RESEARCH DATA ALLIANCE

Data Citation
Working Group Mtg @ P10
Sep. 20 2017, Montreal

research data sharing without barriers

rd-alliance.org

Agenda

- 15:30 Introduction, Welcome
- 15:40 Short description of the WG recommendations
- 16:00 Report on new issues discussed / lessons learned
- 16:30 Brief reports on use cases
- 17:20 Other issues, next steps

Welcome and Intro

Welcome!

to the maintenance meeting

of the

WGDC

Agenda

- 15:30 Introduction, Welcome
- 15:40 Short description of the WG recommendations
  - Goals / challenges
  - Recommendations
  - Benefits
- 16:00 Report on new issues discussed / lessons learned
- 16:30 Brief reports on use cases
- 17:20 Other issues, next steps

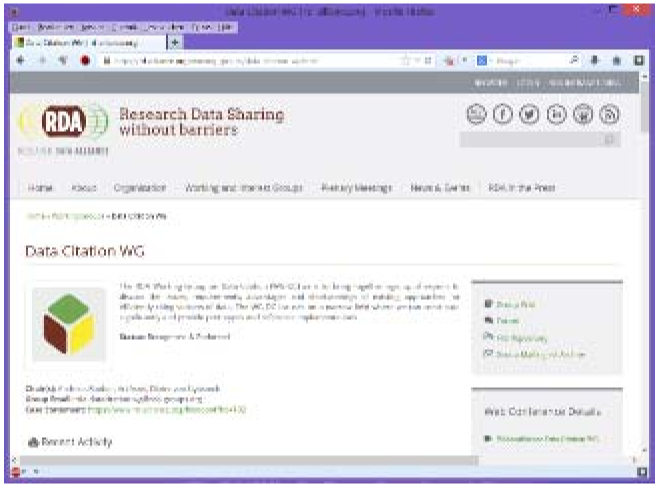
Identification of Dynamic Data

- 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

- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citation text (e.g. BibTeX)

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!!!

- 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

- ubset of data
- er gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citation text (e.g. BibTeX)
- PID resolves t
  - Provides det
  - Option to ret
- 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

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!!!

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

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 of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use

Data Citation – Recommendations

Preparing Data & Query Store

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

When Resolving a PID

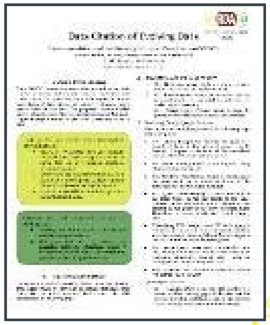
- R11 – Landing Page
- R12 – Machine Actionability

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 of Evolving Data

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

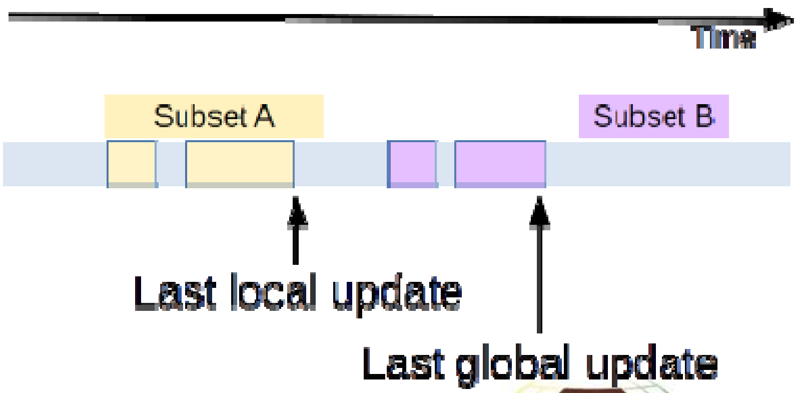
- 15:30 Introduction, Welcome
- 15:40 Short description of the WG recommendations
- 16:00 Report on new issues discussed / lessons learned
    - Update on ICT standardization process
    - R7: timestamping, distributed settings
    - R10: dual PIDs: dynamic dataset + subset
    - R10: automatic generation of citation text
- 16:30 Brief reports on use cases
- 17:20 Other issues, next steps

Standardization

- RDA applied for WGDC recommendations to become **ICT Technical Specification**
- Application to be considered by MSP (European Multi Stakeholder Platform) on 7 March
- Hillary Hanahoe presented to MSP on 23 March
- Evaluation group investigated market acceptance (basically the adoption & implementation) in April
- Status on Sep 17
  - Preliminary evaluation was very positive
  - Final evaluation will be discussed by MSP on Sep 28

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)



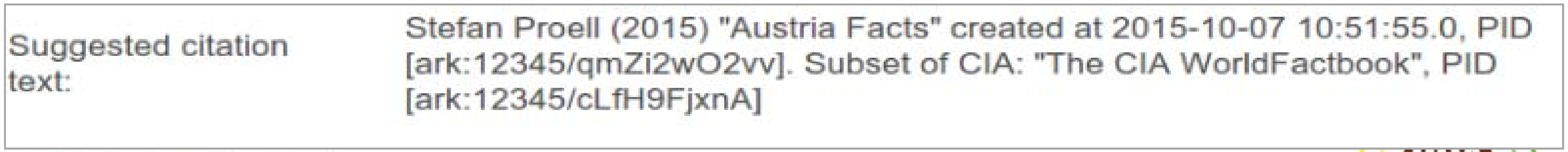
Q&A: R7: Query Timestamping – Distributed Settings

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

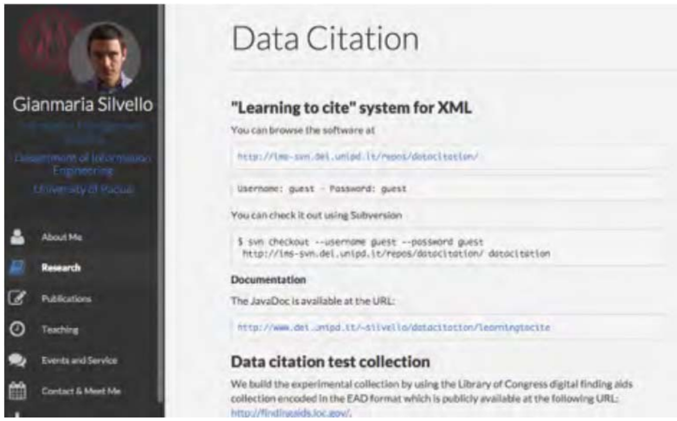
- **Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing and sharing the data.**
  Include the PID in the citation text snippet.
- **2 PIDs!**
  - **Superset**: the "database" and it's holder (repository, data center)
    - Changing / evolving
  - **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



Data Citation

"Learning to cite" system for XML

You can browse the software at

```
http://ims-svn.dei.unipd.it/repos/datacitation/
```

```
Username: guest - Password: guest
```

You can check it out using Subversion

```
$ svn checkout --username guest --password guest
 http://ims-svn.dei.unipd.it/repos/datacitation/ datacitation
```

Documentation

The JavaDoc is available at the URL:

```
http://www.dei.unipd.it/~silvello/datacitation/learningtocite
```

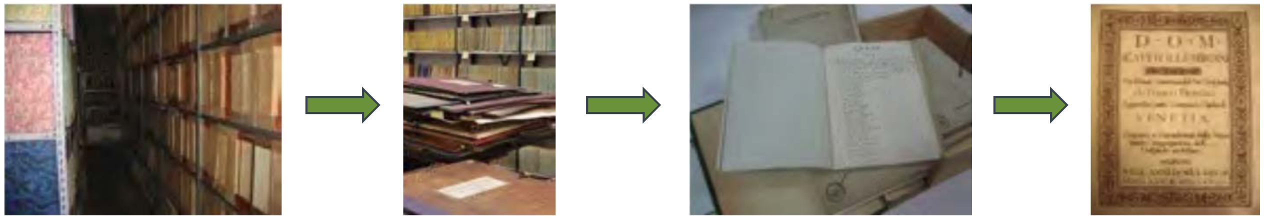
Data citation test collection

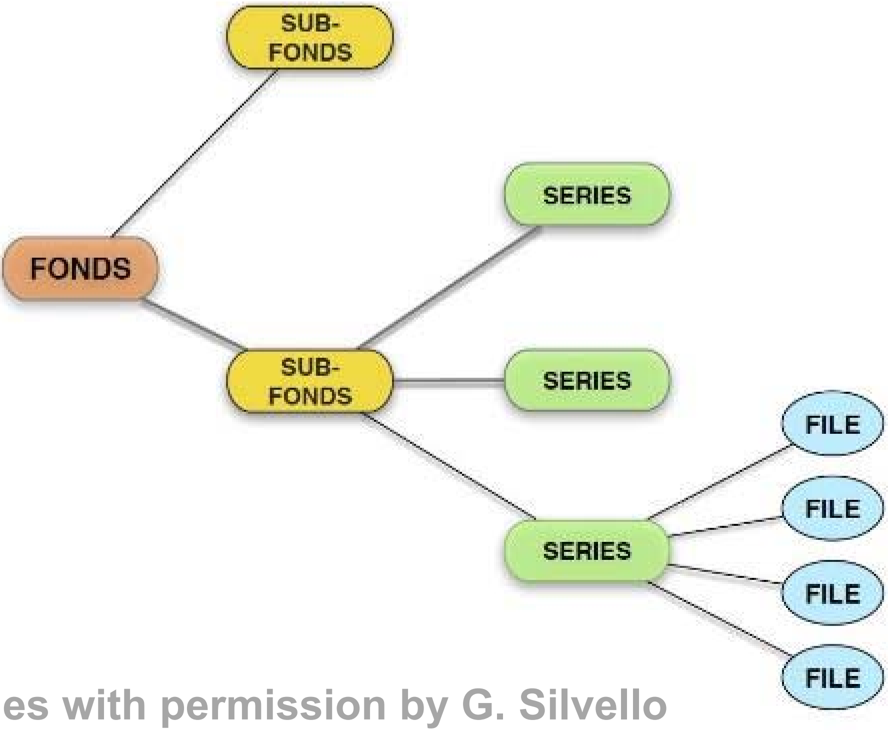
We build the experimental collection by using the Library of Congress digital finding aids collection encoded in the EAD format which is publicly available at the following URL: http://findingaids.loc.gov/.

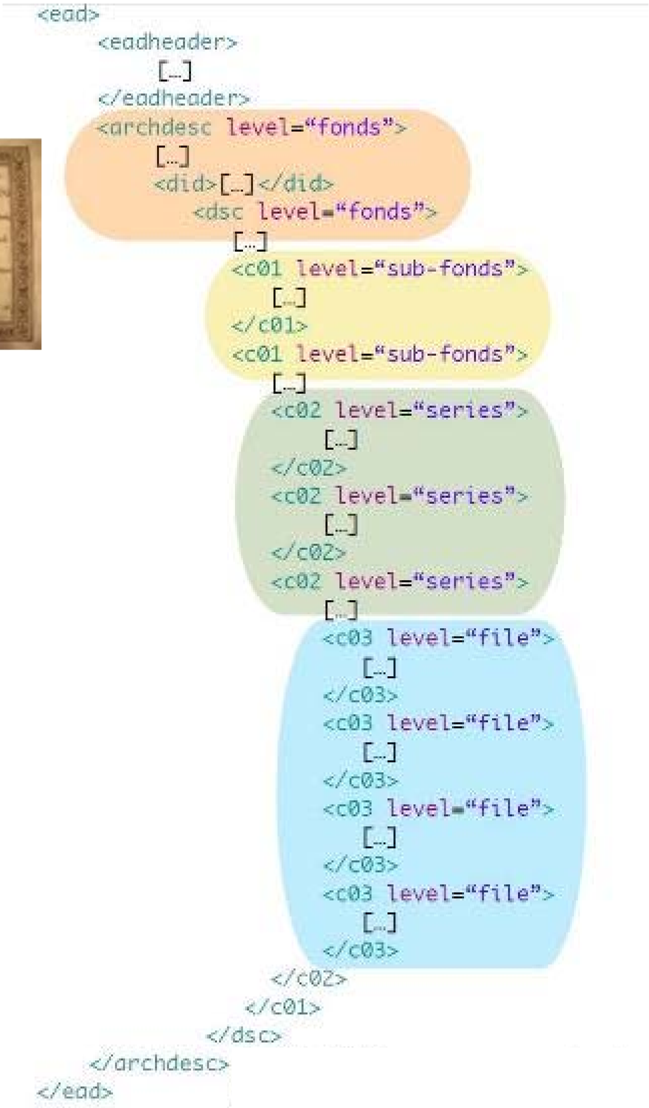
- Learning to Cite:
  - Gianmaria Silvello. Learning to Cite Framework: How to Automatically Construct Citations for Hierarchical Data. Journal of the Association for Information Science and Technology (JASIST), Volume 68 issue 6, pp. 1505-1524, June 2017.
  - http://www.dei.unipd.it/~silvello/datacitation

R10: Automated Citation Texts

- EAD: Encoded Archival Description





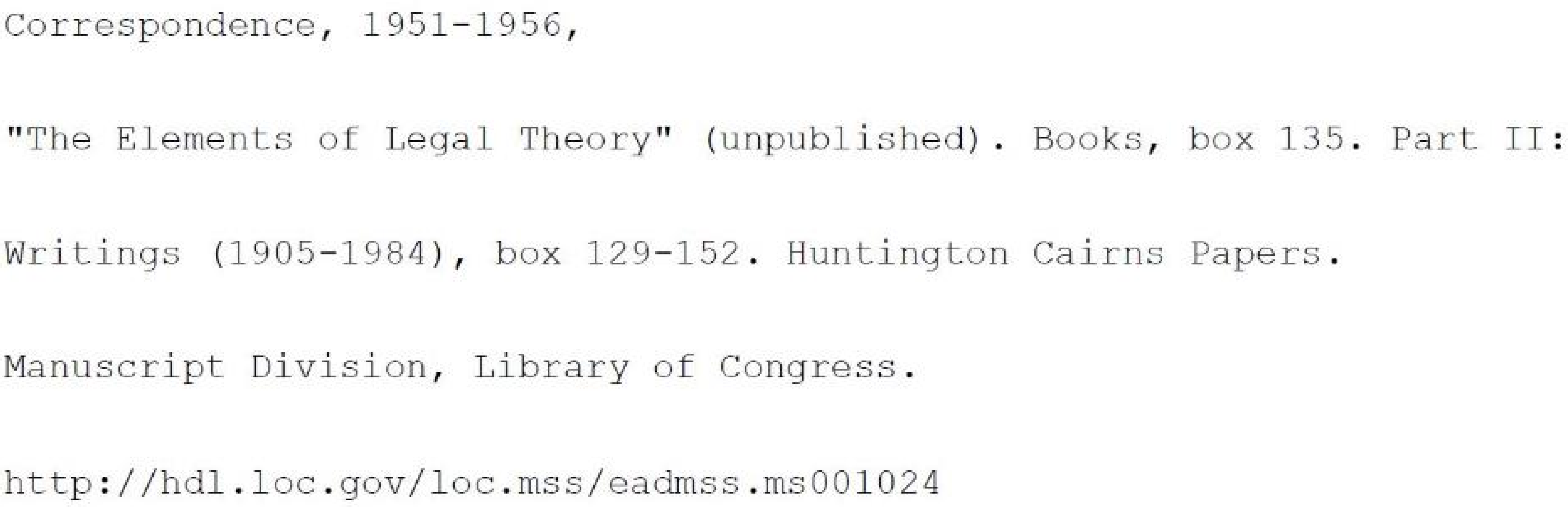


```
<ead>
    <eadheader>
        […]
    </eadheader>
    <archdesc level="fonds">
        […]
        <did>[…]</did>
            <dsc level="fonds">
                […]
                <c01 level="sub-fonds">
                    […]
                </c01>
                <c01 level="sub-fonds">
                    […]
                    <c02 level="series">
                        […]
                    </c02>
                    <c02 level="series">
                        […]
                    </c02>
                    <c02 level="series">
                        […]
                        <c03 level="file">
                            […]
                        </c03>
                        <c03 level="file">
                            […]
                        </c03>
                        <c03 level="file">
                            […]
                        </c03>
                        <c03 level="file">
                            […]
                        </c03>
                    </c02>
                </c01>
            </dsc>
    </archdesc>
</ead>
```

Slides with permission by G. Silvello

R10: Automated Citation Texts

- A human-readable citation:

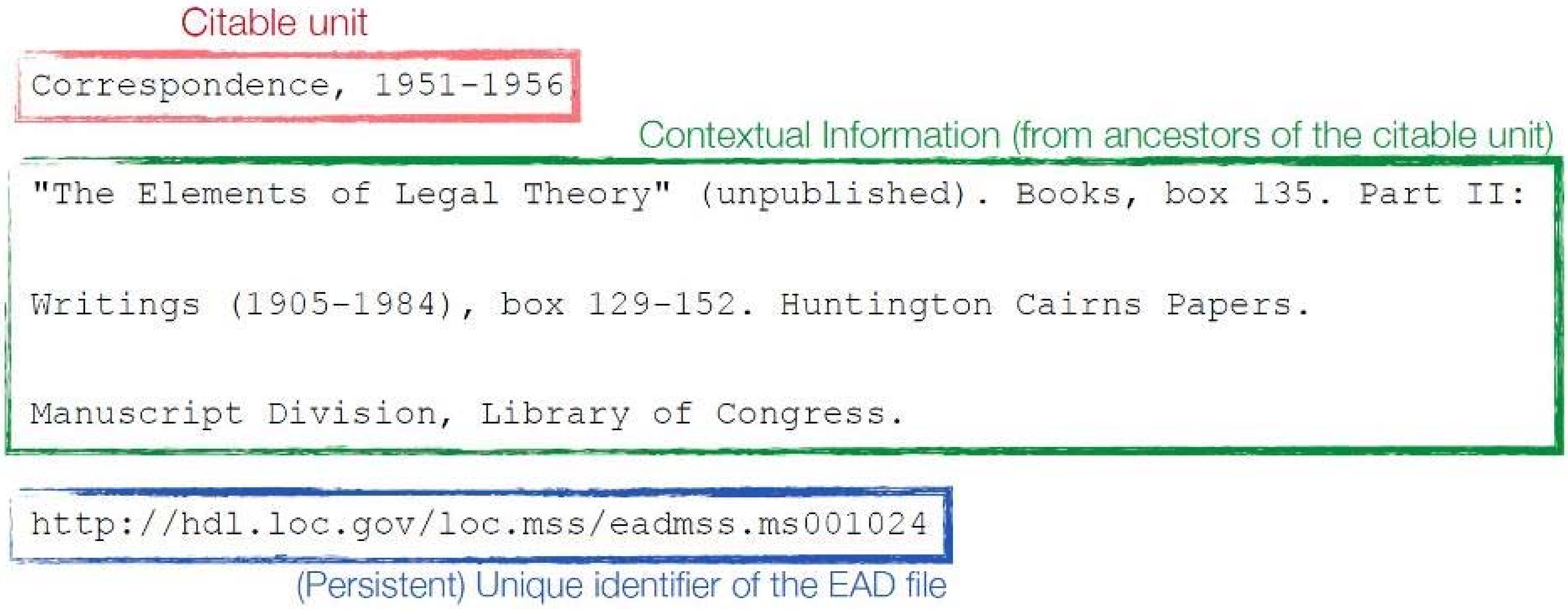


Correspondence, 1951-1956, "The Elements of Legal Theory" (unpublished). Books, box 135. Part II: Writings (1905-1984), box 129-152. Huntington Cairns Papers. Manuscript Division, Library of Congress. http://hdl.loc.gov/loc.mss/eadmss.ms001024

Slides with permission by G. Silvello

R10: Automated Citation Texts

- A human-readable citation:



Citable unit
Contextual Information (from ancestors of the citable unit)
(Persistent) Unique identifier of the EAD file

Correspondence, 1951-1956

"The Elements of Legal Theory" (unpublished). Books, box 135. Part II:

Writings (1905-1984), box 129-152. Huntington Cairns Papers.

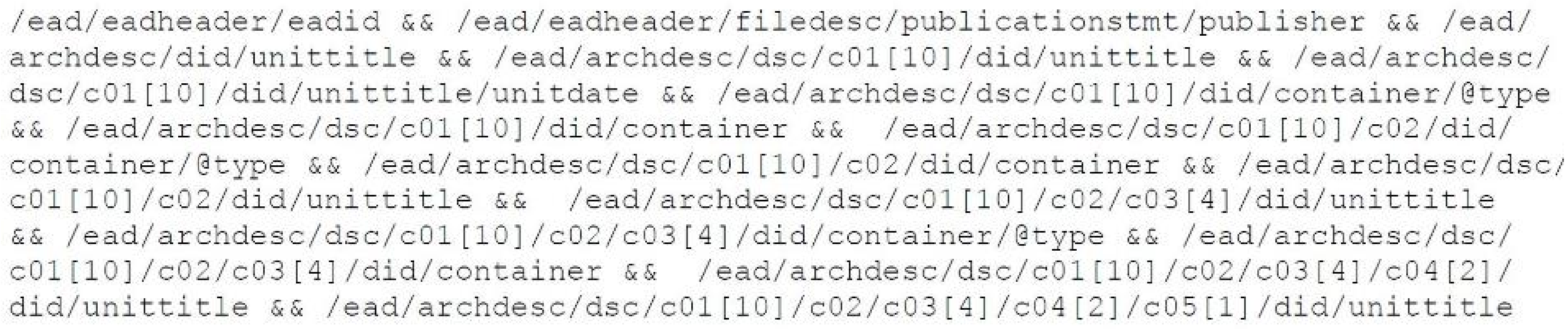
Manuscript Division, Library of Congress.

http://hdl.loc.gov/loc.mss/eadmss.ms001024

Slides with permission by G. Silvello

R10: Automated Citation Texts

- A machine-readable citation:
  - Conjunction of XML paths

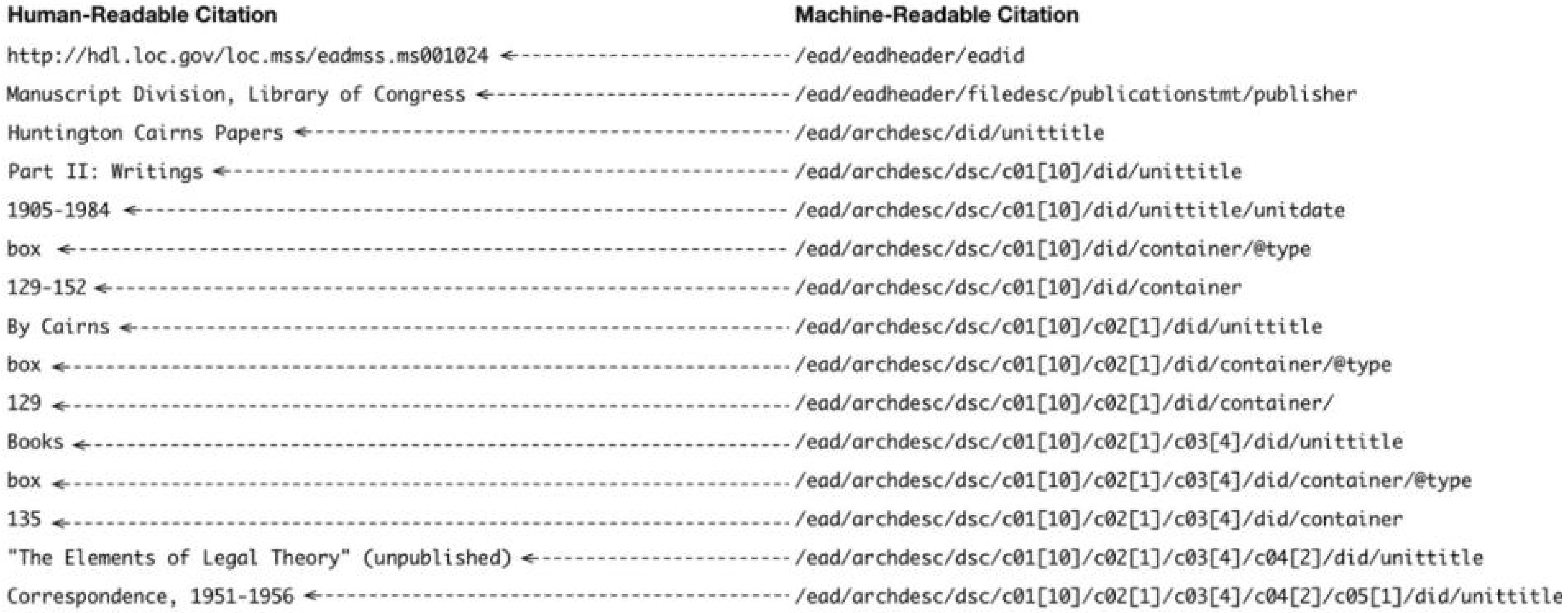


```
/ead/eadheader/eadid && /ead/eadheader/filedesc/publicationstmt/publisher && /ead/
archdesc/did/unittitle && /ead/archdesc/dsc/c01[10]/did/unittitle && /ead/archdesc/
dsc/c01[10]/did/unittitle/unitdate && /ead/archdesc/dsc/c01[10]/did/container/@type
&& /ead/archdesc/dsc/c01[10]/did/container &&  /ead/archdesc/dsc/c01[10]/c02/did/
container/@type && /ead/archdesc/dsc/c01[10]/c02/did/container && /ead/archdesc/dsc/
c01[10]/c02/did/unittitle &&  /ead/archdesc/dsc/c01[10]/c02/c03[4]/did/unittitle
&& /ead/archdesc/dsc/c01[10]/c02/c03[4]/did/container/@type && /ead/archdesc/dsc/
c01[10]/c02/c03[4]/did/container &&  /ead/archdesc/dsc/c01[10]/c02/c03[4]/c04[2]/
did/unittitle && /ead/archdesc/dsc/c01[10]/c02/c03[4]/c04[2]/c05[1]/did/unittitle
```

Slides with permission by G. Silvello

R10: Automated Citation Texts

- Mapping machine-readable to human-readable:

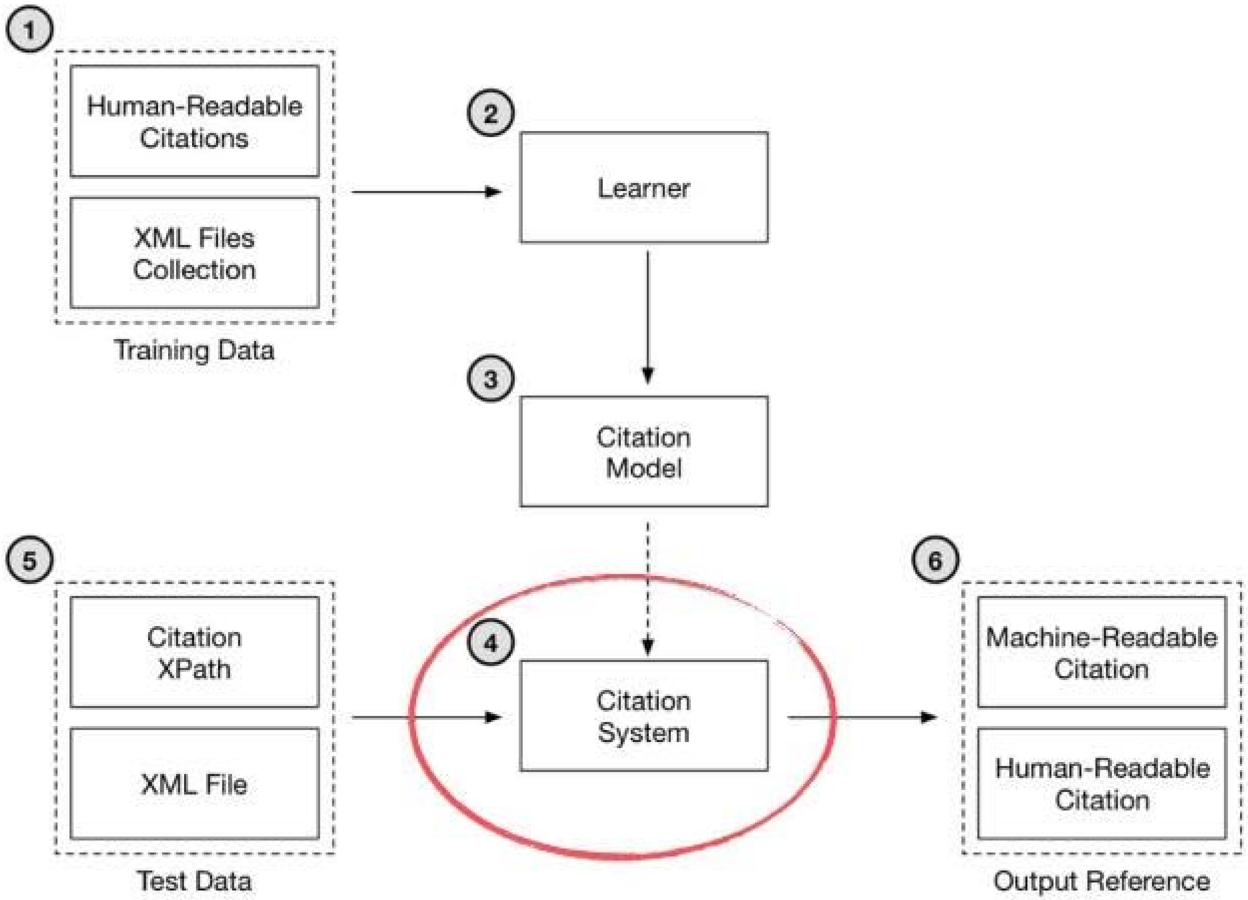


Human-Readable Citation	Machine-Readable Citation
http://hdl.loc.gov/loc.mss/eadmss.ms001024	/ead/eadheader/eadid
Manuscript Division, Library of Congress	/ead/eadheader/filedesc/publicationstmt/publisher
Huntington Cairns Papers	/ead/archdesc/did/unittitle
Part II: Writings	/ead/archdesc/dsc/c01[10]/did/unittitle
1905-1984	/ead/archdesc/dsc/c01[10]/did/unittitle/unitdate
box	/ead/archdesc/dsc/c01[10]/did/container/@type
129-152	/ead/archdesc/dsc/c01[10]/did/container
By Cairns	/ead/archdesc/dsc/c01[10]/c02[1]/did/unittitle
box	/ead/archdesc/dsc/c01[10]/c02[1]/did/container/@type
129	/ead/archdesc/dsc/c01[10]/c02[1]/did/container/
Books	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/unittitle
box	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container/@type
135	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/did/container
"The Elements of Legal Theory" (unpublished)	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/did/unittitle
Correspondence, 1951-1956	/ead/archdesc/dsc/c01[10]/c02[1]/c03[4]/c04[2]/c05[1]/did/unittitle

Slides with permission by G. Silvello

R10: Automated Citation Texts

- Learning citation models



Slides with permission by G. Silvello

Agenda

- 15:30 Introduction, Welcome
- 15:40 Short description of the WG recommendations
- 16:00 Report on new issues discussed / lessons learned
- 16:30 Brief reports on use cases
  - Climate Change Centre Austria (CCCA)
  - Virtual and Atomic Molecular Data Center (C.M. Zwölf)
  - Pilot project in Japan (Yasuhiro Muyarama)
  - River Flow Archive (Matt Fry)
  - Others
- 17:20 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
  - 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
  - Gianmaria Silvello: Automatic generation of citation texts (Nov. 8 2017?)
  - Yasuhiro Muyarama: Citing dynamic datasets at NICT?

RDA

RESEARCH DATA ALLIANCE

Climate Change Centre Austria (CCCA)

Chris Schubert

chris.Schubert@ccca.ac.at

research data sharing without barriers

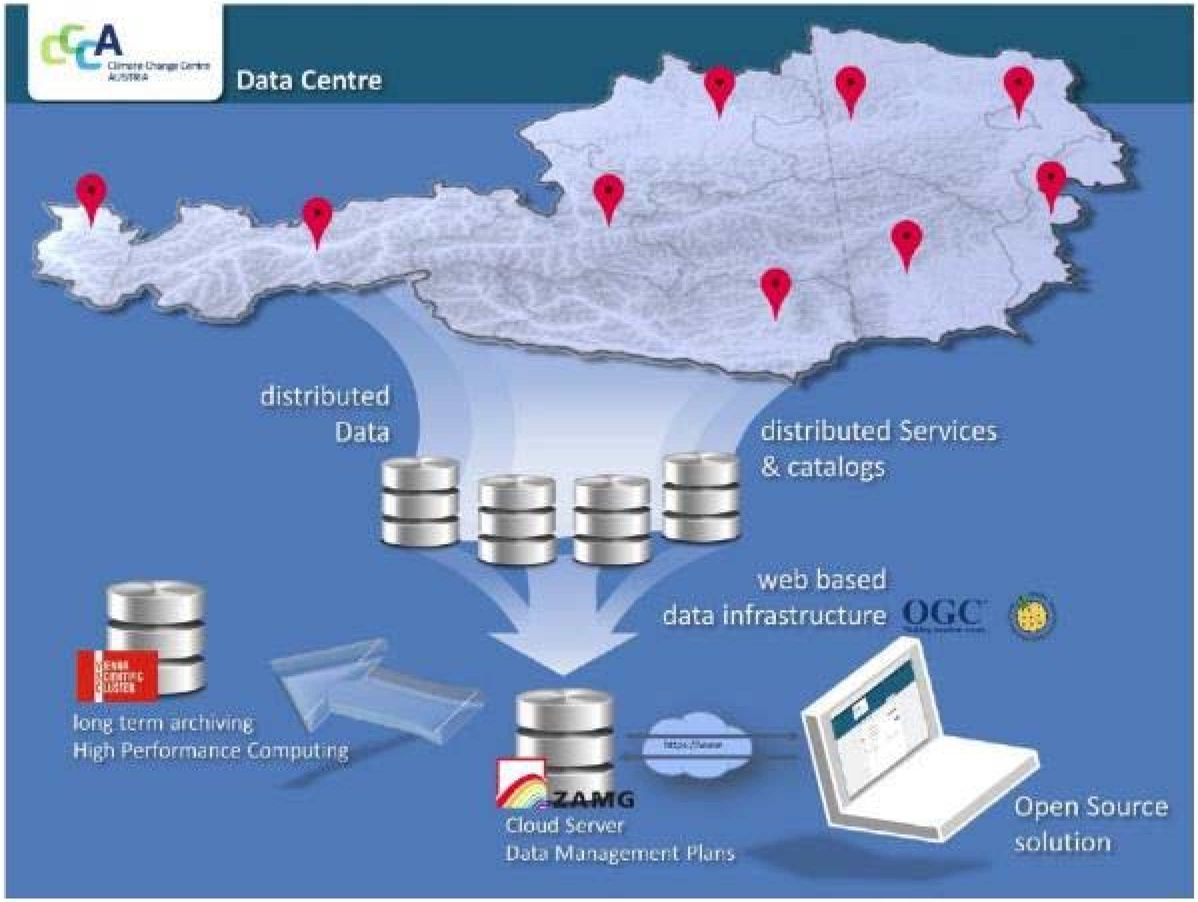
rd-alliance.org

CCCA

- Climate research network for sustained, high-quality Austrian climate research.
- 28 members (11 universities, 13 non-univ. institutions, 4 supporting members)
- Structure: Coordination Office (Vienna, BOKU), Service Centre (Univ. Graz), Data Centre (ZAMG, Vienna)
- Service available http://data.ccca.ac.at

WIFO
AGES
umweltbundesamt
AIT AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY
alp-S
ESSL
UNI GRAZ
BFW.
BOKU
IIASA
UNIVERSITÄT SALZBURG
Z_GIS
universität wien
INSTITUT FÜR INDUSTRIELLE ÖKOLOGIE
universität innsbruck
ÖGM
JOANNEUM RESEARCH
ZSI
TU Graz
ZAMG
ÖAW
vetmeduni vienna
WU
TU WIEN TECHNISCHE UNIVERSITÄT WIEN
UMA
ALPEN-ADRIA UNIVERSITÄT
Weather park
Geologische Bundesanstalt

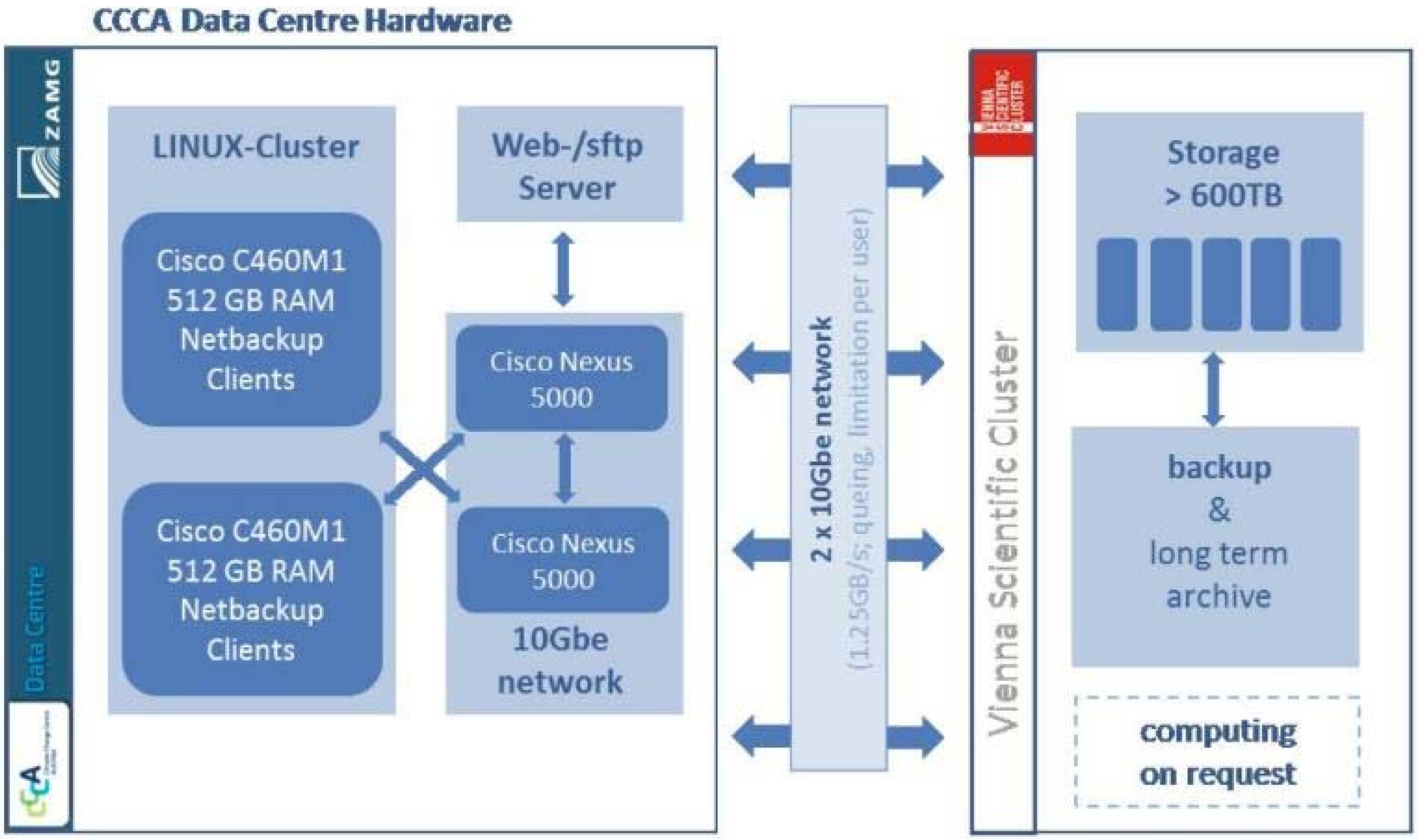
CCCA


Climate Change Centre AUSTRIA
Data Centre
distributed Data
distributed Services & catalogs
web based data infrastructure
OGC
long term archiving High Performance Computing
ZAMG
Cloud Server Data Management Plans
Open Source solution

CCCA Data Centre

- **provision** of climate-relevant information, data, algorithms, reports
- **interoperable interfaces** to international portals, standards, legislation (e.g. INSPIRE)
- conception for **long term archiving** of research data & repositories
- capacity building, consultancy and **support** for **data sharing**

CCCA



CCCA Data Centre Hardware

ZAMG

Data Centre

LINUX-Cluster

Cisco C460M1
512 GB RAM
Netbackup
Clients

Cisco C460M1
512 GB RAM
Netbackup
Clients

Web-/sftp
Server

Cisco Nexus
5000

Cisco Nexus
5000

10Gbe
network

2 x 10Gbe network
(1.25GB/s; queing, limitation per user)

Vienna Scientific Cluster

Storage
> 600TB

backup
&
long term
archive

computing
on request

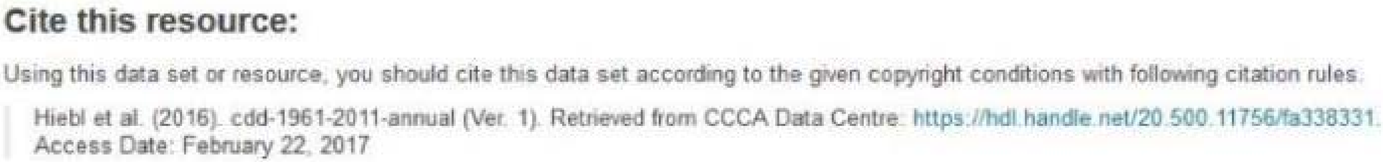
CCCA

… a data portal among many others?

FEATURE No. 4 & 5

- handle® Service implemented to serve persistent identifier (PID) -> fundamental for DataCitation

hdl.handle.net/20.500.11756/7b9374de



Cite this resource:

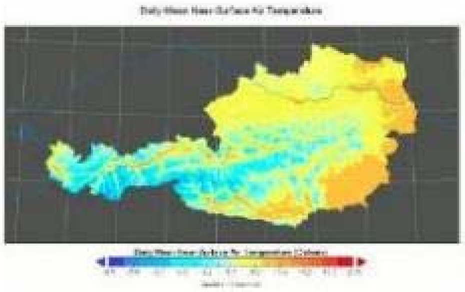
Using this data set or resource, you should cite this data set according to the given copyright conditions with following citation rules.

> Hiebl et al. (2016). cdd-1961-2011-annual (Ver. 1). Retrieved from CCCA Data Centre: https://hdl.handle.net/20.500.11756/fa338331. Access Date: February 22, 2017

Your Publication

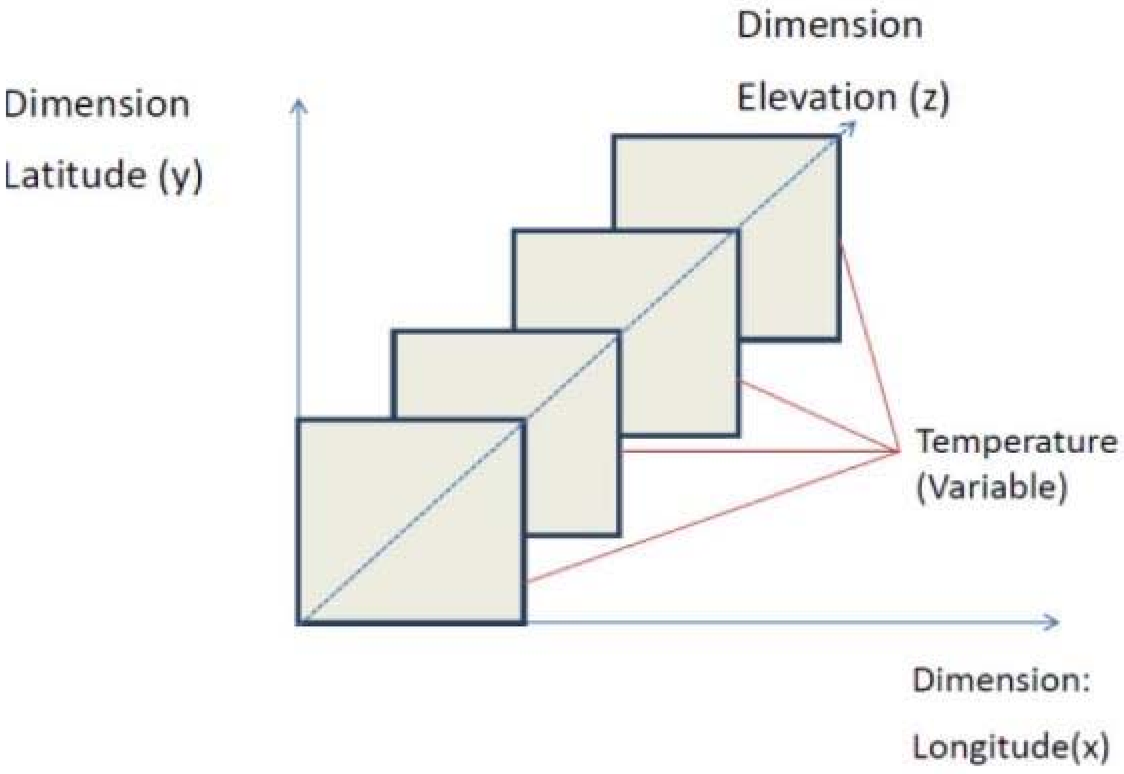
formal
Data Citation

Your Data

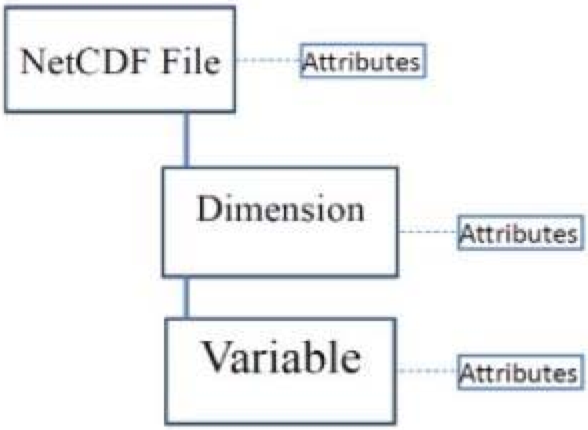


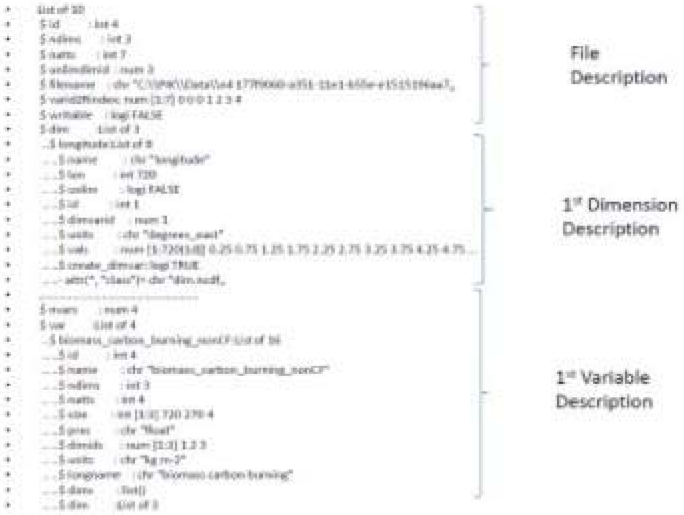
CCCA

- **NetCDF Files:**



modified and based on UCAR Unidata, www.unidata.ucar.edu/





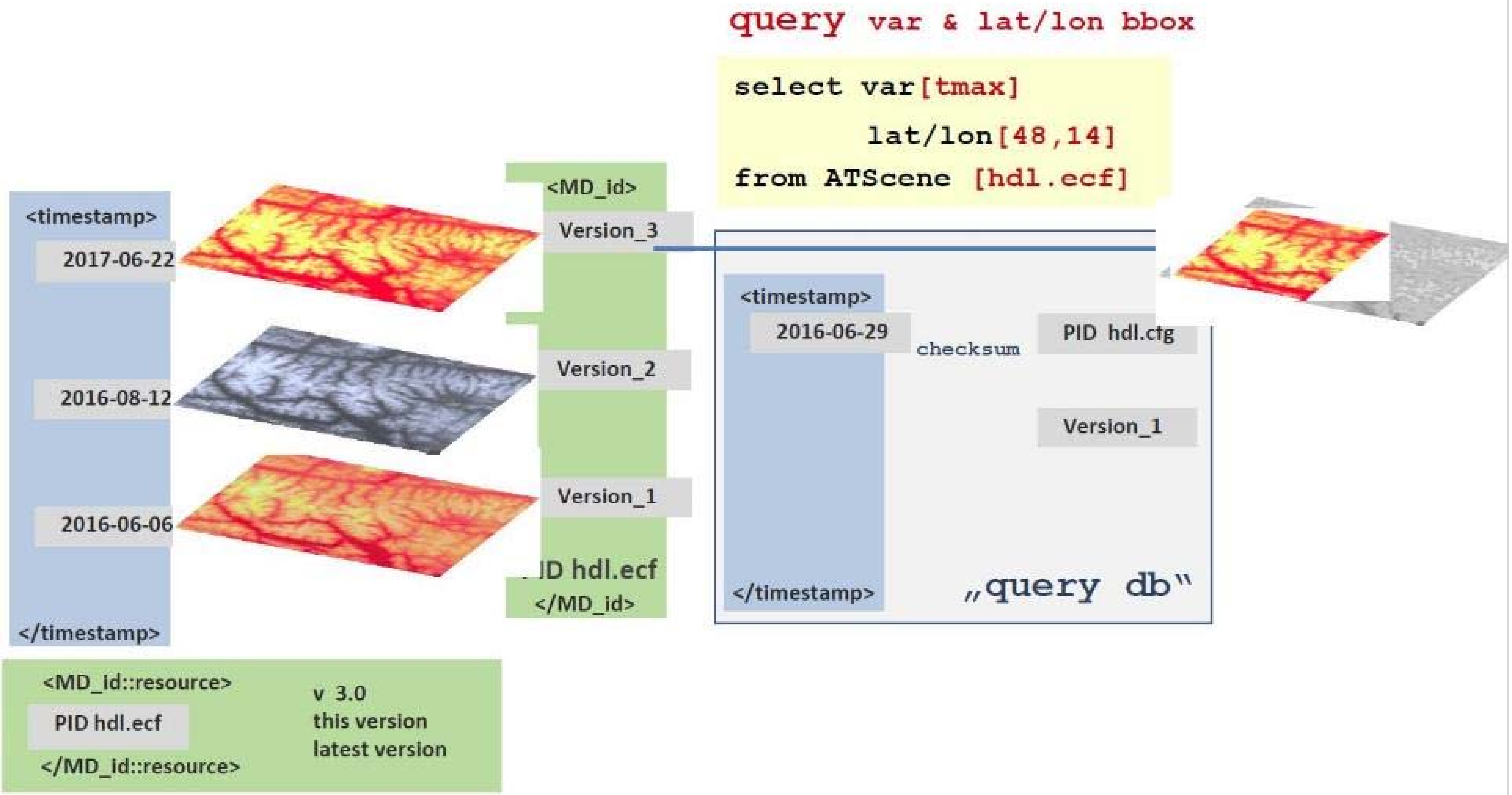
```
List of 10
 $ id          : int 4
 $ ndims       : int 3
 $ natts       : int 7
 $ unlimdimid  : num 3
 $ filename    : chr "C:\\IPW\\Data\\n4 1779000-a351-11e1-b5fe-e151519ae7"|
 $ varid2Rindex: num [1:7] 0 0 0 1 2 3 4
 $ writable    : logi FALSE
 $ dim         :List of 3
 ..$ longitude:List of 8
 .. ..$ name         : chr "longitude"
 .. ..$ len          : int 720
 .. ..$ unlim        : logi FALSE
 .. ..$ id           : int 1
 .. ..$ dimvarid     : num 1
 .. ..$ units        : chr "degrees_east"
 .. ..$ vals         : num [1:720(1d)] 0.25 0.75 1.25 1.75 2.25 2.75 3.25 3.75 4.25 4.75 ...
 .. ..$ create_dimvar: logi TRUE
 .. ..- attr(*, "class")= chr "dim.ncdf"
 ------------------------------------
 $ nvars       : num 4
 $ var         :List of 4
 ..$ biomass_carbon_burning_nonCF:List of 16
 .. ..$ id      : int 4
 .. ..$ name    : chr "biomass_carbon_burning_nonCF"
 .. ..$ ndims   : int 3
 .. ..$ natts   : int 4
 .. ..$ size    : int [1:3] 720 270 4
 .. ..$ prec    : chr "float"
 .. ..$ dimids  : num [1:3] 1 2 3
 .. ..$ units   : chr "kg m-2"
 .. ..$ longname: chr "biomass carbon burning"
 .. ..$ dims    : list()
 .. ..$ dim     :List of 3
```

File Description

1st Dimension Description

1st Variable Description

CCCA

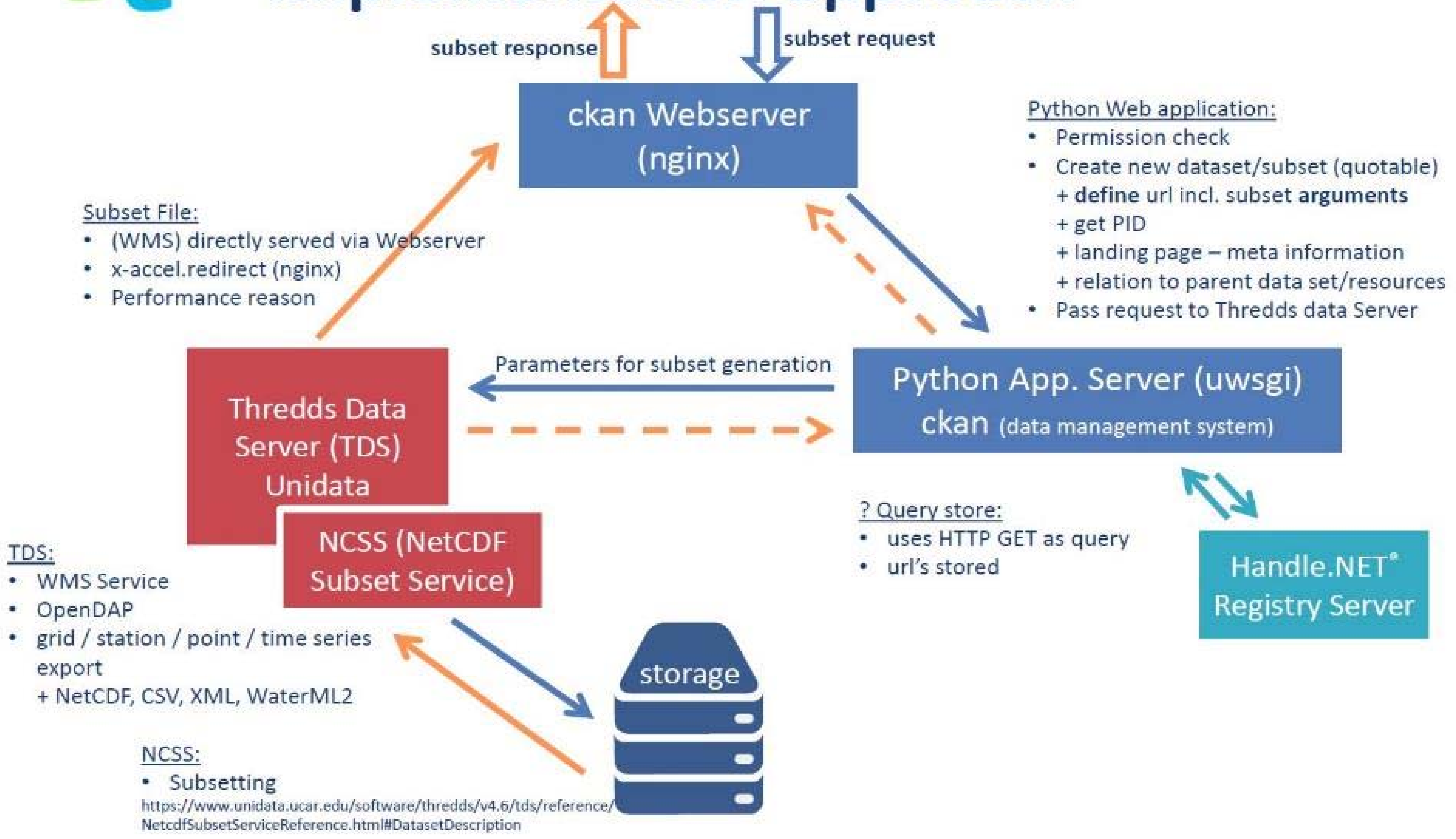


query var & lat/lon bbox

```
select var[tmax]
       lat/lon[48,14]
from ATScene [hdl.ecf]
```

CCCA

Implementation approach



subset response
subset request
ckan Webserver (nginx)
Python App. Server (uwsgi)
ckan (data management system)
Parameters for subset generation
Thredds Data Server (TDS) Unidata
NCSS (NetCDF Subset Service)
storage
Handle.NET® Registry Server

Python Web application:

- Permission check
- Create new dataset/subset (quotable)
  + **define** url incl. subset **arguments**
  + get PID
  + landing page – meta information
  + relation to parent data set/resources
- Pass request to Thredds data Server

Subset File:

- (WMS) directly served via Webserver
- x-accel.redirect (nginx)
- Performance reason

? Query store:

- uses HTTP GET as query
- url's stored

TDS:

- WMS Service
- OpenDAP
- grid / station / point / time series export
  + NetCDF, CSV, XML, WaterML2

NCSS:

- Subsetting

https://www.unidata.ucar.edu/software/thredds/v4.6/tds/reference/NetcdfSubsetServiceReference.html#DatasetDescription

CCCA

NCSS (NetCDF Subset Service)

- for subsetting datasets
- uses **HTTP GET** as query in following scheme:

  http://{host}/{context}/{service}/{**dataset**}[/**dataset.html** | {**?query**}]

Subsetting parameter used:

- **var** - names of our layer
- **north, south, east, west** - for the geographical extend, the bounding box
- **time_start, time_end, time_duration** - for time extend, limited only on 5 years interval
- **accept** - specify the returned format

query store (uwsgi) ckan

All "http get" stored as url in our ckan data store

PID:
hdl.handle.net/20.500.11756/93887ecf

https://data.ccca.ac.at/tds_proxy/ncss/**1dba52b2-4fd0-4fa1-a3ac-cfb0b94a7670**?**north**=47.73168822550699**&west**=9.021605998277664&**accept**=netCDF&**var**=tas**&east**=12.031859904527664**&south**=46.77724203092812

RESEARCH DATA ALLIANCE

VAMDC Query Store Usage feedbacks

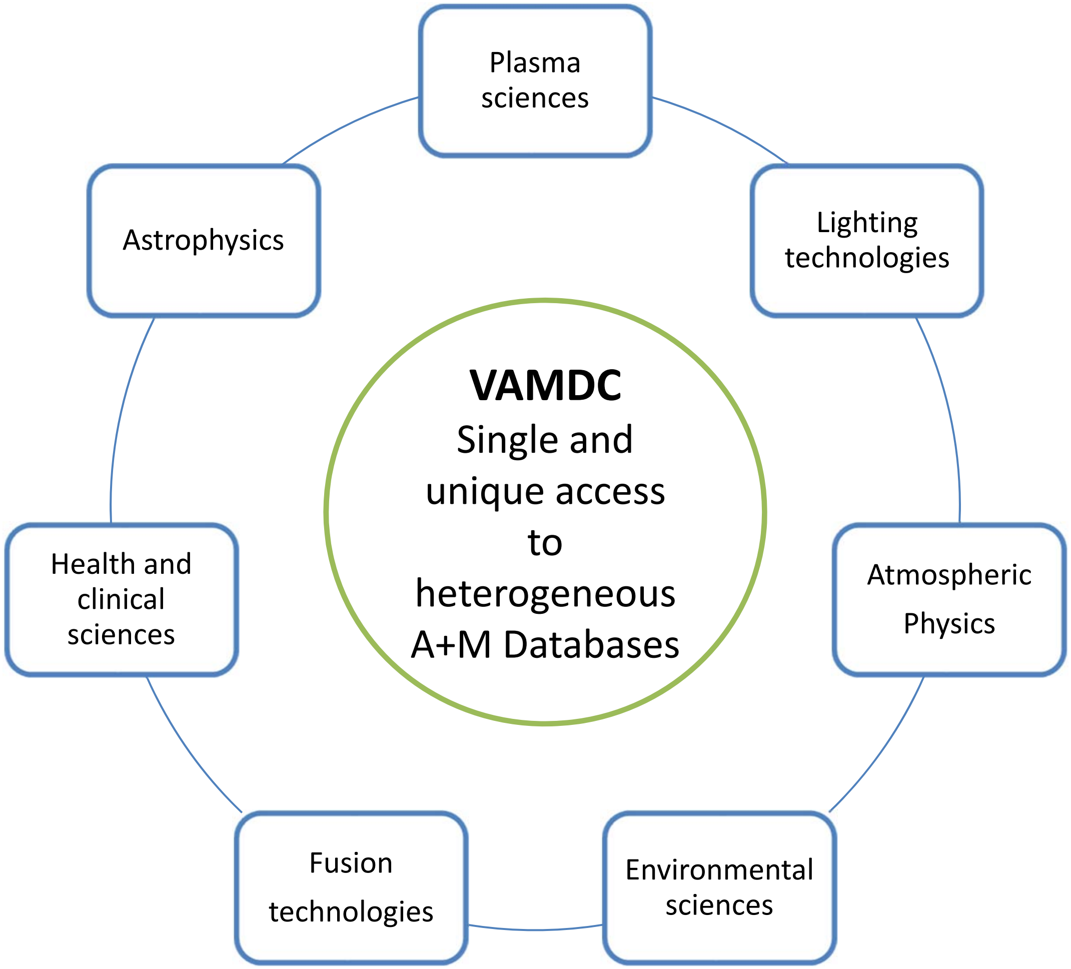
C.M. Zwölf, N. Moreau,
VAMDC Consortium
carlo-maria.zwolf@obspm.fr

research data sharing without barriers
rd-alliance.org

VAMDC Query Store Usage feedbacks

C.M. Zwölf, N. Moreau and VAMDC consortium

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.

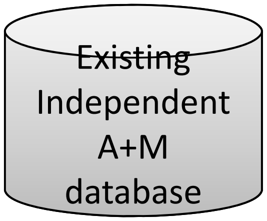
- The consortium is politically organized around a Memorandum of understanding (15 international members have signed the MoU, 1 November 2014)

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

- Provides data producers with a large dissemination platform

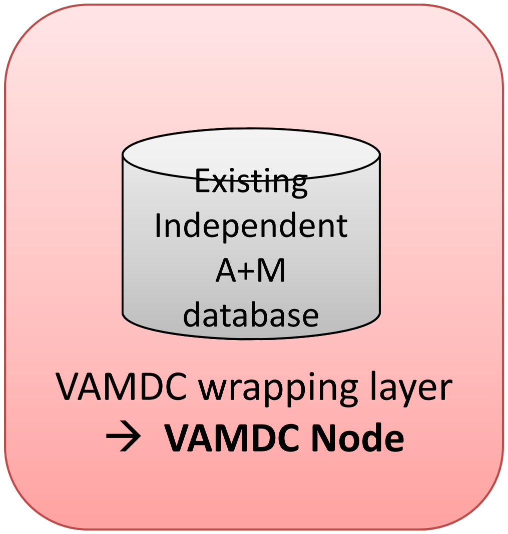
- Removes bottleneck between data-producers and wide body of users

The VAMDC infrastructure technical architecture

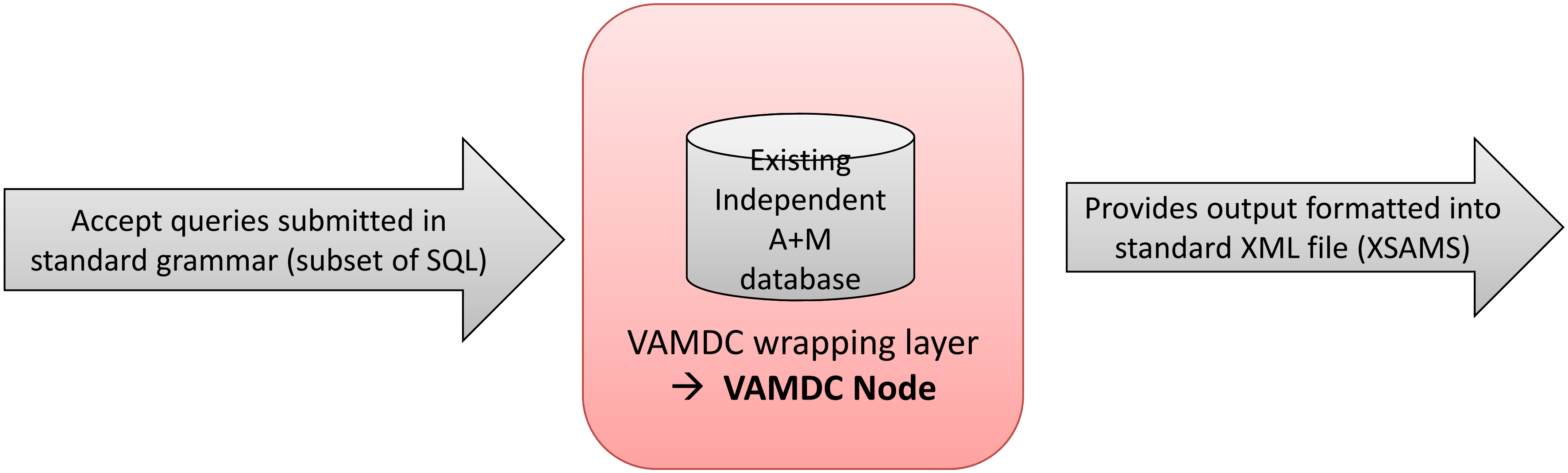


Existing Independent A+M database

The VAMDC infrastructure technical architecture



The VAMDC infrastructure technical architecture



Accept queries submitted in standard grammar (subset of SQL)

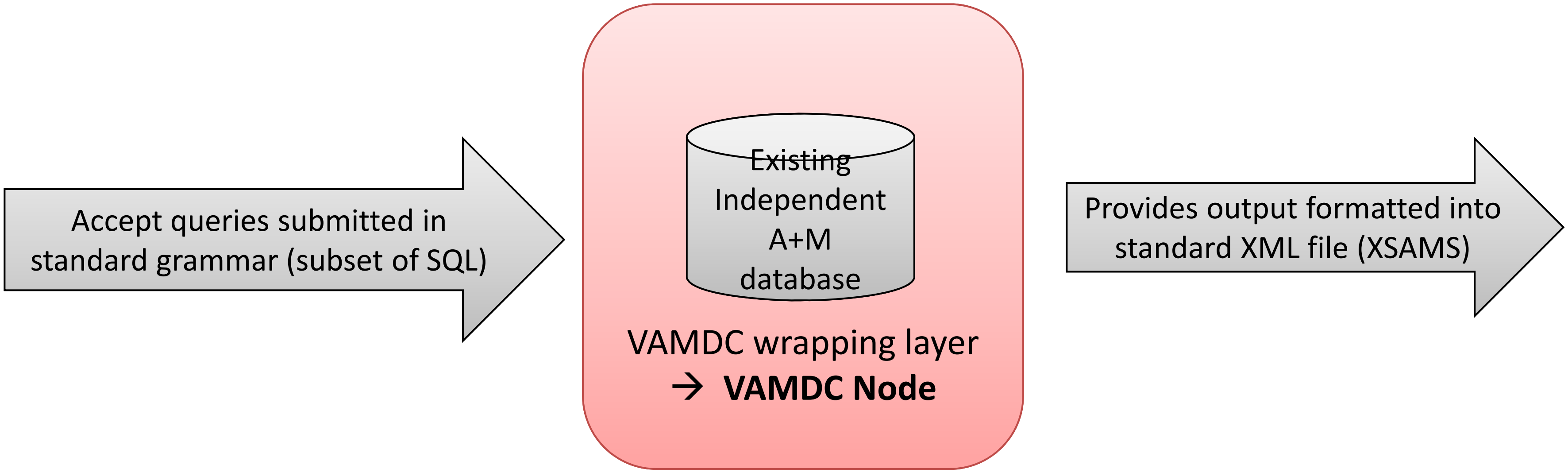
Existing Independent A+M database

VAMDC wrapping layer

→ **VAMDC Node**

Provides output formatted into standard XML file (XSAMS)

The VAMDC infrastructure technical architecture



Accept queries submitted in standard grammar (subset of SQL)

Existing Independent A+M database

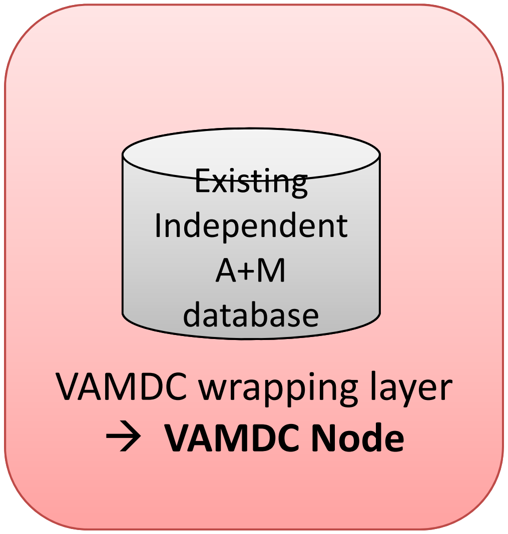
VAMDC wrapping layer

→ **VAMDC Node**

Provides output formatted into standard XML file (XSAMS)

For further details, cf. http://standards.vamdc.eu

The VAMDC infrastructure technical architecture

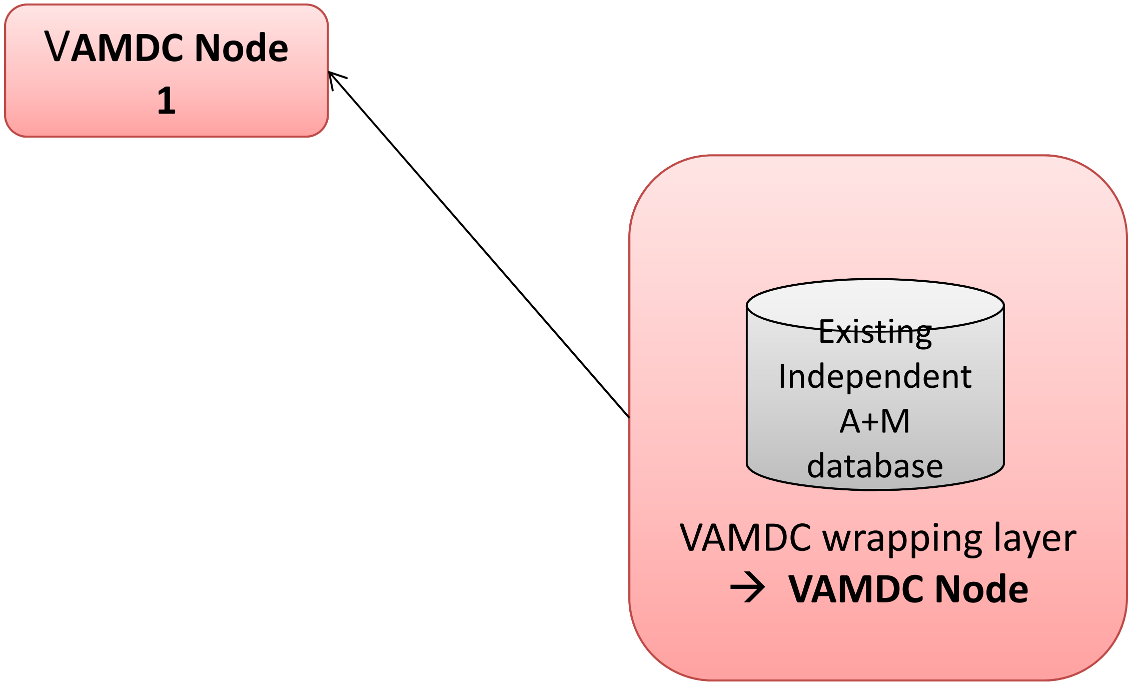


Existing Independent A+M database

VAMDC wrapping layer

→ **VAMDC Node**

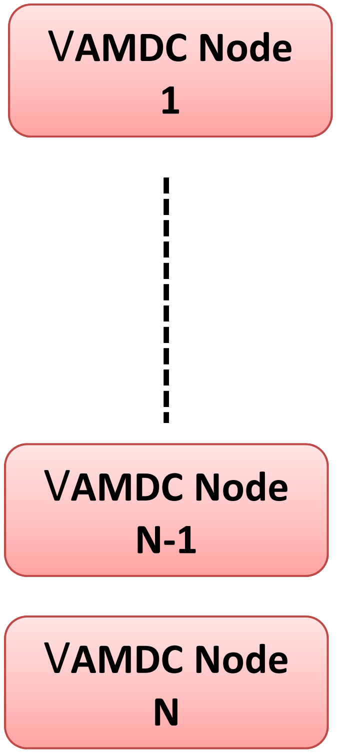
The VAMDC infrastructure technical architecture



The VAMDC infrastructure technical architecture

VAMDC Node 1

The VAMDC infrastructure technical architecture

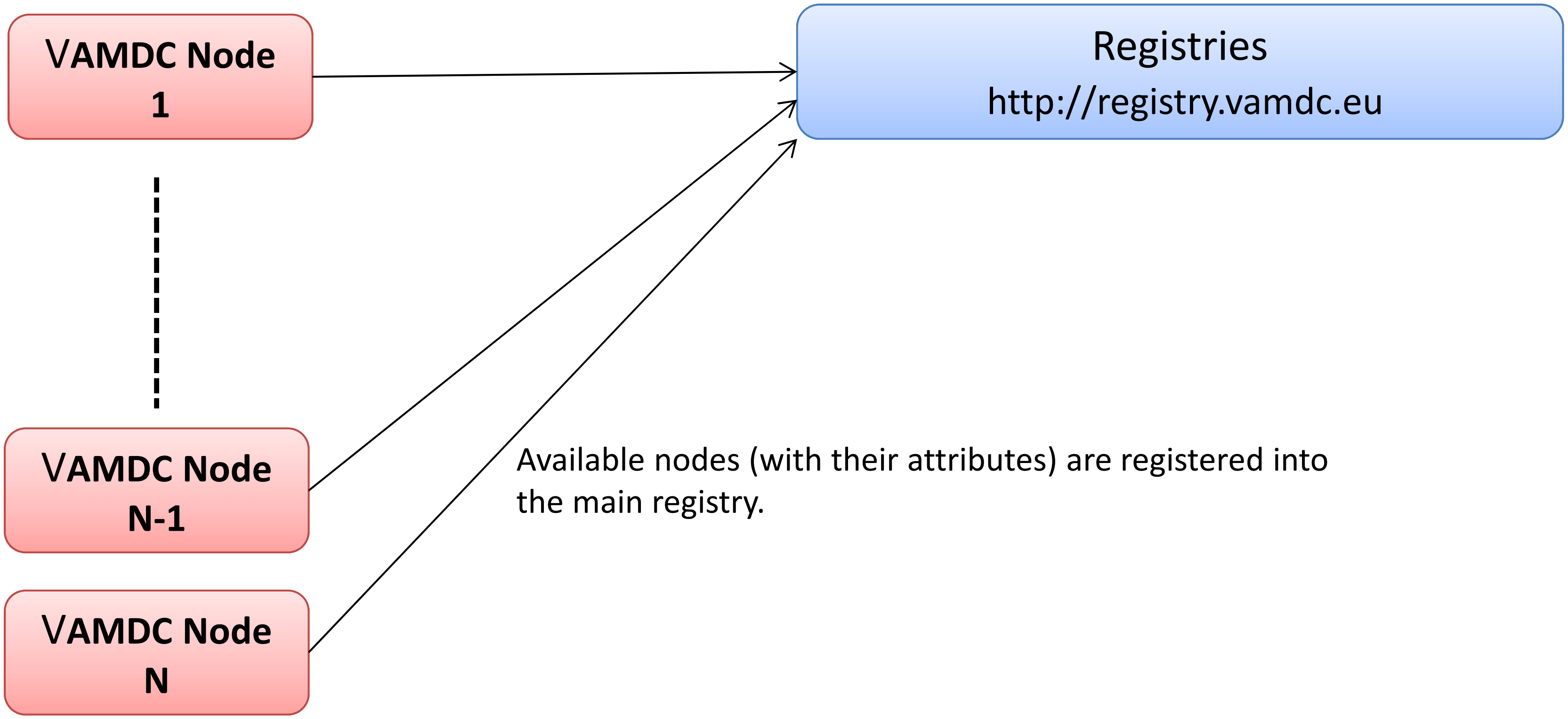


VAMDC Node 1

VAMDC Node N-1

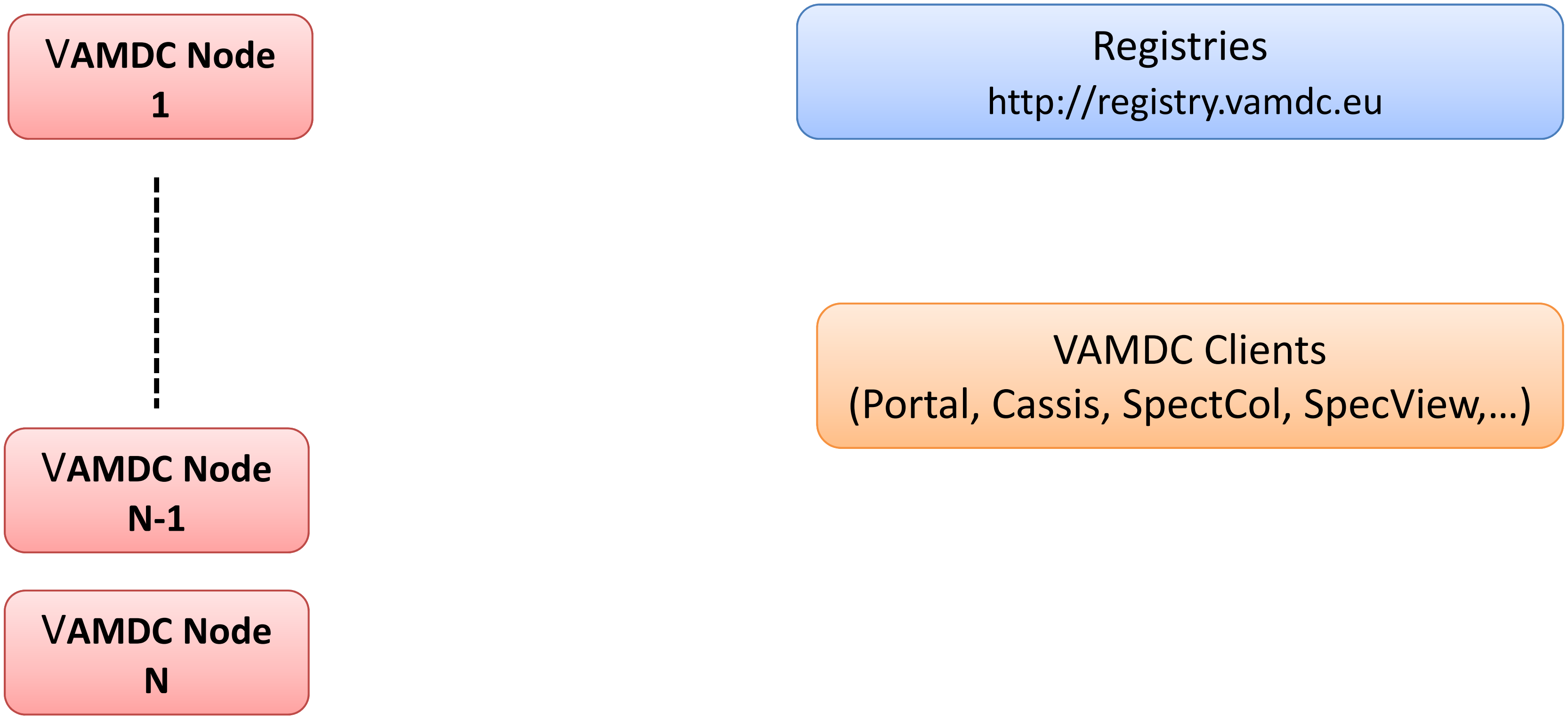
VAMDC Node N

The VAMDC infrastructure technical architecture



Available nodes (with their attributes) are registered into the main registry.

The VAMDC infrastructure technical architecture



VAMDC Node 1

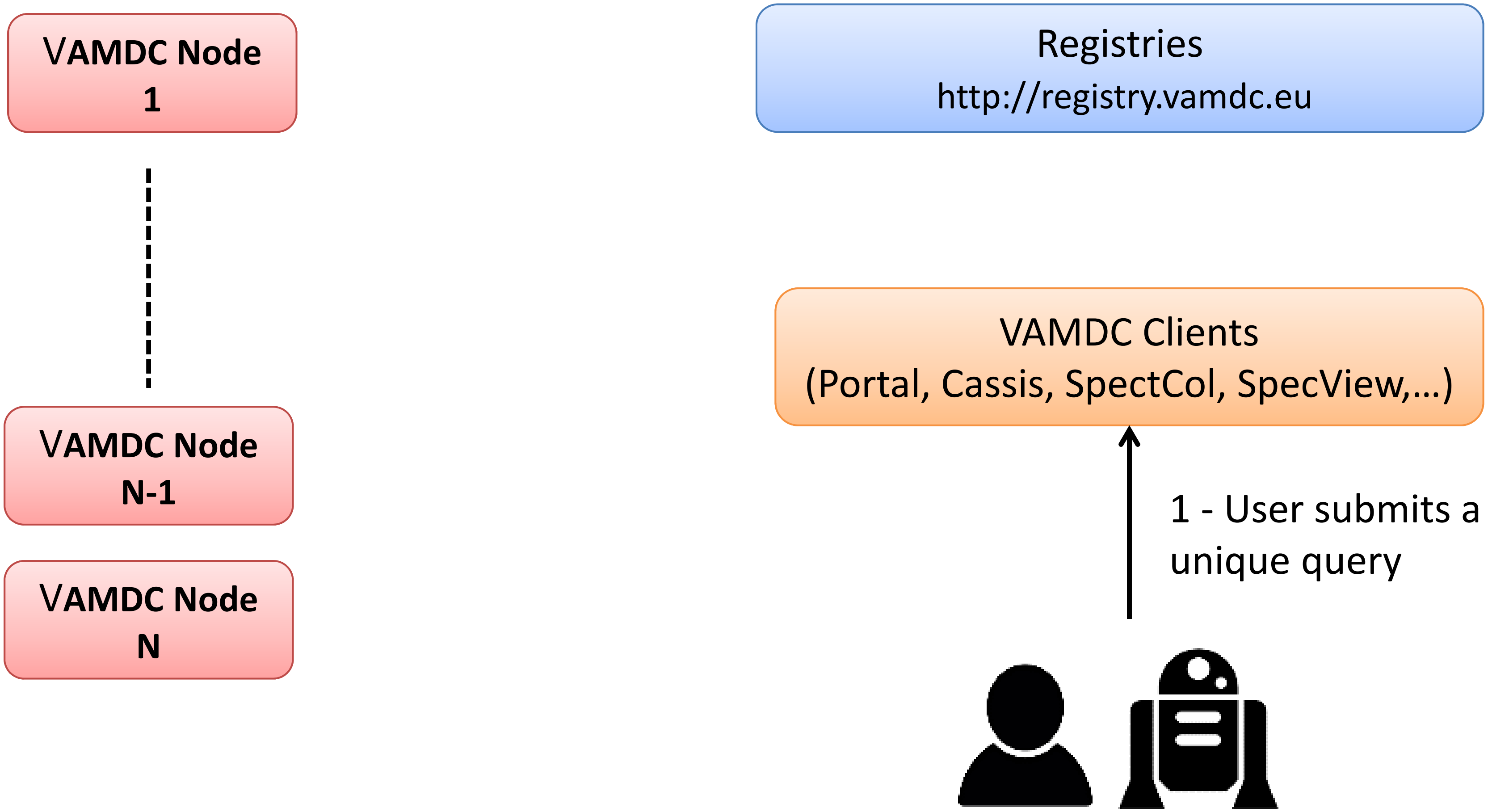
VAMDC Node N-1

VAMDC Node N

Registries
http://registry.vamdc.eu

VAMDC Clients
(Portal, Cassis, SpectCol, SpecView,…)

The VAMDC infrastructure technical architecture



VAMDC Node 1

VAMDC Node N-1

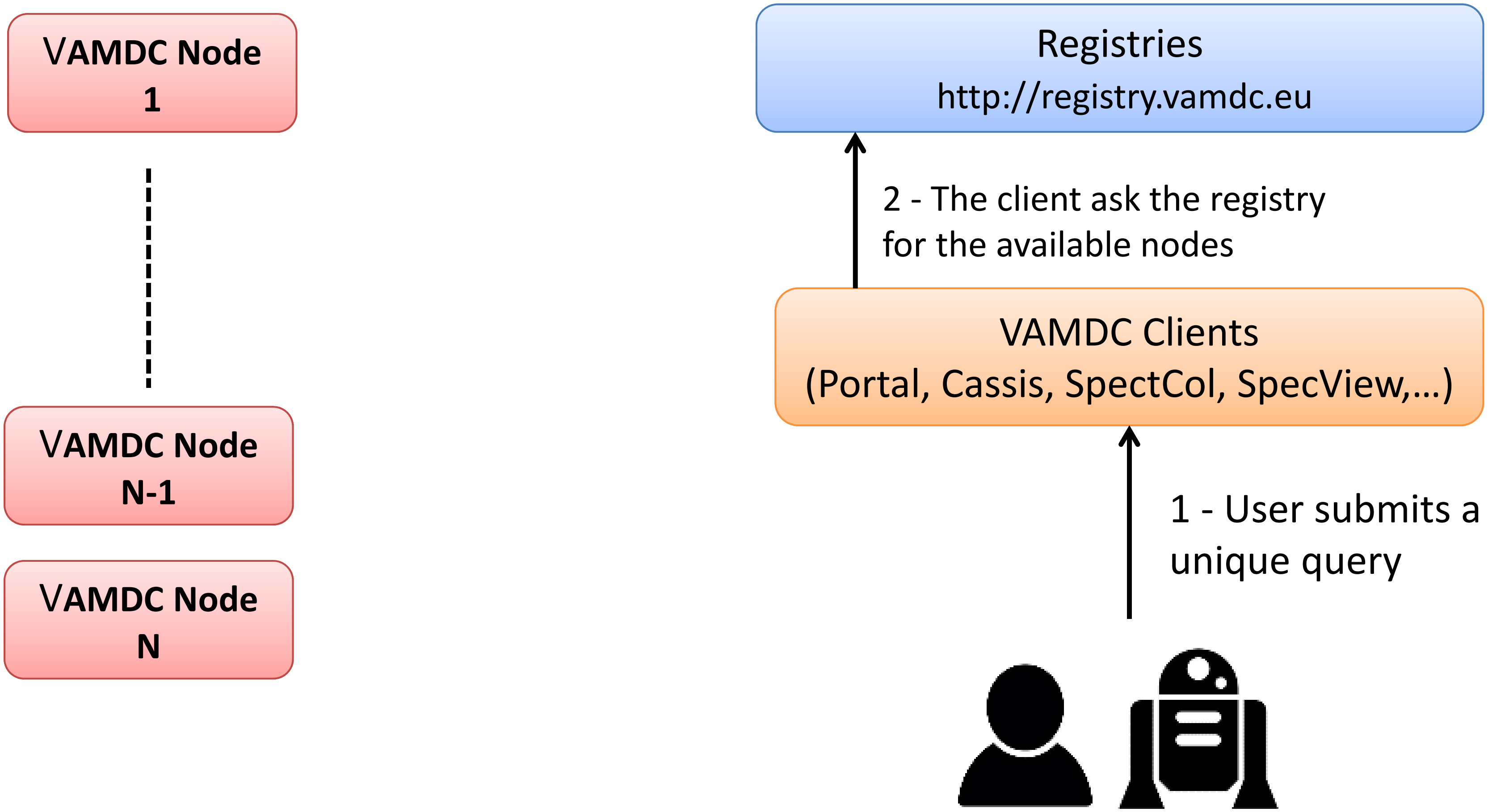
VAMDC Node N

Registries
http://registry.vamdc.eu

VAMDC Clients
(Portal, Cassis, SpectCol, SpecView,…)

1 - User submits a unique query

The VAMDC infrastructure technical architecture



VAMDC Node
1

VAMDC Node
N-1

VAMDC Node
N

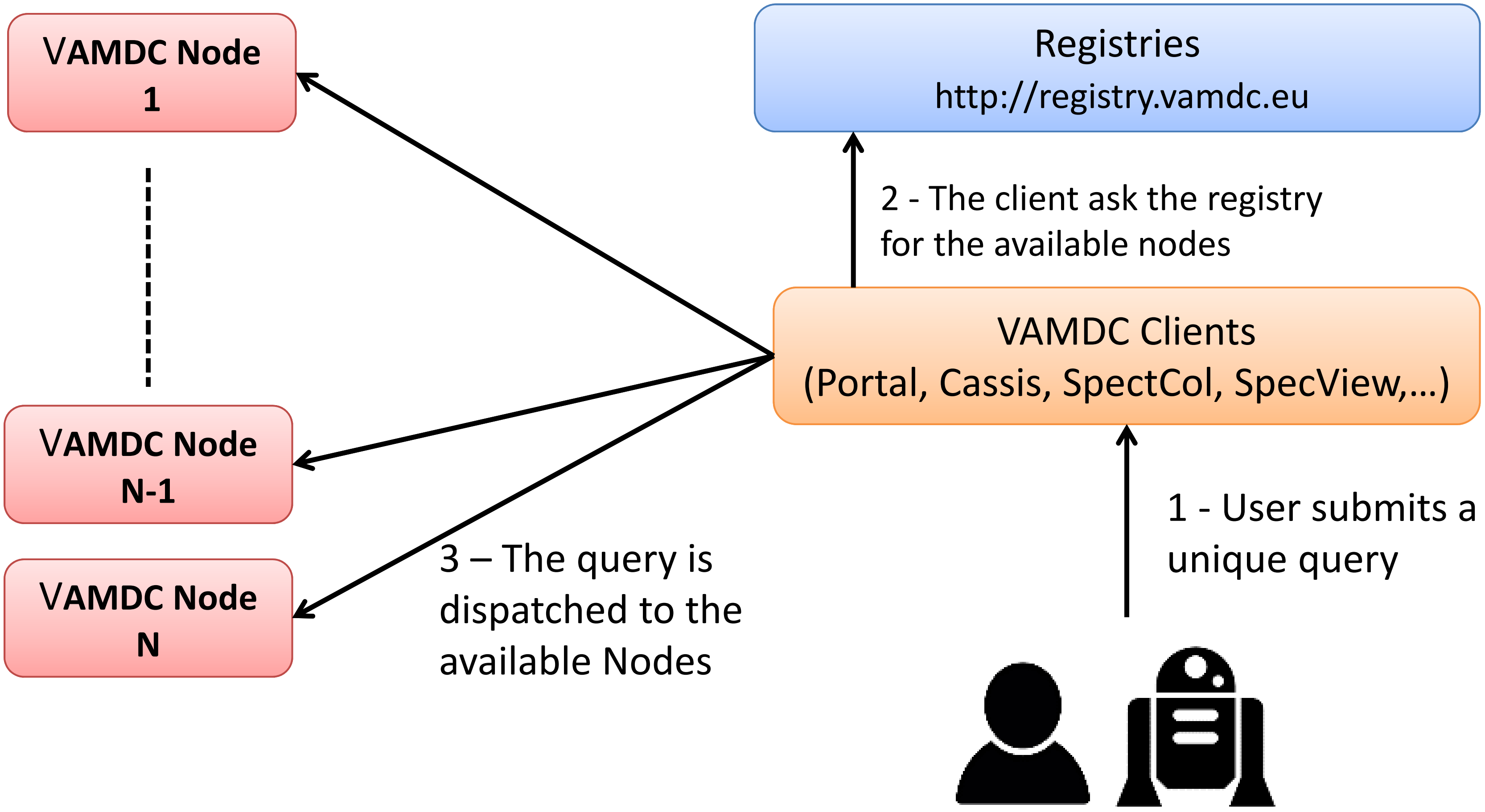
Registries
http://registry.vamdc.eu

2 - The client ask the registry
for the available nodes

VAMDC Clients
(Portal, Cassis, SpectCol, SpecView,...)

1 - User submits a
unique query

The VAMDC infrastructure technical architecture



VAMDC Node 1

VAMDC Node N-1

VAMDC Node N

Registries
http://registry.vamdc.eu

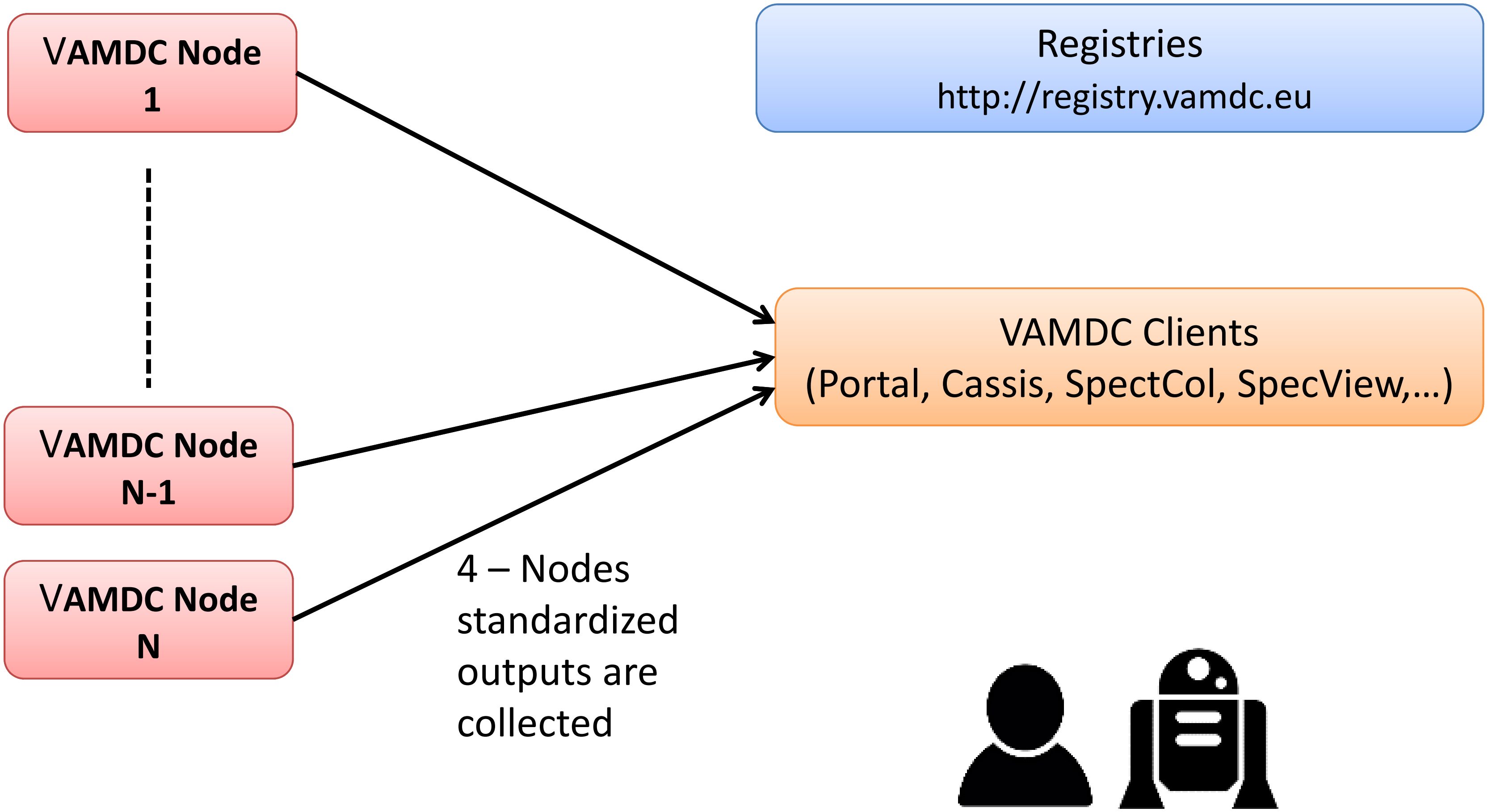
2 - The client ask the registry for the available nodes

VAMDC Clients
(Portal, Cassis, SpectCol, SpecView,…)

1 - User submits a unique query

3 – The query is dispatched to the available Nodes

The VAMDC infrastructure technical architecture



VAMDC Node
1

VAMDC Node
N-1

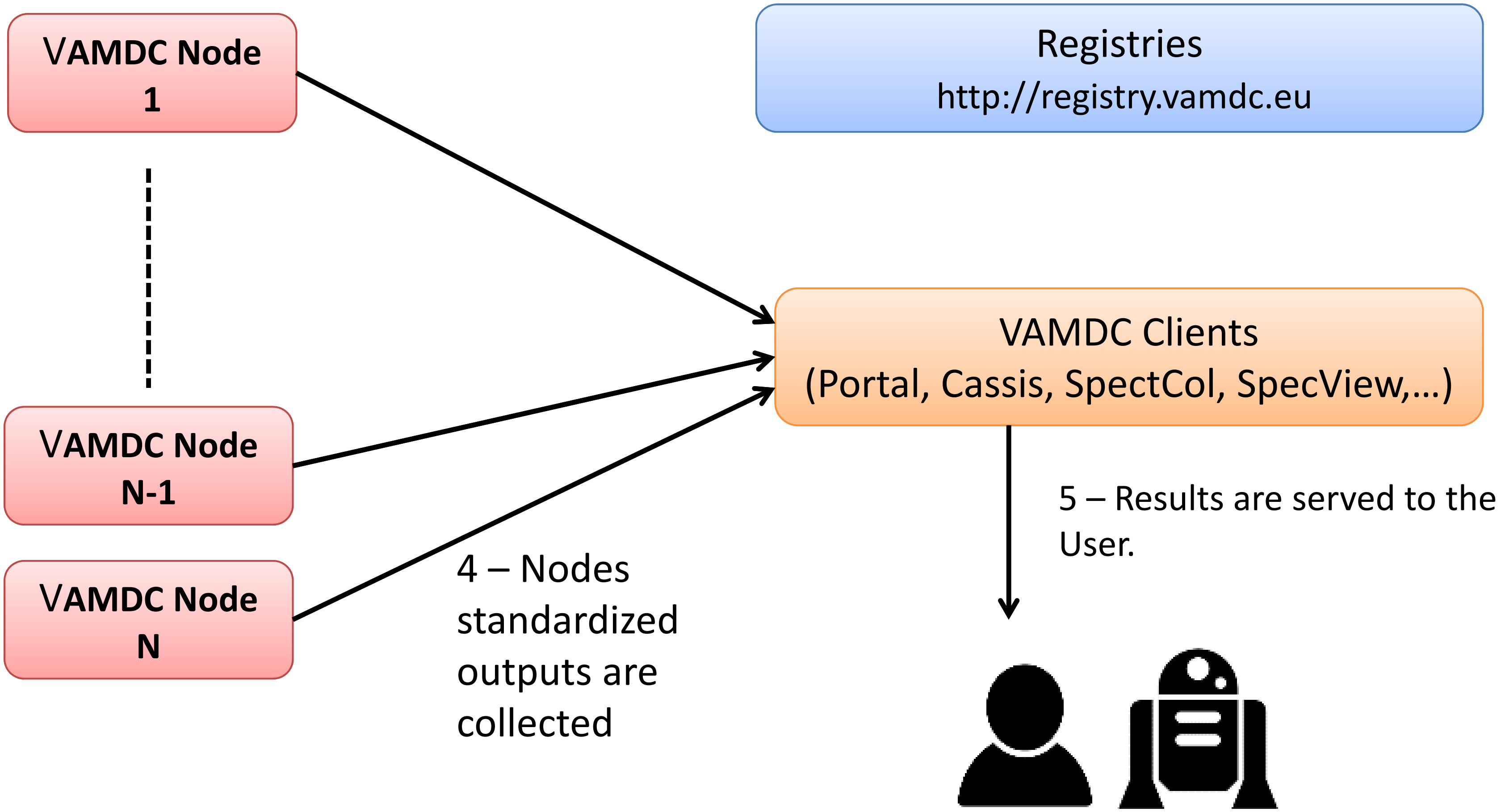
VAMDC Node
N

Registries
http://registry.vamdc.eu

VAMDC Clients
(Portal, Cassis, SpectCol, SpecView,…)

4 – Nodes standardized outputs are collected

The VAMDC infrastructure technical architecture



VAMDC Node 1

VAMDC Node N-1

VAMDC Node N

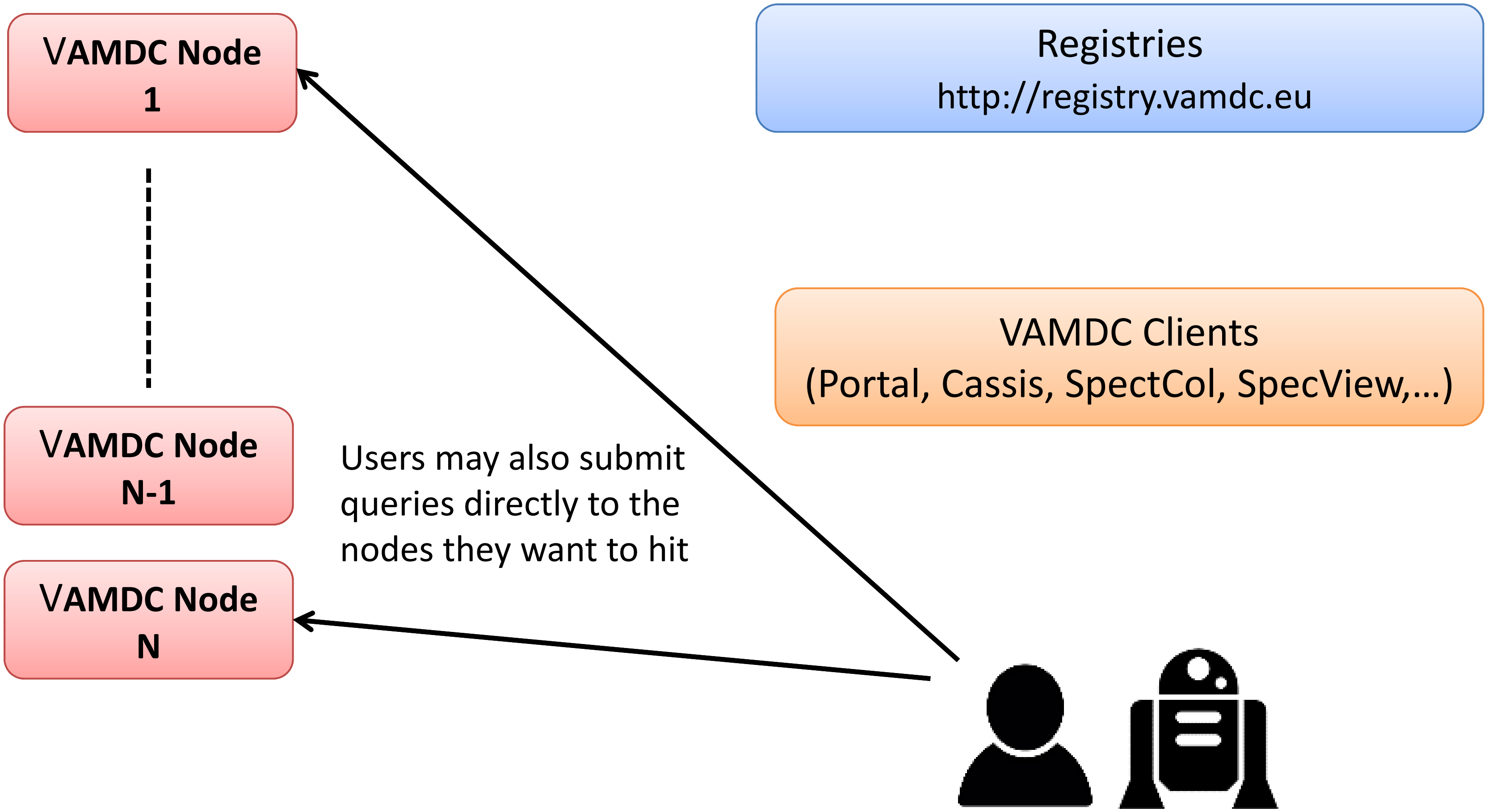
Registries
http://registry.vamdc.eu

VAMDC Clients
(Portal, Cassis, SpectCol, SpecView,…)

4 – Nodes standardized outputs are collected

5 – Results are served to the User.

The VAMDC infrastructure technical architecture



VAMDC Node
1

VAMDC Node
N-1

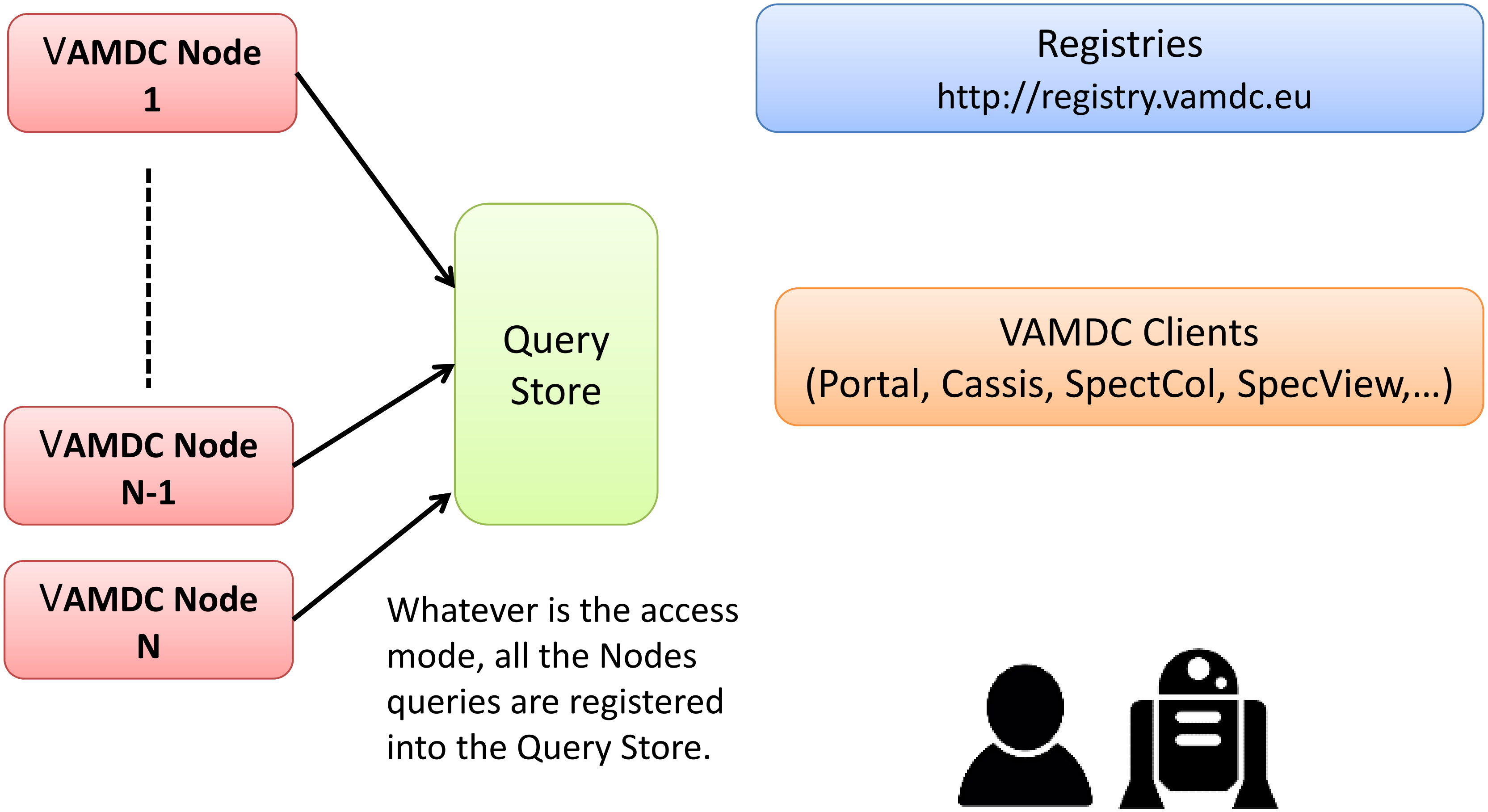
VAMDC Node
N

Registries
http://registry.vamdc.eu

VAMDC Clients
(Portal, Cassis, SpectCol, SpecView,…)

Users may also submit queries directly to the nodes they want to hit

The VAMDC infrastructure technical architecture



VAMDC Node 1

VAMDC Node N-1

VAMDC Node N

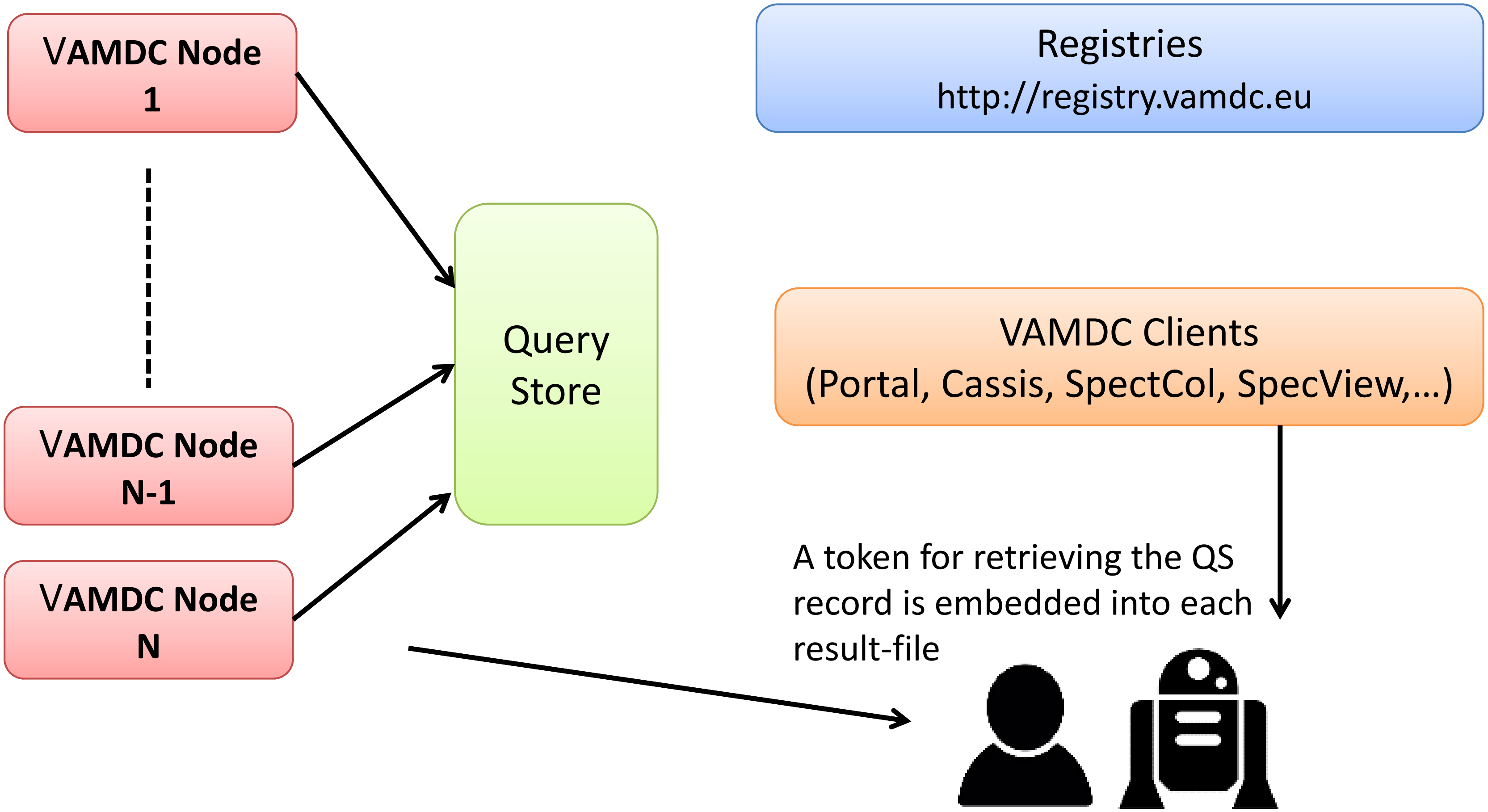
Query Store

Registries
http://registry.vamdc.eu

VAMDC Clients
(Portal, Cassis, SpectCol, SpecView,…)

Whatever is the access mode, all the Nodes queries are registered into the Query Store.

The VAMDC infrastructure technical architecture

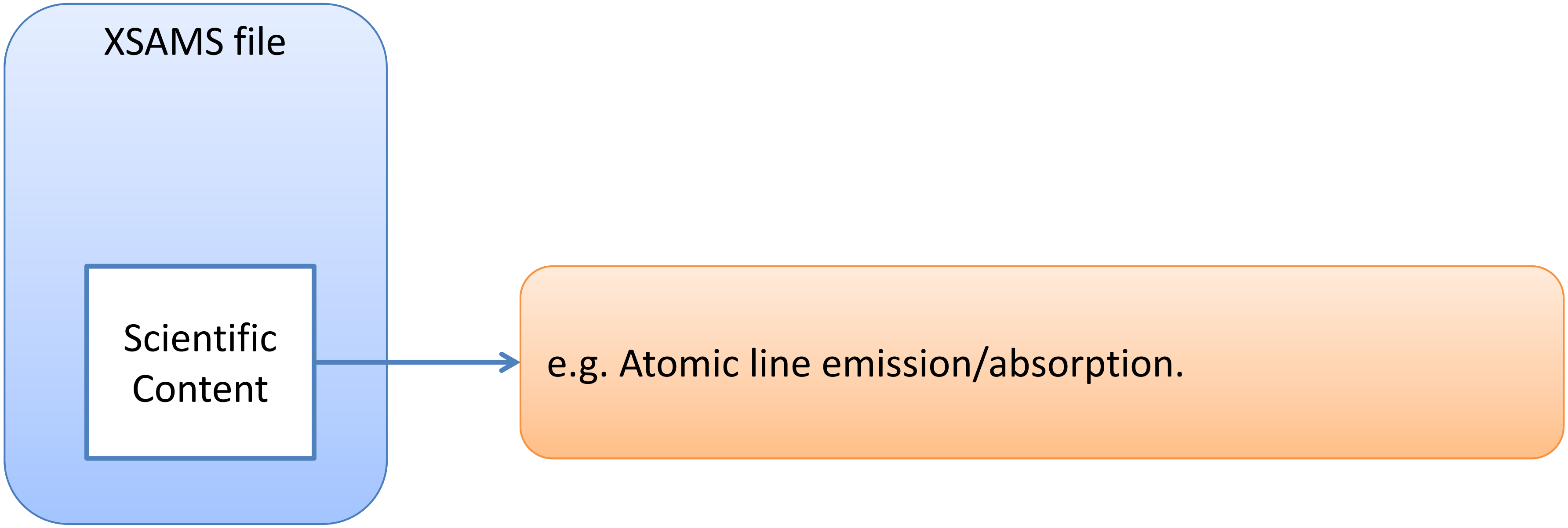


A token for retrieving the QS record is embedded into each result-file

Implementation details at : https://youtu.be/OLe-qcqCcCw

Feedbacks after some months of tests – A first issue

- VAMDC output files are formatted using the international XSAMS standard (XML schema for Atoms Molecules and Solids), endorsed by IAEA.
- For not breaking the backward compatibility the QS-feature is an overlay to these existing standards
- Each VAMDC query-result is an XSAMS file



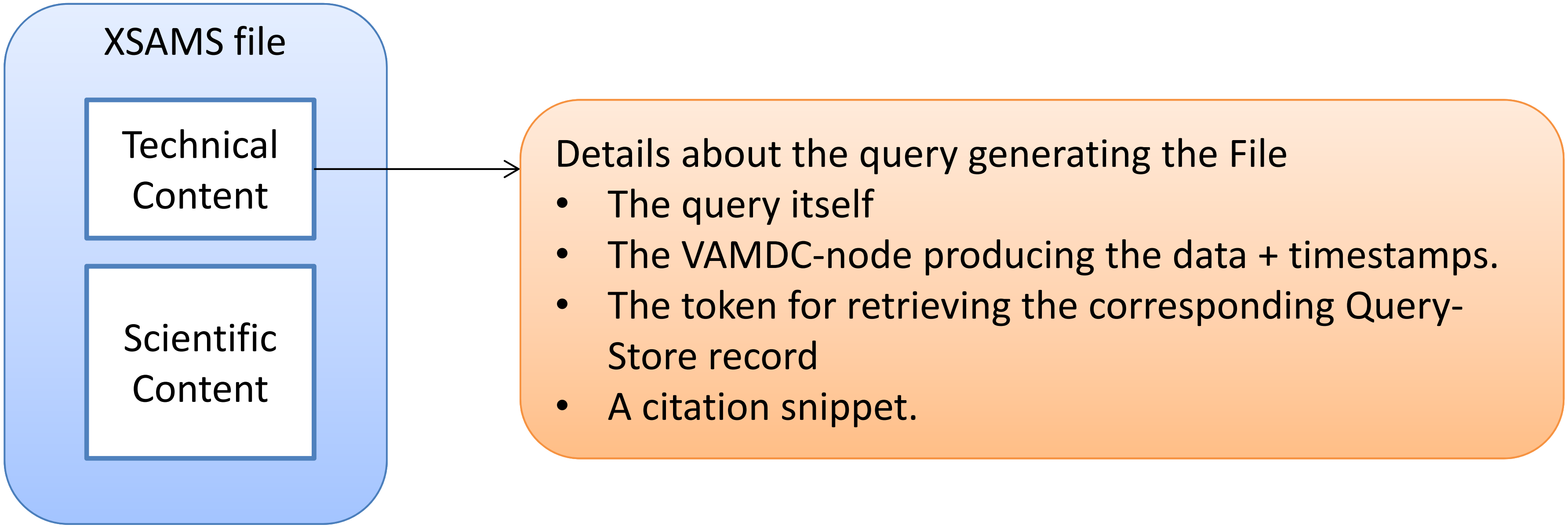
XSAMS file

Scientific Content

e.g. Atomic line emission/absorption.

Feedbacks after some months of tests…

- VAMDC output files are formatted using the international XSAMS standard (XML schema for Atoms Molecules and Solids), endorsed by IAEA.
- For not breaking the backward compatibility the QS-feature is an overlay to these existing standards
- Each VAMDC query-result is an XSAMS file



XSAMS file

Technical Content

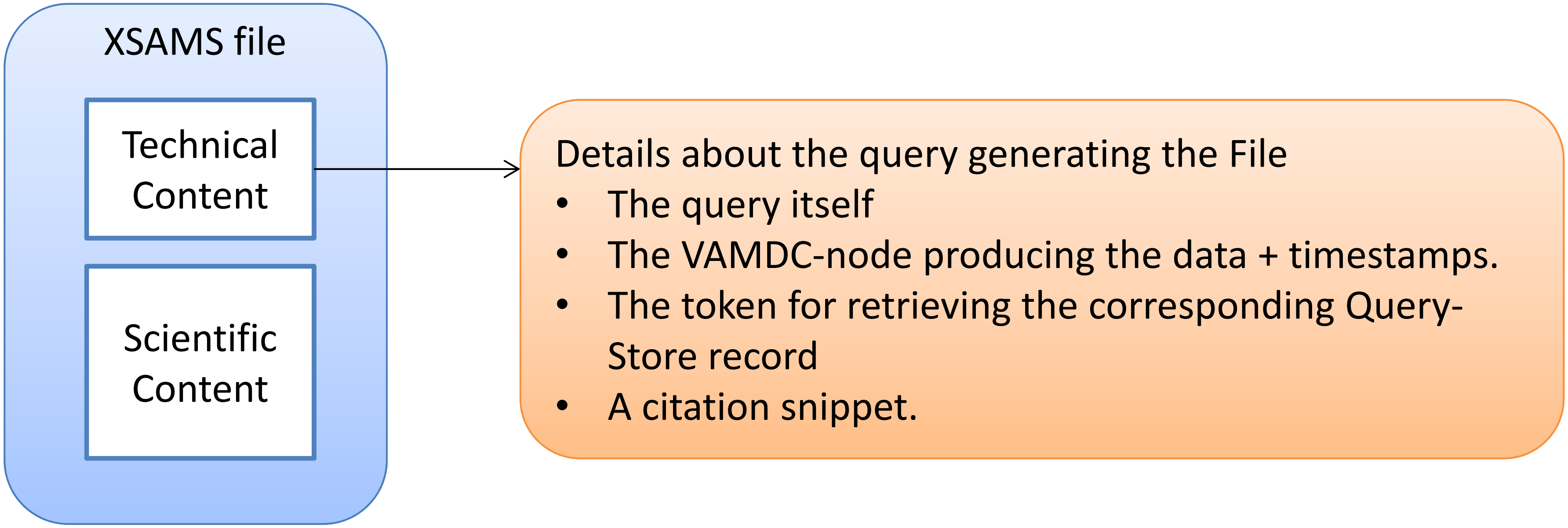
Scientific Content

Details about the query generating the File

- The query itself
- The VAMDC-node producing the data + timestamps.
- The token for retrieving the corresponding Query-Store record
- A citation snippet.

Feedbacks after some months of tests – A first issue

- All the clients provided by VAMDC deal with the new QS feature.
- Many clients “out in the nature” are not produced by VAMDC
  - All the submitted queries are tracked in the Query Store
  - But the clients do not necessarily process the technical part for getting and presenting to the users the citation information.



XSAMS file

Technical Content

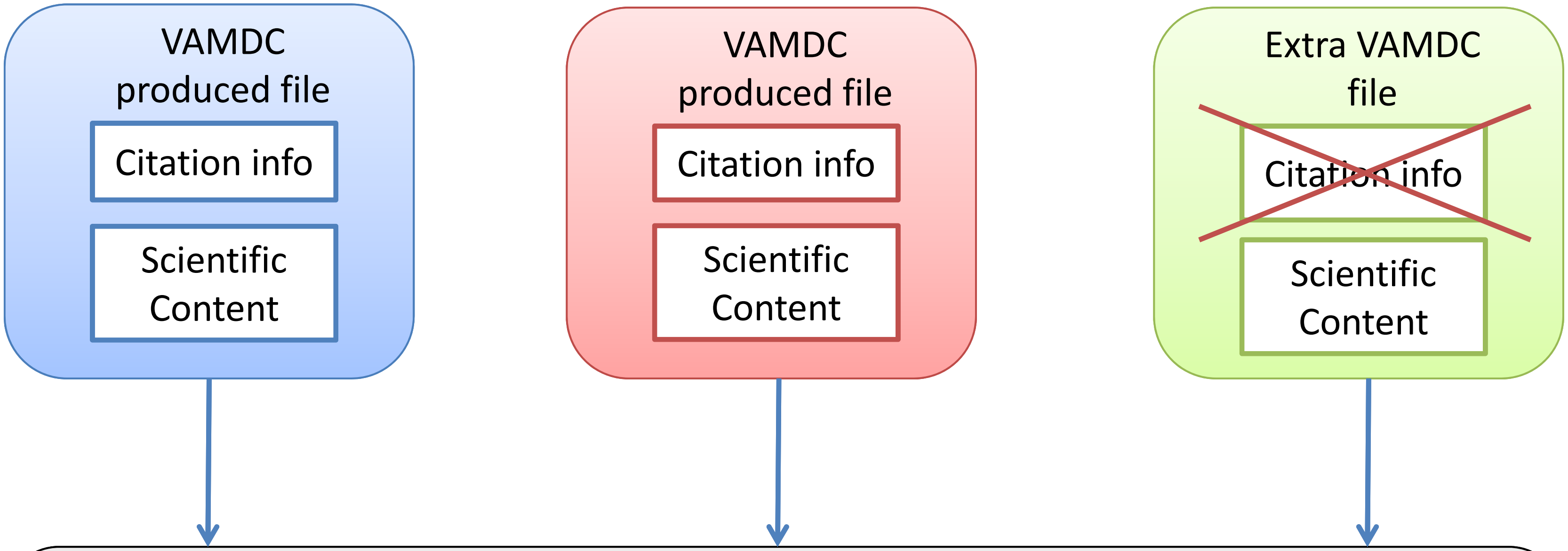
Scientific Content

Details about the query generating the File

- The query itself
- The VAMDC-node producing the data + timestamps.
- The token for retrieving the corresponding Query-Store record
- A citation snippet.

Feedbacks after some months of tests – A second issue

- The output file standard is not designed for being human readable. It is for computer based tools and processors.
- Let us consider a tool aware of QS features, working with different files:

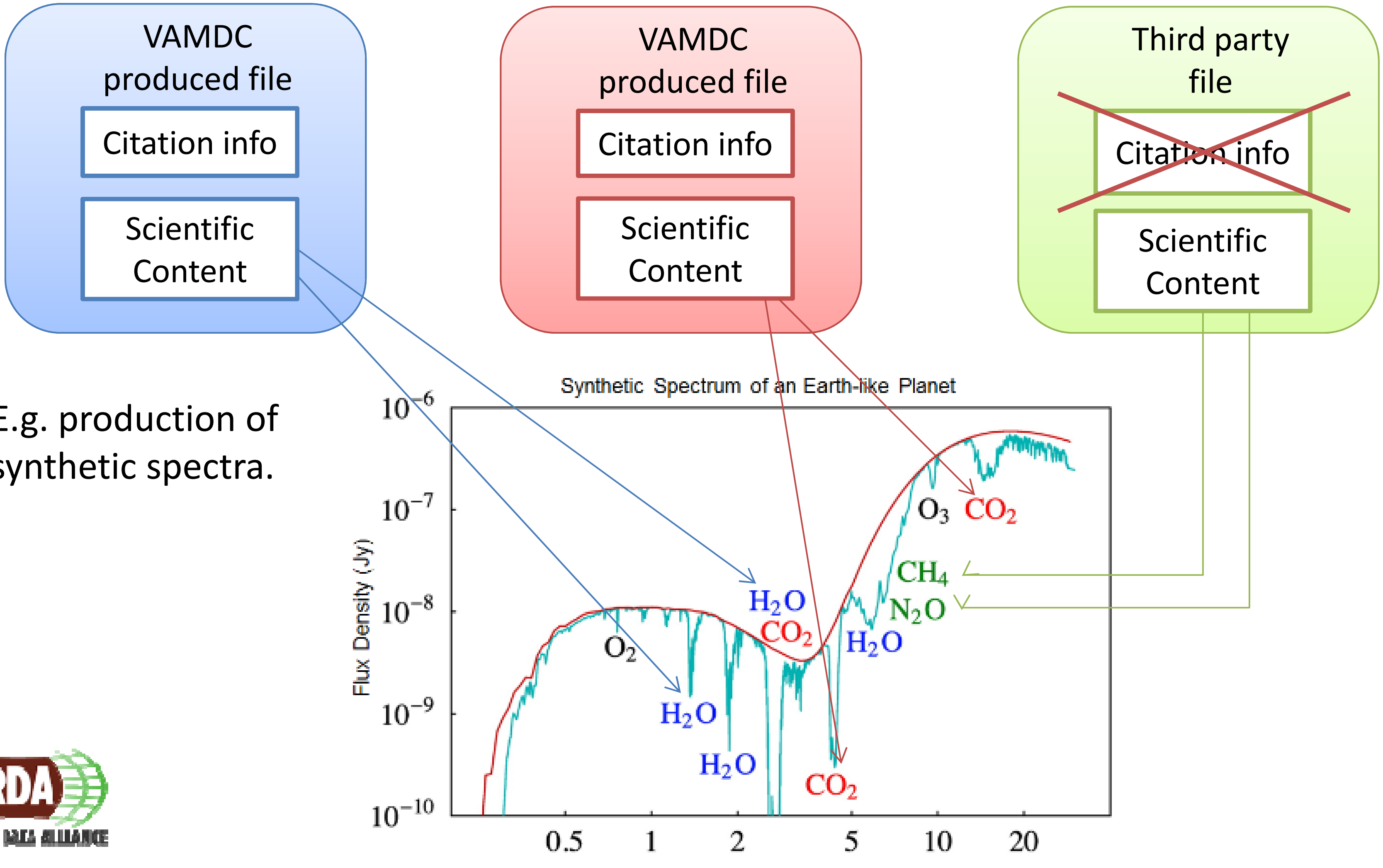


Data may be compared, crossmatched, merged

- The citation information may be lost/diluted
- The processing tools should propagate the citation information while processing
  - The data citation issues is mingled with provenance issues.

Feedbacks after some months of tests – A second issue

- The output file standard is not designed for being human readable. It is for computer based tools and processors.
- Let us consider a tool aware of QS features, working with different files:



VAMDC produced file

Citation info

Scientific Content

VAMDC produced file

Citation info

Scientific Content

Third party file

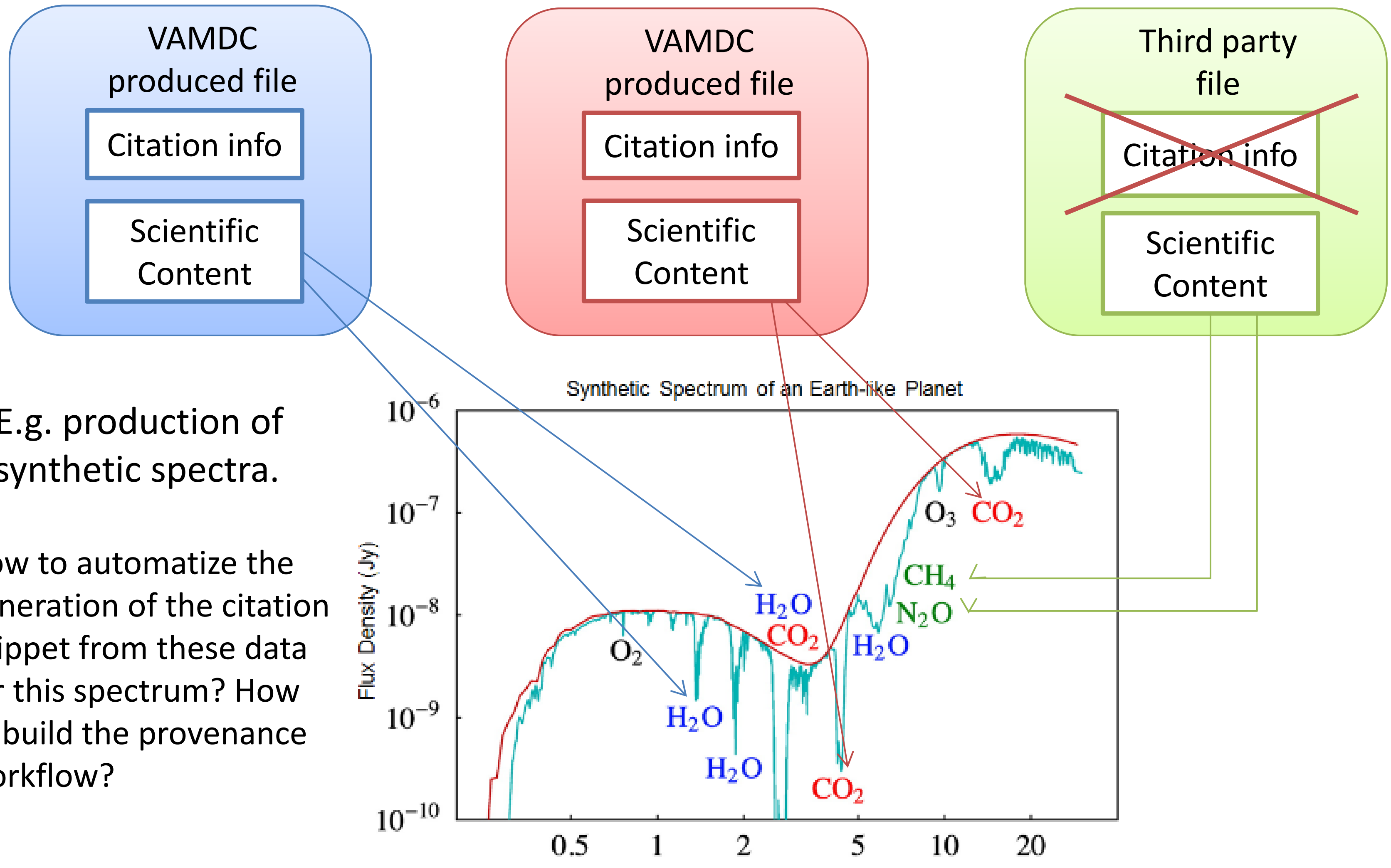
~~Citation info~~

Scientific Content

E.g. production of synthetic spectra.

Feedbacks after some months of tests – A second issue

- The output file standard is not designed for being human readable. It is for computer based tools and processors.
- Let us consider a tool aware of QS features, working with different files:



VAMDC produced file

Citation info

Scientific Content

VAMDC produced file

Citation info

Scientific Content

Third party file

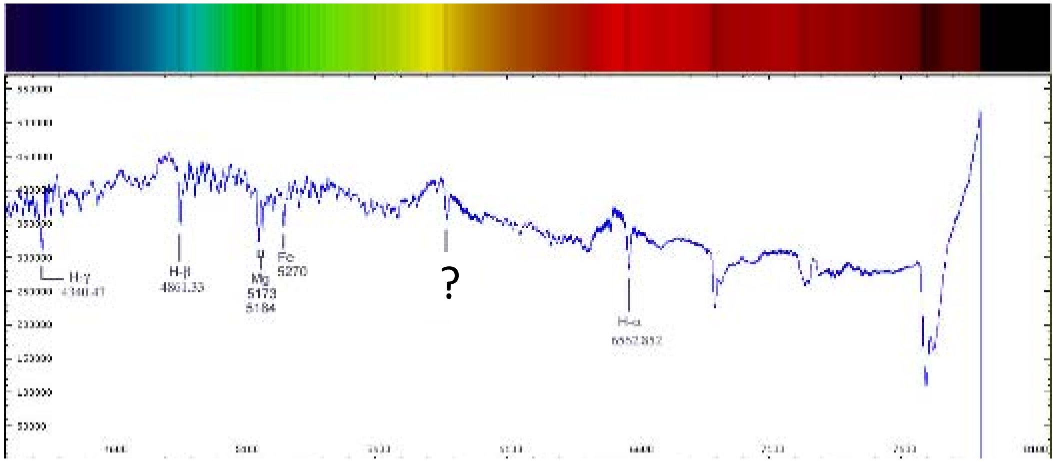
~~Citation info~~

Scientific Content

E.g. production of synthetic spectra.

How to automatize the generation of the citation snippet from these data for this spectrum? How to build the provenance workflow?

Feedbacks after some months of tests – A third issue



Scientist may need to identify an unknown absorption line on a natural spectrum

•This is a “guess and try” work where users extract lots of data (multiple queries)

- Each query may contain hundreds of lines.

The Experience shown:

•This is an iterative work.

•During this work, scientists do not pay attention to the citation information contained into the files, since they are not sure that the data content will be useful (lot of files in this try and guess procedures are deleted).

•When they finally find a match between the line in the spectrum and the VAMDC datum, it is too late! The citation information is lost during the process

Concluding remarks, questions

- The experience shown that the adoption of the Query Store by the community is not straightforward

- How to solve these adoption issues?
  - An ad hoc education to data citation for the community members?
  - Provide the community with new tools? Which ones?

- Are other communities facing the same problems?
  - May we find trans-community solutions (in the RDA spirit)?

RDA

RESEARCH DATA ALLIANCE

Citing Dynamic Datasets at NICT

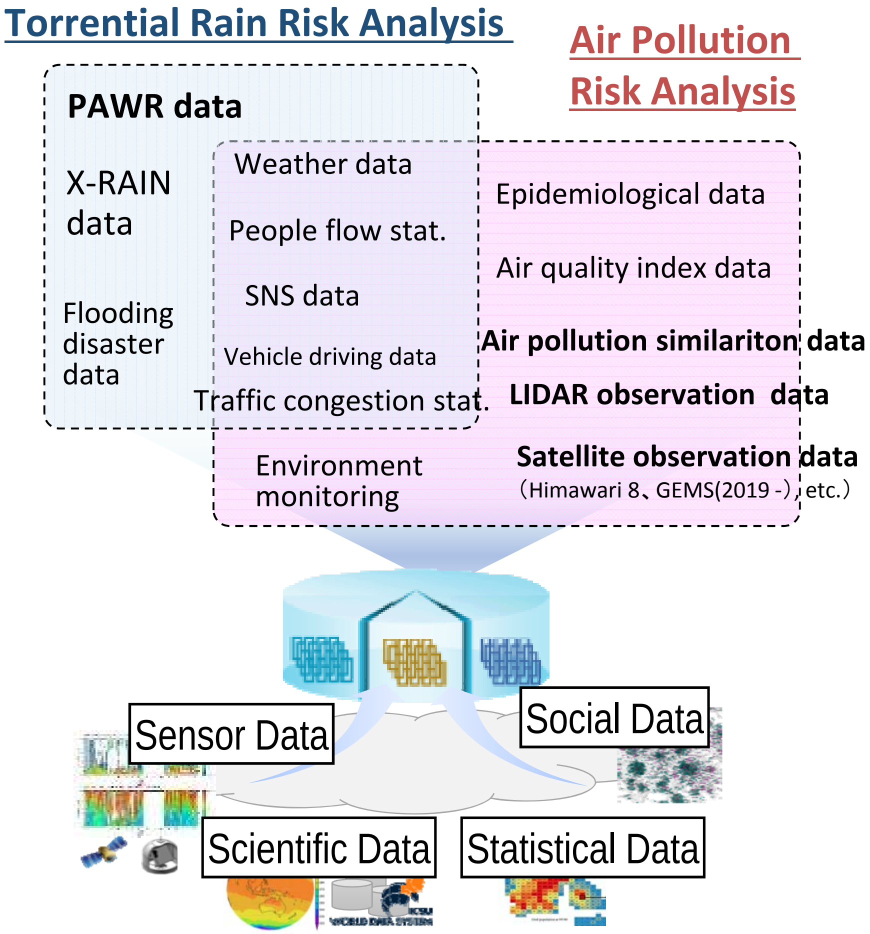
Yasuhiro Murayama
murayama@nict.go.jp

research data sharing without barriers
rd-alliance.org

Citing dynamic datasets for Sensing Big Data at NICT

Koji Zettsu and Yasuhiro Murayama
National Institute of Information & Communications Technology (NICT), Japan

- Discoverability of individually-disseminated sensing data for natural disaster events
- Analysis of data reputation for incentive reward in participatory sensing
- Data provenance for quality assurance of data mining results



Torrential Rain Risk Analysis

Air Pollution Risk Analysis

PAWR data

X-RAIN data

Flooding disaster data

Weather data

People flow stat.

SNS data

Vehicle driving data

Traffic congestion stat.

Epidemiological data

Air quality index data

Air pollution similariton data

LIDAR observation data

Satellite observation data
（Himawari 8、GEMS(2019 -) , etc.）

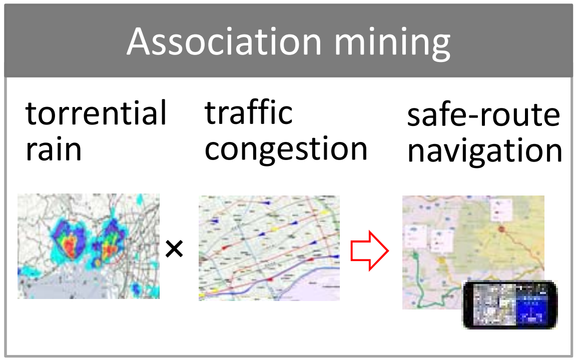
Environment monitoring

Sensor Data

Social Data

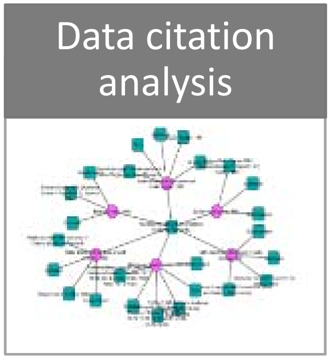
Scientific Data

Statistical Data



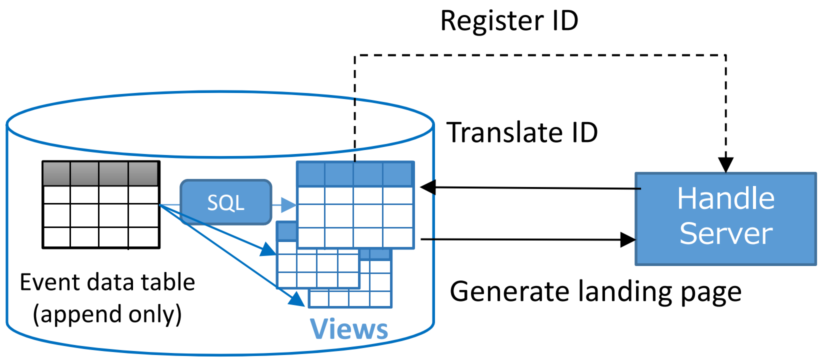
Association mining

torrential rain × traffic congestion ⇨ safe-route navigation



Data citation analysis

Citable mechanism and experiment



Register ID

Translate ID

SQL

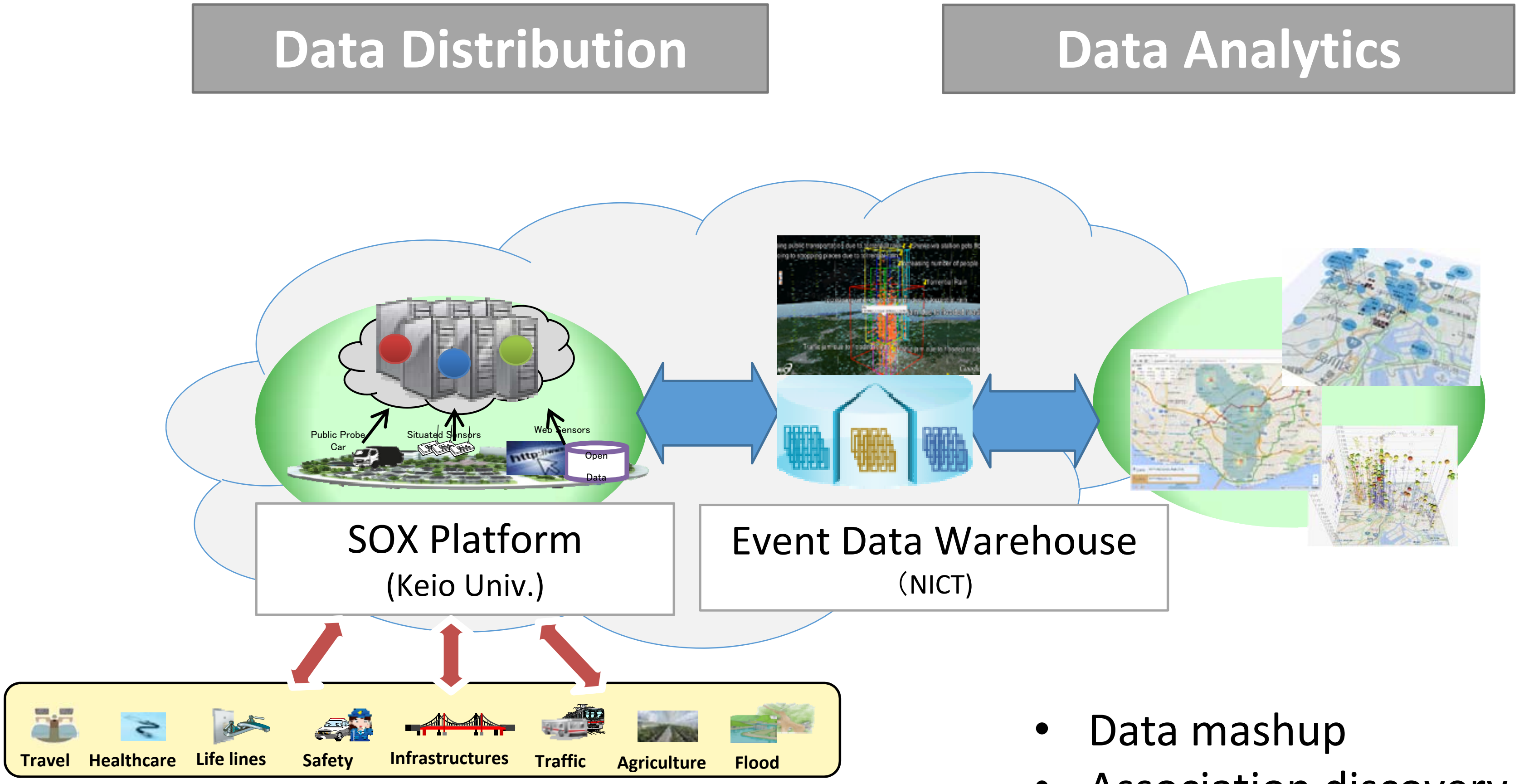
Handle Server

Generate landing page

Event data table (append only)

Views

NICT Social Big Data Platform



Data Distribution

Data Analytics

SOX Platform
(Keio Univ.)

Event Data Warehouse
（NICT）

Travel Healthcare Life lines Safety Infrastructures Traffic Agriculture Flood

Heterogeneous sensing data created by contract research partners

- Data mashup
- Association discovery
- Visual analysis

NICT Event Data Warehouse

Torrential Rain Risk Analysis

PAWR data

X-RAIN data

Flooding disaster data

Weather data

People flow stat.

SNS data

Vehicle driving data

Traffic congestion stat.

Environment monitoring

Epidemiological data

Air quality index data

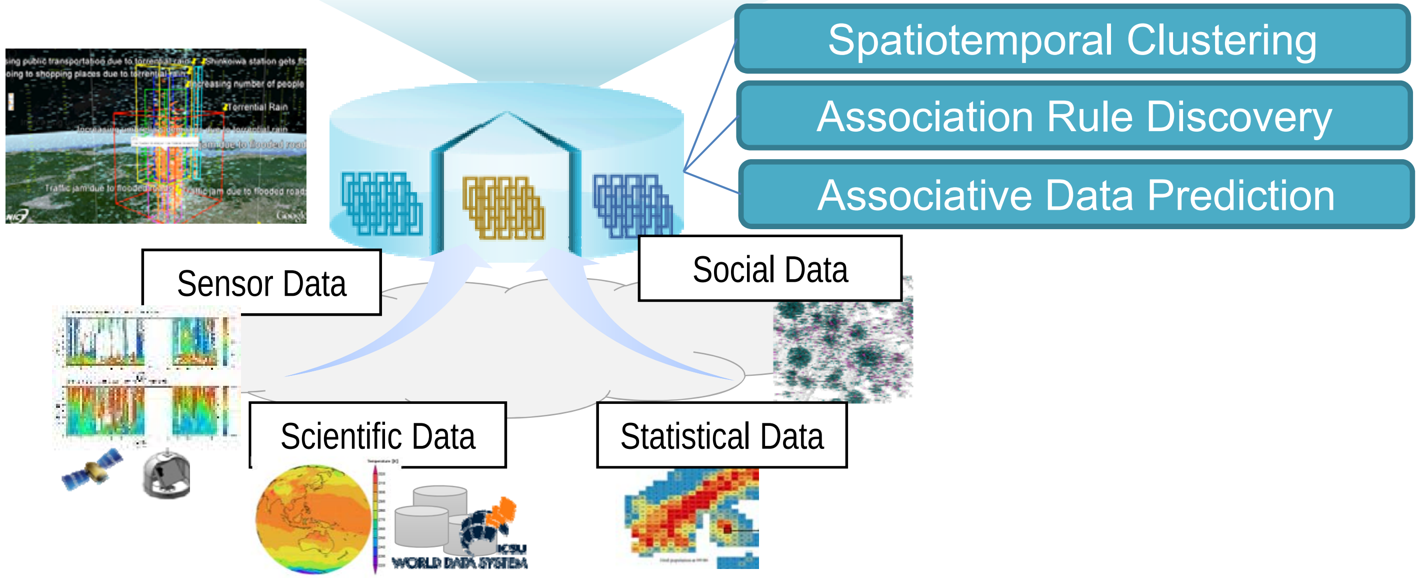
Air pollution similariton data

RIDAR observation data

Satellite observation data
（Himawari 8、GEMS(2019 -), etc.）

Air Pollution Risk Analysis

Bold: NICT-owned data



Spatiotemporal Clustering

Association Rule Discovery

Associative Data Prediction

Sensor Data

Social Data

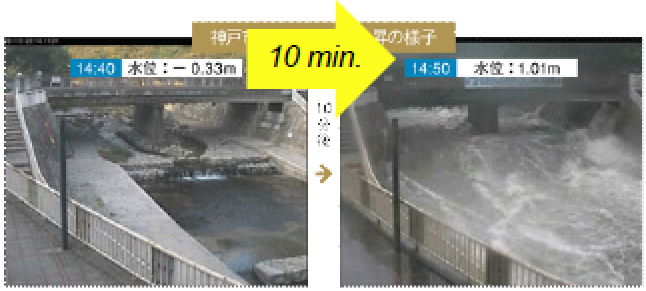
Scientific Data

Statistical Data

Case Study: Urban Natural Disaster

Localized Torrential Rain Disaster

- < 100 mm/h heavy rain within $10km^2$ area in short time (< 1 hour).
- Unpredictable. Urban heat island as a major cause
- 200+ times per year in Japan. 180 billion JPY damage for past 10 years in Tokyo



Flash flooding (Kobe, 2008)

Flood in subway
(Fukuoka, 1999)

Landslide
(Hiroshima, 2014)

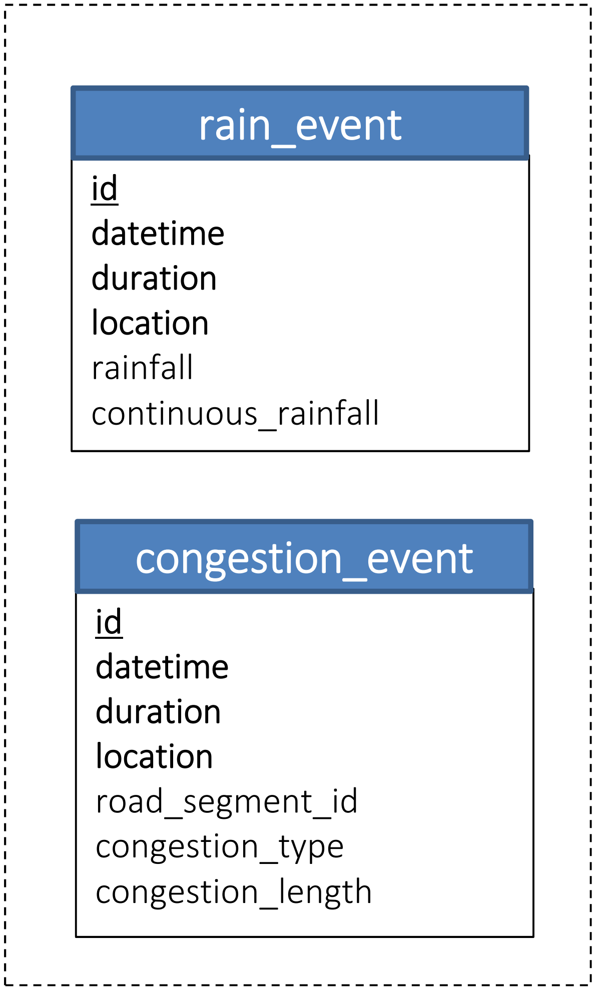
Sensing Data Statistics (Part)

Sensing Data	Type	Records	Bytes	Period
PAWR precipitation data (2km grid)	Live	19.4 billion 16.8 million/day	8.7TB (6.9GB/day)	2013.5 - current
PAWR precipitation data (250m grid)	Live	(2,800/day)	(3.5GB/day)	(preparing)
XRAIN precipitation data (250m grid)	Archive	(14,400/day)	(5.9GB/day)	(preparing)
Soratena weather station data (3,000 stations nationwide)	Live	1.3 billion	479GB	2013.9 – current
AEROS atmospheric environmental observation data（Soramame-kun)	Live	36.4 million	37GB	2013.9 – current
Twitter SNS data (25 keywords) 2014	Archive	88 million	23GB	2014.4 - 2015.3
Twitter SNS data (geo-tagged) 2014	Archive	95 milion	25GB	(preparing)
Twitter SNS 2015	Archive			(preparing)

- PAWR precipitation data, http://pawr.nict.go.jp/index_en.html
- Soratena weather sensing data, http://soratena.live-e.org/?file=top
- AEROS (Soramame-kun), http://soramame.taiki.go.jp/

Event Data Warehouse Management System

- Storing environmental and social sensing data sets from both data archives and live streams
- Event data model
  - A table := an event type
  - Common event attributes
    - Event record ID
    - Occurrence time and duration
    - Occurrence location (geographical shape)
- Implemented on Greenplum DWH system
  - PostgreSQL-compatible
  - GIS data types and operations by PostGIS
  - DBMS with HPC enhancement



rain_event

id
datetime
duration
location
rainfall
continuous_rainfall

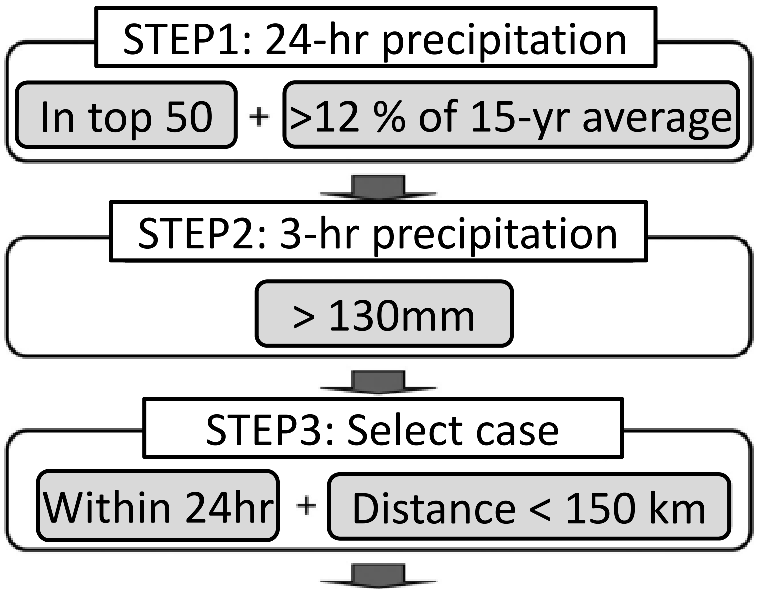
congestion_event

id
datetime
duration
location
road_segment_id
congestion_type
congestion_length

Query example of Torrential Rainfall Event

○Extract event data from Weather Radar databases

◇Creating metadata based on query



STEP1: 24-hr precipitation

In top 50 + >12 % of 15-yr average

STEP2: 3-hr precipitation

> 130mm

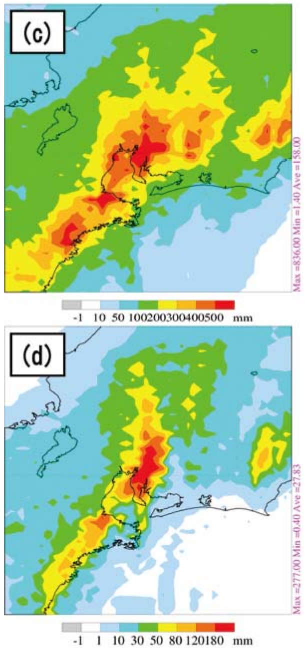
STEP3: Select case

Within 24hr + Distance < 150 km

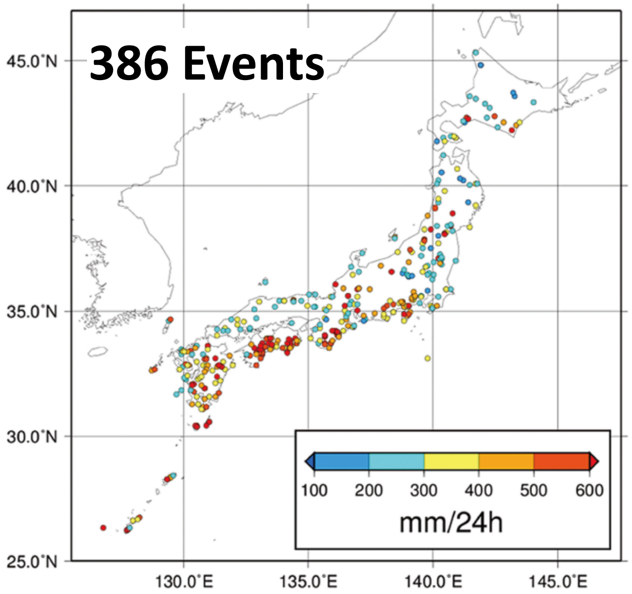
Events of "Torrential Rainfall"

Example of Extracted Torrential Rainfall Events

Top: 24-hr precipitation
Bottom: 3-hr precipitation


(c)
-1 10 50 100200300400500 mm
Max =836.00 Min =1.40 Ave =158.00
(d)
-1 1 10 30 50 80 120180 mm
Max =277.00 Min =0.40 Ave =27.83

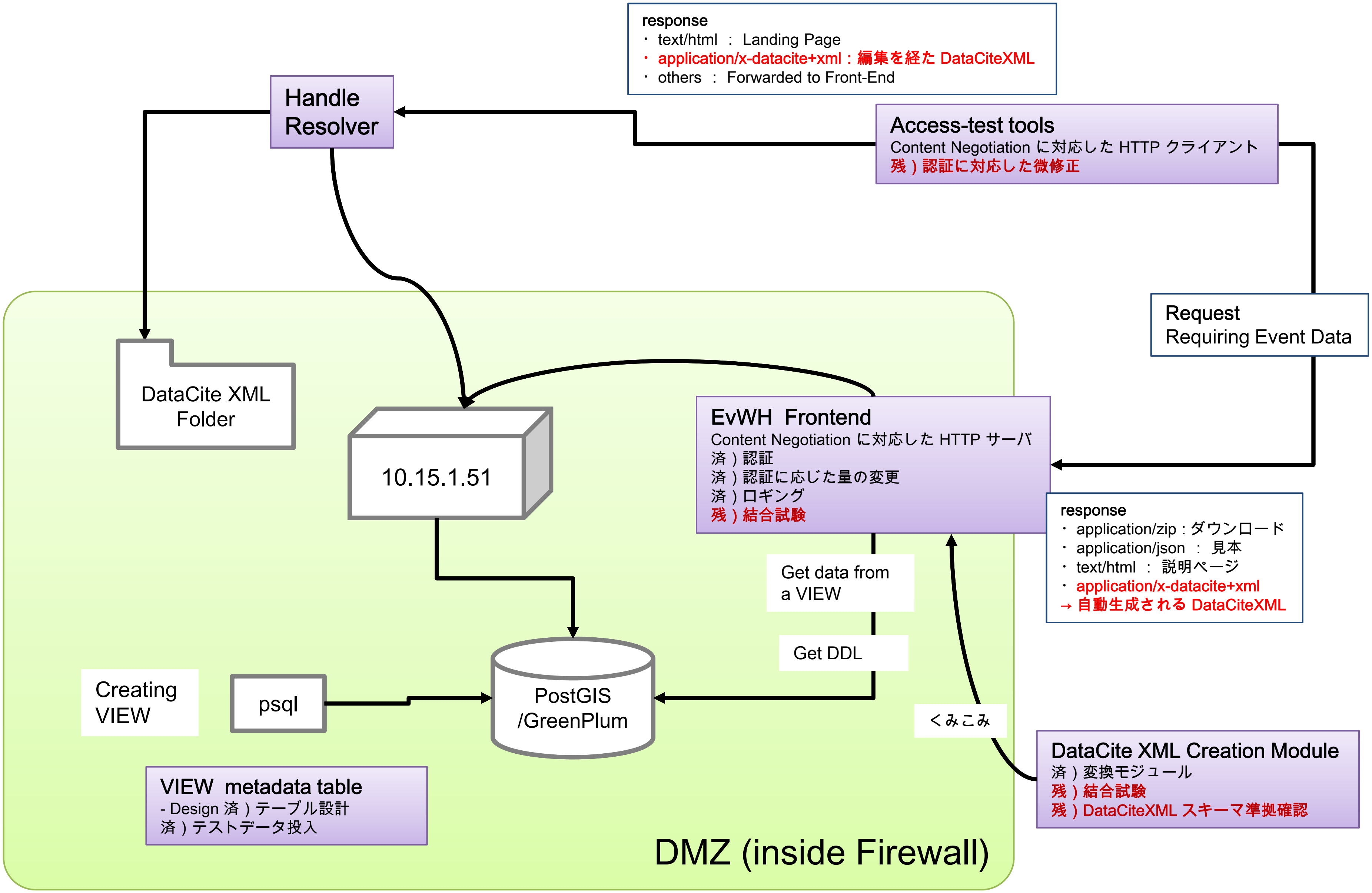
Events extracted database for a period of April-Nov. of 1995 2009



第５図　抽出された集中豪雨事例の分布．各点は，各集中豪雨事例における最大前24時間積算降水量の点を示す．陰影は前24時間積算降水量を表す．

出典：津口裕茂, 加藤輝之:集中豪雨事例の客観的な抽出とその特性・特徴に関する統計解析,日本気象学会, 天気61. 6, pp.455-469（2014年6月）

System Development



response
・ text/html ： Landing Page
・ application/x-datacite+xml：編集を経た DataCiteXML
・ others ： Forwarded to Front-End

Handle Resolver

Access-test tools
Content Negotiation に対応した HTTP クライアント
残）認証に対応した微修正

Request
Requiring Event Data

DataCite XML Folder

10.15.1.51

EvWH Frontend
Content Negotiation に対応した HTTP サーバ
済）認証
済）認証に応じた量の変更
済）ロギング
残）結合試験

response
・ application/zip：ダウンロード
・ application/json ： 見本
・ text/html ： 説明ページ
・ application/x-datacite+xml
→ 自動生成される DataCiteXML

Get data from a VIEW

Get DDL

くみこみ

Creating VIEW

psql

PostGIS /GreenPlum

DataCite XML Creation Module
済）変換モジュール
残）結合試験
残）DataCiteXML スキーマ準拠確認

VIEW metadata table
- Design 済）テーブル設計
済）テストデータ投入

DMZ (inside Firewall)

Identifier and Event Table Design

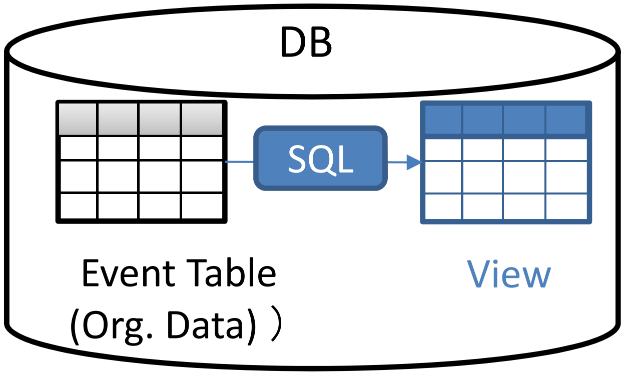
(doi):《project_prefix》/《DB_ID》/《Dataset_ID》

NameOfEventTable+Version ＋ ViewName+Versoin

NB: For a View integrating multiple event tables, multiple sets of "NameOfEventTable+Versoin" may be combined..

Versioning in Event Table

- Sensing data → Update is "append only" in principle
- No versioning for appending data
- Versioning transaction for change in past data, and change in schema
  - Recording transaction history
- Reproduce schema/data by rollback by specifying a version

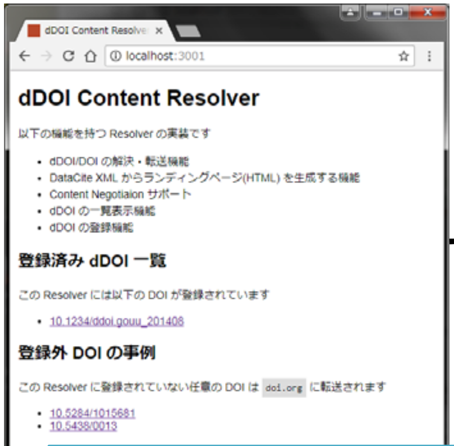


Event Data Ware House
(PostgreSQL/Greenplum)

Versioning of View

- Metadata on Landing Page is generated from View definition. SQL executed in response to data request
- New version is given when changing a SQL query to define a View
  - Recording the transaction history of updating views
- Reproduce schema/data by rollback by specifying a version

Web I/F transition



dDOI Content Resolver

localhost:3001

dDOI Content Resolver

以下の機能を持つ Resolver の実装です

- dDOI/DOI の解決・転送機能
- DataCite XML からランディングページ(HTML) を生成する機能
- Content Negotiaion サポート
- dDOI の一覧表示機能
- dDOI の登録機能

登録済み dDOI 一覧

この Resolver には以下の DOI が登録されています

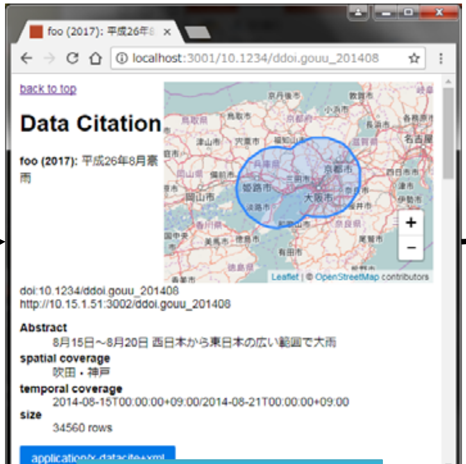
- 10.1234/ddoi.gouu_201408

登録外 DOI の事例

この Resolver に登録されていない任意の DOI は `doi.org` に転送されます

- 10.5284/1015681
- 10.5438/0013

ID Resolver Homepage



foo (2017): 平成26年8

localhost:3001/10.1234/ddoi.gouu_201408

back to top

Data Citation

foo (2017): 平成26年8月豪雨

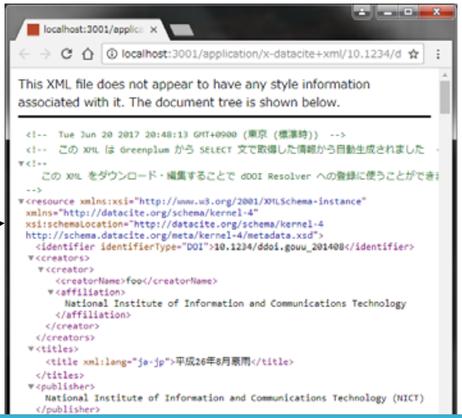
Leaflet | © OpenStreetMap contributors

doi:10.1234/ddoi.gouu_201408
http://10.15.1.51:3002/ddoi.gouu_201408

Abstract
8月15日～8月20日 西日本から東日本の広い範囲で大雨
spatial coverage
吹田・神戸
temporal coverage
2014-08-15T00:00:00+09:00/2014-08-21T00:00:00+09:00
size
34560 rows

application/x-datacite+xml

Landing Page



localhost:3001/applic

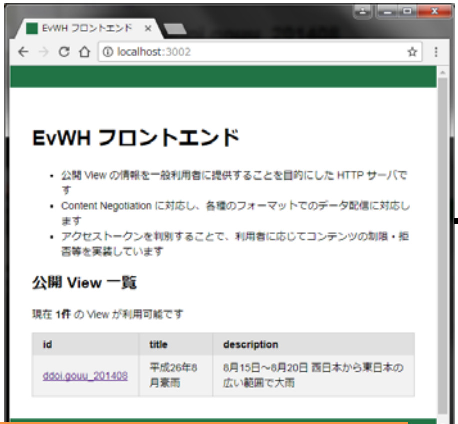
localhost:3001/application/x-datacite+xml/10.1234/d

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<!--  Tue Jun 20 2017 20:48:13 GMT+0900 (東京 (標準時))  -->
<!--  この XML は Greenplum から SELECT 文で取得した情報から自動生成されました
▼<!--
   この XML をダウンロード・編集することで dDOI Resolver への登録に使うことができ
-->
▼<resource xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns="http://datacite.org/schema/kernel-4"
 xsi:schemaLocation="http://datacite.org/schema/kernel-4
 http://schema.datacite.org/meta/kernel-4/metadata.xsd">
   <identifier identifierType="DOI">10.1234/ddoi.gouu_201408</identifier>
  ▼<creators>
    ▼<creator>
       <creatorName>foo</creatorName>
      ▼<affiliation>
         National Institute of Information and Communications Technology
       </affiliation>
     </creator>
   </creators>
  ▼<titles>
     <title xml:lang="ja-jp">平成26年8月豪雨</title>
   </titles>
  ▼<publisher>
     National Institute of Information and Communications Technology (NICT)
   </publisher>
```

Metadata：DataCiteXML

Data download



EvWH フロントエンド

localhost:3002

EvWH フロントエンド

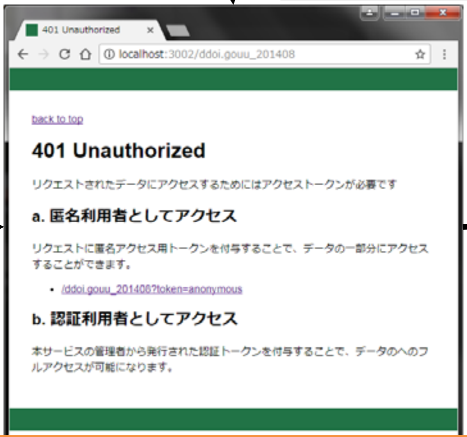
- 公開 View の情報を一般利用者に提供することを目的にした HTTP サーバです
- Content Negotiation に対応し、各種のフォーマットでのデータ配信に対応します
- アクセストークンを判別することで、利用者に応じてコンテンツの制限・拒否等を実装しています

公開 View 一覧

現在 1件 の View が利用可能です

id	title	description
ddoi.gouu_201408	平成26年8月豪雨	8月15日～8月20日 西日本から東日本の広い範囲で大雨

“list of views” service Web



401 Unauthorized

localhost:3002/ddoi.gouu_201408

back to top

401 Unauthorized

リクエストされたデータにアクセスするためにはアクセストークンが必要です

a. 匿名利用者としてアクセス

リクエストに匿名アクセス用トークンを付与することで、データの一部分にアクセスすることができます。

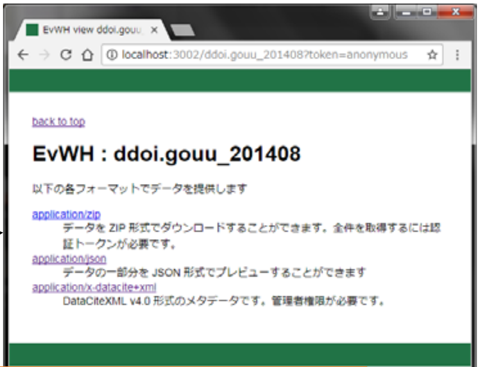
- /ddoi.gouu_201408?token=anonymous

b. 認証利用者としてアクセス

本サービスの管理者から発行された認証トークンを付与することで、データのへのフルアクセスが可能になります。

Individual VIEW (for public)

Token



EvWH view ddoi.gouu

localhost:3002/ddoi.gouu_201408?token=anonymous

back to top

EvWH : ddoi.gouu_201408

以下の各フォーマットでデータを提供します

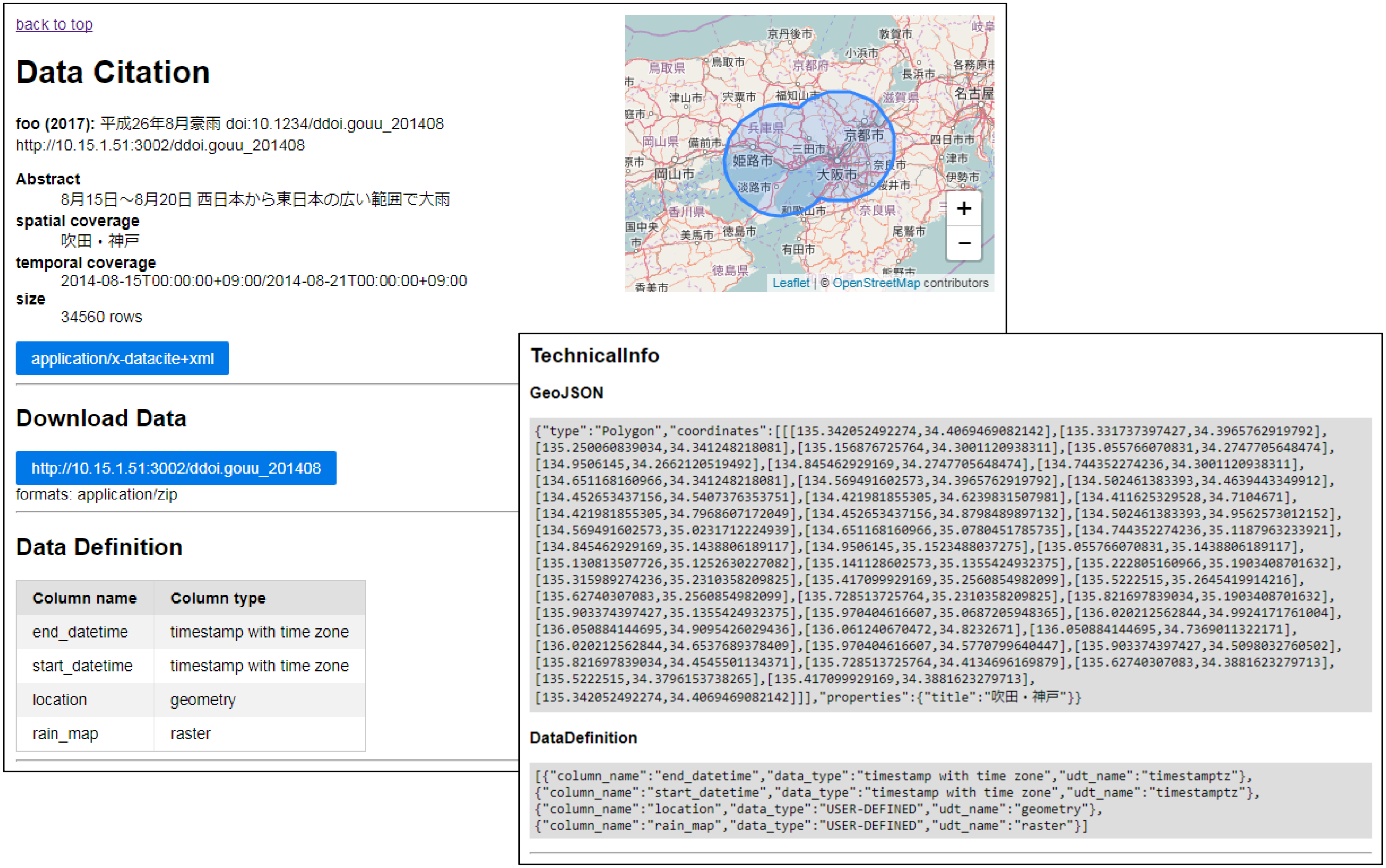
application/zip
データを ZIP 形式でダウンロードすることができます。全件を取得するには認証トークンが必要です。
application/json
データの一部分を JSON 形式でプレビューすることができます
application/x-datacite+xml
DataCiteXML v4.0 形式のメタデータです。管理者権限が必要です。

Individual VIEW
(for authenticated user)

Download
ZIP files

内部者向け
ここで編集用の DataCiteXML を
動的に生成してダウンロードすることができる

Landing Page Example



back to top

Data Citation

foo (2017): 平成26年8月豪雨 doi:10.1234/ddoi.gouu_201408
http://10.15.1.51:3002/ddoi.gouu_201408

Abstract
8月15日～8月20日 西日本から東日本の広い範囲で大雨

spatial coverage
吹田・神戸

temporal coverage
2014-08-15T00:00:00+09:00/2014-08-21T00:00:00+09:00

size
34560 rows

Download Data

formats: application/zip

Data Definition

Column name	Column type
end_datetime	timestamp with time zone
start_datetime	timestamp with time zone
location	geometry
rain_map	raster

TechnicalInfo

GeoJSON

```
{"type":"Polygon","coordinates":[[[135.342052492274,34.4069469082142],[135.331737397427,34.3965762919792],
[135.250060839034,34.341248218081],[135.156876725764,34.3001120938311],[135.055766070831,34.2747705648474],
[134.9506145,34.2662120519492],[134.845462929169,34.2747705648474],[134.744352274236,34.3001120938311],
[134.651168160966,34.341248218081],[134.569491602573,34.3965762919792],[134.502461383393,34.4639443349912],
[134.452653437156,34.5407376353751],[134.421981855305,34.6239831507981],[134.411625329528,34.7104671],
[134.421981855305,34.7968607172049],[134.452653437156,34.8798489897132],[134.502461383393,34.9562573012152],
[134.569491602573,35.0231712224939],[134.651168160966,35.0780451785735],[134.744352274236,35.1187963233921],
[134.84546292916 9,35.1438806189117],[134.9506145,35.1523488037275],[135.055766070831,35.1438806189117],
[135.130813507726,35.1252630227082],[135.141128602573,35.1355424932375],[135.222805160966,35.1903408701632],
[135.31598927423 6,35.2310358209825],[135.417099929169,35.2560854982099],[135.5222515,35.2645419914216],
[135.62740307083,35.2560854982099],[135.728513725764,35.2310358209825],[135.821697839034,35.1903408701632],
[135.903374397427,35.1355424932375],[135.970404616607,35.0687205948365],[136.02021256284 4,34.9924171761004],
[136.050884144695,34.9095426029436],[136.061240670472,34.8232671],[136.050884144695,34.7369011322171],
[136.020212562844,34.6537689378409],[135.970404616607,34.5770799640447],[135.903374397427,34.5098032760502],
[135.821697839034,34.4545501134371],[135.728513725764,34.4134696169879],[135.62740307083,34.3881623279713],
[135.5222515,34.3796153738265],[135.417099929169,34.3881623279713],
[135.342052492274,34.4069469082142]]],"properties":{"title":"吹田・神戸"}}
```

DataDefinition

```
[{"column_name":"end_datetime","data_type":"timestamp with time zone","udt_name":"timestamptz"},
{"column_name":"start_datetime","data_type":"timestamp with time zone","udt_name":"timestamptz"},
{"column_name":"location","data_type":"USER-DEFINED","udt_name":"geometry"},
{"column_name":"rain_map","data_type":"USER-DEFINED","udt_name":"raster"}]
```

Data Citation – Recommendations

Preparing Data & Query Store

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

When Resolving a PID

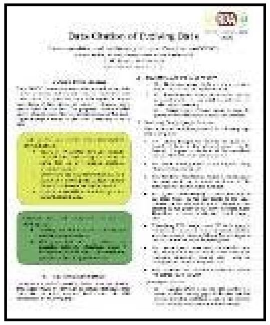
- R11 – Landing Page
- R12 – Machine Actionability

When Data should be persisted

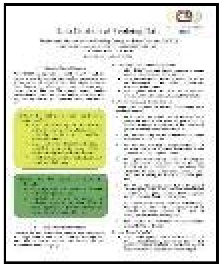
- R4 – Query Uniqueness
- R5 – Stable Sorting
- R6 – Result Set Verification
- R7 – Query Timestamping
- R8 – Query PID
- R9 – Store Query
- R10 – Citation Text

Upon Modifications to the Data Infrastructure

- R13 – Technology Migration
- R14 – Migration Verification



Data Citation – Recommendations



A) Preparing the Data and the Query Store

- **R1 – Data Versioning:** Apply versioning to ensure earlier states of data sets the data can be retrieved

Transaction history of Event table, rollback

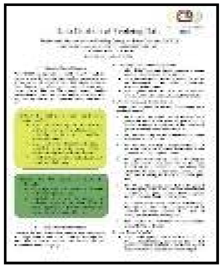
- **R2 – Timestamping:** Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp

Version control on Event table

- **R3 – Query Store:** Provide means to store the queries and metadata to re-execute them in the future

Using VIEW

Data Citation – Recommendations



B) Persistently Identify Specific Data sets (1/2)

When a data set should be persisted:

- **R4 – Query Uniqueness:** Re-write the query to a normalized form so that identical queries can be detected. Compute a checksum of the normalized query to efficiently detect identical queries.

Using VIEW

- **R5 – Stable Sorting:** Ensure an unambiguous sorting of the records in the data set

Time attribution of each event

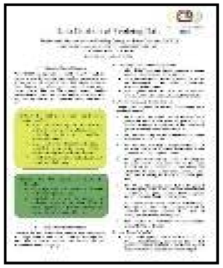
- **R6 – Result Set Verification:** Compute fixity information/checksum of the query result set to enable verification of the correctness of a result upon re-execution

→ Not secure for appended data

- **R7 – Query Timestamping:** Assign a timestamp to the query based on the last update to the entire database (or the last update to the selection of data affected by the query or the query execution time). This allows retrieving the data as it existed at query time

Version control of VIEW

Data Citation – Recommendations



B) Persistently Identify Specific Data sets (2/2)

When a data set should be persisted:

- **R8 – Query PID:** Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID

Target: the internal ID to DOI when published

- **R9 – Store Query:** Store query and metadata (e.g. PID, original and normalized query, query & result set checksum, timestamp, superset PID, data set description and other) in the query store

By using VIEW

- **R10 – Citation Text:** Provide citation text including the PID in the format prevalent in the designated community to lower barrier for citing data.

RDA

RESEARCH DATA ALLIANCE

UK National River Flow Archive

Matthew Fry

mfry@ceh.ac.uk

research data sharing without barriers

rd-alliance.org

The dataset

- RDBMS of river flow time series and metadata (~20M daily flow records), only 5-10GB
- Currently most edits are audited, but reconstruction is complex
- Many users downloading small subsets via an API – 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

Current progress

- Workflow defined for creating new versions at appropriate intervals, expect ~2 per year (based on our data update schedule)
- Database table definitions created
- Automation of processes to created versions
- Implementation of versioning throughout data access code, including API
- Working to further “clean” the structure prior to goi-live, as structural changes will be complex in future
- Still questions over checksum mechanism and how versions will be citable

Agenda

- 15:30 Introduction, Welcome
- 15:40 Short description of the WG recommendations
- 16:00 Report on new issues discussed / lessons learned
  - dual PIDs: dynamic dataset + subset
  - distributed settings
  - automatic generation of citation text
- 16:30 Brief reports on use cases
- 17:20 Other issues, next steps

Next Steps

- IG on Data Versioning, Citation Metadata
- Work with 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, …
- 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