RESEARCH DATA ALLIANCE

Data Citation
Working Group Mtg @ P14
October 23 2019, Helsinki

research data sharing without barriers

rd-alliance.org

Agenda

- 16:30 Introduction, Welcome
- 16:40 Short description of the WG recommendations
- 17:00 Reports by adopters / pilots
- 17:50 Paper on adoption stories
- 17:55 Other issues, next steps

Welcome and Intro

Welcome!

to the maintenance meeting of the

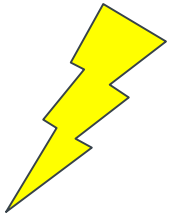
WGDC

Agenda

- 16:30 Introduction, Welcome
- 16:40 Short description of the WG recommendations
- 17:00 Reports by adopters / pilots
- 17:50 Paper on adoption stories
- 17:55 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!)



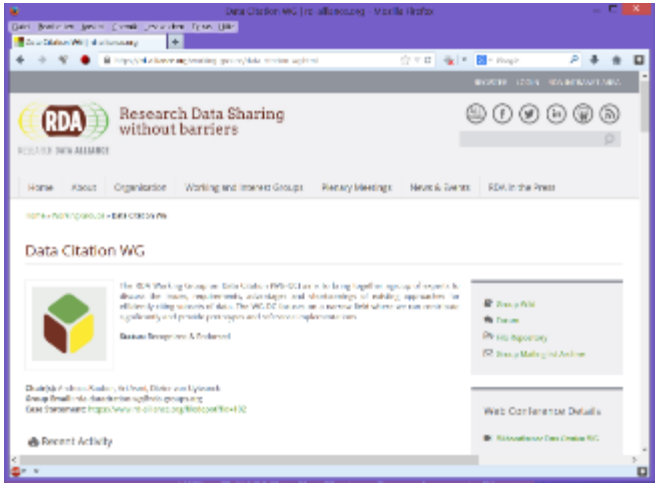
- Would like to identify precisely the **data as it existed at a specific point in time**

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

- 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:

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!

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

- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citat text (e.g. BibTeX)
- 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 citat[illegible] text (e.g. BibTeX)
- PID resolves
  - Provides det
  - Option to ret
- Upon activating PID associated w[illegible]n a data citation
  - Query is re-executed against time-st[illegible]mped and versioned DB
  - Results as above are returned
- Query store aggregates data usage

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

Identify which parts of the data are used. If data changes, identify which queries (studies) are affected

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

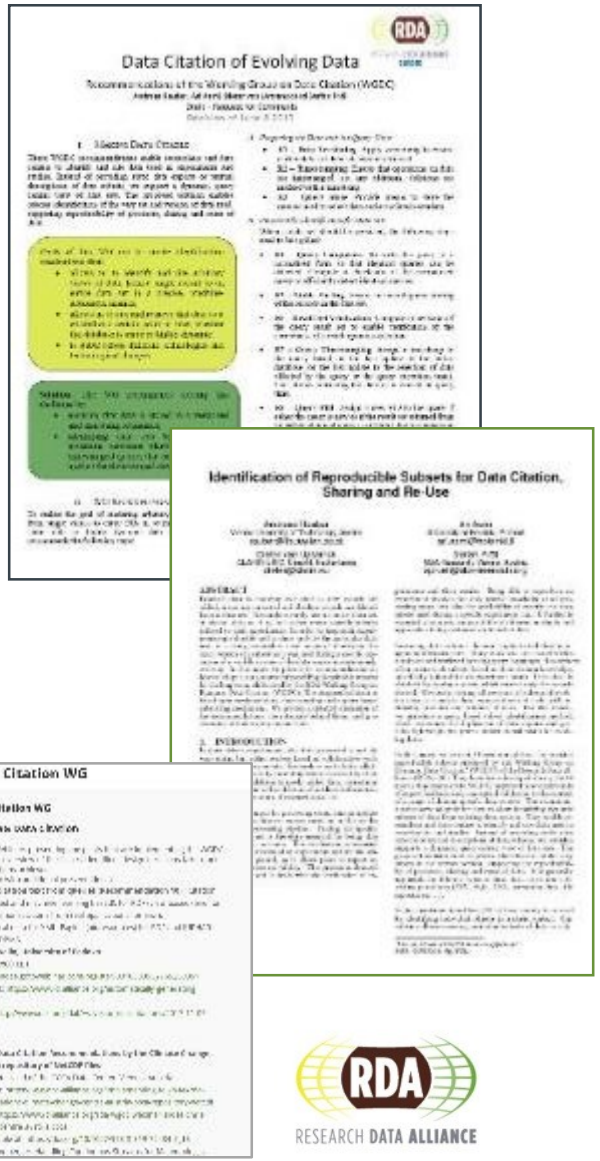
- R4 – Query Uniqueness
- R5 – Stable Sorting
- R6 – Result Set Verification
- R7 – Query Timestamping
- R8 – Query PID
- R9 – Store Query
- R10 – Citation Text

Upon Modifications to the Data Infrastructure

- R13 – Technology Migration
- R14 – Migration Verification

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
- **Adopter's presentations, webinars and reports**
  https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html



WGDC Webinar Series

- https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html
  - **Implementation of the RDA Data Citation Recommendations by the Earth Observation Data Center (EODC) for the openEO platform Wed, Nov 20 2019, 17:00 CET**
  - **Automatically generating citation text from queries 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**

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

- 16:30 Introduction, Welcome
- 16:40 Short description of the WG recommendations
- 17:00 Reports by adopters / pilots
- 17:50 Paper on adoption stories
- 17:55 Other issues, next steps

Large Number of Adoptions

- **Standards / Reference Guidelines / Specifications:**
  - Joint Declaration of Data Citation Principles:
    Principle 7: Specificity and Verifiability (https://www.force11.org/datacitation)
  - ESIP:Data Citation Guidelines for Earth Science Data Vers. 2 (P14)
  - ISO 690, Information and documentation - Guidelines for bibliographic references and citations to information resources (P13)
  - EC ICT TS5 Technical Specification (pending) (P12)
  - DataCite Considerations (P8)
- **Reference Implementations**
  - MySQL/Postgres (P5, P6)
  - CSV files: MySQL, Git (P5, P6, P8, Webinar)
  - XML (P5)
  - CKAN Data Repository (P13)

Large Number of Adoptions

- **Pilot implementations, Use cases**
  - DEXHELPP: Social Security Records (P6)
  - NERC: ARGO Global Array (P6)
  - LNEC: River dam monitoring (P5)
  - CLARIN: Linguistic resources, XML (P5)
  - MSD: Million Song Database (P5)
  - many further individual ones discussed …

Large Number of Adoptions

- **Adoptions deployed**
  - CBMI: Center for Biomedical Informatics, WUSTL (P8, Webinar)
  - VMC: Vermont Monitoring Cooperative (P8, Webinar)
  - CCCA: Climate Change Center Austria (P10/P11/P12, Webinar)
  - EODC: Earth Observation Data Center (P14, Webinar)
  - VAMDC: Virtual Atomic and Molecular Data Center (P8/P10/P12, Webinar)
- **In progress**
  - NICT Smart Data Platform (P10/P14)
  - Dendro System (P13)
  - Ocean Networks Canada (P12)
  - Deep Carbon Observatory (P12)

RDA WGDC Recommendations in ESIP Guidelines

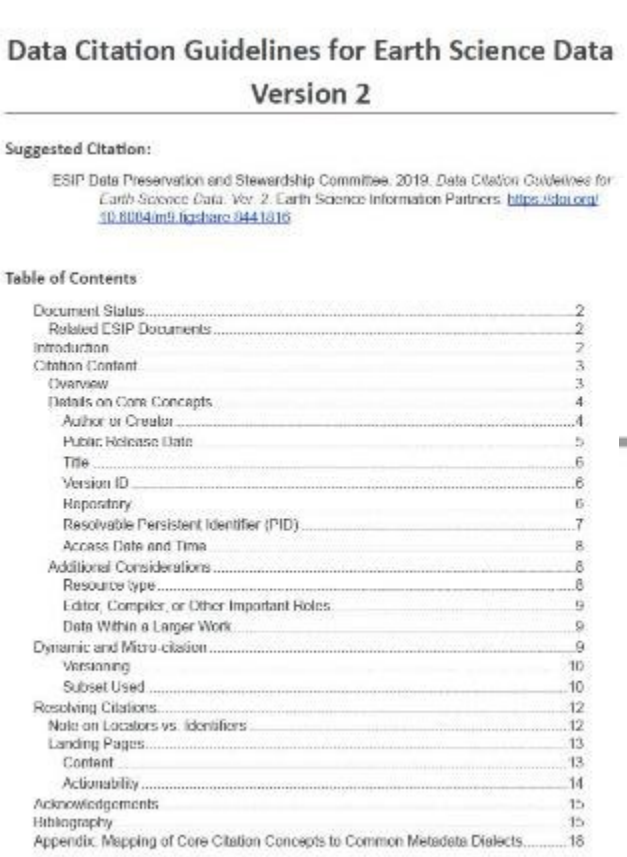
Mark Parsons

research data sharing without barriers
rd-alliance.org

ESIP: Data Citation Guidelines for Earth Science Data Version 2

Official version available:

https://doi.org/10.6084/m9.figshare.8441816



Data Citation Guidelines for Earth Science Data Version 2

Suggested Citation:

ESIP Data Preservation and Stewardship Committee. 2019. *Data Citation Guidelines for Earth Science Data, Ver. 2.* Earth Science Information Partners. https://doi.org/10.6084/m9.figshare.8441816

Table of Contents

Dynamic and Micro-citation

This may be the most challenging aspect of data citation. It is necessary to enable "micro-citation" or the ability to refer to the specific data used—the exact files, granules, records, etc.

9

from a particular version. Scientifically, this is to enable reproducibility by providing a precise reference to the data used. It may, however, impact the credit or attribution functions of a citation. Different subsets of a larger collection may have been created by different people. As discussed in Data Within a Larger Work, mechanisms for crediting at finer granularity are still being developed.

Mechanisms for referencing and providing access to precise subsets of data are more established. Ideally, the repository should provide a PID that resolves to the precise subset and version of the data used. We recommend that repositories implement the Research Data Alliance (RDA) Recommendation on Scalable Dynamic Data Citation, which provides a PID for a particular query.

We recognize, however, that not all repositories have the ability to implement the RDA Recommendation so other approaches that can work reasonably well, at least for human interpretation, may be used.

RESEARCH DATA ALLIANCE

research data sharing without barriers
rd-alliance.org

ESIP: Data Citation Guidelines for Earth Science Data Version 2

Official version available:
https://doi.org/10.6084/m9.figshare.8441816

Data Citation Guidelines for Earth Science Data
Version 2

Suggested Citation:

ESIP Data Preservation and Stewardship Committee. 2019. *Data Citation Guidelines for Earth Science Data, Ver. 2*. Earth Science Information Partners. https://doi.org/10.6084/m9.figshare.8441816

ESIP
The Federation of Earth Science Information Partners
Fostering connections to make data matter

Mechanisms for referencing and providing access to precise subsets of data are more established. Ideally, the repository should provide a PID that resolves to the precise subset and version of the data used. We recommend that repositories implement the Research Data Alliance (RDA) Recommendation on Scalable Dynamic Data Citation, which provides a PID for a particular query.

Mechanisms for referencing and providing access to precise subsets of data are more established. Ideally, the repository should provide a PID that resolves to the precise subset and version of the data used. We recommend that repositories implement the Research Data Alliance (RDA) Recommendation on Scalable Dynamic Data Citation, which provides a PID for a particular query.

We recognize, however, that not all repositories have the ability to implement the RDA Recommendation so other approaches that can work reasonably well, at least for human interpretation, may be used.

research data sharing without barriers
rd-alliance.org

RESEARCH DATA ALLIANCE

Designing Dynamic Data Citation for Data Provenance on Smart Data Platform

Koji Zettsu, Yasuhiro Murayama
National Institute of Information and Communications Technology
Japan

research data sharing without barriers
rd-alliance.org

Designing Dynamic Data Citation for Data Provenance on Smart Data Platform

Koji Zettsu, Yasuhiro Murayama

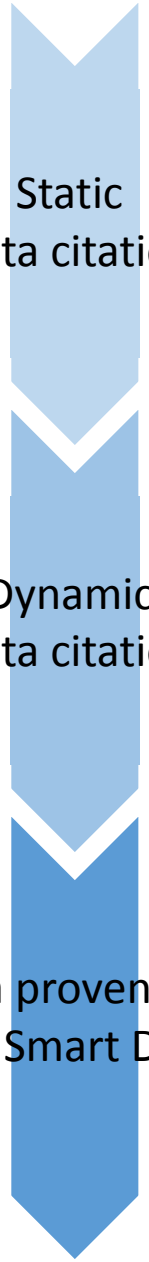
National Institute of Information and Communications Technology
(NICT), Japan

RDA Plenary 14

October 23, 2019

History

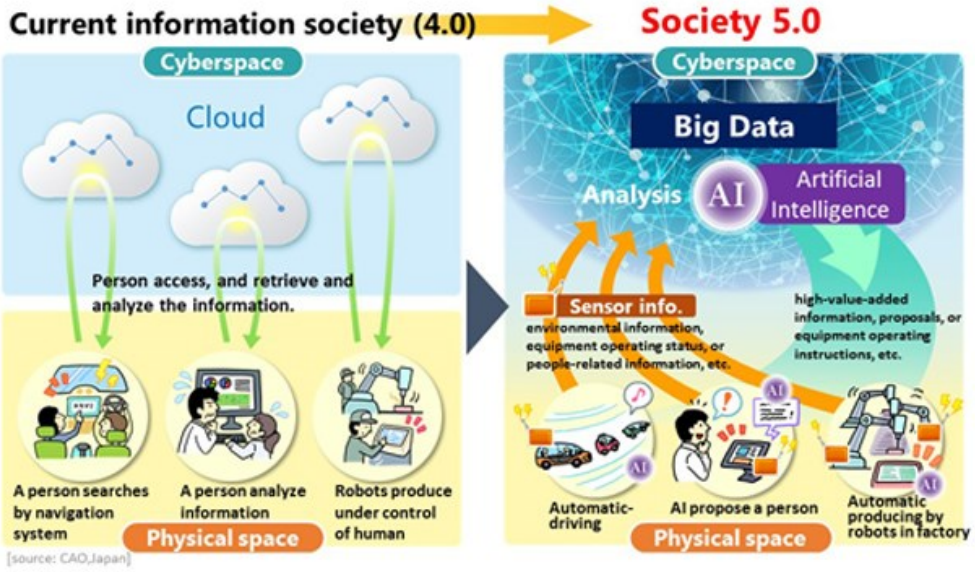
Sep. 2013	Contributed article in *"out of cite, out of mind"* *(Data Science Journal 12(13))* published by CODATA-ICSTI Task Group on Data Citation Standards and Parities
Sep. 2013	**[RDA-P2]** Started discussion with A. Rauber (RDA Data Citation WG chair)
Nov. 2014	Research presentation at SciDataCon 2014 (New Delhi): ***"Mining Data Citation for Usage Analysis of Open Science Data"***
Oct. 2016	RDA Data Citation WG recommendation published (DOI: http://dx.doi.org/10.15497/RDA00016)
Apr. 2017	**[RDA-P9]** Kick-off presentation of Japanese adoption (***dynamic data citation***) at RDA Data Citation WG
Sep. 2017	**[RDA-P10]** Interim report talk of the *dynamic data citation* work
Oct. 2017	Research presentation at CODATA 2017 (St. Petersburg) : "***A Data Citation System Framework for Identification of Evolving Data***"
Apr. 2018	Start *present work* **data provenance for a Smart Data Platform** in NICT Real Space Information Analysis project
Mar. 2019	**[RDA-P13]** Preliminary report at RDA Data Citation WG



Background: Super Smart Society

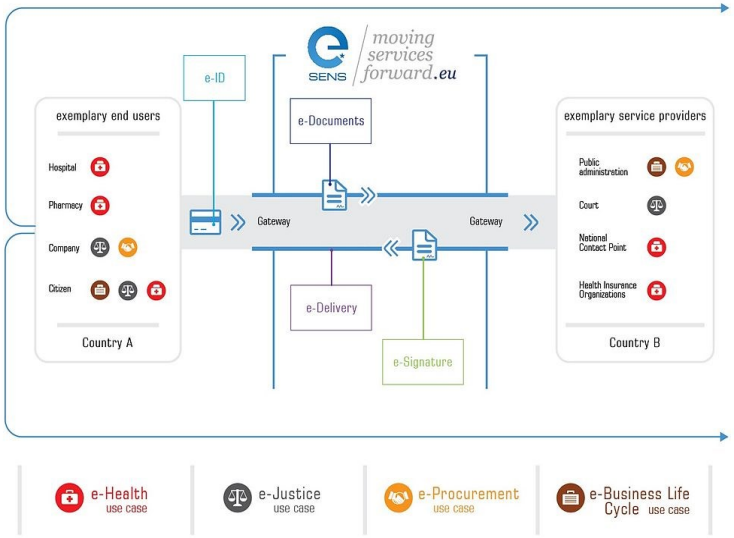
- High degree of convergence between cyberspace (virtual space) and physical space (real space) through IoT
- Interdisciplinary data collaboration for complex problem solving in smart societies
- Data-driven AI with Smart Data
  - IoT big data collected and processed to be turned into 'actionable information'
  - Fairness, Accountability, Transparency in Machine Learning (FAT/ML)

Society 5.0 (Cabinet Office of Japan)



Source: https://www8.cao.go.jp/cstp/english/society5_0/index.html

EU Digital Single Market



Source: EU Digital Single Market, https://en.wikipedia.org/wiki/Digital_Single_Market#/media/File:E-SENS_architecture.jpg

NICT xData Platform

Collection → Association → Training → Prediction

Sensing data

- Weather (DIAS, PANDA, etc.)
- Atmosphere (AEROS, etc.)
- Traffic (ITARDA, JARTIC, etc.)
- Probe car
- Healthcare (medical recept, wearable sensor, etc.)
- SNS (Twitter)

Data Loader API

Association Mining API

Associative Prediction API

Map Creation API

Route search API

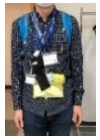
Alert API

Training dataset

Prediction result data

Event data (common format)
11 domains・24 kinds・13.3 billion records・13.8TB (as of Jan. 2019)

Application-specific sensing data collecti



NICT Integrated Testbed

DB servers×8, Storage servers× 2 ,
Analysis server×10, Cluster server×36

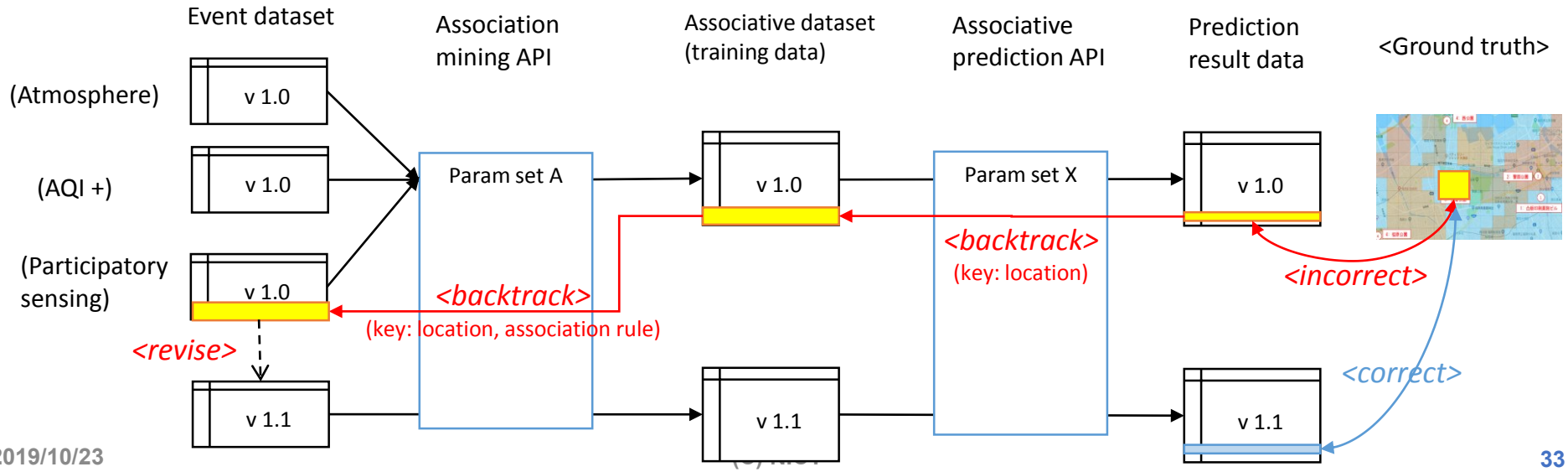
- Smart sustainable mobility

- Smart environmental healthcare

Feedback

- Generate data provenance graph based on API logging and dataset versioning
- Revise datasets and/or API parameters by “back-tracking” a data provenance graph
- A **workbench tool** for supporting trial-and-error revision work
  - Browsing and backtracking a data provenance graph
  - Revision control of datasets

[Example of Smart Environmental Healthcare]

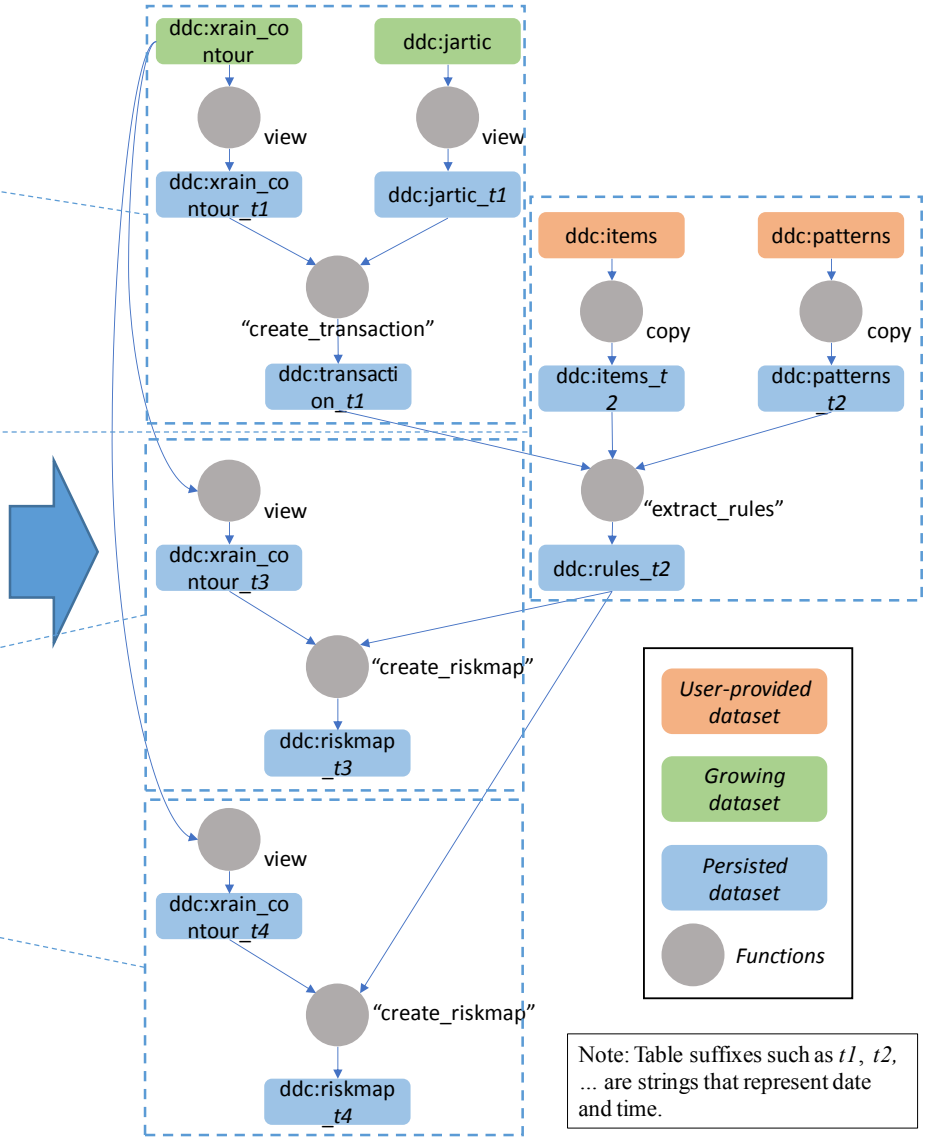


Basic Concepts

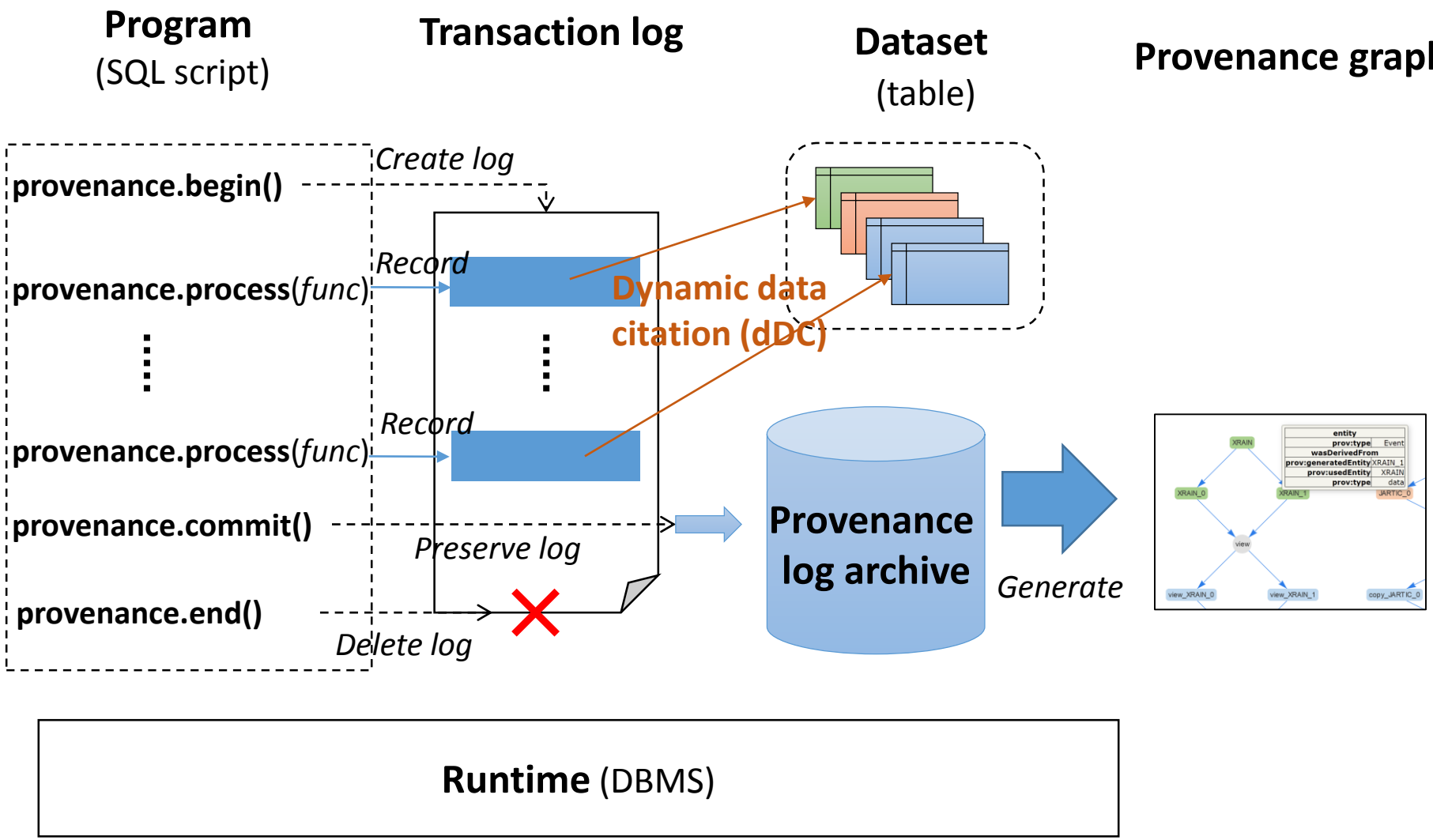
- Workflow provenance at the level of function boundaries
  - Nodes: arbitrary functions with some input, output, and parameters
  - Edges: dataflow or control flow between these functions
- Retrospective
  - Tracing a workflow execution for obtaining function calls and data resources accessed/generated
  - Programming library for provenance capture: **begin/end/commit** and **process**
- Evolutional
  - Keep track of changes made between different versions of data by the dynamic data citation **dDC**
  - Alleviates rapid iterations on various data, parameters, and workflow manipulations
- W3C PROV-based provenance graph
  - 'Entity' → table/view, 'Activity' → processes
  - *ProvJS* library for visualization of a provenance graph

Example of Data Provenance

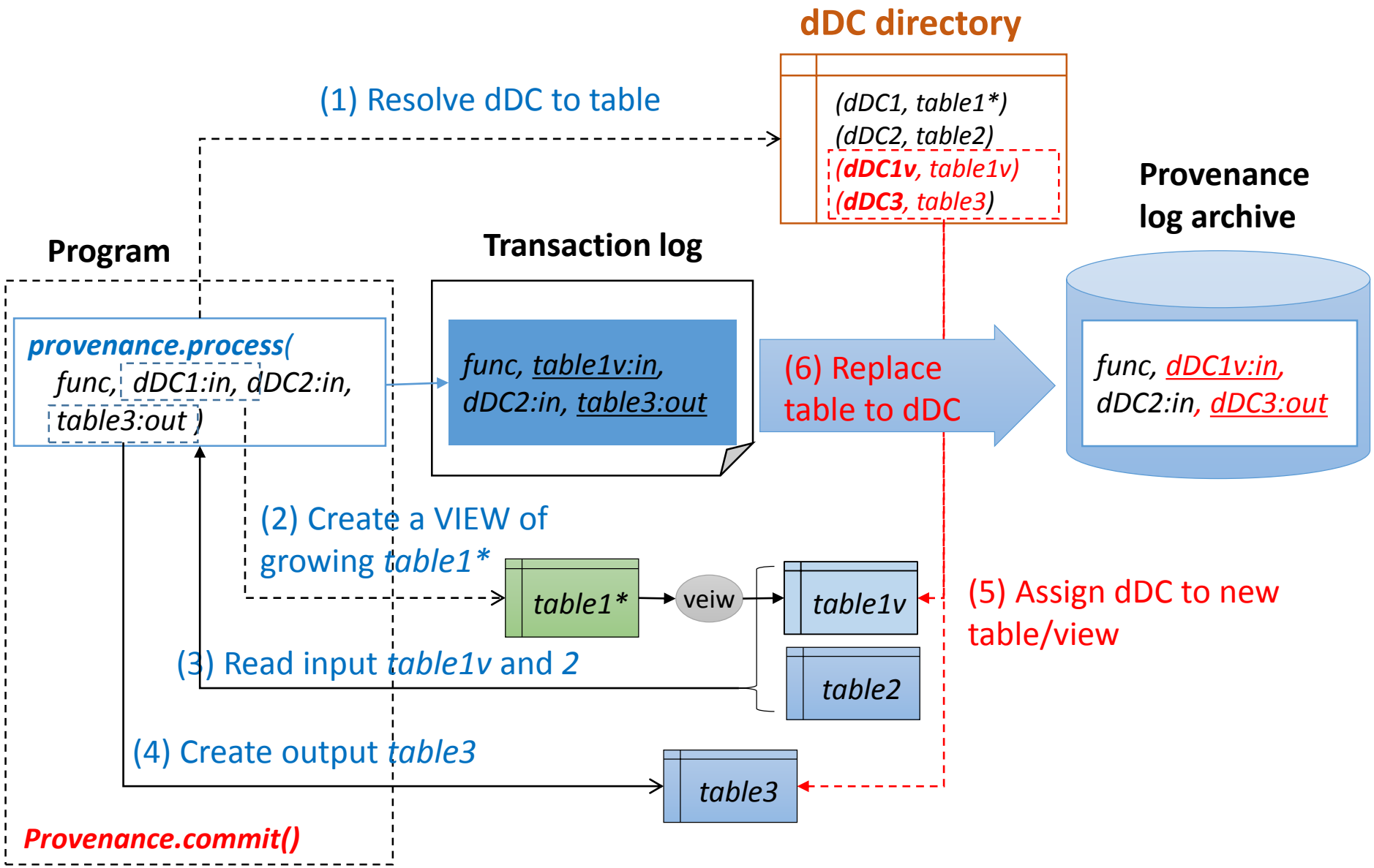
```
SELECT provenance.begin();
SELECT provenance.process(
  'SELECT create_transaction(
     {params}
  )',
  ['ddc:xrain_contour', 'ddc:jartic'], -- in
  'ddc:transaction' -- out
);
SELECT provenance.process(
  'SELECT extract_rules(
    {params}
  )',
  ['ddc:transaction', 'ddc:items', 'ddc:patterns'], -- in
  'ddc:rules' -- out
);
SELECT provenance.process(
  'SELECT create_riskmap(
     {params}
  )',
  ['ddc:rules', 'ddc:xrain_contour'], -- in
  'ddc:riskmap' -- out
);
SELECT provenance.process(
  'SELECT create_riskmap(
     {params}
  )',
  ['ddc:rules', 'ddc:xrain_contour'], -- in
  'ddc:riskmap' -- out
);
SELECT provenance.commit();
SELECT provenance.end();
```



Provenance Capture Mechanism



Step-by-Step Example of Provenance Capture

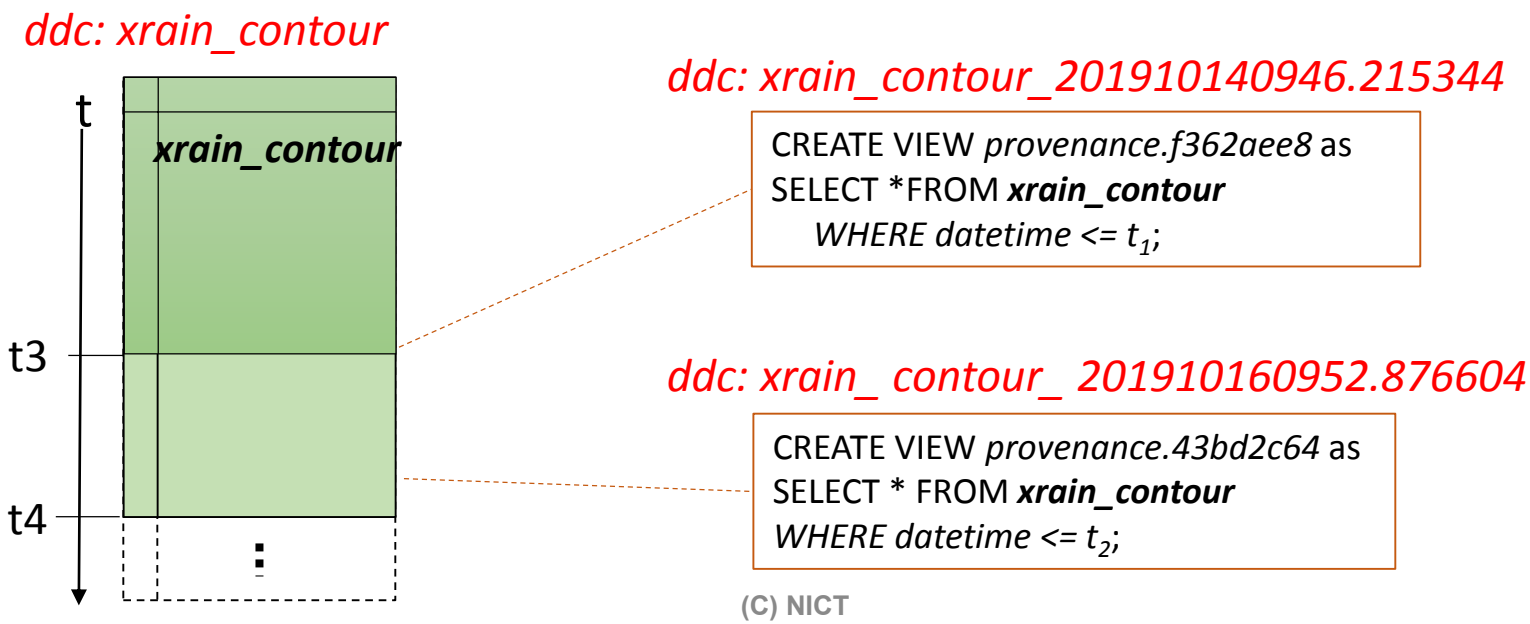


Issues on Dynamic Data Citation for Data Provenance (1)

- **Reproducibility:** Ensure access to the datasets for a specific transaction
  - Create a VIEW (query) of a growing dataset at the transaction time (e.g., sensor data archive)
  - Assign a data citation to a dataset/view on transaction
  - Ensure uniqueness of a data citation at the level of runtime environment

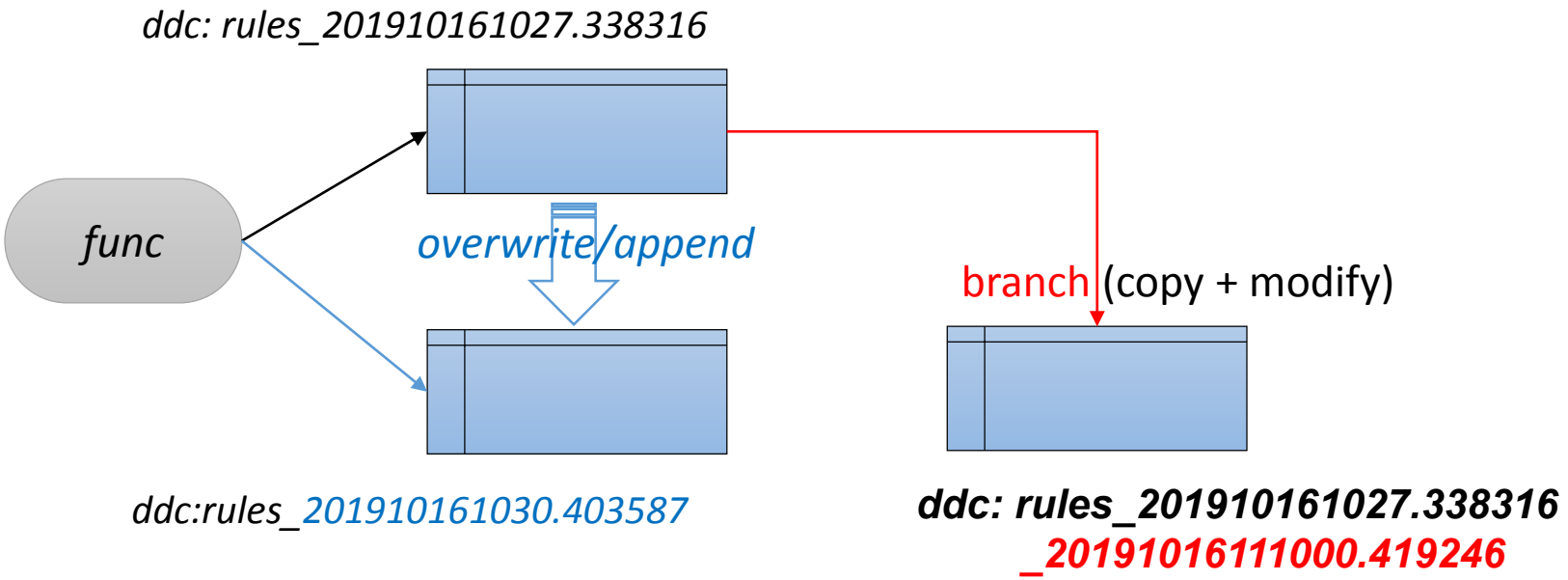
Citation text : "*ddc:* <table name>_<timestamp>"

(omit <timestamp> for growing table)



Issues on Dynamic Data Citation for Data Provenance (2)

- Trace change of datasets in a provenance graph
  - Create a new revision of data citation for (automatic) overwrite/append to an existing dataset by a different transaction
  - Create a branch of data citation for (manual) derivation of an existing dataset/view
  - Manage a cited dataset/view in a secure dataspace protected from unproven (casual) changes



Issues on Dynamic Data Citation for Data Provenance (3)

- **Usage analysis:** Retrieve provenance subgraphs for a specific data citation
  - Associate a node of provenance graph with a data citation
  - Abstract a provenance graph for a same data citation (e.g., node merging)
  - Detect ‘orphan’ citations (not referred from any transaction) for garbage collection

- **Persistency:** Standard format of dynamic data citation with globally unique ID

Future Work

- Performance improvement
  - Reduction of runtime overhead for provenance capture
  - Focus + context view for a large provenance graph
  - VIEW materialization: reproducibility vs. storage
- Provenance graph operations
- Programming language support for provenance library
  - SQL, Python
- Technical report & use case summary
  - RDA Data Citation WG
  - Cross Data Collaboration Project, Smart IoT Acceleration Forum Japan

OpenEO

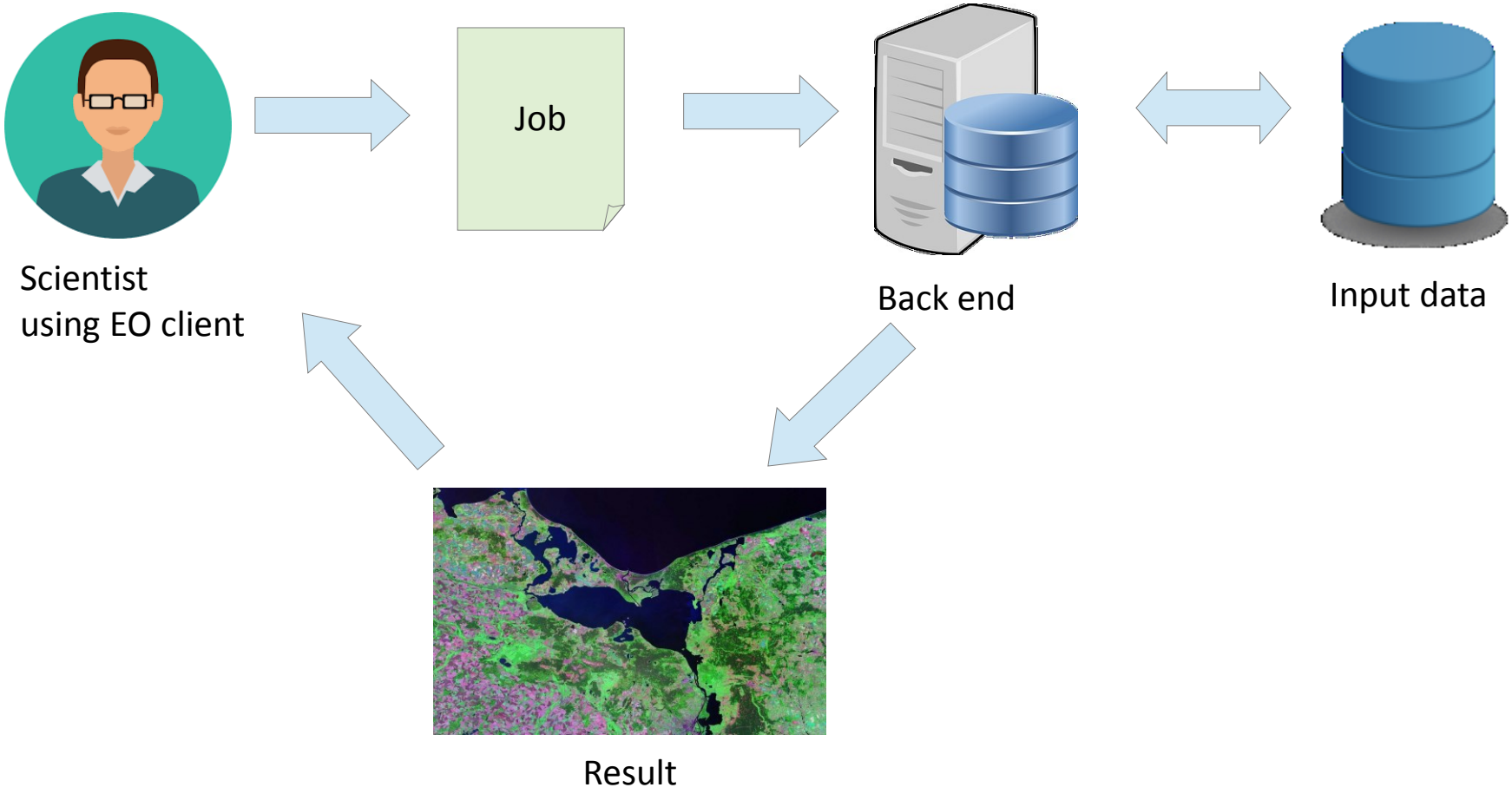
Tomasz Miksa, Bernhard Gößwein

research data sharing without barriers
rd-alliance.org

Designing a Framework Gaining Repeatability for the openEO Platform

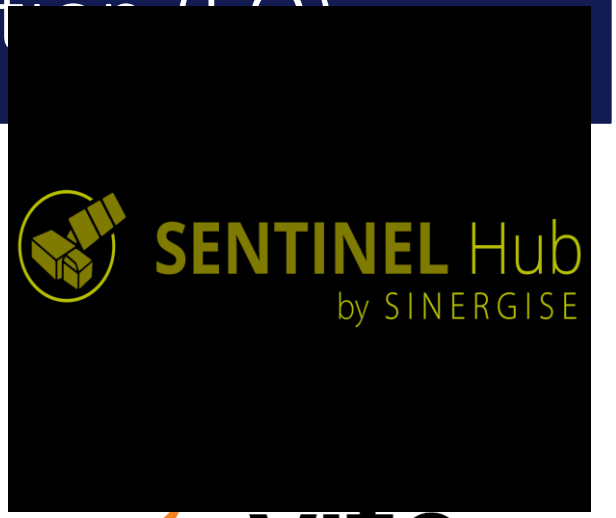
Bernhard Gößwein, Tomasz Miksa, Andreas Rauber, Wolfgang Wagner

Introduction



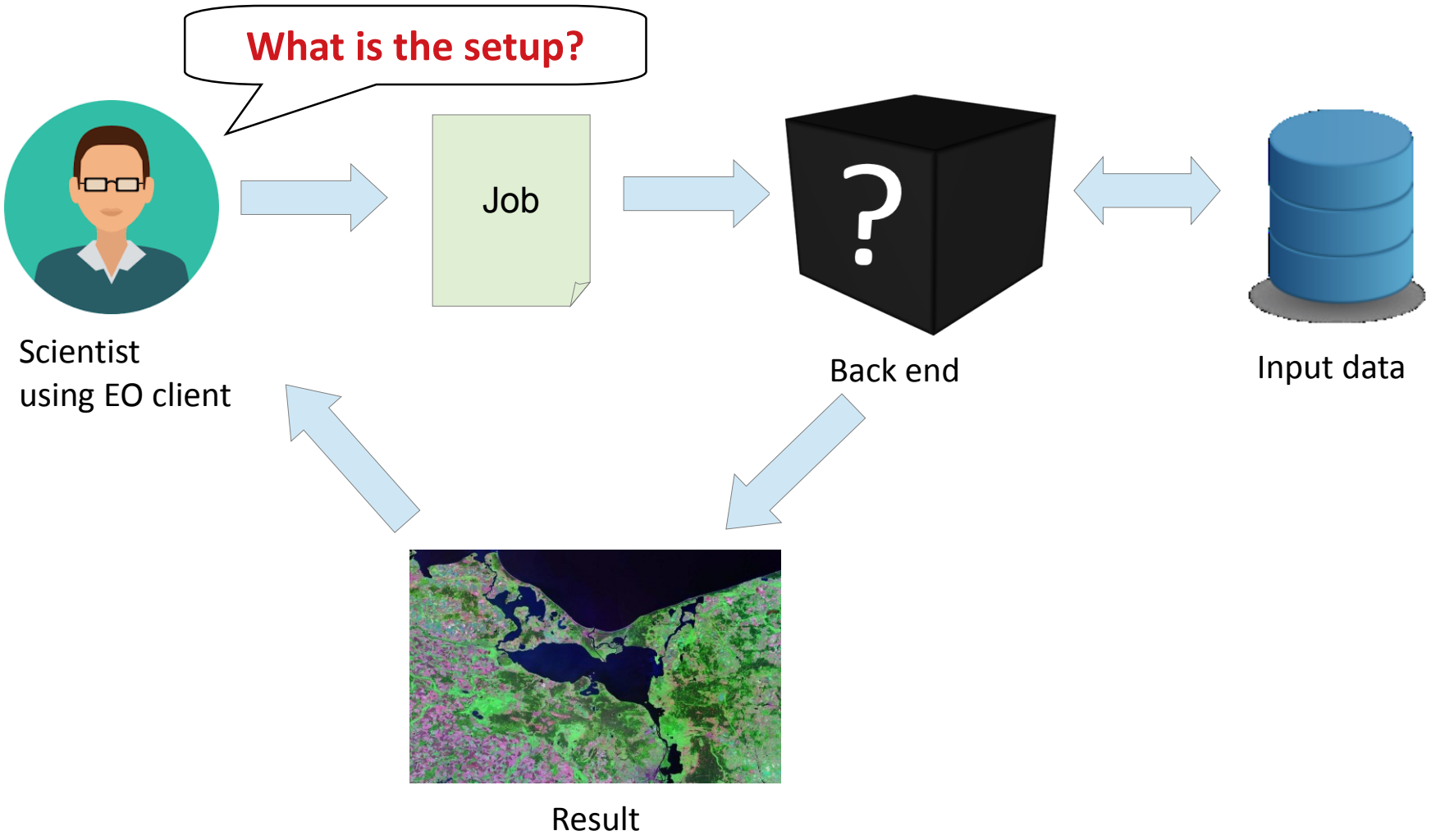
Situation: Earth Observation (EO)

- Diverse set of data provider
- Processing happens at the data provider
- Backends provide data from similar sources e.g. ESA
- openEO provides a standardized API to access multiple backends

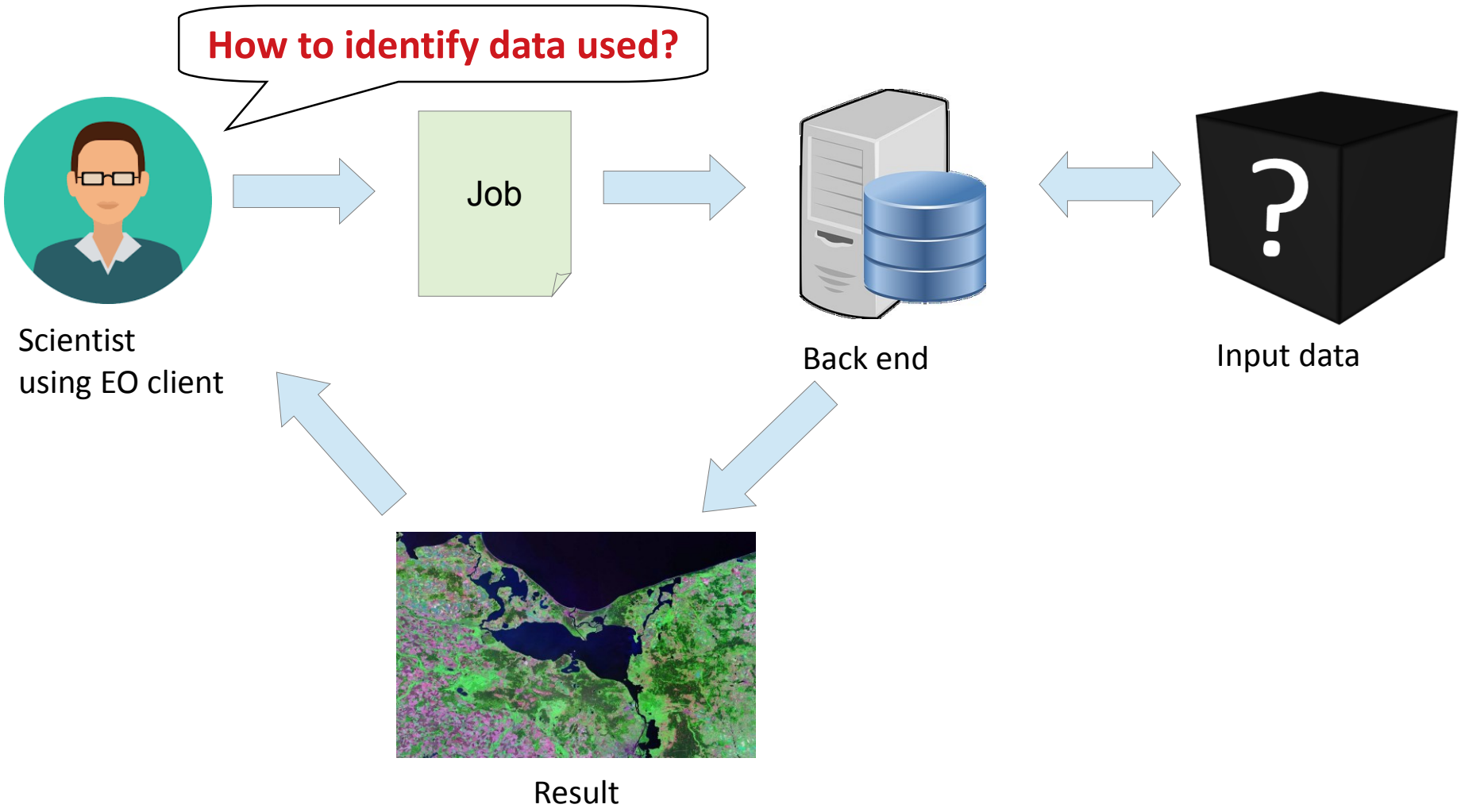


Google Earth Engine

Problem #1 – backend is a black box



Problem #2 –input data identification



Problem #2: Input data changed

Query arguments

- **Temporal extent**: Date range of interest e.g. May of 2018
- **Spatial extent**: Geographic area of interest e.g. rectangle over Los Angeles
- **Spectral bands**: Bands of interest e.g. near infra red

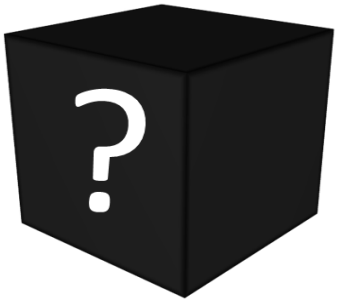
Query result

- **Subset** of the backends satellite data storage
- Input data is usually big, since dimensions have not been reduced yet

Problem #3

Scientist
using EO client

Job



Back end

Input data

How to reproduce?
- what data was used?
- which software was used in processing?

Result

Another scientist

Aim

Back end

Input data

Result

- **Document** relevant software involved in processing, e.g. GDAL
- **NOT enable to** restore previous versions of the backend

- Enable identification of **CHANGING** data without making copies of subsets
- Provide easy way to cite and re-use input data

- **Comparable** - Enable to identify whether differences come from data / environment or a **real** scientific phenomena

Methodology

- **RDA – Research Data Alliance**
  - Recommendations on data identification including citation and retrieval of data that existed at a certain point of time.

[DOI: 10.15497/RDA00016]

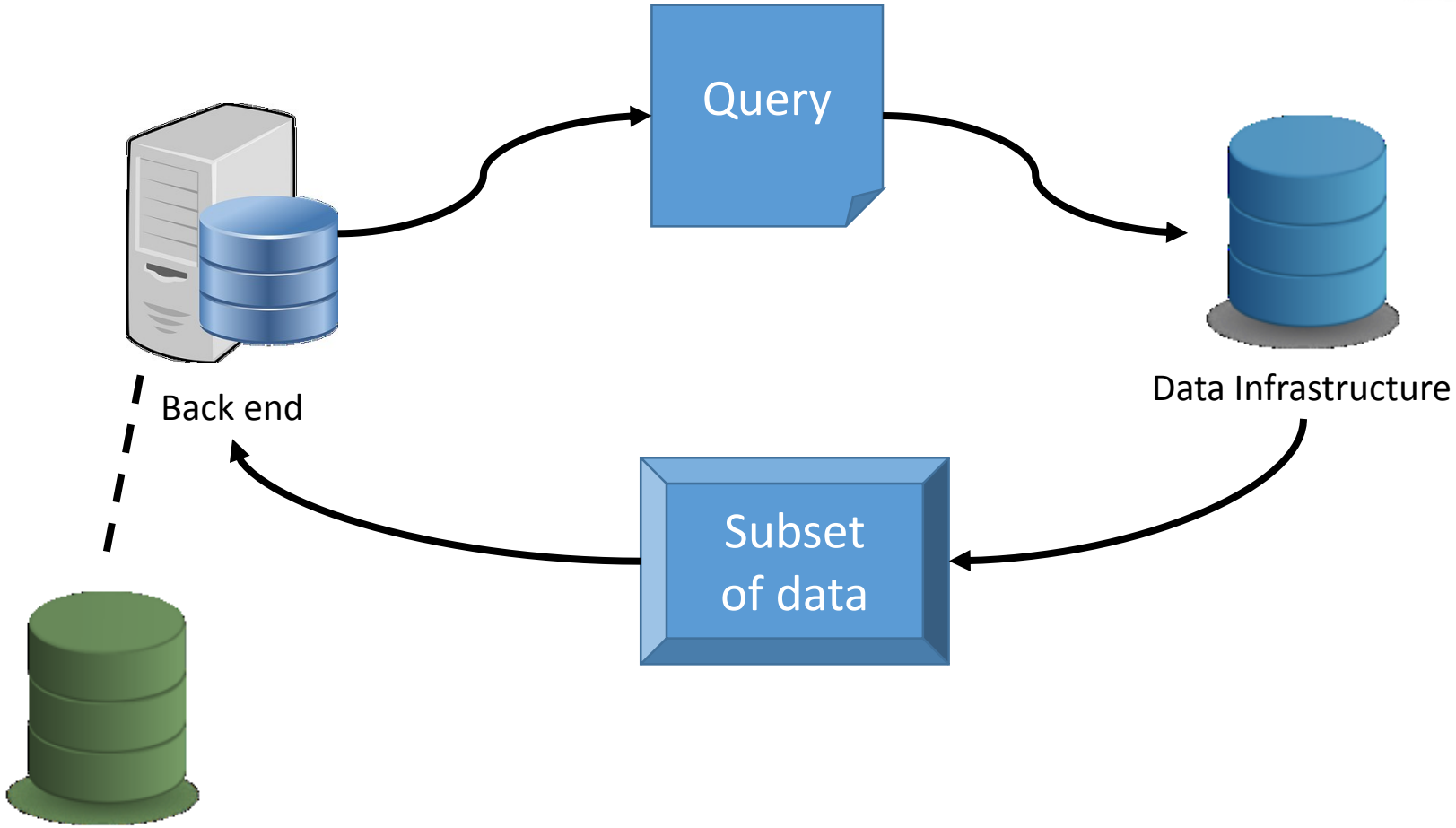
- **VFramework and Context Model**
  - Automatically document execution environments and enable their comparison.

[DOI:10.1016/j.jbi.2016.10.011]

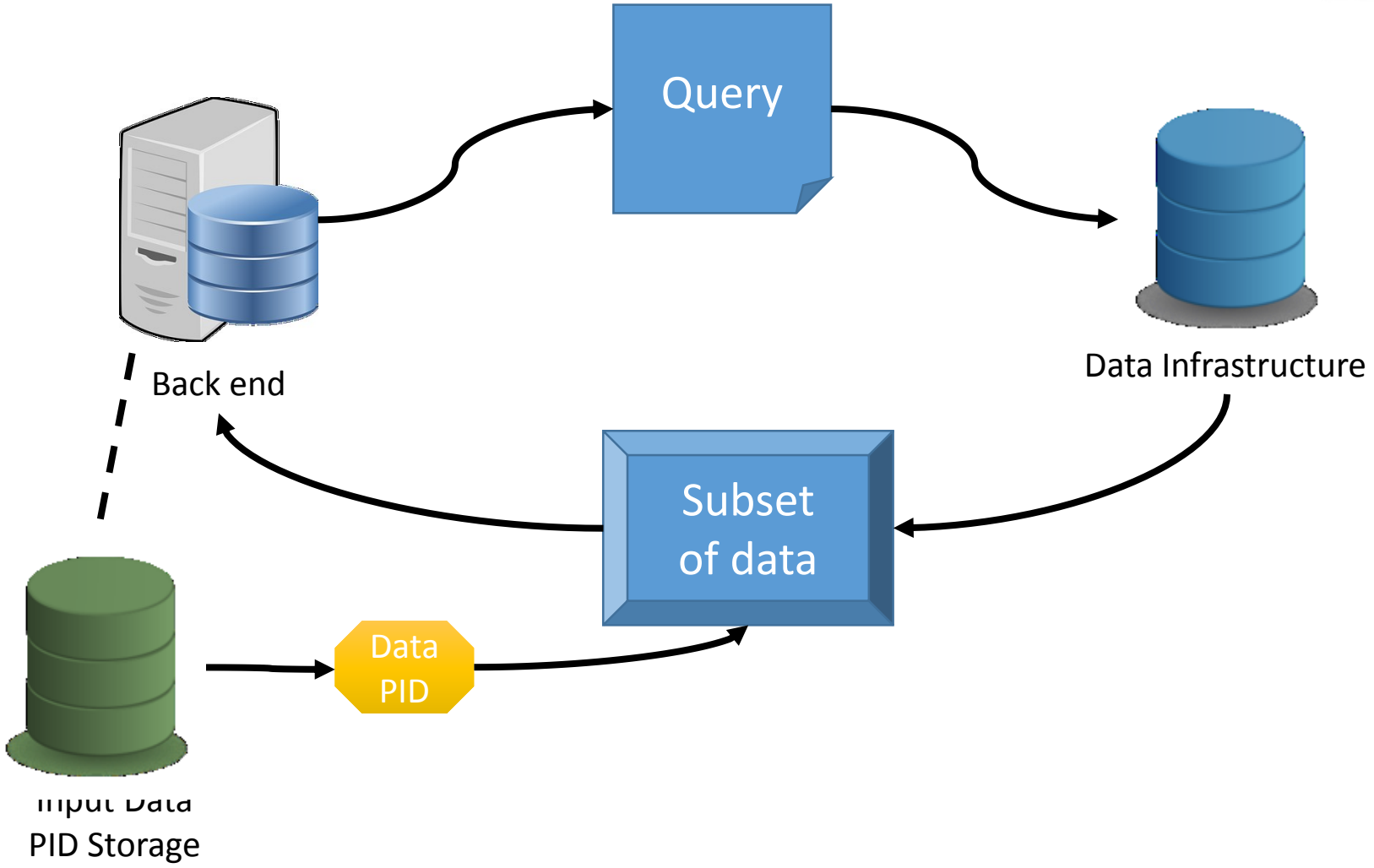
- **OpenEO Project**
  - Common EO interface enabling interoperability of EO backends. Allows researchers to run the same code on different backends.

[DOI:10.5281/zenodo.1065474]

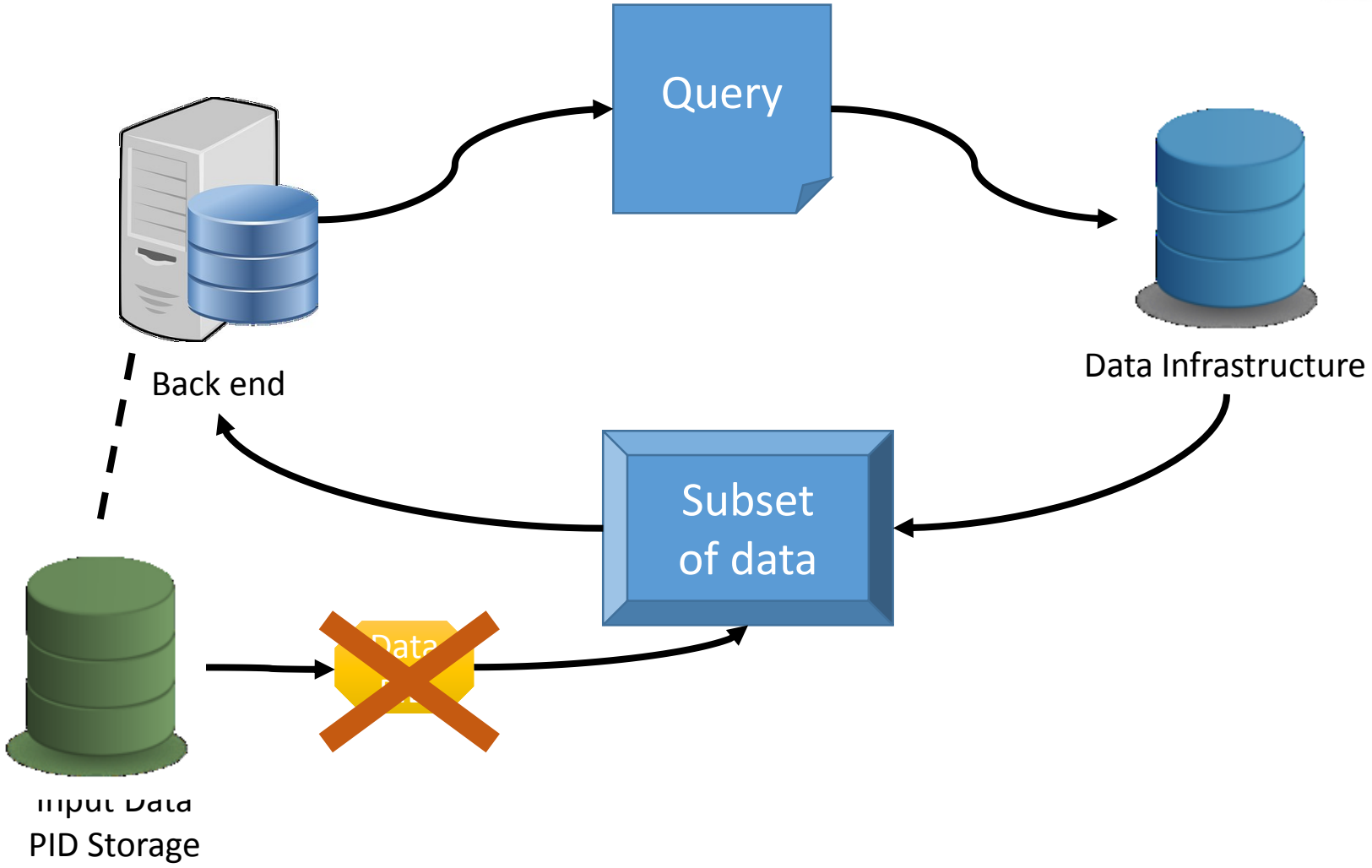
RDA - Data Identification



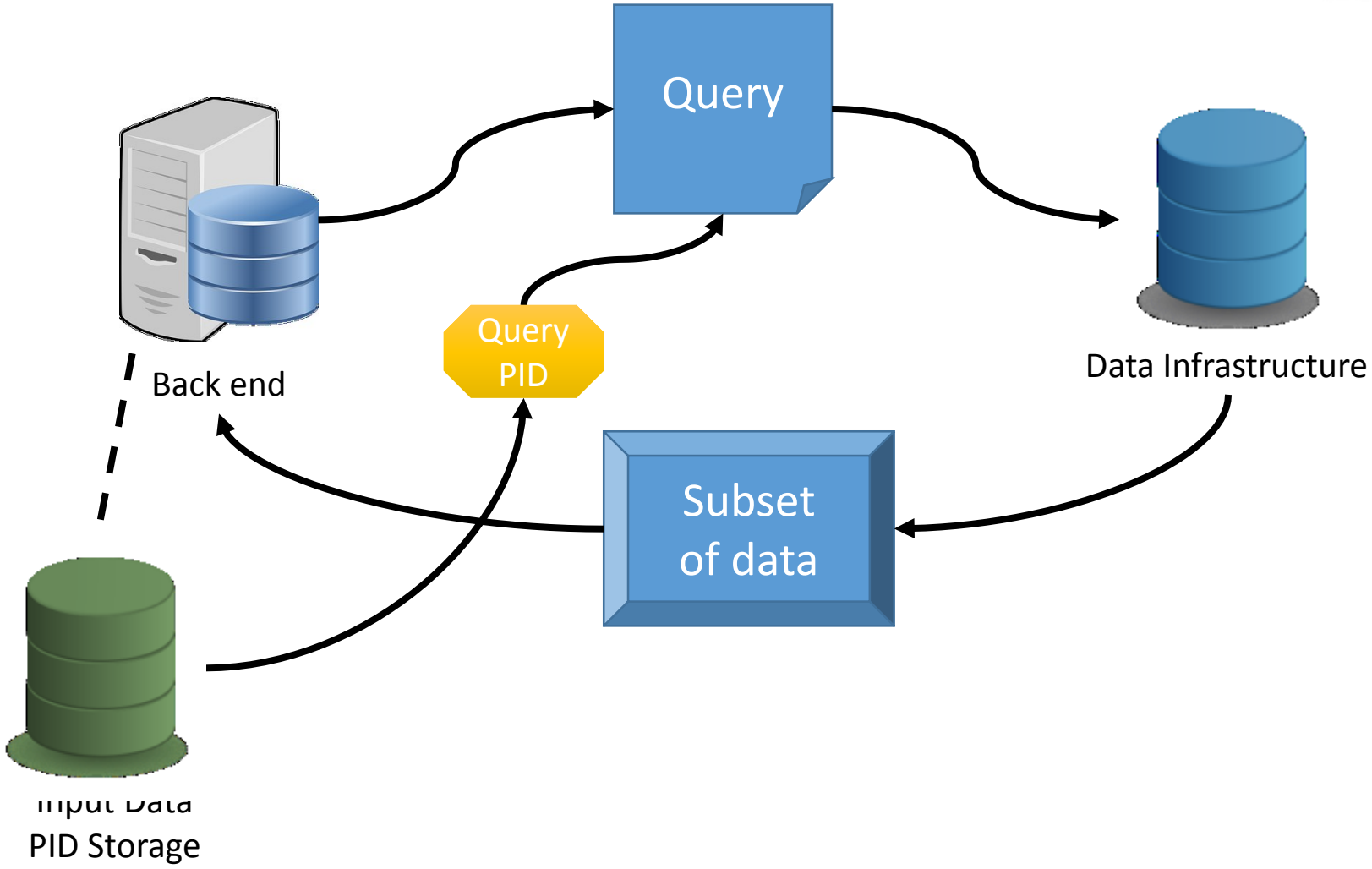
RDA - Data Identification



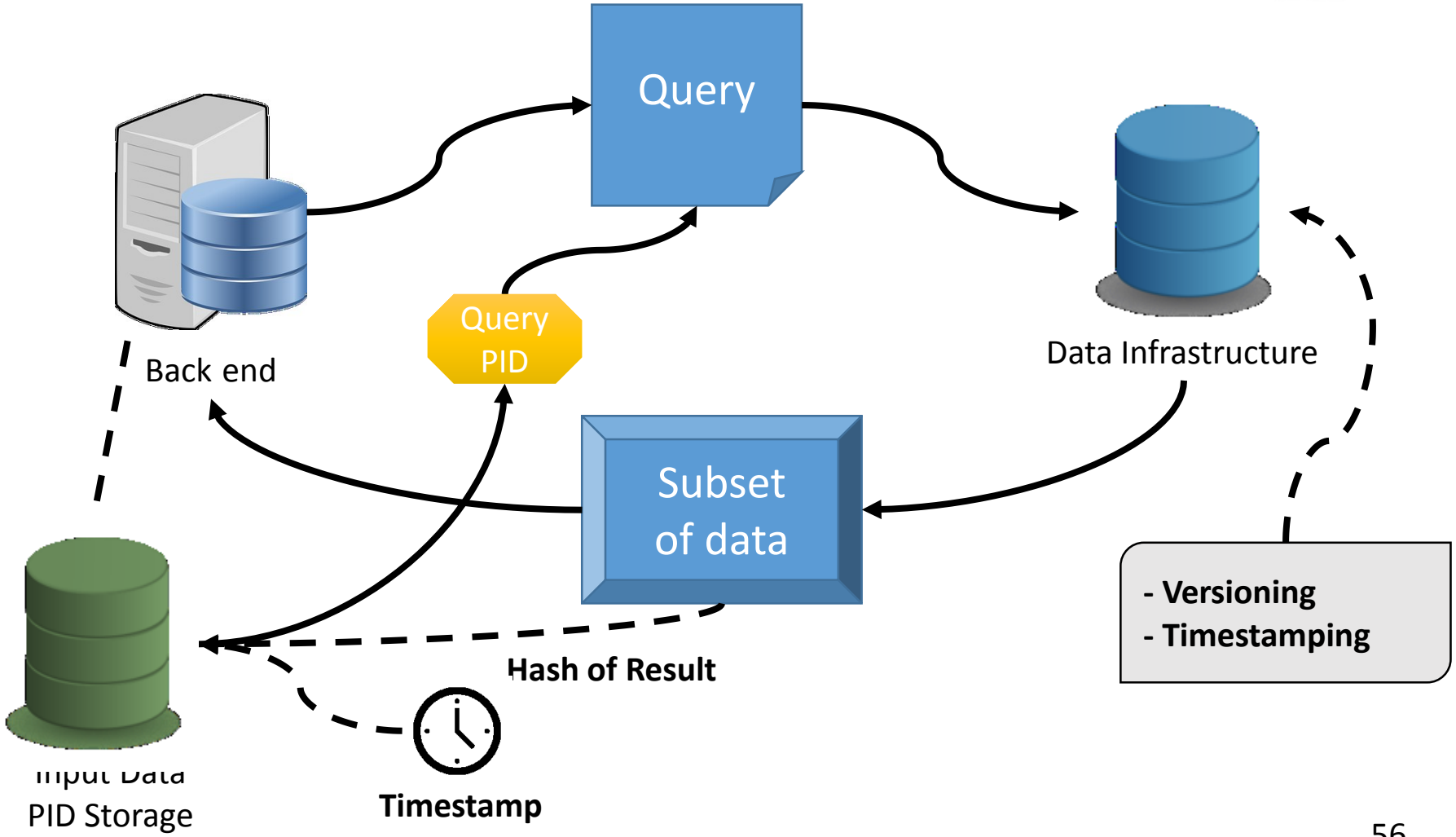
RDA - Data Identification



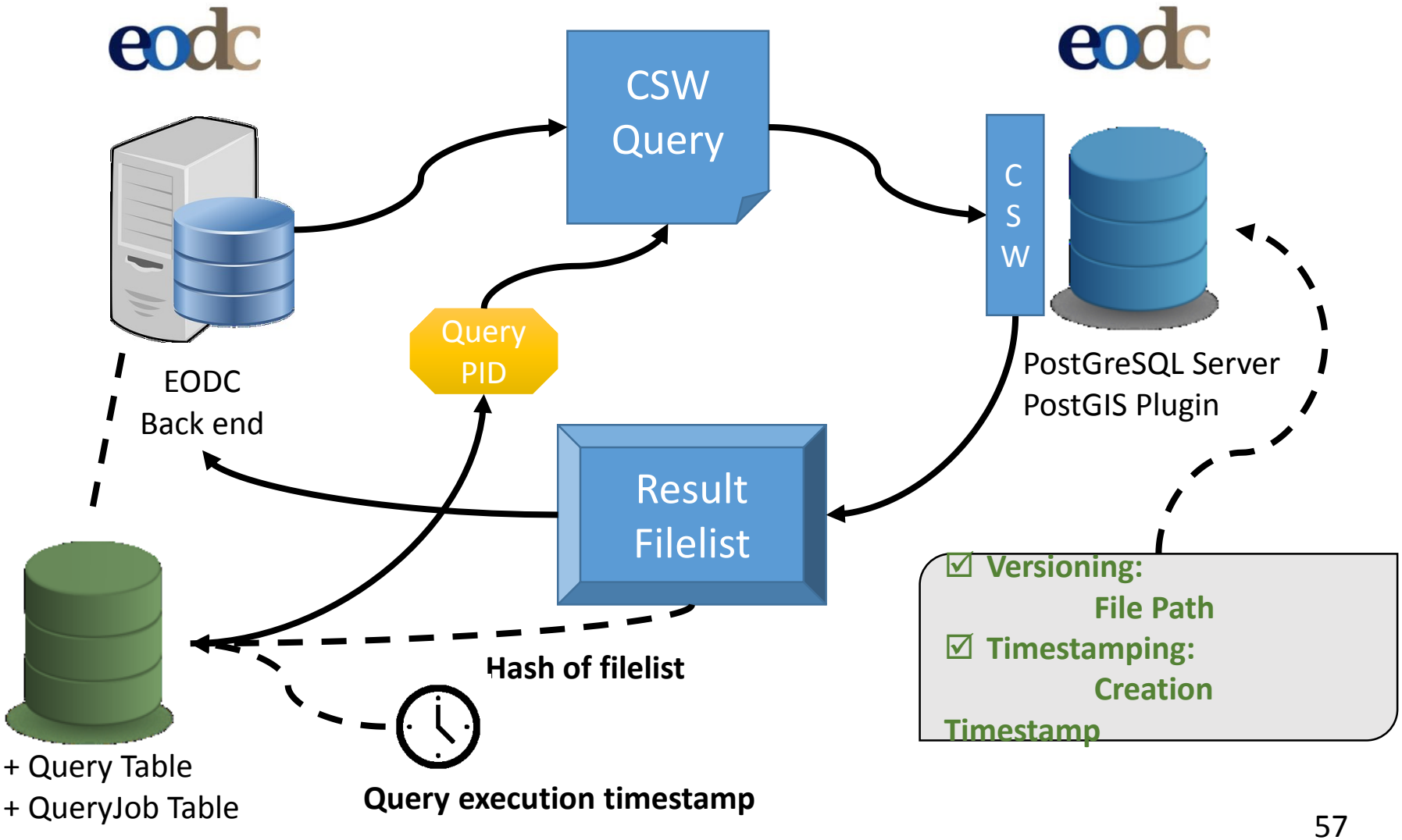
RDA - Data Identification



RDA - Data Identification



Solution - Data Identification



Solution – Example Query

```
<?xml version="1.0" encoding="UTF-8"?>
<csw:GetRecords xmlns:csw="http://www.opengis.net/cat/csw/2.0.2" xmlns:apiso="http://www.opengis.net/cat/csw/apiso/1.0" xmlns:gmd="http://www.isotc211.org/2005/gmd"
                xmlns:gml="http://www.opengis.net/gml" xmlns:ogc="http://www.opengis.net/ogc" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" service="CSW"
                version="2.0.2" resultType="results" startPosition="1" maxRecords="1000" outputFormat="application/json" outputSchema="http://www.isotc211.org/2005/gmd"
                xsi:schemaLocation="http://www.opengis.net/cat/csw/2.0.2 http://schemas.opengis.net/csw/2.0.2/CSW-discovery.xsd">
   <csw:Query typeNames="csw:Record">
      <csw:ElementSetName>full</csw:ElementSetName>
      <csw:Constraint version="1.1.0">
         <ogc:Filter>
            <ogc:And>
               <ogc:PropertyIsEqualTo>
                  <ogc:PropertyName>apiso:ParentIdentifier</ogc:PropertyName>
                  <ogc:Literal>s2a_prd_msil1c</ogc:Literal>
               </ogc:PropertyIsEqualTo>
               <ogc:PropertyIsGreaterThanOrEqualTo>
                  <ogc:PropertyName>apiso:TempExtent_begin</ogc:PropertyName>
                  <ogc:Literal>2017-05-01T00:00:00Z</ogc:Literal>
               </ogc:PropertyIsGreaterThanOrEqualTo>
               <ogc:PropertyIsLessThanOrEqualTo>
                  <ogc:PropertyName>apiso:TempExtent_end</ogc:PropertyName>
                  <ogc:Literal>2017-05-31T23:59:59Z</ogc:Literal>
               </ogc:PropertyIsLessThanOrEqualTo>
               <ogc:BBOX>
                  <ogc:PropertyName>ows:BoundingBox</ogc:PropertyName>
                  <gml:Envelope>
                     <gml:lowerCorner>46.905246 10.288696</gml:lowerCorner>
                     <gml:upperCorner>45.935871 12.189331</gml:upperCorner>
                  </gml:Envelope>
               </ogc:BBOX>
               <ogc:PropertyIsLessThanOrEqualTo>
                  <ogc:PropertyName>apiso:Modified</ogc:PropertyName>
                  <ogc:Literal>2019-03-31 17:36:43.064445</ogc:Literal>
               </ogc:PropertyIsLessThanOrEqualTo>
            </ogc:And>
         </ogc:Filter>
      </csw:Constraint>
      <ogc:SortBy>
         <ogc:SortProperty>
            <ogc:PropertyName>dc:date</ogc:PropertyName>
            <ogc:SortOrder>ASC</ogc:SortOrder>
         </ogc:SortProperty>
      </ogc:SortBy>
   </csw:Query>
</csw:GetRecords>
```

```
{
    "filter_bbox": {
        "left": 650000,
        "right": 672000,
        "srs": "EPSG:32632",
        "top": 5161000
        },
    "filter_daterange": {
        "from": "2018-01-01",
        "to": "2018-01-08"
        },
    "product_id": "s1a_csar_grdh_iw"

}
```

Unique Query

Solution: openEO Python client example

```
con = openeo.connect("http://openeo.local.127.0.0.1.nip.io")
# Choose dataset
processes = con.get_processes()
pgA = processes.get_collection(name="s2a_prd_msil1c")
pgA = processes.filter_daterange(pgA, extent=["2017-05-01", "2017-05-31"])
pgA = processes.filter_bbox(pgA, west=10.288696, south=45.935871,
east=12.189331, north=46.905246, crs="EPSG:4326")
# Choose processes
pgA = processes.ndvi(pgA, nir="B08", red="B04")
pgA = processes.min_time(pgA)
# Create and start job A out of the process graph A (pgA)
jobA = con.create_job(pgA.graph)
jobA.start_job()
# Get data PID of jobA
pidA = jobA.get_data_pid()
# Re-execute the query to print the
file_listA = con.get_filelist(pidA)
# Get state of the resultfiles, so i
# the original execution
file_listA["input_files"]["state"] #
```

```
# Take input data of job A by using the input data PID A of job A
pgC = processes.get_data_by_pid(data_pid=pidA)
# Choose processes
pgC = processes.ndvi(pgC, nir="B08", red="B04")
pgC = processes.min_time(pgC)
# Create and start Job C
jobC = con.create_job(pgC.graph)
jobC.start_job()
# re-execute query and get the resulting file list from the backend
pidC = jobC.get_data_pid()
file_listC = con.get_filelist(pidC)
# Compare resulting files with the original execution of jobA
(file_listA == file_listC) # Returns True
```

Solution: Data PID - Landing Page

Earth Observation Data Centre
for Water Resources Monitoring
An open and international cooperation

Cite this dataset:

Using this data set or resource, you should cite it with the following citation text:

Copernicus Sentinel data (2017). Retrieved from EODC, Austria [2019-04-17], processed by ESA. PID:
http://openeo.local.127.0.0.1.nip.io/data/qu-d1701f4e-e7c5-4a83-92e0-9facbd401a06

Show Result **JSON**

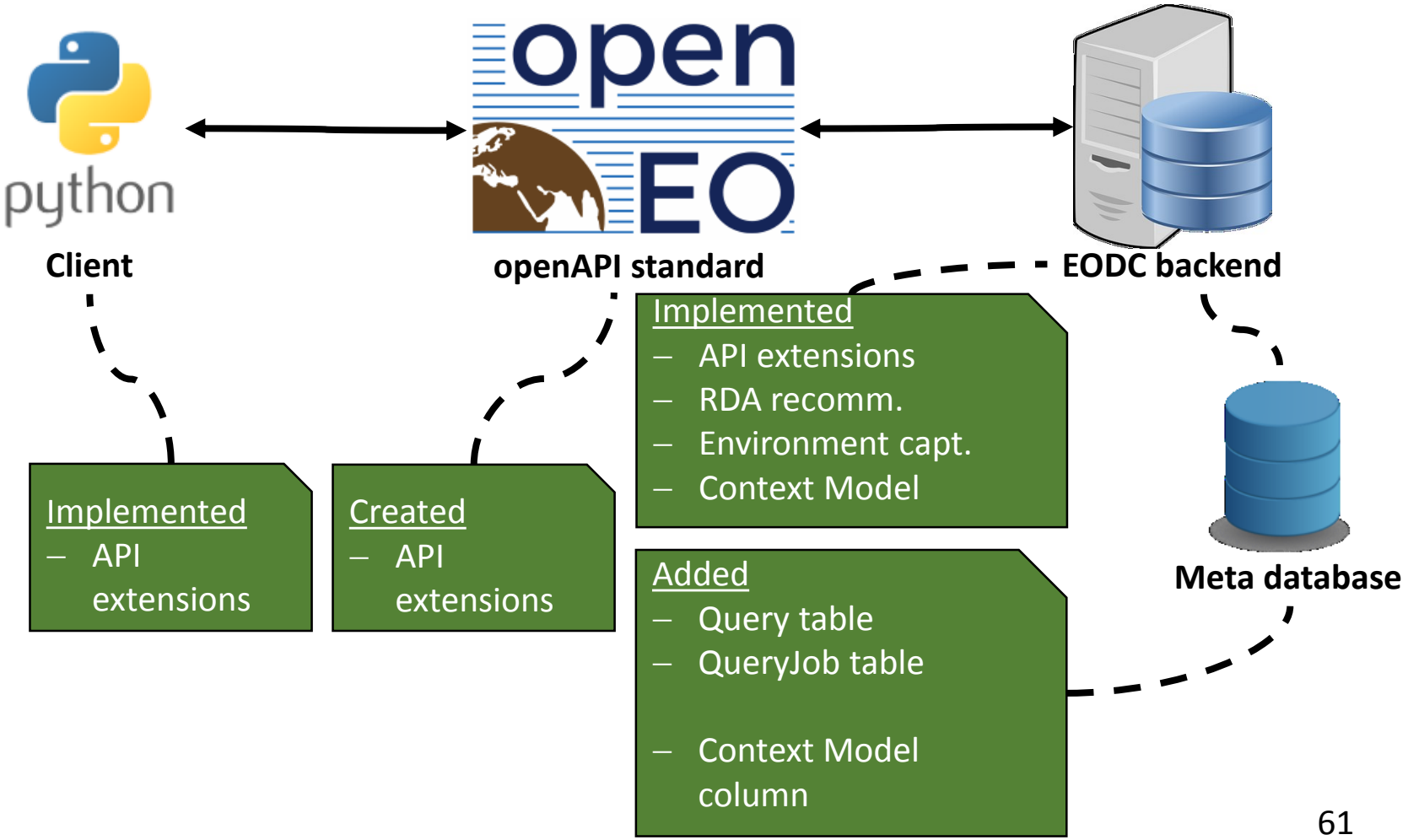
Source data description

Sentinel-2 is a multispectral, high-resolution, optical imaging mission, developed by the European Space Agency (ESA) in the frame of the Copernicus program of the European Commission.

Dataset Metadata

Organization	EODC
Data source (Author)	ESA - Copernicus Program
Data source Identifier	s2a_prd_msil1c
Date of creation	2019-04-17 15:46:11.728540
Spatial Extent	BoundingBox CRS: EPSG:4326 WEST: 10.288696, EAST: 45.935871 SOUTH: 45.935871, NORTH: 46.905246
Temporal Extent	from 2017-05-01 to 2017-05-31

Solution: Overview



Data identification and process monitoring for reproducible earth observation research

Bernhard Gößwein
TU Wien
Vienna, Austria

Tomasz Miksa
TU Wien & SBA Research
Vienna, Austria

Andreas Rauber
TU Wien
Vienna, Austria

***Abstract*—Earth observation researchers use specialised computing services for satellite image processing offered by various data backends. The source of data is often the same, for example Sentinel-2 satellites operated by the European Space Agency, but the way how data is pre-processed, corrected, updated, and later analysed may differ among the backends. Backends often lack mechanisms for data versioning, for example, data corrections are not tracked. Furthermore, an evolving software stack used for data processing remains a black box to researchers. Researchers have no means to identify why executions of the same code deliver different results. This hinders repeatability and reproducibility of earth observation experiments. In this paper, we present how infrastructure of existing earth observation data backends can be modified to support reproducibility. The proposed extensions are based on recommendations of the Research Data Alliance regarding data identification and the VFramework for process capturing. We implemented our approach at the Earth Observation Data Centre, which is a partner within the openEO project. We evaluated the solution on typical usage scenarios. We also provide performance and storage measures to evaluate the impact of the modifications on performance. The results indicate reproducibility can be supported with minimal performance and storage overhead.**

I. Introduction

Earth Observation (EO) data consists mostly of satellite images. Similar as in the other eScience disciplines, data is too big to be downloaded for local analysis. The solution is to store it in high-performance computational backends, process it there, and browse the results or download resulting figures or numbers [13].

Such an approach addresses the performance issues, but does not allow researchers to take a full control of the environment in which their experiments are executed. The backends act as black boxes to the researchers with no possibility of getting information on environment configuration, e.g. software libraries used in processing and their versions. Studies in different domains show that environment can have impact on reproducibility of scientific experiments and must be documented in order to ensure reproducibility [4] [1] [8]. Still the vast majority of backend providers do not share the environment information.

Another problem deals with a precise identification of data used for processing. EO backends in Europe usually obtain data from the same source, for example from the Sentinel-2 satellites operated by the European Space Agency (ESA). The ESA releases updates and corrections to data in cases when one of the instruments used for observation was wrongly calibrated or broken and raw data had to be pre-processed again. Updated data is released to backend operators. Usually there is no versioning mechanism for data. Researchers do not know which version of data was used in their study, i.e. before or after the correction was made available at the backend. This leads to a problem that scientists are not capable of precisely identifying the input data of their experiments, which hinders reproducibility and in turn undermines trust in the results.

Research Data Alliance (RDA) has identified 14 general rules [2] for identification of data used in computation that allows to cite and retrieve that data as it existed at a certain point in time. The VFramework [8] and context model [10] were proposed to automatically document environments in which computational workflows execute and to enable their comparison. The openEO project [7] works on creating a common EO interface to enable interoperability of EO backends by allowing researchers to run their experiments on different backends without reimplementing their code.

In this paper, we build on top of these developments and present a solution improving reproducibility of earth observation experiments executed at the openEO compliant backends. We follow the RDA recommendations for data identification and present how data provided by backends is made identifiable by assigning identifiers to queries made by researchers. We discuss which specific information must be captured, which interfaces must be modified, and which software components must be implemented. We also show how jobs executed at backends can be captured and compared using the VFramework to identify whether any differences in software dependencies among two executions exist. We implemented our solution for the backend of the Earth Observation Data Centre for Water Resources Monitoring (EODC). In evaluation we simulated typical use cases representing updates of data and changes in the backend environment. We also measured the performance and storage impact on the backend, which turned out to be minimal.

The remainder of this paper is structured as follows. Section II presents related work that is a basis of our solution and provides earth observation context. Section III presents architecutre of the proposed solution. Section IV presents implementation of the prototype at the EODC backend. Section V presents methods offered to researchers enhancing reproducibility. Section VII describes the experimental evaluation and discussion. Conclusion appears in Section VIII.

15th eScience 2019
INTERNATIONAL CONFERENCE

Bernhard Gößwein, Tomasz Miksa, Andreas Rauber, Wolfgang Wagner. Data Identification and Process Monitoring for Reproducible Earth Observation Research. IEEE eScience 2019, San Diego, USA.

RESEARCH DATA ALLIANCE

Climate Change Center Austria

Chris Schubert

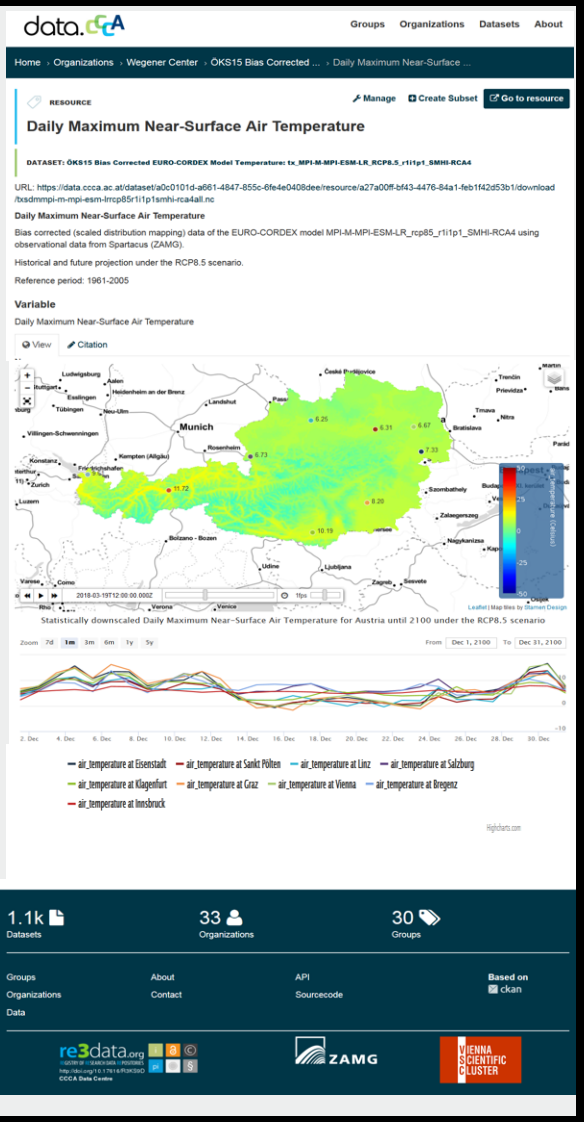
research data sharing without barriers

rd-alliance.org

DYNAMIC DATA CITATION FOR FREQUENTLY MODIFIED HIGH RESOLUTION CLIMATE DATA

Chris Schubert
Head of CCCA – Data Centre
data.ccca.ac.at
1190 Vienna, Austria

chris.schubert[at]ccca.ac.at



CONNECT

We promote interoperability and collaboration between different science and research communities.

Forschungsinfrastruktur

re3

RE RE RE

http://doi.org/10.17616/R3K98D

CCCA Data Centre

SERVICES

Publish and cite resources & data

Centralized access to relevant meta-information

Storage, Server, VM & HPC facilities

On the fly preview of NetCDF files

Create subsets of large NetCDF files

RESPONSIBLE FOR A BETTER

Data Access & Reuse

Data Preservation

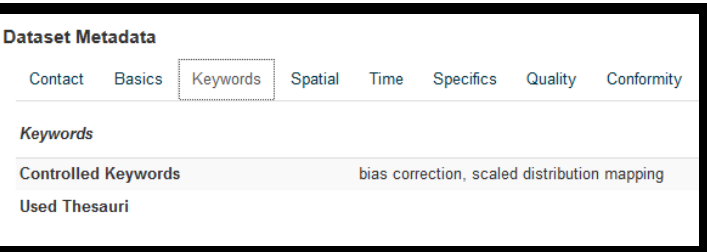
Data Processing and analysis

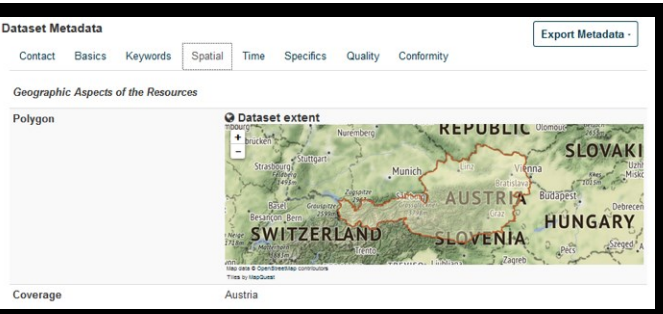
Domain tailored Data Management

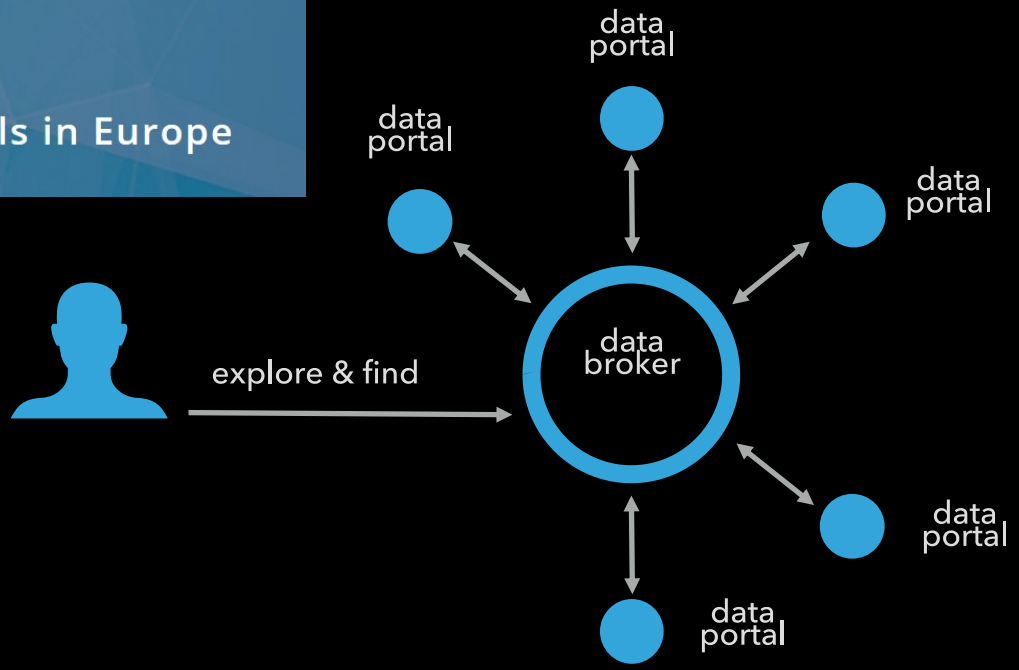
Data Life Cycle, Data Provenance

DCAT-AP
FOR
DATA PORTALS
IN EUROPE

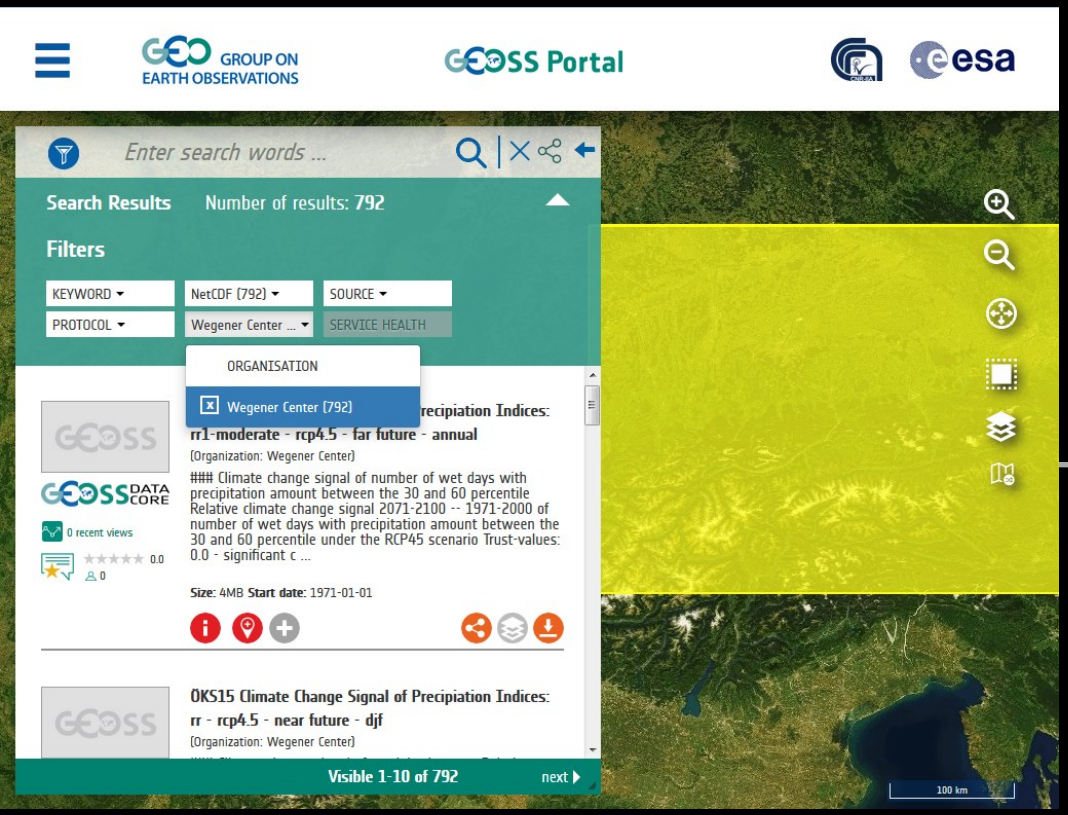
DCAT application profile for data portals in Europe

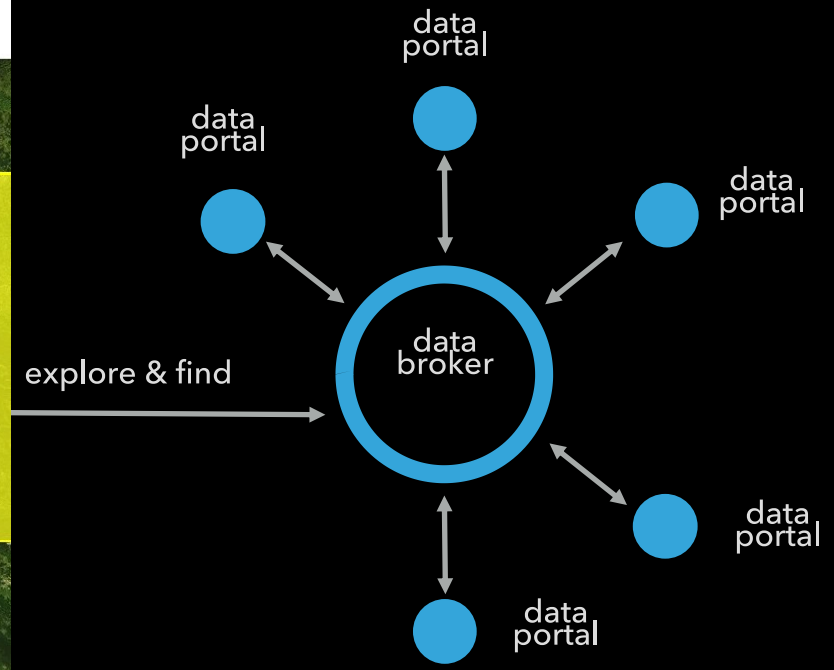




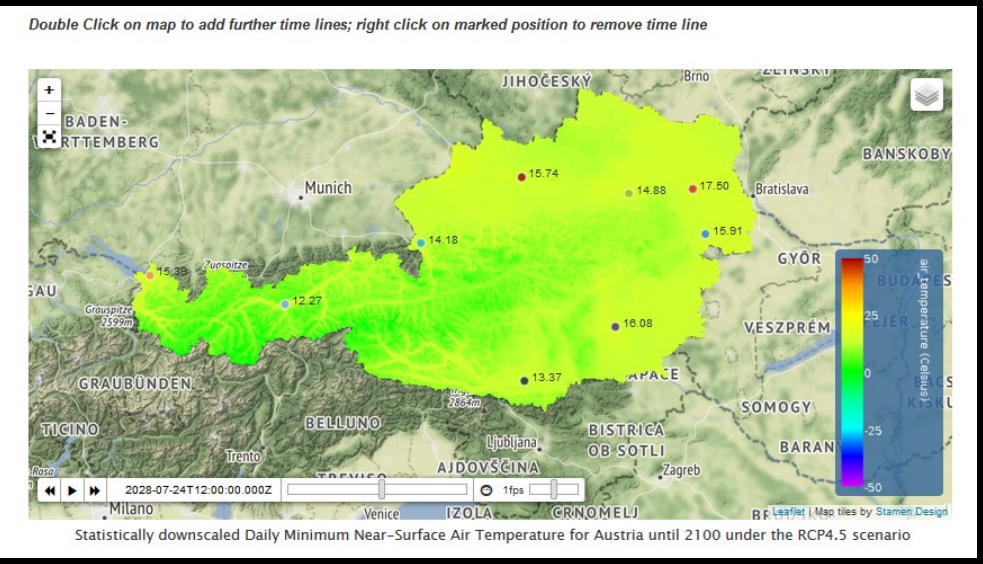


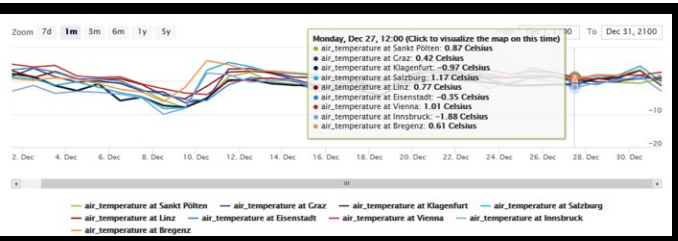
METADATA DCAT - AP

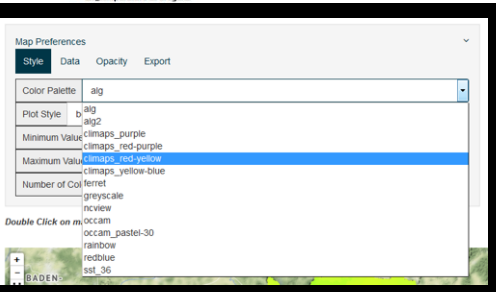




METADATA DCAT - AP







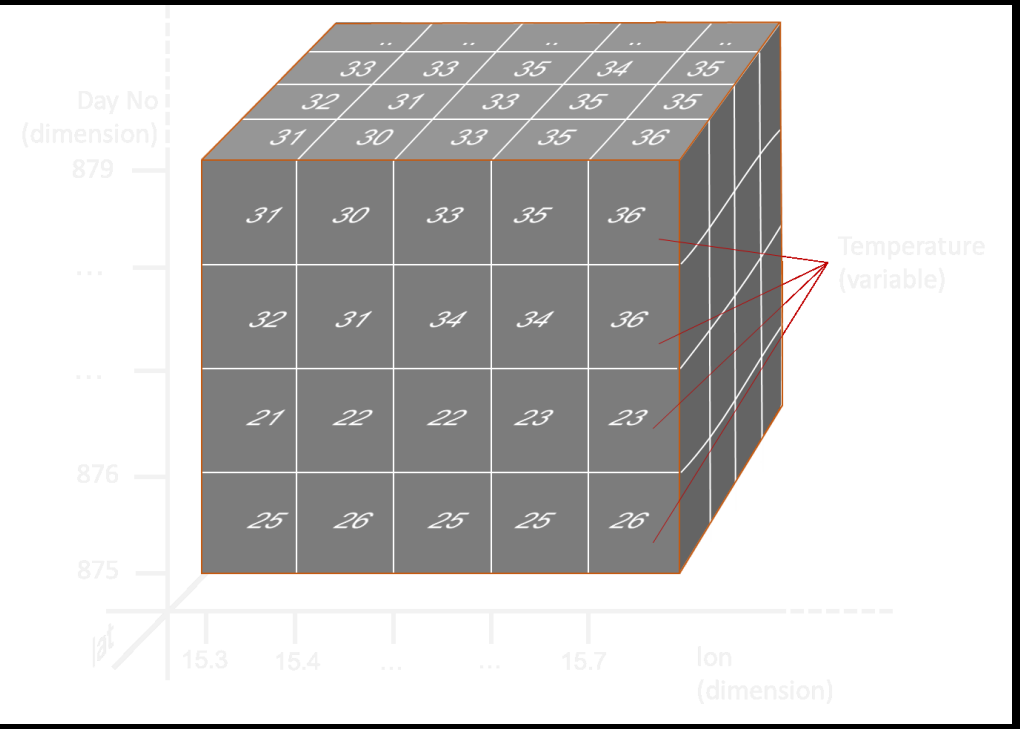
visual context

easier exposed and

recognized information

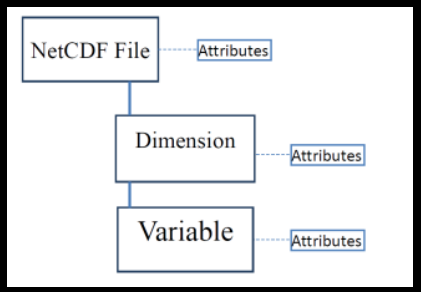
VISUALISATION

SHOW your data



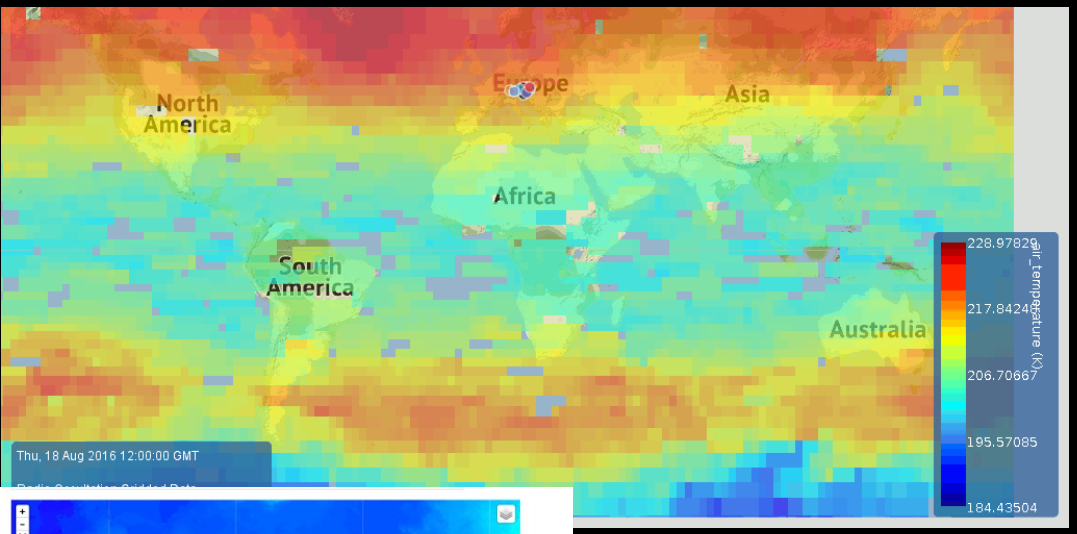
net*work Common Data Form*

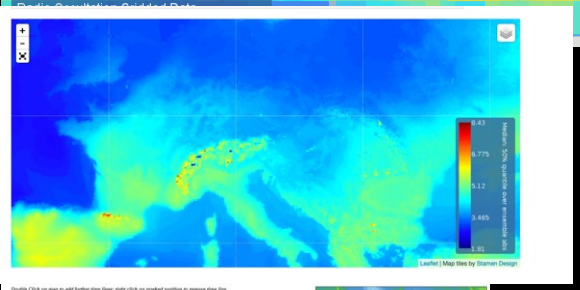
... more than a data format

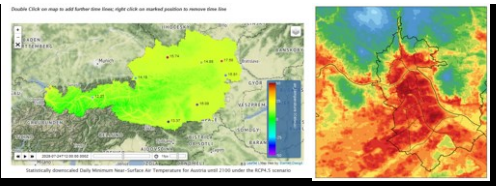


STANDARDISED DATA FORMAT

SHOW your data



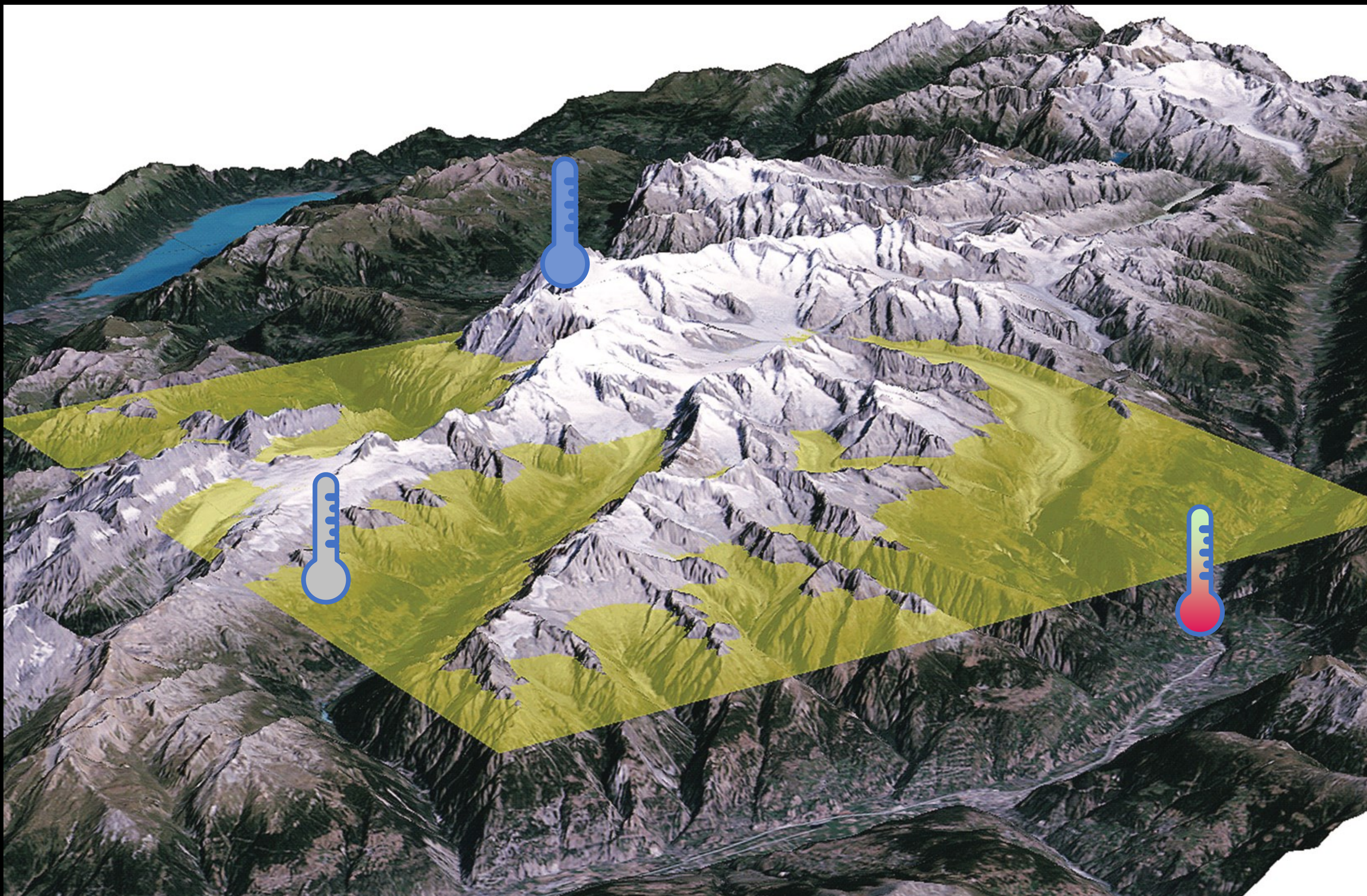




different extent

STANDARDISED DATA FORMAT

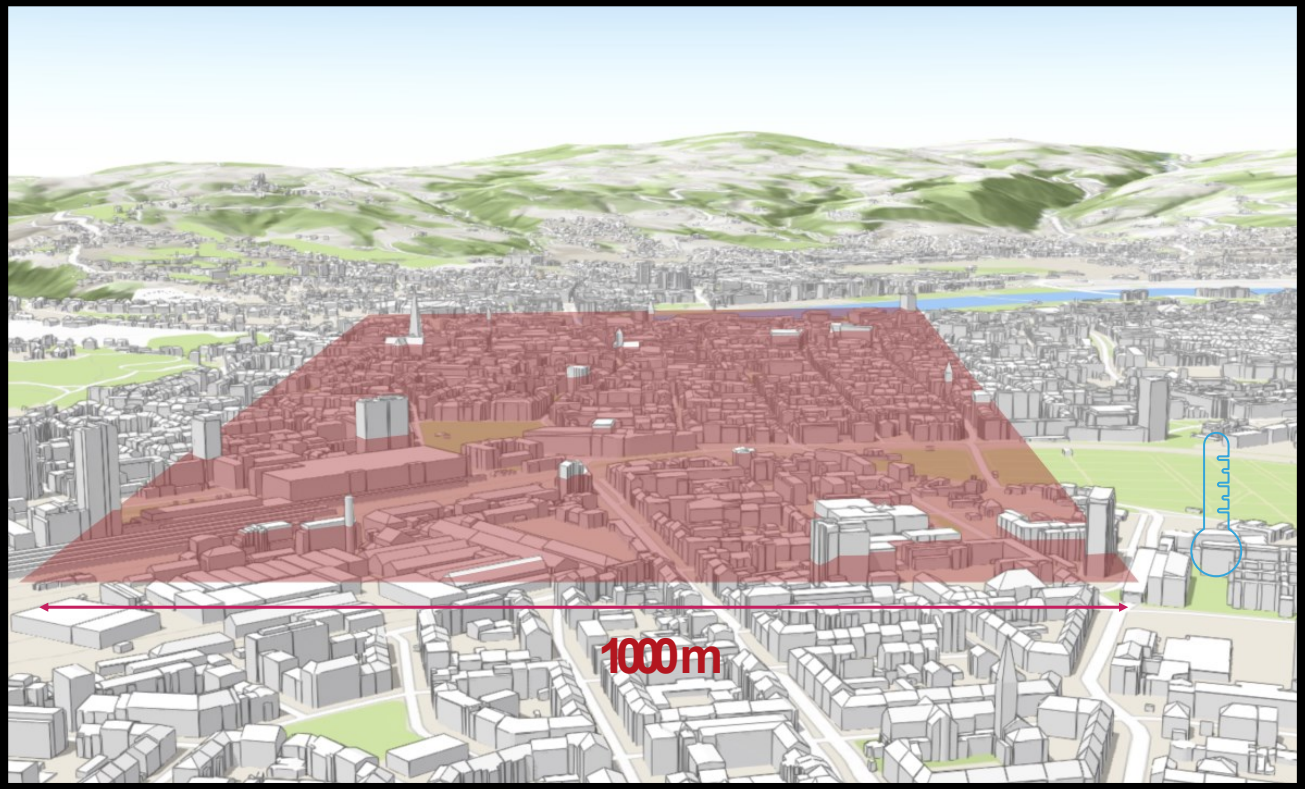
SHOW your data



The issue on the level of detail

Gruber, Stephan. (2011). Derivation and analysis of a high-resolution estimate of global permafrost zonation. The Cryosphere Discussions. 5. 1547-1582. 10.5194/tcd-5-1547-2011.

Urban Resilience to Extreme Weather



https://arcg.is/OnKnf

Sketch on grid dimension of Very High Resolution Climate Data needed for Urban Planning

Clarity
http://clarity-h2020.eu/

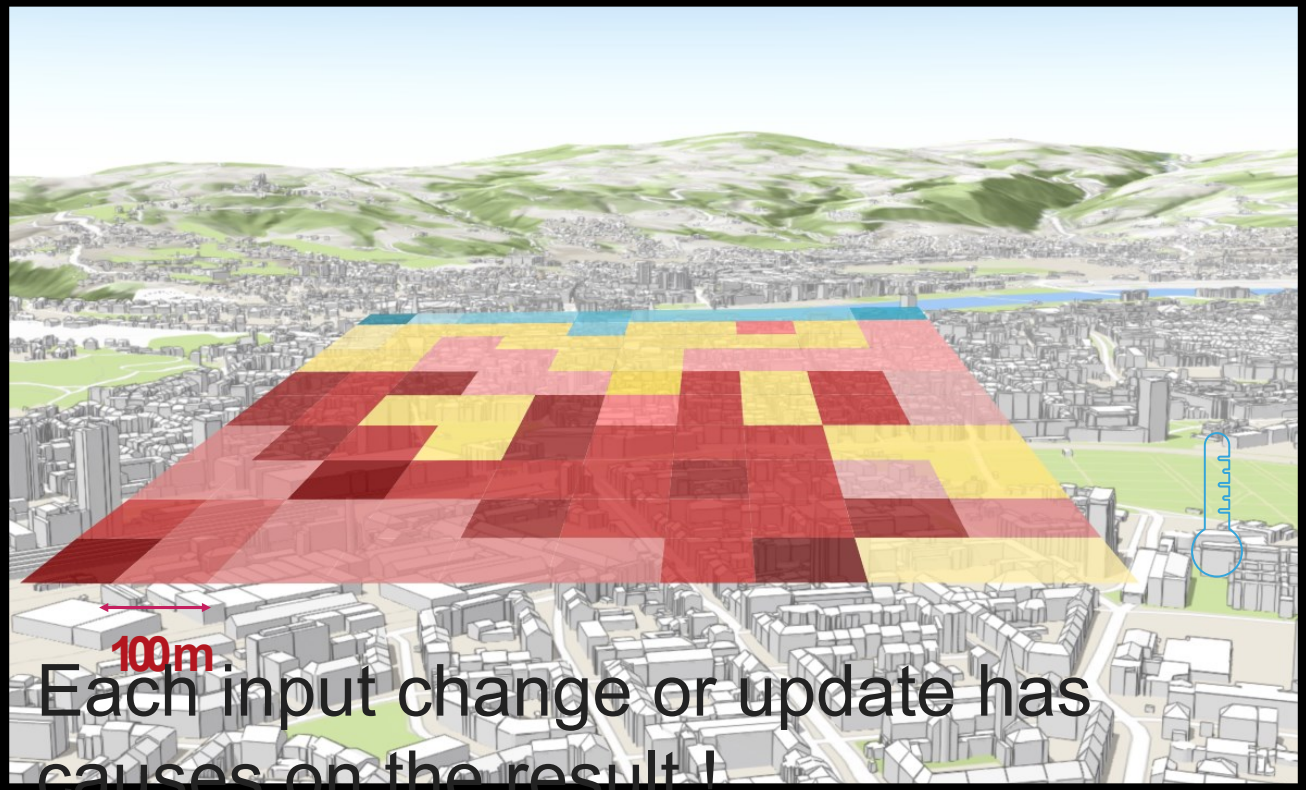
e-shape
disaster climate

The issue on the level of detail

Disaster Risk Management Knowledge Centre

data.CCA

Urban Resilience to Extreme Weather



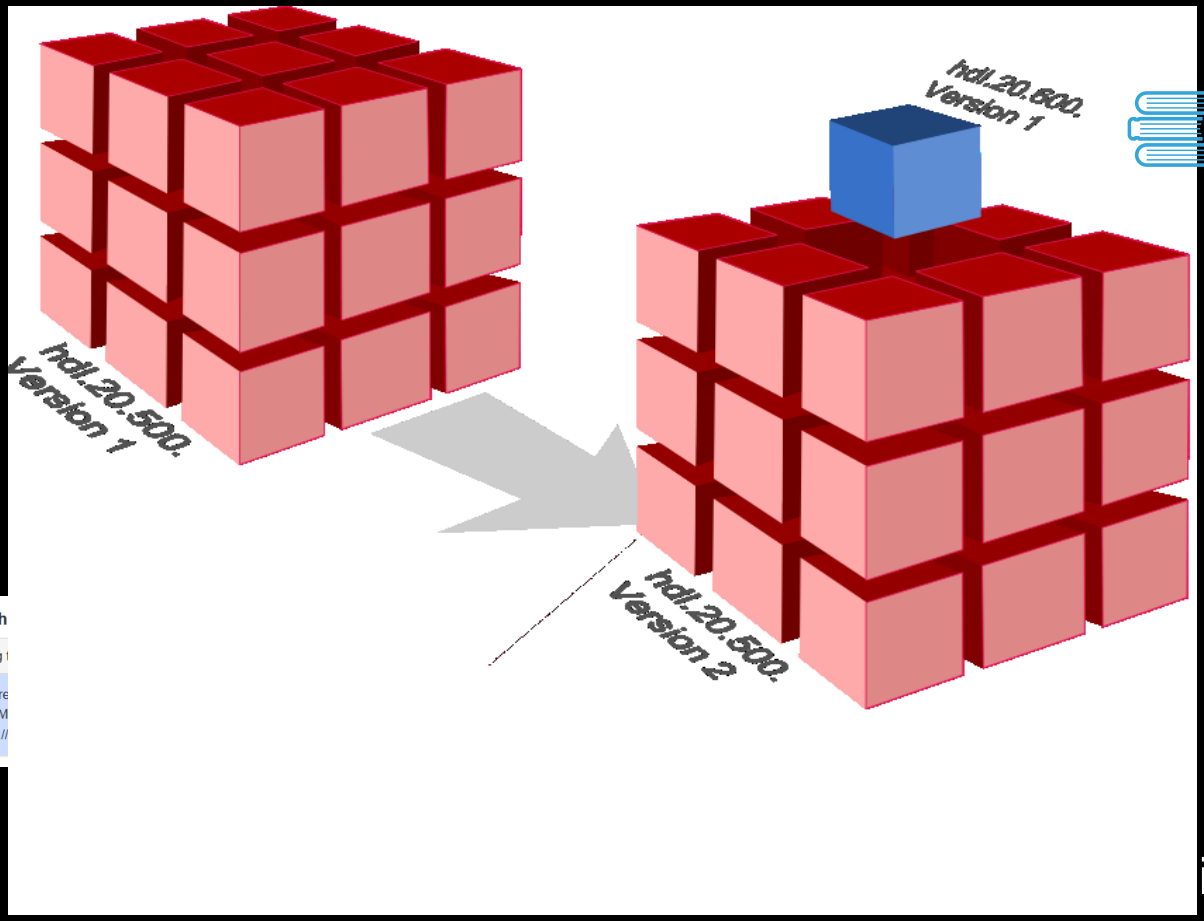
Each input change or update has causes on the result !

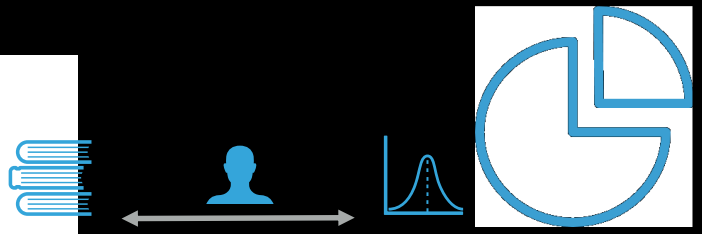
https://arcg.is/OnKnf

Sketch on grid dimension of High Resolution Climate Data

model input data:

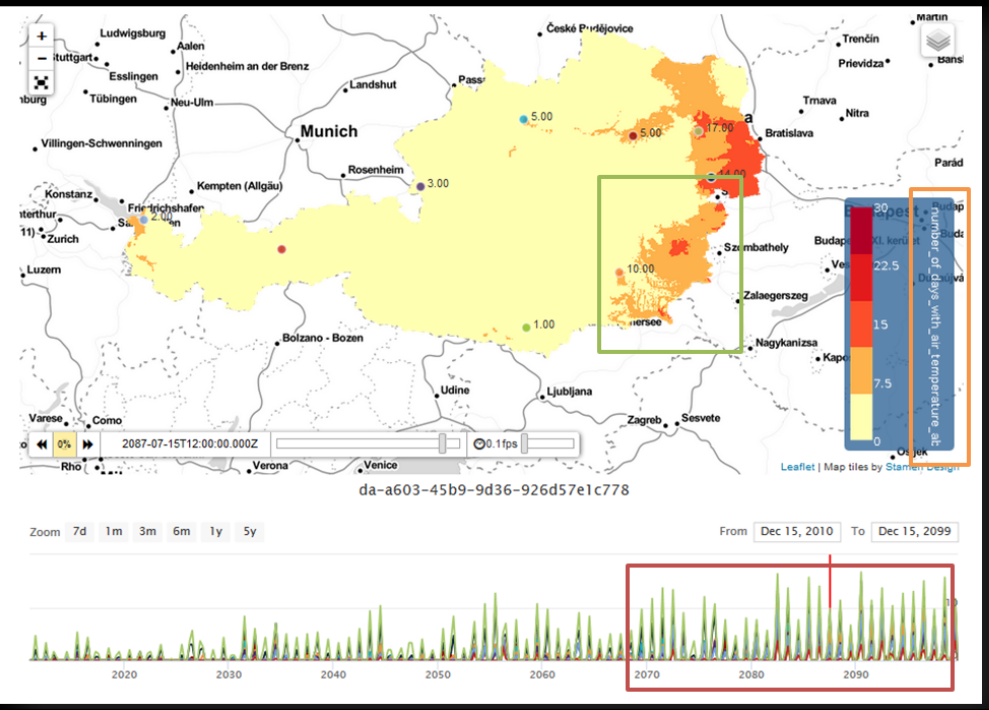
- Land Use
- Roughness, surface structure
- Stations, Sensor
- Citicen Science Sensor, e.g. NetATMO
- etc.

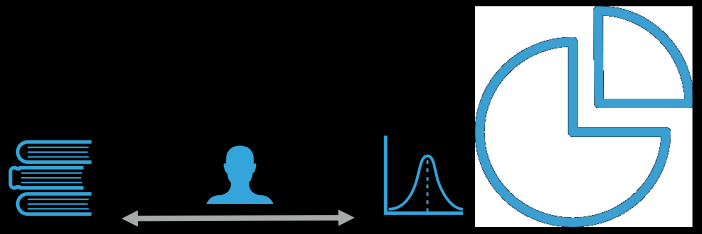




ic data citation

Cite your Data





(research) data is dynamic

identify precisely the data at a specific point in time

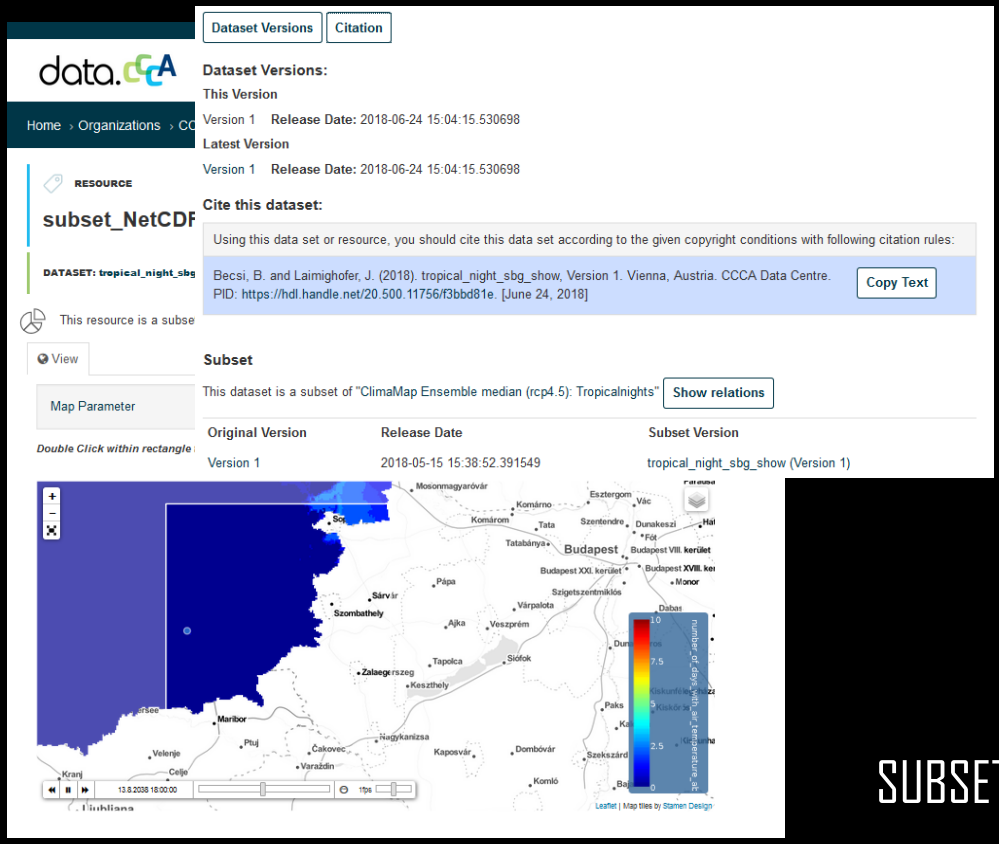
identify precisely the subset of (dynamic) in a process

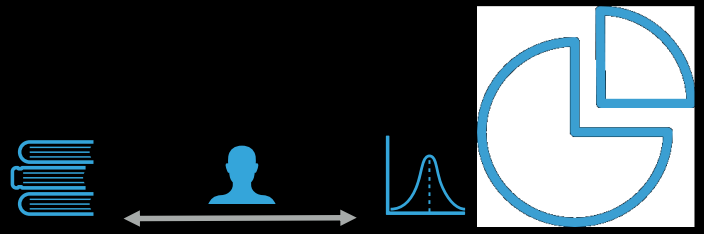
Choose a:

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

SUBSETTING + dynamic data citation

Cite your Data





(research) data is dynamic

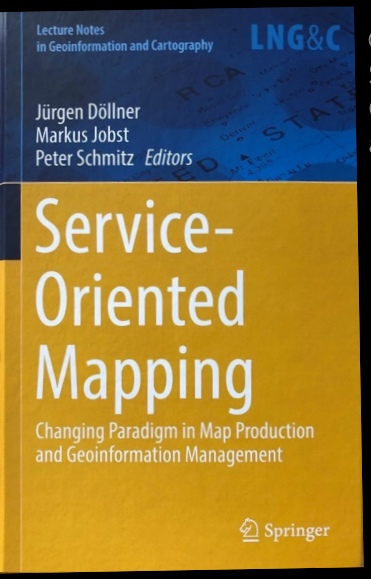
Re-published

avoid redundant storage consumption

keep all relations between updates, original sources & subsets

SUBSETTING + dynamic data citation

Cite your Data



Service Oriented Mapping
Changing Paradigm in Map Production
and Geoinformation Management

Handling Continuous Streams for Meteorological Mapping

Chris Schubert[1], Harald Bamberger[2]
[1] CCCA Data Centre, Vienna, Austria, hosted by ZAMG,
[2] ZAMG, Dep. Software Application development and Data Management

Special Issue "Earth Observation Data

Open Access Article

Dynamic Data Citation Service—Subset Tool for Operational Data Management

by Chris Schubert [1,*], Georg Seyerl [1] and Katharina Sack [2]
[1] Data Centre—Climate Change Centre Austria, 1190 Vienna, Austria
[2] Institute for Economic Policy and Industrial Economics, WU—Vienna University of Economics and Business, 1020 Vienna, Austria
* Author to whom correspondence should be addressed.

Data **2019**, *4*(3), 115; https://doi.org/10.3390/data4030115
Received: 31 May 2019 / Revised: 29 July 2019 / Accepted: 30 July 2019 / Published: 1 August 2019
(This article belongs to the Special Issue Earth Observation Data Cubes)

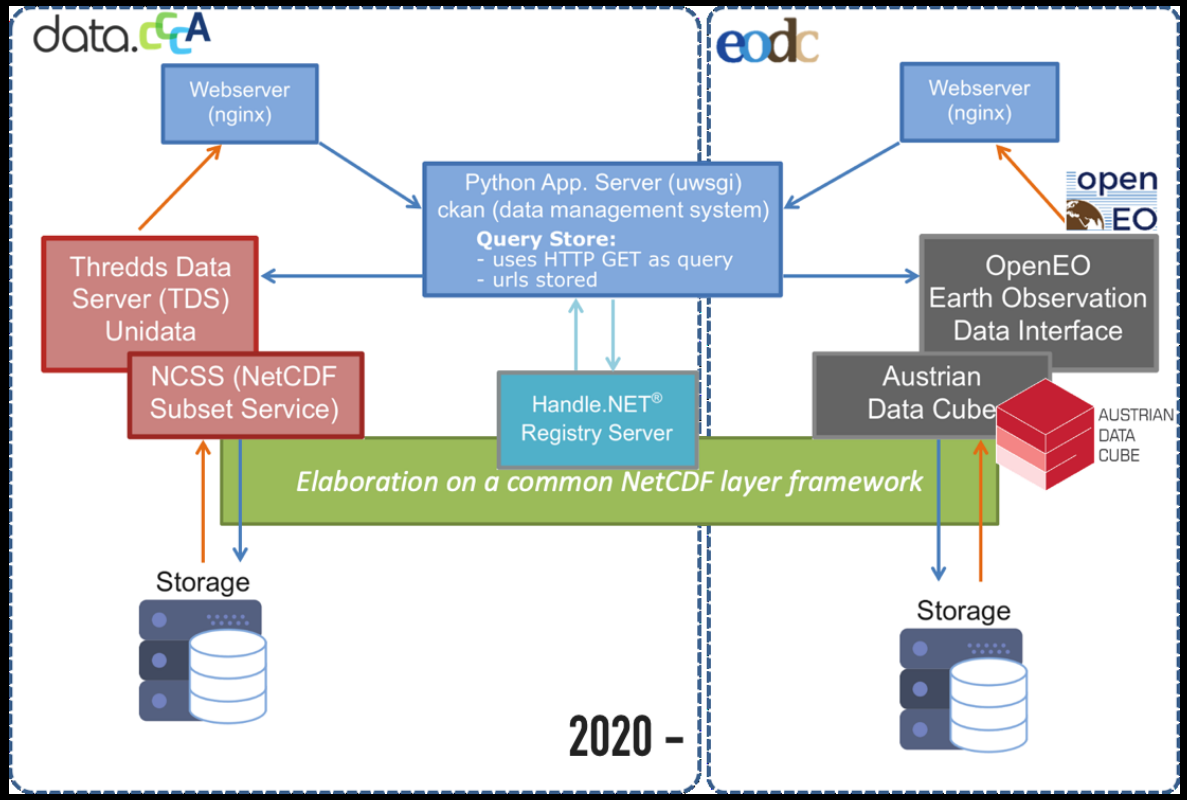
View Full-Text | Download PDF | Browse Figures | Review Reports

https://doi.org/10.3390/data4030115

SUBSETTING + dynamic data citation

PUBLICATION

- CCCA Data Center wants to extend services to the Earth Observation domain w/o collecting redundant data
- Extent the user community
- Improvement of software architecture
- Find best synergies to OpenEO, OGC Standards
- Find best synergies for the OpenDataCube initiatives
- Developed software modules are still Open Source and free available

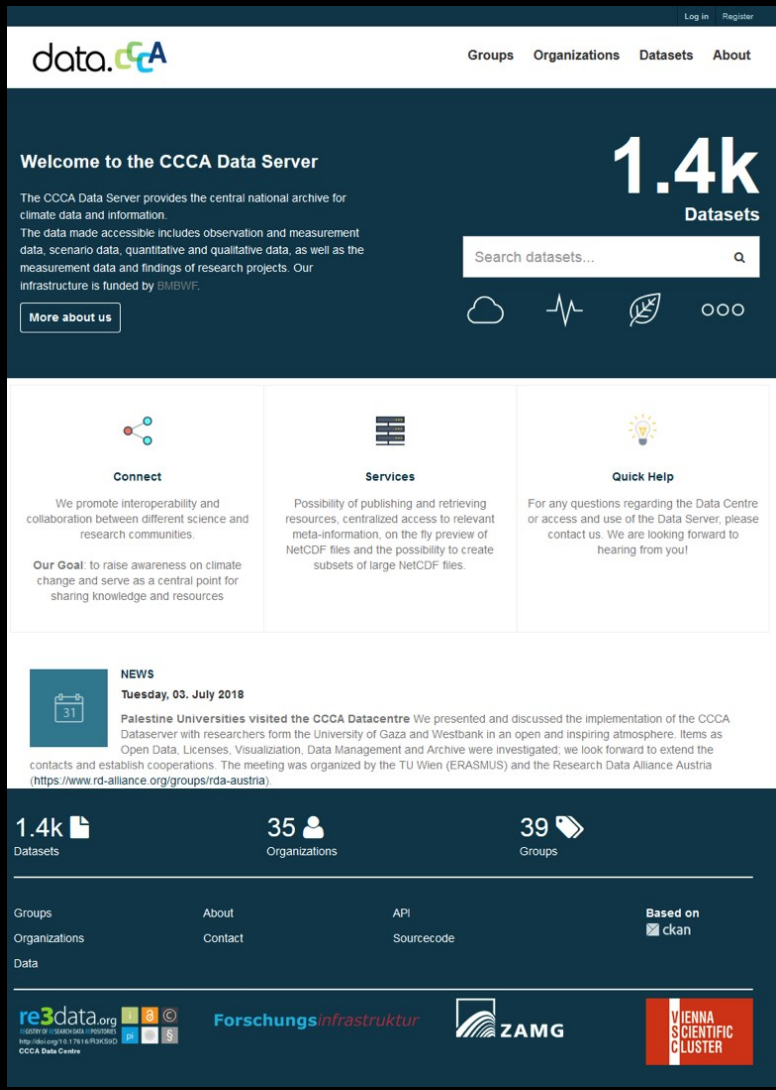


SUBSETTING + dynamic data citation

RDA ADOPTION

Thank you for attention !

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



RESEARCH DATA ALLIANCE

Others?
Plans, On-going, Feedback

Anybody

research data sharing without barriers
rd-alliance.org

Adoption Stories

- Let us know if you are (planning to) implement (part of) the recommendations

- Submit your adoption story to the RDA Webpage:

https://www.rd-alliance.org/recommendations-outputs/adoption-stories

Agenda

- 16:30 Introduction, Welcome
- 16:40 Short description of the WG recommendations
- 17:00 Reports by adopters / pilots
- 17:50 Paper on adoption stories
- 17:55 Other issues, next steps

WGDC Paper

- Paper summarizing adoptions & lessons learned
- 1 Section per adoption with description of
  - data center, data & data dynamics
  - solution architecture
  - versioning / timestamping approach
  - query store set-up
  - lessons learned, issues identified
- Finalizing paper
- Other forms of summary?

Precisely and Persistently Identifying Arbitrary subsets of Dynamic Data:
A Review of Operational Deployments and Lessons Learned from Implementing the Recommendations of the RDA Working Group on Data Citation

Name1 Surname, Name2 Surname, Name3 Surname, Name4 Surname, Name5 Surname, Name6 Surname, Name7 Surname,

Abstract

XXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXX XXXXXXXXXXXXXXX

1 Introduction

As the importance of data in research, industry and business settings increases, new requirements towards proper research data management (RDM) are arising. Accountability and Transparency in automated decision making [?] have important implications on the way we perform studies, analyze data, and prepare the basis for data-driven decision making. In this context, reproducibility in various forms, i.e. the ability to re-compute analyses, arriving at the same conclusions / insights is gaining importance. This has impact on the way analyses are being performed, requiring processes to be documented and code to be deposited. Additionally, data – being the basis of such analyses and thus likely the most relevant ingredient in any data-driven decision making process – needs to be findable and accessible if any result is to be verified.

However, identifying precisely which data was used in a specific analysis is a non-trivial challenge in most settings: Rather than relying on static, acrhived data collected and frozen for analysis, today's decision making processes rely increasingly on continuous data streams tha should be available and useable for decision making on a continuous basis. working on last year's (or last week's) data is not an acceptable alternative in many settings. Additionally, data is undergoing complex pre-processing routines, being re-calibrated, data quality is being improved by correcting error, keeping data in a constant flux.

Additionally, data is getting "big": enormous amounts of data are being collected, of which specific subsets are selected for analysis, from a small number of individual values to massive subsets of even bigger data sets. Describing which subset was actually being used – and trying to re-create the exactly same subset at a later point in time based on descriptions provided in the methods sections of papers and reports – may constitute a daunting challenge due to the complexity

Next steps

- Finalizing paper
- Webinar
  - **Implementation of the RDA Data Citation Recommendations by the Earth Observation Data Center (EODC) for the openEO platform**
    Wed, Nov 20 2019, 17:00 CET
  - https://www.rd-alliance.org/group/data-citation-wg/webconference/webconference-data-citation-wg.html
- Which other forms of experience sharing would be helpful?

Thanks

Thanks!

And hope to see you at the next meeting

of the

WGDC