated Spectroscopic Database	2018-11-1 6:34:15	9ca8e6a8-b5d7-4f9e-a510-13cb8c4b759
select * where ( atomsymbol = ...	VALD (atoms)	2018-10-23 14:40:20	9113e23c-e38b-40e8-8b49-4289d184390f
select * where ( atomsymbol = ...	TIPbase : VAMDC TAP interface	2018-10-23 14:39:48	7a11bcbf-8e5d-4a79-b115-04c8a2a65aea
select * where ( atomsymbol = ...	TOPbase : VAMDC TAP interface	2018-10-23 14:39:45	3c99b777-0531-46bf-9e24-bd69d078f075
select * where ( atomsymbol = ...	VALD (atoms)	2018-10-23 13:35:15	1ce2c1b5-f1dc-493c-ae3a-a36a9a32cde9
select * where ( ...	TOPbase : VAMDC TAP		337b1379-

On the right side of the screenshot, there is a detailed view of a query result. It includes the 'Data source' (MeCaSDa - Methane Calculated Spectroscopic Database), 'Data source version' (2018-04-04), 'Query' (select \* where ( atomsymbol = 'H' )), 'Query identifier' (5113e23c-e38b-40e8-8b49-4289d184390f), 'Query result' (33ANS.R), 'XRAMS version' (17.07), and 'Query result downloaded on (UTC+1)' (2018-4-17 10:44:02, 2018-4-17 10:44:01, 2018-6-20 14:11:10, 2018-7-19 13:57:41, 2018-7-23 14:11:00, 2018-7-23 14:13:55, 2018-10-15 14:35:37, 2018-10-23 14:40:20). Below this, there are 'References' and 'Bibliography' sections with links to scientific papers.

research data sharing without barriers  
rd-alliance.org

## Further developments (1/2)

68

- We interlinked our Query Store with the Zenodo open science repository:
  - The
    - Query +
    - produced data +
    - metadata +
    - References (i.e. the set of paper used for compiling the data) are stored into Zenodo and assigned a DOI.
- Zenodo is indexed in OpenAIRE which implements Scholix
  - With the QS-Zenodo interlinking, we indirectly benefit from Scholix capabilities
  - Each time a VAMDC Query is cited by its DOI, all the authors referenced in the dataset receive credits automatically.

## Further developments (2/2)

69

- We started a collaboration with *fireblock.io* for certifying each entry of the Query Store using a block-chain (Ethereum).
  - This addresses the issues linked with data integrity and/or certified provenance (which are crucial issues in a data driven science and/or FAIR contexts).
  - Blockchain ensures a greater sustainability of certification than what a single e-infrastructure or repository may provide (VAMDC nodes may disappear, the infrastructure may migrate or, in the worst case, disappear in few decades).
  - A first demonstrator of integration between the Query Store and Ethereum should be available by December 2018.
- Implementing the Data Citation recommendation opened the door to further interesting innovations.



**Climate Change Centre Austria  
(CCCA)**

**Chris Schubert**

***chris.Schubert@ccca.ac.at***

**research data sharing without barriers**  
**[rd-alliance.org](http://rd-alliance.org)**



data.ccca

## DYNAMIC DATA CITATION

### FROM (WG) PILOT TO OPERATIONAL SERVICE OFFER

---

*Chris Schubert, Head of Data Centre at Climate Change Centre Austria*

GEO Coordinator for Austria, Member of EuroGEOSS Coordination Group,

E-Mail: [chris.schubert@ccca.ac.at](mailto:chris.schubert@ccca.ac.at)

[www.data.ccca.ac.at](http://www.data.ccca.ac.at)

BMBW

**Forschungsinfrastruktur**

re3data.org  
REGISTRY OF RESEARCH DATA REPOSITORIES



<http://dx.doi.org/10.17616/R3K590>

CCCA Data Centre



Dataset Versions Citation

#### Cite this dataset:

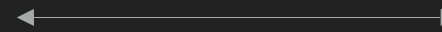
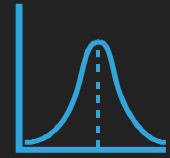
Using this data set or resource, you should cite this data set according to the given copyright conditions with following citation rules:

Leuprecht et al (2016). ÖKS15 Bias Corrected EURO-CORDEX Model Precipitation: pr\_CNRM-CERFACS-CNRM-CM5\_RCP4.5\_r111p1\_CLMcom-CCLM4-8-17, Version 1. Vienna, Austria. CCA Data Centre. PID: <https://hdl.handle.net/20.500.11756/9df12611>. [April 20, 2018]

Copy Text

[hdl.handle.net/20.500.11756/9df12611](https://hdl.handle.net/20.500.11756/9df12611)

*use PID persistent identifier, adequate for doi*



reuse of data

reproducibility

proper attribution and credit

## DYNAMIC DATA CITATION

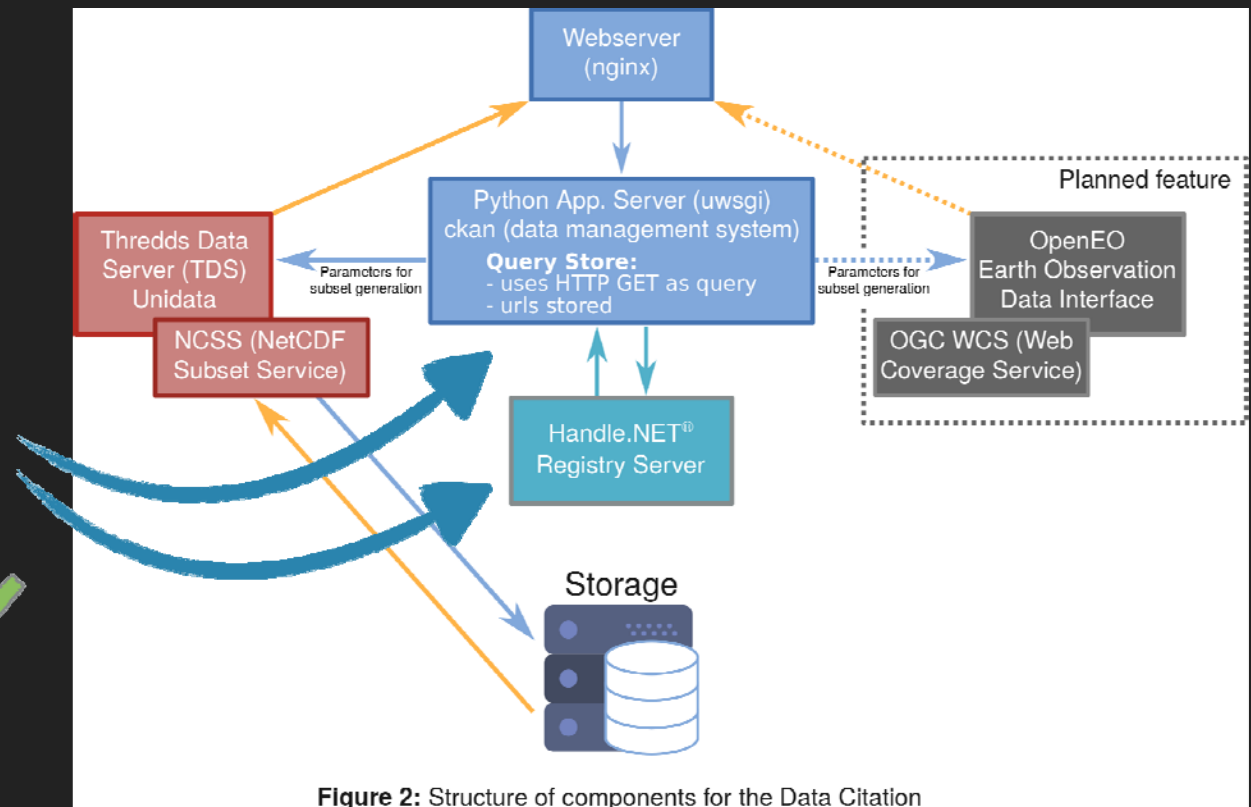
**CITE YOUR DATA**



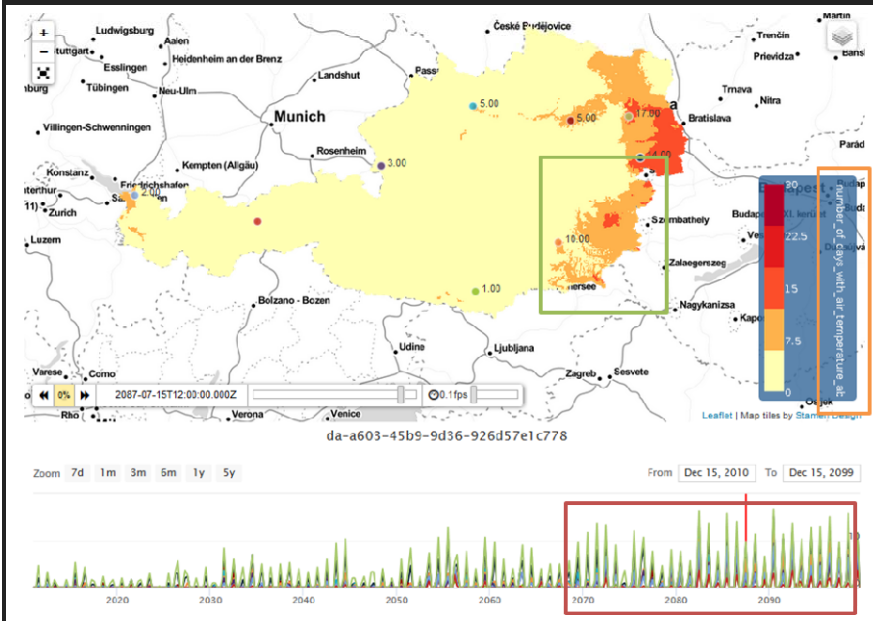
# APPROACH & IMPLEMENTATION

Data Store	R1 - Data Versioning	<b>R13 - Technology Migration</b> <b>R14 - Migration Verification</b>
	R2 - Event Time-stamping	
Query	R4 - Unique Queries	
	R7 - Query Time-stamping	
	R8 - Query PID	
	R9 - Query Metadata	
R3 - Query Store	R5 - Stable Sorting	
	R6 - Result Verification	
	R10 - Citation Text	
	R11 - Human Readable	
Landing Page	R12 - Machine Actionable	

**Figure 1: Recommendation of RDA Guidelines, making data citable (modified after Rauber et al.)**



**Figure 2: Structure of components for the Data Citation**



(research) data is dynamic

identify precisely the data at a specific point in time

identify precisely the subset of (dynamic) in a process

Citing entire dataset,  
providing textual  
description of subset  
-> imprecise

Storing a copy  
of subset as  
used in study

based on Rauber et al.

## SUBSETTING + DYNAMIC DATA CITATION



- Choose a:
- PARAMETER
  - AREA OF INTEREST
  - TIME RANGE
  - @KEEP VERSIONING
  - @KEEP TIMESTAMPS
  - @KEEP & ADAPT METADATA

Dataset Versions Citation

Dataset Versions:

This Version

Version 1 Release Date: 2018-06-24 15:04:15.530698

Latest Version

Version 1 Release Date: 2018-06-24 15:04:15.530698

Cite this dataset:

Using this data set or resource, you should cite this data set according to the given copyright conditions with following citation rules:

Becsi, B. and Laimighofer, J. (2018). tropical\_night\_sbg\_show, Version 1. Vienna, Austria. CCA Data Centre. PID: <https://hdl.handle.net/20.500.11756/f3bbd81e>. [June 24, 2018]

Copy Text

RESOURCE

subset\_NetCDF

DATASET: tropical\_night\_sbg

This resource is a subset of

View

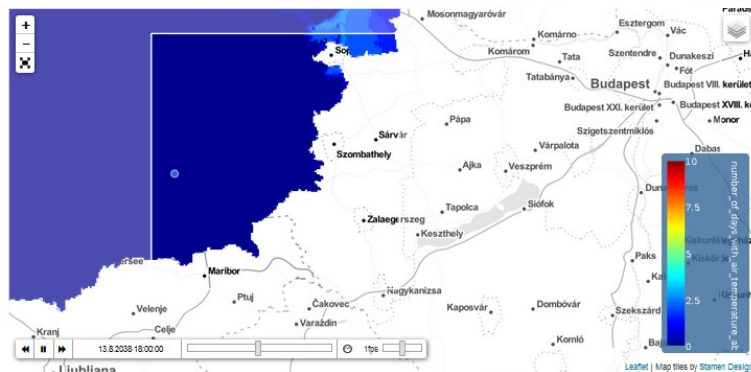
Subset

This dataset is a subset of "ClimaMap Ensemble median (rcp4.5): Tropicalnights" [Show relations](#)

Map Parameter

Double Click within rectangle

Original Version	Release Date	Subset Version
Version 1	2018-05-15 15:38:52.391549	tropical_night_sbg_show (Version 1)



(research) data is dynamic

Re-published

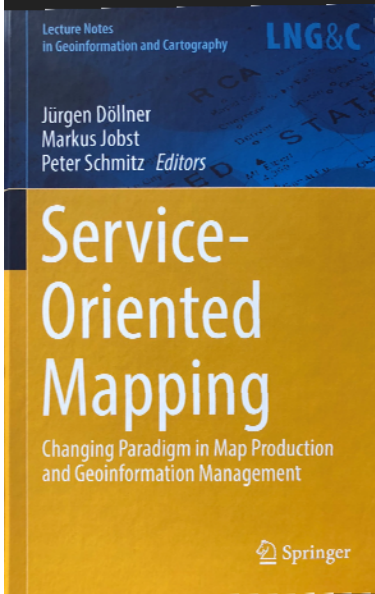
avoid redundant storage consumption

keep all relations between updates, original sources & subsets

## SUBSETTING + DYNAMIC DATA CITATION

CREATE

data.CCA



© 2018, June  
Service Oriented Mapping  
Changing Paradigm in Map Production and  
Geoinformation Management

Handling Continuous Streams for Meteorological Mapping

Chris Schubert<sup>1</sup>, Harald Bamberger<sup>2</sup>

<sup>1</sup> CCCA Data Centre, Vienna, Austria, hosted by ZAMG,

<sup>2</sup> ZAMG, Dep. Software Application development and Data  
Management

## SUBSETTING + DYNAMIC DATA CITATION

Short description on how to deal with  
large data files by subsetting & Dynamic  
Citation Tool

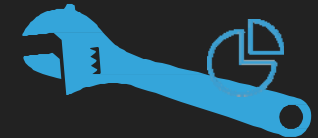


(research) data is dynamic

identify precisely the data at a specific point in time

identify precisely the subset of (dynamic) in a  
process

# PUBLICATION



communication between various software components

Jupyter Notebook for individual processing, visualization, analytics data collection etc.

Remove monthly mean from timeseries to calculate anomalies

```
In [35]: tile_dimensions = (1, int(ls_oks15_monthly_anom.coord('time').shape[0]/ls_oks15_monthly_anom.coord('time').shape[0]))
ls_oks15_monthly_anom.data = ls_oks15_monthly_anom.data - np.tile(ls_oks15_monthly_data, (1,30))
```

Slice over realization dimension and plot monthly means for all ensemble members.

```
In [36]: for oks15_monthly in ls_oks15_monthly.slices_over('realization'):
qp.plot(oks15_monthly)
qp.show()
```

Contourplot of the first timeslice in first cube of the loaded CubeList

```
qp.contourf(ls_oks15_test[0][0,:,:])
ax = plt.gca()
ax.gridlines()
qp.show()
```

Slice over realization dimension and plot monthly anomalies for all ensemble members.

```
In [37]: for oks15_monthly_anom in ls_oks15_monthly_anom.slices_over('realization'):
qp.plot(oks15_monthly_anom)
qp.show()
```

### OpenDAP Dataset Access Form

Action: [Get ASCII](#) [Get Binary](#) [Show Help](#)

Data URL: <http://data.cca.ac.uk/briddo/iodc/char/9d/126/11-4367-4594-9c37-c178/>

Global Attributes: comment: Bias corrected (scaled distribution mapping) data of the EURO-CORDEX model COR6-CERFACS-COR6-CM5\_rcp45\_r11p1\_CIMCO3-CC164-8-17 using observational data from GISSM (IAP).  
Historical and future projection under the RCP4.5 scenario.  
Reference period: 1961-2005  
contact: Anna.Leung@met.rdg.ac.uk, leung@met.rdg.ac.uk

Variables:  pr: Grid

time: y: S:

pr: FillValue: 1.0E20  
standard\_name: precipitation\_amount  
long\_name: Daily precipitation amount  
cell\_method: time: min  
pr: pr: mapping: Lambert\_conformal  
coordinates: lat lon

Lambert\_conformal: 32 bit Integer  
Lambert\_conformal = \*

standard\_parallel: 49.0, 66.0  
longitude\_of\_central\_meridian: 13.33  
latitude\_of\_projection\_origin: 67.5  
false\_northing: 400000.0  
false\_easting: 400000.0  
pr: mapping\_name: Lambert\_conformal\_ossic

time: Array of 64 bit Reals [time = 0.54786]

time:

axis: T  
boundary\_line\_bnds  
units: days since 1949-12-01T00:00:00Z  
standard\_name: time  
long\_name: time  
calendar: proleptic\_gregorian

time\_bnds: Array of 64 bit Reals [time = 0.54786][bnds = 0.1]

time\_bnds:

ChunkSize: 1, 2

OpenDAP Interface

API - SUBSETTING + DYNAMIC DATA CITATION

### Everything is an object

In IPython you can get the list of object's methods and attributes by typing dot and pressing TAB:

```
c.
```

```
File "<ipython-input-20-fcdd94312687>", line 1
c.
^
SyntaxError: invalid syntax
```



data.cca



data.cca Groups Organizations Datasets About

### Create Subset

**Select Layers/Parameters**

daily precipitation amount

**Reuse Query**

Select Query Template -

Use spatial and time arguments from queries that were stored as a resource or skip that point and fill them out by yourself.

**Choose Geographical Extent**

Hi! Cool, you want to create a subset!  
Let me explain you how this works.  
So first, you have to decide which  
layers you want to include in your  
subset. In this case the dataset has  
only one layer so you will obviously  
include this one layer.

.....

Skip ← Back Next →

Stepwise introduction , ....

IMPROVED DOCUMENTATION  
SUBSETTING + DYNAMIC DATA CITATION

DOCUMENTATION



Stepwise introduction , ....

Before you fill out all the arguments below you can also decide to just reuse a query that has already been used on our site. Check out the dropdown menu to see if there is something that is of interest to you.

Skip Back Next

### Reuse Query

Select Query Template

### Choose Geographical Extent



If you want to add a new dataset on this site click this button. Here you need to define a dataset title, resource title and an organization. You can still rename your dataset and resource afterwards. Furthermore, you must define if you want your dataset to be public or private. Keep in mind that once you set the dataset public you cannot set it private again nor delete it. So if you are not sure simply select private. You can still set the dataset public afterwards. Don't be shy, the dataset will not yet be created, so please press the button after this tour!

Done Back Next

Yes

Submit

If you just want to download the file you need to decide in which format you want to have it. Bounding boxes are only available in NetCDF, points are available in all 3 formats.

Skip Back Next

No, just download the subset

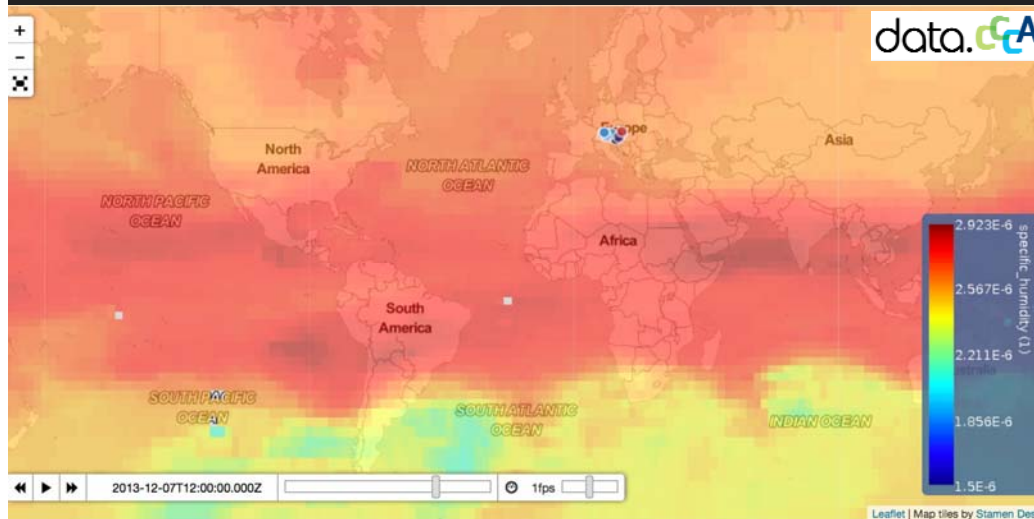
Format  
 netCDF  CSV  XML

XML and CSV can only be chosen if the subset coordinates were drawn as a point.

IMPROVED DOCUMENTATION  
SUBSETTING + DYNAMIC DATA CITATION

DOCUMENTATION

data.cca



Global data sets available

Subset arguments:

*Multi Parameter*

*Altitude Range*

*Time Range*

*Geographical Extend*

Radio Occultation Data :: Specific Humidity

SUBSETTING + DYNAMIC DATA CITATION

**SPATIAL EXTENSION**



data.cca



CCA Subset Service in use ?  
not really ...

- ▶ Is Austrian Research community to small ?
- ▶ Barriers (conscious , scruples, etc.) to re-publish not own results ?
- ▶ Trust ?

*Operational Service Chain is waiting*  
e.g. in Austria, Climate Services gives analyses on impacts, economically, health or decision making, etc. on Regional Level (status: ongoing)



Citation & Subsetting Service planned for West Balkan & Carpathian Region in 2019

SUBSETTING + DYNAMIC DATA CITATION

**SPATIAL EXTENSION & USAGE**

data.cca



<https://github.com/cca-dc/dkanext-cca>



Citation



Visualisation



Relation



Information

Contact:

Chris Schubert

[chris.schubert@cca.ac.at](mailto:chris.schubert@cca.ac.at)

---

**THANK YOU**



**Others?  
Plans, On-going, Feedback**

**Anybody**

**research data sharing without barriers**  
**[rd-alliance.org](http://rd-alliance.org)**

# Agenda

- 14:00 Introduction, Welcome
- 14:10 Short description of the WG recommendations
- 14:30 Report on new issues discussed / lessons learned
- 14:45 Reports on use cases
- 15:20 Other issues, next steps

# Next Steps

- 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, ...
  - Further sessions at plenaries?
- Dissemination of information from on-going pilots
  - Structuring: contact, descriptions, results, lessons learned
  - Outcomes: reports, slides, publications, code, discussions
  - Summary paper on pilots
- New Webinars?
- Anything else? AOB? Wishes?

Thanks

86

Thanks!

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