Data Citation
Working Group Mtg @ P11
Mar. 21 2017, Berlin

research data sharing without barriers

rd-alliance.org

Agenda

- 12:00 Introduction, Welcome
- 12:10 Short description of the WG recommendations
- 12:30 Report on new issues discussed / lessons learned
- 12:45 Reports on use cases
- 13:20 Other issues, next steps

Welcome and Intro

Welcome!

to the maintenance meeting of the

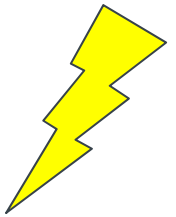
WGDC

Agenda

- 12:00 Introduction, Welcome
- 12:10 Short description of the WG recommendations
  - Goals / challenges
  - Recommendations
  - Benefits
- 12:30 Report on new issues discussed / lessons learned
- 12:45 Brief reports on use cases
- 13:20 Other issues, next steps

Identification of Dynamic Data

- 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!)



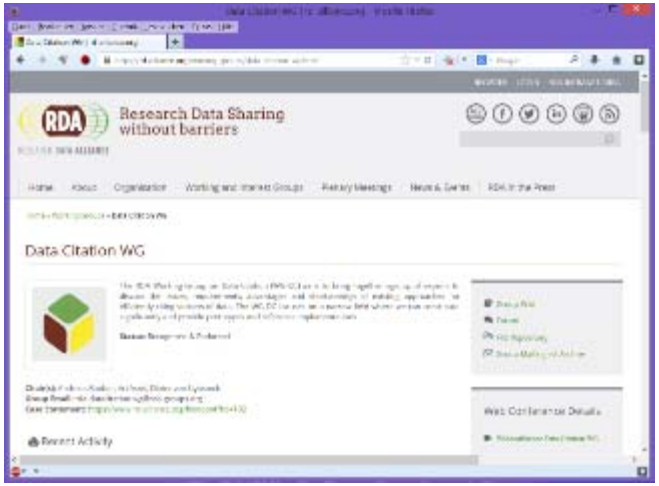
- Would like to identify precisely the **data as it existed at a specific point in time**

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

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:

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:

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

- Researcher uses workbench to identify subset of data
- Upon executing selection („download“) user gets
  - Data (package, access API, …)
  - PID (e.g. DOI) (Query is time-stamped and stored)
  - Hash value computed over the data for local storage
  - Recommended citation text (e.g. BibTeX)
- PID resolves to landing page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

Note: query string provides excellent provenance information on the data set!

- ubset of data
- er gets
  - Data (package, acce s API, …)
  - PID (e.g. DOI) (Que is time-stamped and stored)
  - Hash value compute over the data for local storage
  - Recommended citati n text (e.g. BibTeX)
- PID resolves to land g page
  - Provides detailed metadata, link to parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

- Note: query string provides excellent provenance information on the data set!
- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citat Text (e.g. BibTeX)
- This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!
- PID resolves to land g page
  - Provides detailed metadata, link o parent data set, subset,…
  - Option to retrieve original data OR current version OR changes
- Upon activating PID associated with a data citation
  - Query is re-executed against time-stamped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Deployment

- Note: query string provides excellent provenance information on the data set!
- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citati Text (e.g. BibTeX)
- This is an important advantage over traditional approaches relying on, e.g. storing a list of identifiers/DB dump!!!
- PID resolves
  - Provides det
  - Option to ret
- Identify which parts of the data are used. If data changes, identify which queries (studies) are affected
- Upon activating PID associated w h a data citation
  - Query is re-executed against time-st mped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

Data Citation – Output

- **14 Recommendations grouped into 4 phases:**
  - Preparing data and query store
  - Persistently identifying specific data sets
  - Resolving PIDs
  - Upon modifications to the data infrastructure
- **2-page flyer**
  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

Data Citation – Recommendations

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

RDA Recommendations - Summary

- ***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

- 12:00 Introduction, Welcome
- 12:10 Short description of the WG recommendations
  - Goals / challenges
  - Recommendations
  - Benefits
- 12:30 Report on new issues discussed / lessons learned
- 12:45 Brief reports on use cases
- 13:20 Other issues, next steps

Standardization

- 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

New contacts

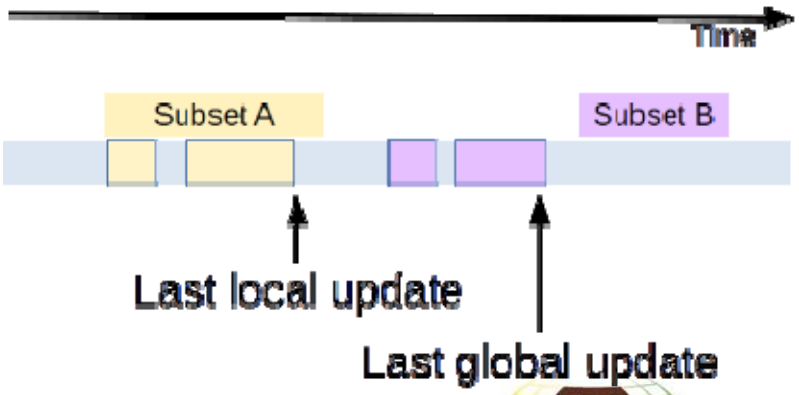
- New project on adopting data citation functionality for an open-access repository of chemical substances, supported by the Austrian National Science foundation started in Feb 2018
- H2020 project discussing adoption of the recommendations for medical data sharing
- Meeting with Ocean Networks Canada to discuss options for implementing the recommendations in Jan 2018

Q&A: R13: Technology Migration

- R13 – Technology Migration: When data is migrated to a new representation (e.g. new database system, a new schema or a completely different technology), migrate also the queries and associated fixity information
- Detailed study on how to address schema evolution in RDBMS
- Different types of schema modifying operations
- Different approaches to address these for integrated, separated and hybrid versioning approaches
- Reference implementation of these for smaller databases
- Specification of query rewriting for massive-scale tables

Q&A: R7: Query Timestamping

- **Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time).**
- Allows to map the execution of a query to a state of the database
  - Execution time: default solution, simple, potentially privacy concerns?
  - Last global update: simple, **recommended**
  - Last update to affected subset: complex to implement
- All equivalent in functionality! (transparent to user)



Q&A: R7: Query Timestamping – Distributed Settings

Distributed Setting

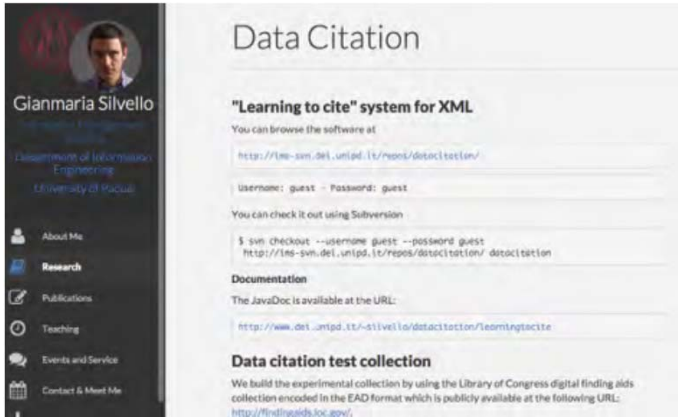
- No need for synchronized timestamps across nodes
- Each node keeps local time
- Solution with one central query store (master node):
  - Master node distributes queries
  - Distributed nodes return query result with local execution timestamp
  - Master stores timestamps per node where response received
- 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

R10: Automated Citation Texts

- **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]

R10: Automated Citation Texts



- Can be created automatically
  - relatively simple for relational
  - more complex for hierarchical/XML
- Learning to Cite:
  - Gianmaria Silvello. Learning to Cite Framework: How to Automatically Construct Citations for Hierarchical Data. Journal of the Association for Information Science and Technology (JASIST), Volume 68 issue 6, pp. 1505-1524, June 2017.
  - http://www.dei.unipd.it/~silvello/datacitation

Agenda

- 12:00 Introduction, Welcome
- 12:10 Short description of the WG recommendations
  - Goals / challenges
  - Recommendations
  - Benefits
- 12:30 Report on new issues discussed / lessons learned
- 12:45 Brief reports on use cases
- 13:20 Other issues, next steps

Adopters

- 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**

Adoption

- **Series of Webinars**
  https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html

- All webinars available for off-line viewing
- More webinars to come
  - Yasuhiro Muyarama: Citing dynamic datasets at NICT
  - Further pilots as they emerge (let us know)

Deep Carbon Observatory Adoption of RDA Recommendations

Ahmed Eleish, Brenda Thomson, Mark Parsons, Peter Fox

research data sharing without barriers

rd-alliance.org

Deep Carbon Observatory Adoption of RDA Recommendations

Ahmed Eleish, Brenda Thomson, Mark Parsons, Peter Fox
Tetherless World Constellation
Rensselaer Polytechnic Institute

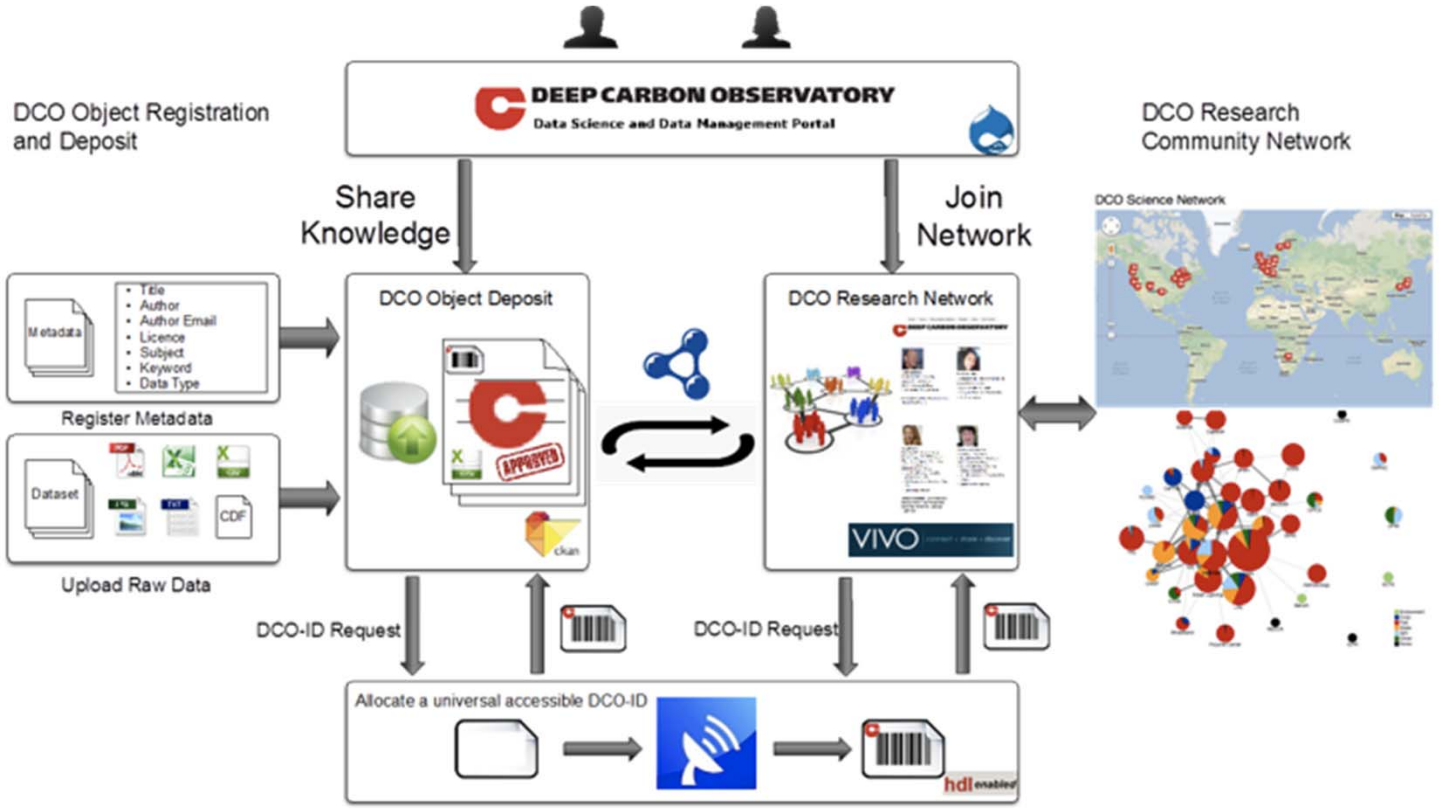
Deep Carbon Observatory (DCO)

- A 10-year project launched in September 2009
- More than 1,000 diverse researchers from >40 countries
- 4 distinct research communities (Extreme Physics and Chemistry, Reservoirs and Fluxes, Deep Life, and Deep Energy)

DCO Data Portal

- VIVO - web-based Java application utilizing an RDF datastore
- Custom DCO ontology
  - Modified to incorporate the Data Type Registry concept
- Multiple referenced ontologies (VIVO, BIBO, DCT, DCAT, FOAF, SKOS,...)
- Data organized into related entities (Person, Publication, Project, Dataset, …)
- Uses DCO ID as identifier and handle server to resolve

DCO Data Portal



DCO Data Portal - Faceted Browsers

- RDF of major entities (Persons, Publications, etc.) is routinely extracted and loaded into Elasticsearch instance
  - Elasticsearch is a distributed, RESTful search and analytics engine
- Data in Elasticsearch is accessible through faceted browsers with search/filtering enabled including a data type facet.

Data Citation of Evolving Data

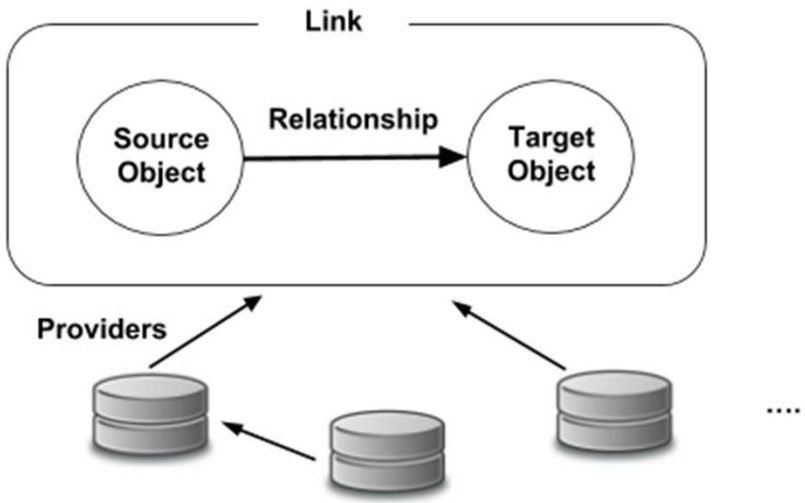
- Working Group on Data Citation
- 13 Recommendations
- Making Data Citable
  - Preparing the Data and the Query Store
  - Persistently Identify Specific Data Sets
  - Resolving PIDs and Retrieving the Data
  - Upon Modifications to the Data Infrastructure

DCO alignment with Dynamic Data Citation

- Dataset versioning not implemented, needs further investigation (possible intersection with Scholix)
- Modifications are timestamped (not logged) in RDF backend (MySQL DB)
- Faceted browser queries contain URL query string parameters; potential for PID, storage, timestamping..
- Outstanding question of how to enable with cooperating repositories.

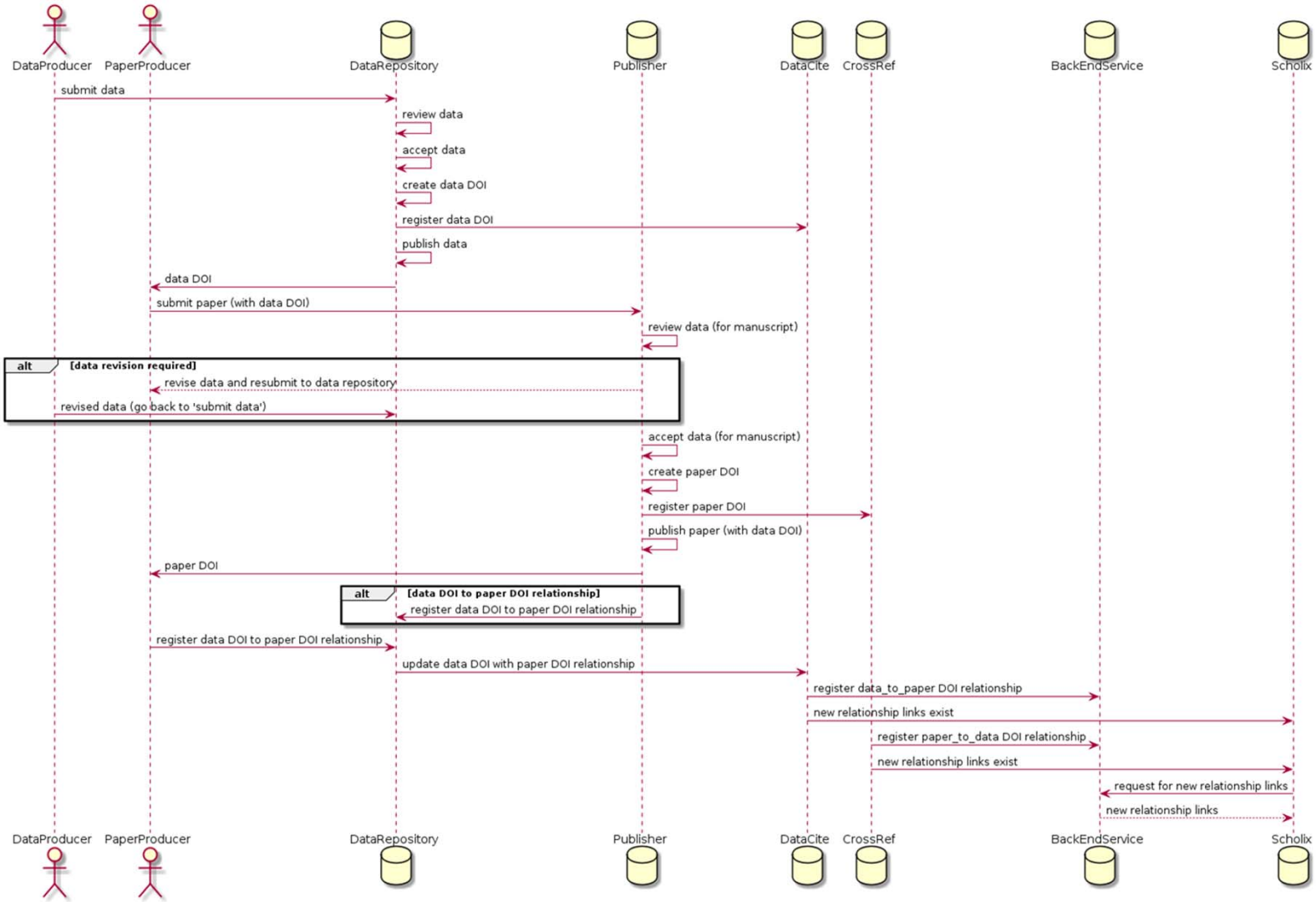
Scholarly Link Exchange

- Scholarly Link Exchange (Scholix) WG
- “a high level interoperability framework for exchanging information about the links between scholarly literature and data”



DCO alignment with Scholix

- DCO does not currently implement explicit links between Publications & Datasets, there is potential for implementation..
- DCO stores Crossref DOI for publications
- DCO references BIBO (Bibliographic Ontology) for publication and DCAT (Data Catalog Vocabulary) for dataset related terms
- Question of whether deepcarbon.net acts as a Scholix hub
  - How to automate the workflow and capture “unknown” article/data links
  - Coordination with other repositories in the DCO data system


DataProducer
PaperProducer
DataRepository
Publisher
DataCite
CrossRef
BackEndService
Scholix
submit data
review data
accept data
create data DOI
register data DOI
publish data
data DOI
submit paper (with data DOI)
review data (for manuscript)
alt
[data revision required]
revise data and resubmit to data repository
revised data (go back to 'submit data')
accept data (for manuscript)
create paper DOI
register paper DOI
publish paper (with data DOI)
paper DOI
alt
[data DOI to paper DOI relationship]
register data DOI to paper DOI relationship
register data DOI to paper DOI relationship
update data DOI with paper DOI relationship
register data_to_paper DOI relationship
new relationship links exist
register paper_to_data DOI relationship
new relationship links exist
request for new relationship links
new relationship links
DataProducer
PaperProducer
DataRepository
Publisher
DataCite
CrossRef
BackEndService
Scholix

Thanks!

Climate Change Centre Austria (CCCA)

Chris Schubert
chris.Schubert@ccca.ac.at

research data sharing without barriers
rd-alliance.org

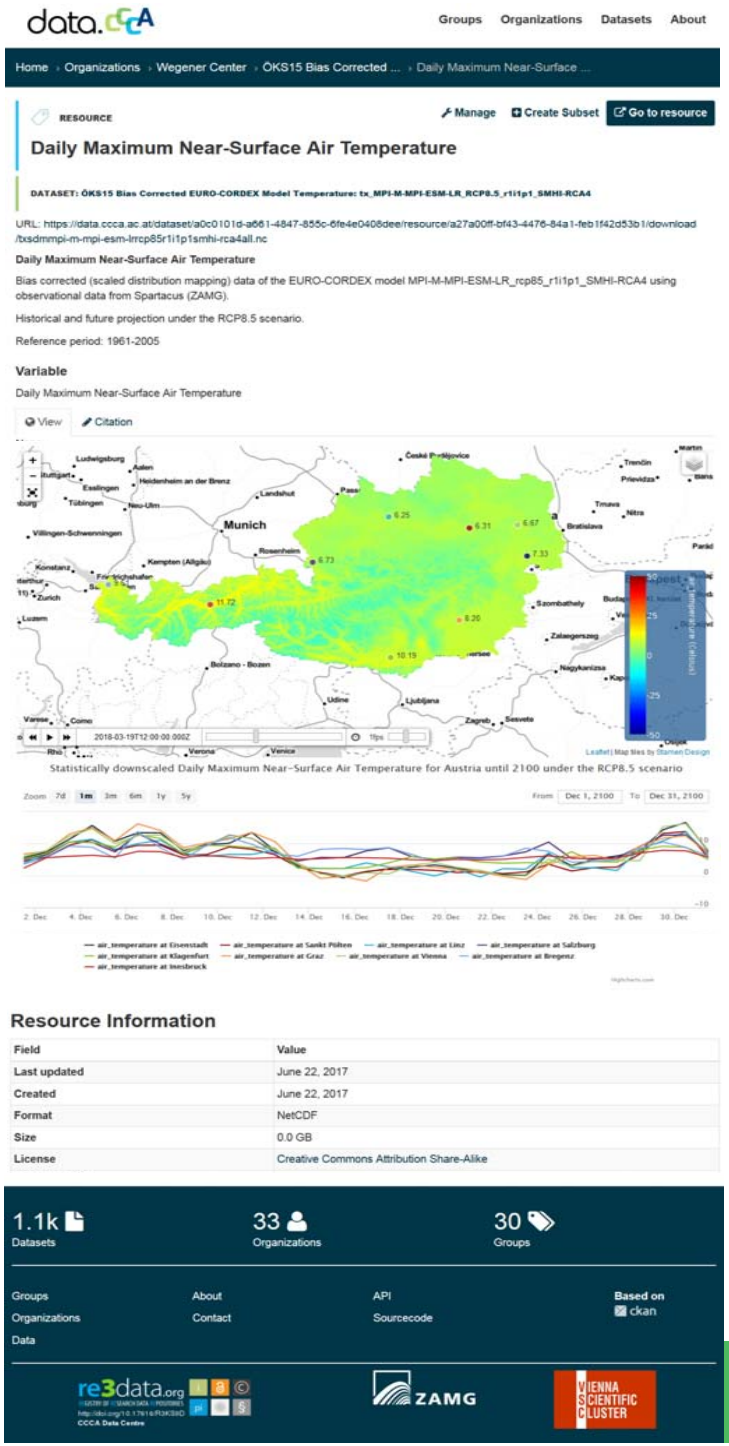
Practical Implementation
RDA Pilot for Dynamic Data Citation for NetCDF files

Chris Schubert, Katharina Sack, Georg Seyerl

CCCA Data Centre

Vienna

Berlin, 23/03/2018



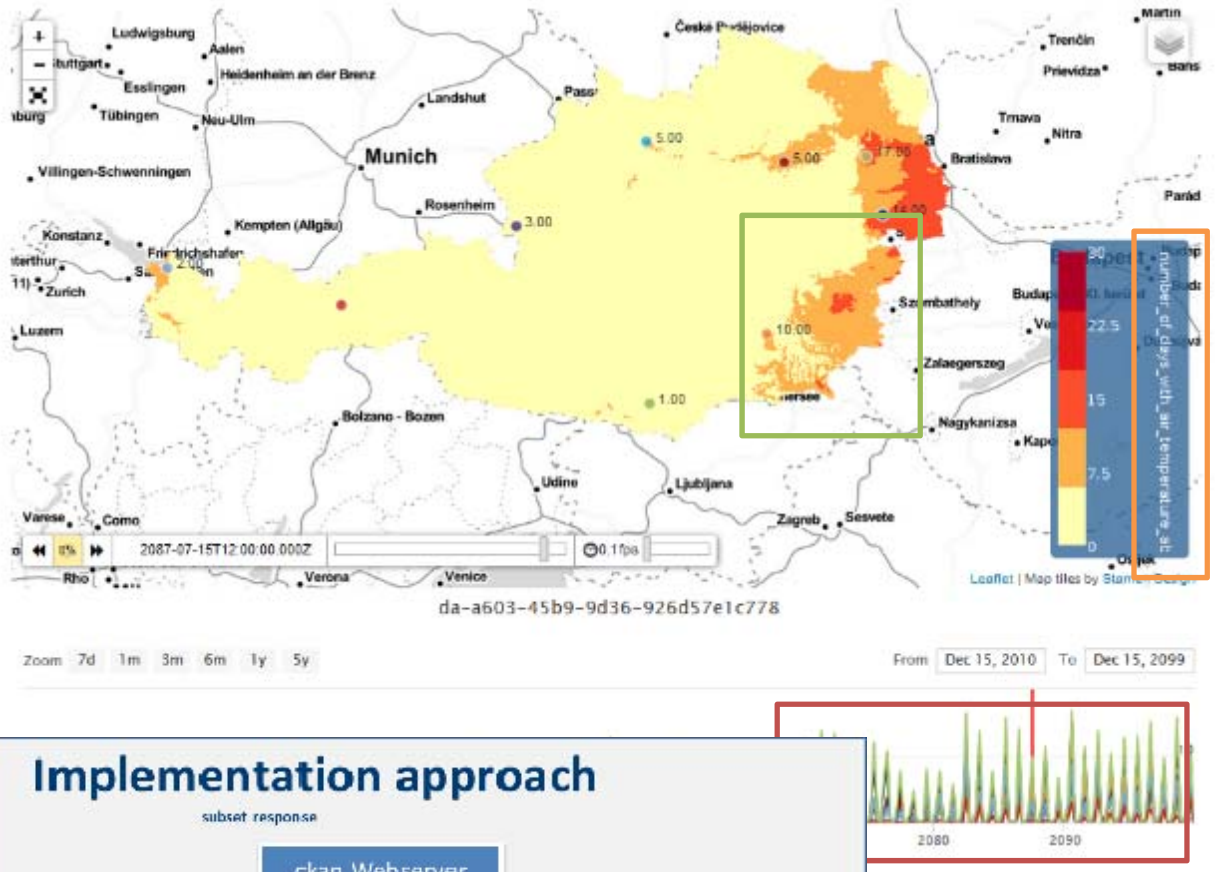
CCCA Data Centre

- provision of distributed climate information and research data in Austria
- sharing large amounts of data
- FAIR data sharing principles implemented
- establishing processes for long term archiving of research data & repositories,
- capacity building, consultancy and support for data sharing, data management & life cycles

Dynamic Data Citation - motivation:

- a proper data management
- no redundant storage consumption
- created subsets available
- citable subsets
- inherited metadata

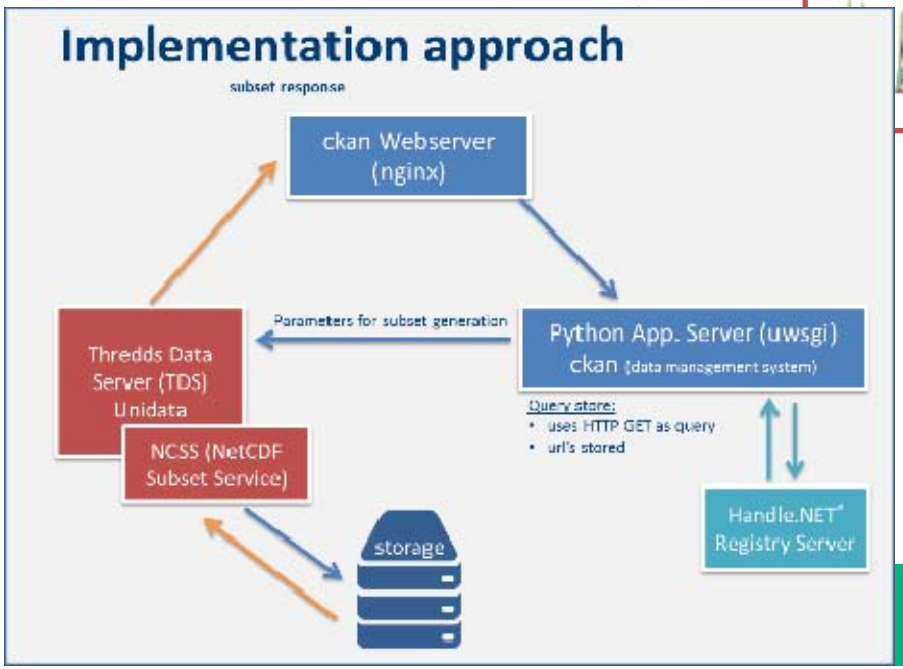
} dynamically



Subset Use Case

Choose a

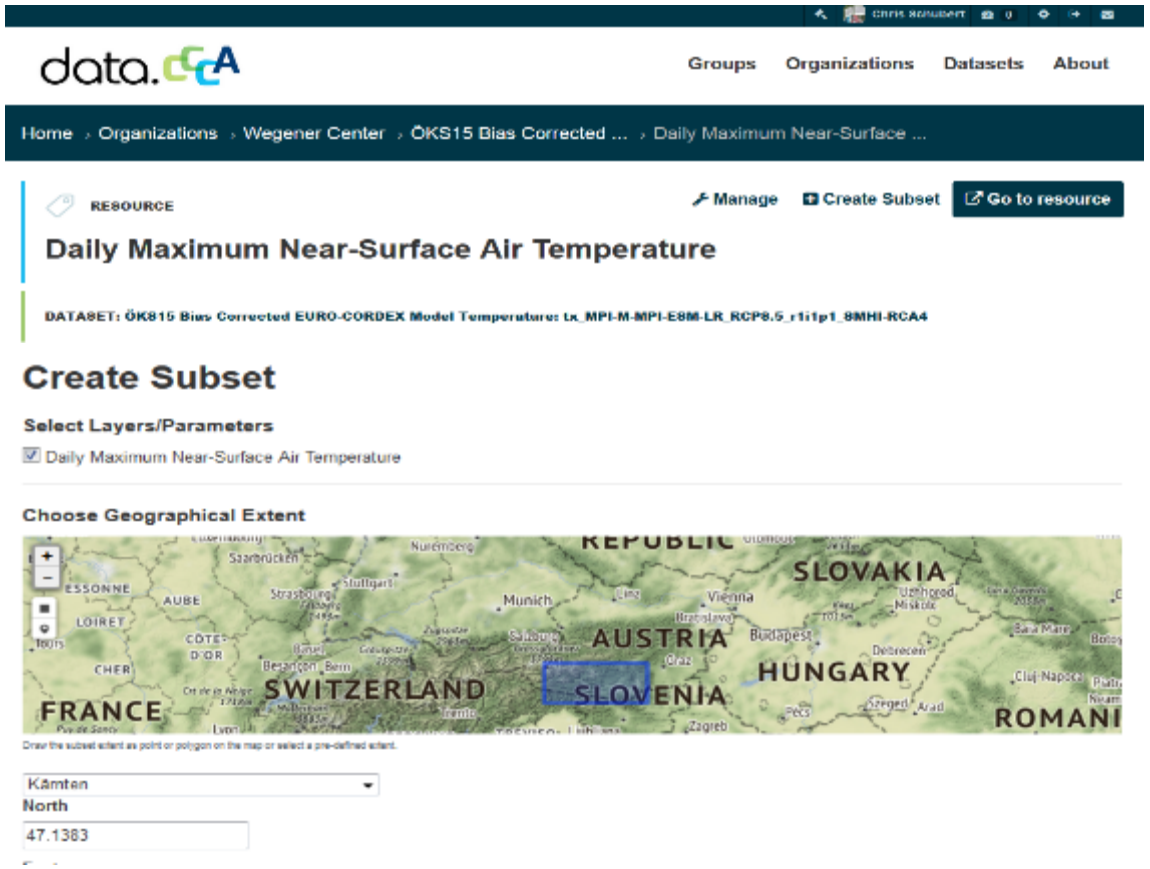
- Parameter
- Area of interest
- Time range
- @keep versioning
- @keep timestamps
- @keep & adapt Metadata
- @no redundant storage consumption



Server architecture:

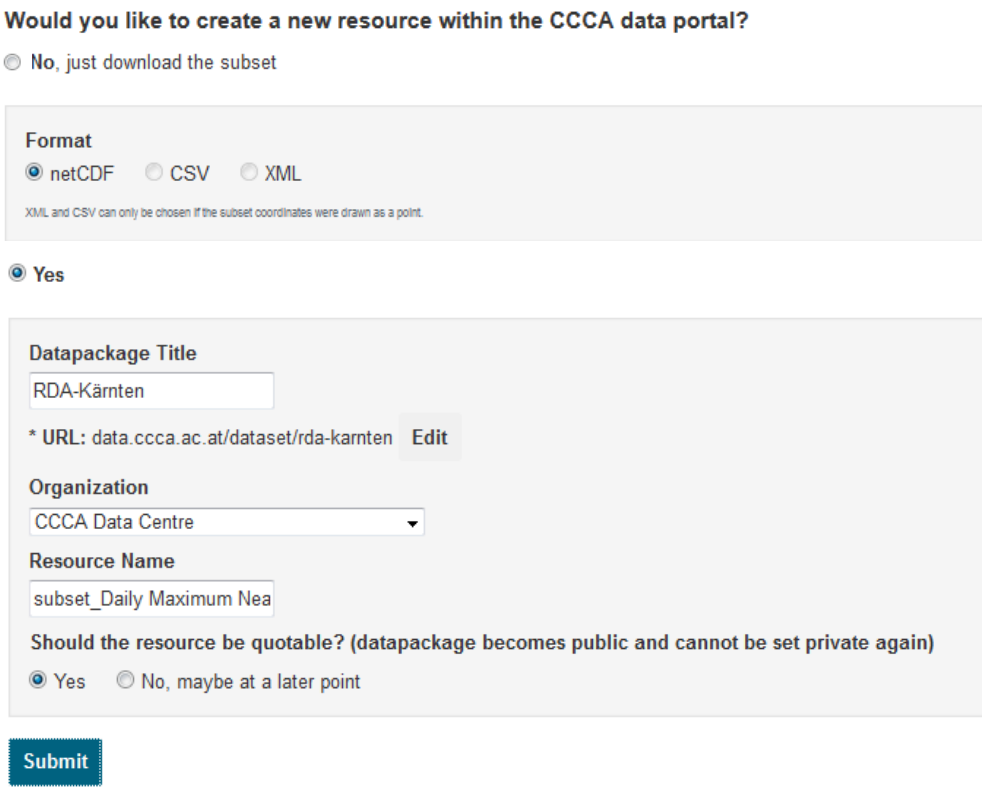
- ckan as data management system
- Thredds Data Server + NCSS Subset Server
- implementation of HDL.NET® Registry

data.ccca - Create Subsets

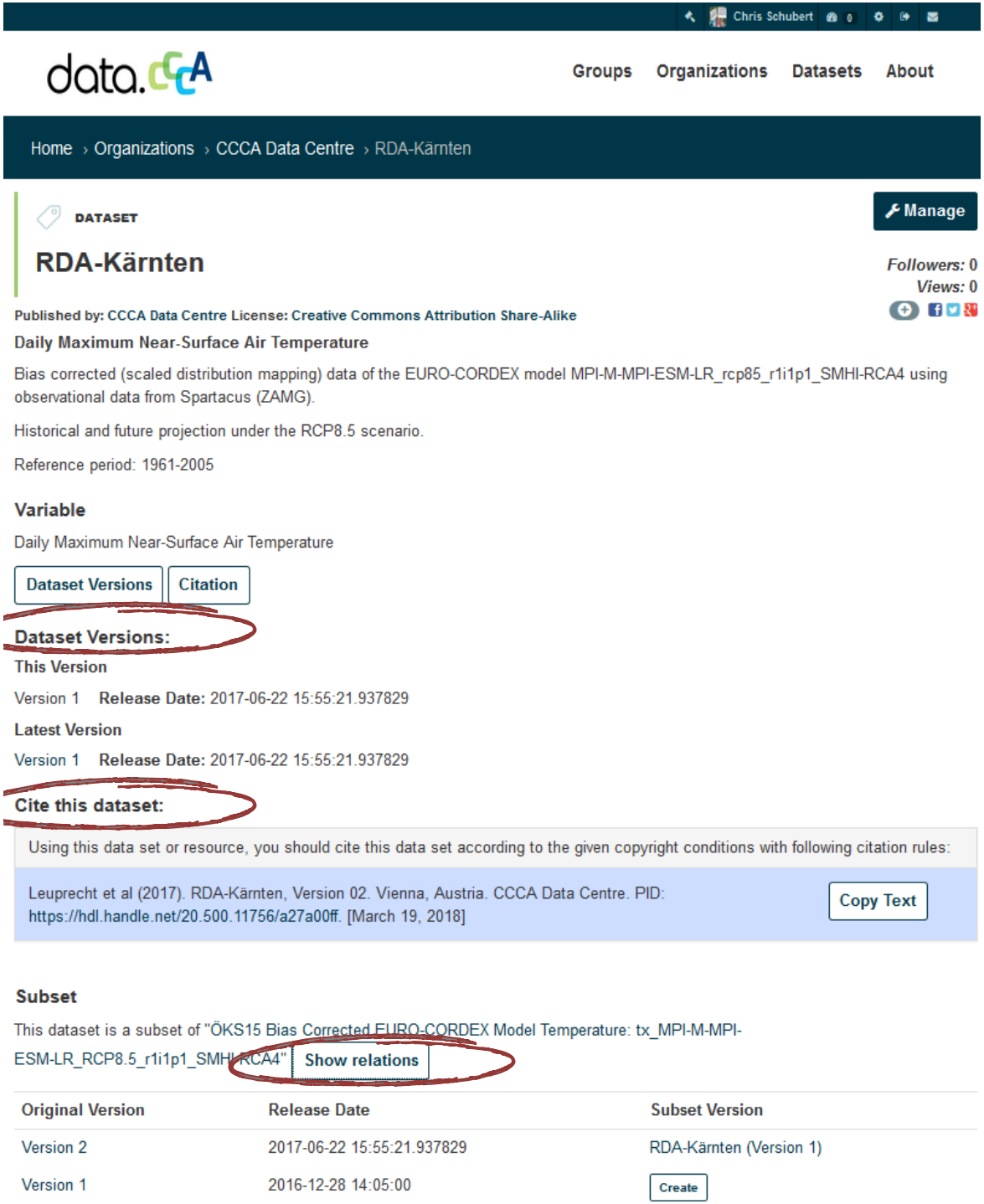


- Login necessary for performance & notification (**step 1a**),
- Push “create Subset” button (**step 1b**),
- Choose a layer (parameter), spatial extent and or a time range (**step 2**)

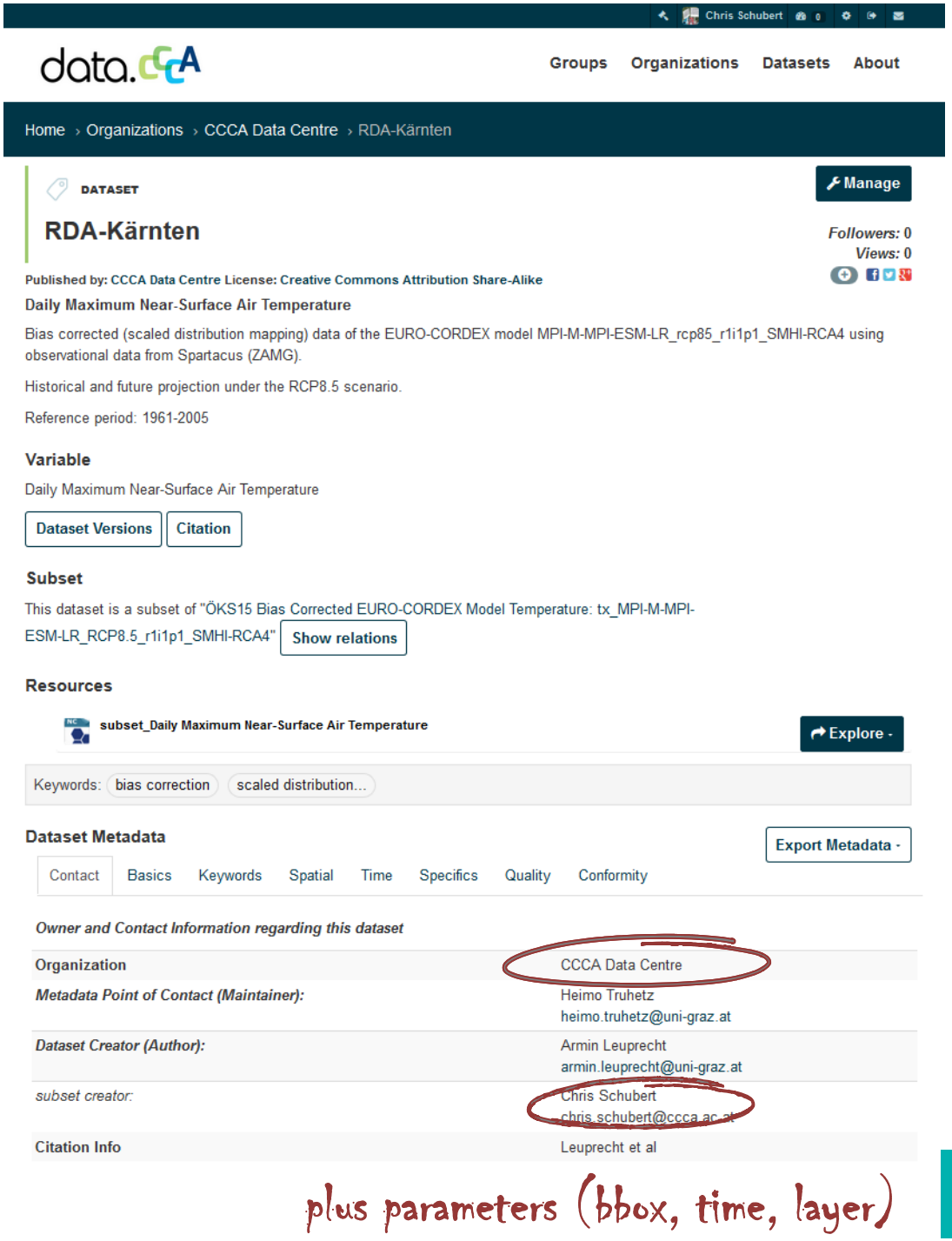
data.ccca - Create Subsets



- Choose a format, adapt the proposal of title (**step 3**)
- Submit! (final step)



- View on Relations & Citation



View on Metadata,
inherited & dyn. adapted

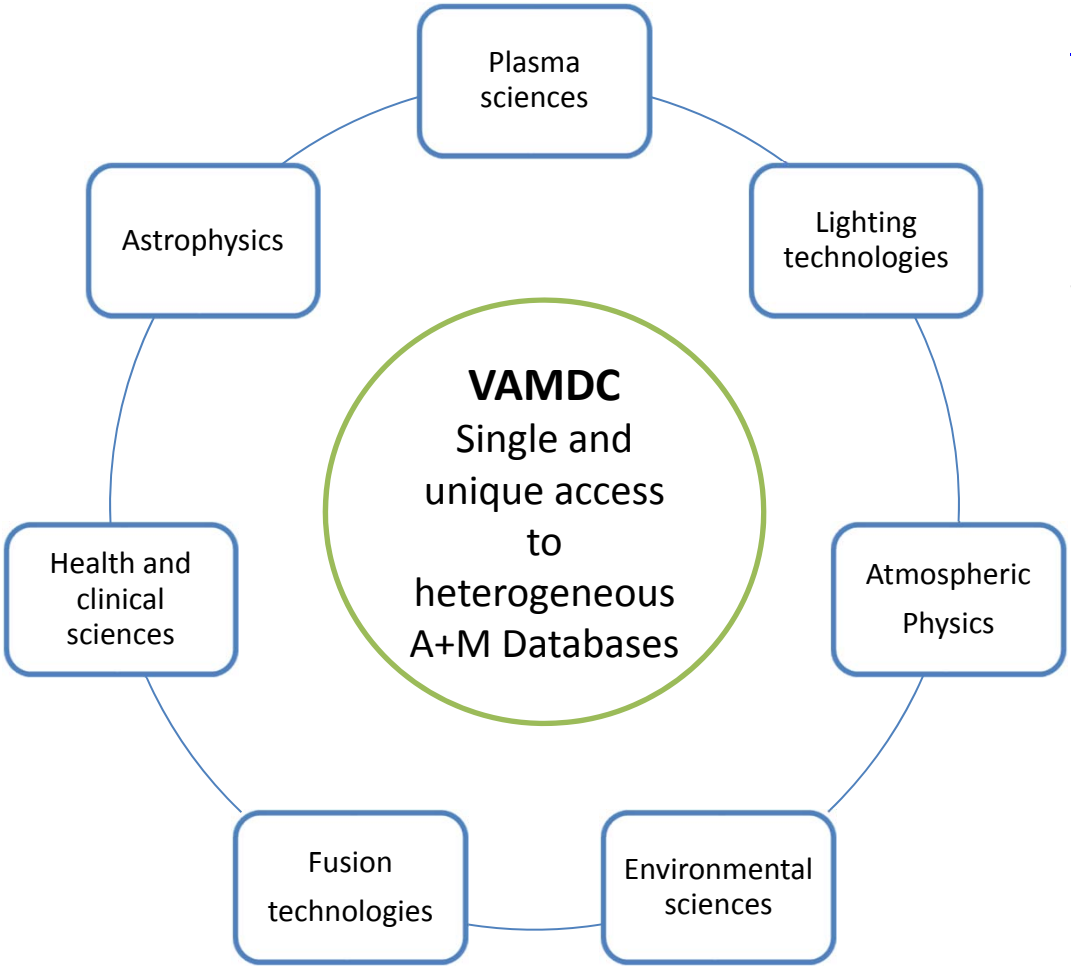
Chris Schubert
Head of CCCA – Data Centre
GEO Coordinator for Austira
1190 Wien, Hohe Warte 38
Tel: +43136026 2519
chris.schubert@ccca.ac.at

VAMDC Query Store implementation

C.M. Zwölf, N. Moreau,
VAMDC Consortium
carlo-maria.zwolf@obspm.fr

research data sharing without barriers
rd-alliance.org

The Virtual Atomic and Molecular Data Centre



- Federates ~30 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
- Removes bottleneck between data-producers and wide body of users

The Virtual Atomic and Molecular Data Centre

VAMDC is distributed e-infrastructure with no central management system

- The different ressources are linked by an interoperable middleware
- The e-infrastructure provide a unique access to all the federated resources

- The implementation of the Query-Store recommendation in this distributed (and not synchronized) context was an amazing challenge:
  - Implementation was performed during a successful cooperation with RDA-EU3 project (2017).
  - The VAMDC Query-Store is now operational at : http://cite.vamdc.eu

Some references about our implementation

- Git-hub repository for all the source-code: https://github.com/VAMDC/QueryStore
  - Implementing note and deployement instruction at: https://github.com/VAMDC/QueryStore/tree/master/documentation
- Link to RDA Data citation webinar : https://youtu.be/rfHfnPvH1r4

Agenda

- 12:00 Introduction, Welcome
- 12:10 Short description of the WG recommendations
  - Goals / challenges
  - Recommendations
  - Benefits
- 12:30 Report on new issues discussed / lessons learned
- 12:45 Brief reports on use cases
- 13:20 Other issues, next steps

Next Steps

- IG on Data Versioning, Citation Metadata, Domain IGs
- 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
- Anything else? AOB? Wishes?

Thanks

Thanks!

And hope to see you at the next meeting

of the

WGDC