

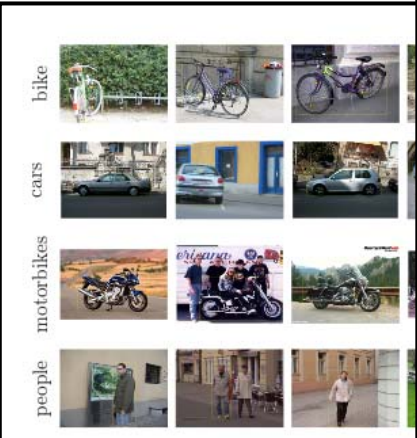


Enabling Precise Identification and Citability of Dynamic Data: Recommendations of the RDA Working Group WGDC

research data sharing without barriers
rd-alliance.org

-
- Recap: Challenges addressed by the WG
 - Recommendation of the RDA Working Group
 - Pilots and Adoption
 - Summary
-

- Citing data may seem easy
 - from providing a URL in a footnote
 - via providing a reference in the bibliography section
 - to assigning a PID (DOI, ARK, ...) to dataset in a repository
- What's the problem?



218 SYSTEMATIC BIOLOGY VOL. 62

SUPPLEMENTARY MATERIAL

Data files and/or other supplementary information related to this paper have been deposited on Dryad at <http://datadryad.org> under doi: 10.5061/dryad.6p76c3pb.

FUNDING

This work was supported by the National Science Foundation [DEB 1011206 to C.E.H.].

ACKNOWLEDGEMENTS

Useful discussions from WSU/UI PuRGe were invaluable, as was specific feedback from Matt Pennell, Luke Harmon, and Jeremiah Busch. Many thanks to Steve Orzell and Edwin Bridges of Avon Park, FL, for companionship and hospitality in the field that mistakenly went unobserved in an earlier paper.

REFERENCES

Aberer A.J., Krompass D., Stamatakis A. 2012. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic Biology* (in press) doi:10.1093/sysbio/sys078.

quantitative analysis of the problem of constructing an NLI we shows: we downloaded a dataset which has been frequently used of natural language interfaces, i.e. the Geobase dataset col- and his student¹. The Geobase dataset describes states, cities, rivers and roads in the U.S., together with attributes such as population (state, city), length (river), height (mountain, lo-

consists of a set of 880 test questions (actually 883 questions) through a web interface hosted at the University of Austin in the 883 test questions for our analysis. After downloading the g), we converted the whole dataset into the ontology languages OWL³. The datasets are available from <http://www.cimiano.de> datasets and other Material → ORAKEL.

Fig. 2. Image examples with ground truth object annotation for different categories of the PASCAL 2005 challenge. The dataset may be obtained from <http://www.pascal-network.org/challenges/VOC>.

¹ This dataset is available from: <http://www.cs.utexas.edu/users/ml/nldata.html>

² There is also a dataset consisting of 250 questions available from the University of Texas but this is merely a subset of the larger dataset.

³ <http://www.w3.org/TR/owl-features/>

Granularity of Data Identification / Citation

- What about the **granularity** of data to be identified/cited?
 - Databases collect enormous amounts of data over time
 - Researchers use specific subsets of data
 - Need to identify precisely the subset used
 - Current approaches
 - Citing entire dataset, providing textual description of subset
-> imprecise (ambiguity)
 - Storing a copy of subset as used in study -> scalability
 - 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 & cite precisely the **subset of (dynamic) data used** in a study

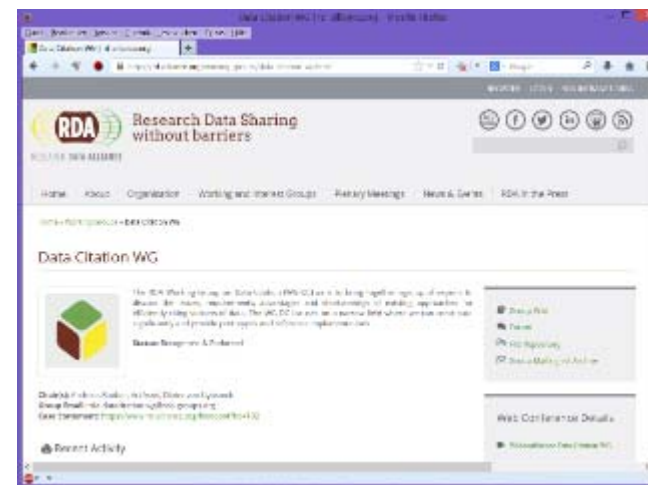
Citation of Dynamic Data

- Citable 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 cite precisely the **data as it existed at certain point in time**, without delaying release of new data

Data Citation – Requirements

- Dynamic data
 - corrections, additions, ...
- Arbitrary subsets of data (granularity)
 - rows/columns, time sequences, ...
 - from single number to the entire set
- Stable across technology changes
 - e.g. migration to new database
- Machine-actionable
 - not just machine-readable,
definitely not just human-readable and interpretable
- Scalable to very large / highly dynamic datasets
 - But: should also work for small and/or static datasets!

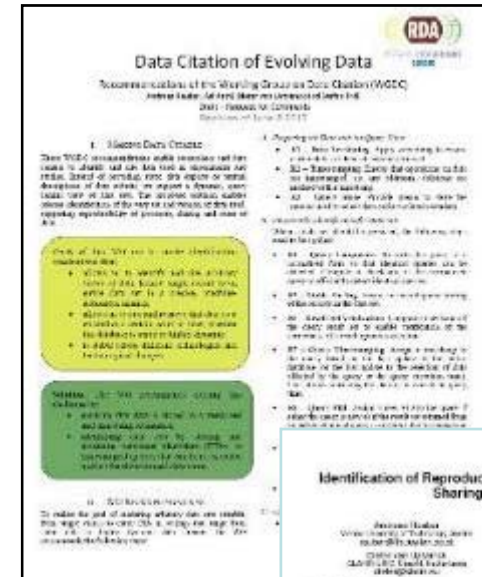
- Research Data Alliance
- WG on **Data Citation: Making Dynamic Data Citeable**
- March 2014 – Sep 2015
 - Concentrating on the problems of **large, dynamic (changing) datasets**
 - Focus! Identification of data!
Not: PID systems, metadata, citation string, attribution, ...
 - Liaise with other WGs and initiatives on data citation (CODATA, DataCite, Force11, ...)
 - Continuing support for adoption



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

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
- Technical Report: draft at https://rd-alliance.org/system/files/documents/RDA-Guidelines_TCDL_draft.pdf
- Reference implementations (SQL, CSV, XML) and Pilots



-
- Recap: Challenges addressed by the WG
 - Recommendation of the RDA Working Group
 - Pilots and Adoption
 - Summary
-

Data Citation: Data + Means-of-access

- Data → time-stamped & versioned (aka history)

Researcher creates working-set via some interface:

- Access → **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

S. Pröll, A. Rauber. **Scalable Data Citation in Dynamic Large Databases: Model and Reference Implementation.**

In IEEE Intl. Conf. on Big Data 2013 (IEEE BigData2013), 2013

http://www.ifs.tuwien.ac.at/~andi/publications/pdf/pro_ieeebigdata13.pdf

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

Data Citation – Deployment

- Note: query string provides excellent subset of data
- provenance information on the data set! 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 subset of data
- provenance information on the data set! user gets
 - Data (package)
 - PID (e.g. DOI)
 - Hash value
 - Recommended citation text (e.g. PID text)
- 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

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

Data Citation – Deployment

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

- Data (package)
- PID (e.g. DOI)
- Hash value
- Recommended citation text (e.g. PID text)

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

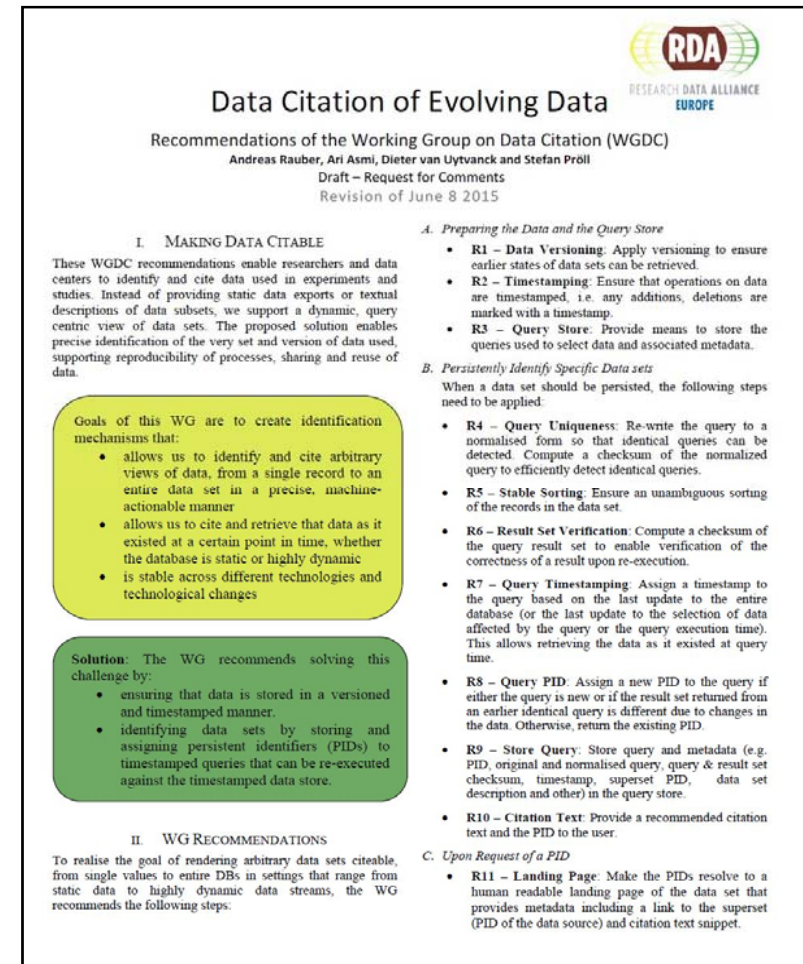
- PID resolves
 - Provides details
 - Option to retrieve
- Upon activating PID associated with a data citation
 - Query is re-executed against time-stamped and versioned DB
 - Results as above are returned
- Query store aggregates data usage

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



Data Citation – Recommendations

- 2-page flyer,
more extensive doc to follow
- **14 Recommendations**
- Grouped into **4 phases**:
 - Preparing data and query store
 - Persistently identifying specific data sets
 - Upon request of a PID
 - Upon modifications to the data infrastructure
- History
 - First presented March 30 2015
 - Major revision after workshop April 20/21
 - 4 workshops & presentations
 - 2 webinars (June 9, June 24)



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



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
- **R2 – Timestamping:** Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp
- **R3 – Query Store:** Provide means to store the queries and metadata to re-execute them in the future

Note:

- **R1 & R2** are already pretty much standard in many (RDBMS-) research databases
- Different ways to implement
- A bit more challenging for some data types (XML, LOD, ...)

A) Preparing the Data and

- **R1 – Data Versioning:** Apply versioning to ensure earlier states of data sets the data can be retrieved
- **R2 – Timestamping:** Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp
- **R3 – Query Store:** Provide means to store the queries and metadata to re-execute them in the future

Note:

- **R3:** query store usually pretty small, even for extremely high query volumes

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

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

When a data set should be persisted:

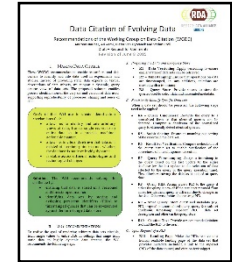
- **R8 – Query PID:** Assign a new PID to the query if either the query is new or if the result set returned from an earlier identical query is different due to changes in the data. Otherwise, return the existing PID
- **R9 – Store Query:** Store query and metadata (e.g. PID, original and normalized query, query & result set checksum, timestamp, superset PID, data set description and other) in the query store
- **R10 – Citation Text:** Provide citation text including the PID in the format prevalent in the designated community to lower barrier for citing data.

C) Resolving PIDs and Retrieving Data

- **R11 – Landing Page:** Make the PIDs resolve to a human readable landing page that provides the data (via query re-execution) and metadata, including a link to the superset (PID of the data source) and citation text snippet
- **R12 – Machine Actionability:** Provide an API / machine actionable landing page to access metadata and data via query re-execution

D) Upon Modifications to the Data Infrastructure

- **R13 – Technology Migration:** When data is migrated to a new representation (e.g. new database system, a new schema or a completely different technology), migrate also the queries and associated checksums
- **R14 – Migration Verification:** Verify successful data and query migration, ensuring that queries can be re-executed correctly



D) Upon Modifications to the Data Infrastructure

- **R13 – Technology Migration:** When data is migrated to a new representation (e.g. new database system, a new schema or a completely different technology), migrate also the queries and associated checksums
- **R14 – Migration Verification:** Verify successful query migration should, ensuring that queries can be re-executed correctly

Benefits

- Retrieval of precise subset with low storage overhead
- Subset as cited or as it is now (including e.g. corrections)
- Query provides provenance information
- Query store supports analysis of data usage
- Checksums support verification
- Same principles applicable across all settings
 - Small and large data
 - Static and dynamic data
 - Different data representations (RDBMS, CSV, XML, LOD, ...)
- Would work also for more sophisticated/general transformations on data beyond select/project

-
- Recap: Challenges addressed by the WG
 - Recommendation of the RDA Working Group
 - Pilots and Adoption
 - Summary
-

WG Pilots

Name	Data	Type	Status	Notes
Timbus	RDBMS	research	finished	Sensor data, pilot
XML-Reference	XML	research	finished	eXist-DB
DEXHELPP	CSV/RDBMS	research	running	Social security data
CSV-Reference	CSV/RDBMS	reference	running - β	Reference implem.
GIT-Reference	<ASCII>	reference	running - α	Reference implem.
VAMDC	SQL/NoSQL/ ASCII -> XML	deployment	running	Distributed data center
CBMI@wustl	RDBMS	deployment	starting	integration into i2b2
CCCA	NetCDF	deployment	starting	climate data
ENVRIplus		deployment	starting	ICOS: Carbon Obs.Infr.
ARGO	NetCDF	deployment	starting	ODIP-II, RDA-Europe
BCO-DMO	CSV	deployment	starting	RDA-US
VMC (Vermont)	VMC data cat.	deployment	starting	Forest Research Data
<a few others>	CSV, RDBMS	deployment	planned	Conceptual evaluation, seeking funding



First Pilots for SQL Data

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

research data sharing without barriers
rd-alliance.org

- LNEC Laboratory of Civil Engineering, Portugal
- Monitoring dams and bridges
- 31 manual sensor instruments
- 25 automatic sensor instruments
- Web portal
 - Select sensor data
 - Define timespans
- Report generation
 - Analysis processes
 - LaTeX
 - publish PDF report



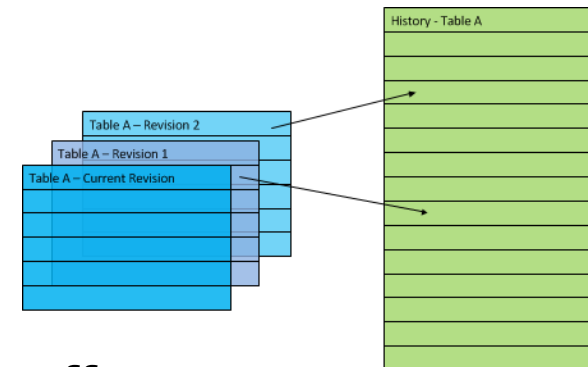
Florian Fuchs [CC-BY-3.0 (<http://creativecommons.org/licenses/by/3.0>)], via Wikimedia Commons



- Million Song Dataset
<http://labrosa.ee.columbia.edu/millionsong/>
- Largest benchmark collection in Music Retrieval
- Original set provided by Echonest
- No audio, only several sets of features
(16 – 1440 measurements/features per song)
- Harvested, additional features and metadata
extracted and offered by several groups
e.g. <http://www.ifs.tuwien.ac.at/mir/msd/download.html>
- Dynamics because of metadata errors, extraction errors
- Research groups select subsets by genre, audio length,
audio quality,...

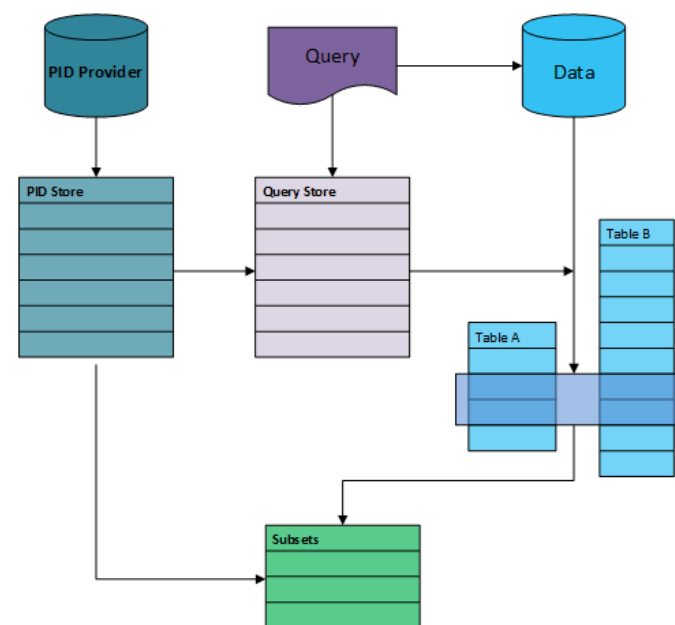
SQL Time-Stamping and Versioning

- Integrated
 - Extend original tables by temporal metadata
 - Expand primary key by record-version column
- Hybrid
 - Utilize history table for deleted record versions with metadata
 - Original table reflects latest version only
- Separated
 - Utilizes full history table
 - Also inserts reflected in history table
- Solution to be adopted depends on trade-off
 - Storage Demand
 - Query Complexity
 - Software adaption



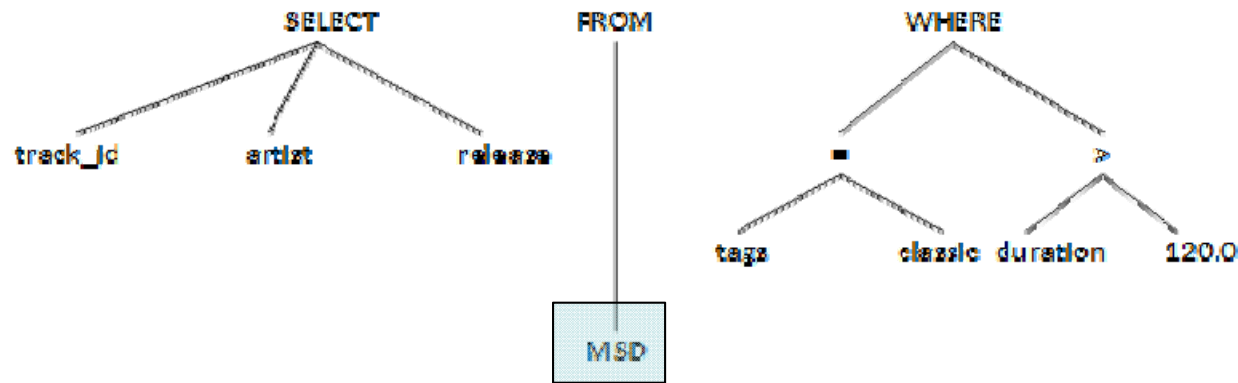
SQL: Storing Queries

- Add query store containing
 - PID of the query
 - Original query
 - Re-written query + query string hash
 - Timestamp
(as used in re-written query)
 - Hash-key of query result
 - Metadata useful for citation /
landing page
(creator, institution, rights, ...)
 - PID of parent dataset
(or using fragment identifiers for query)



SQL Query Re-Writing

- Adapt query to history table



```

SELECT results.track_id, results.artist, results.release
FROM MSD AS results JOIN (
    SELECT track_id, max(timestamp) AS latestTimestamp
    FROM MSD
    WHERE timestamp <= (SELECT @queryExecutionTimestamp)
    AND (track_id NOT IN
        (SELECT track_id FROM MSD AS deletedRecords
         WHERE deletedRecords.status_mark = 'deleted'
         AND (deletedRecords.timestamp < @queryExecutionTimestamp))
    )
    GROUP BY track_id
) AS version ON results.track_id = version.track_id AND results.timestamp = version.latestTimestamp

WHERE
    results.tags = 'classic' AND results.duration > 120
ORDER BY results.track_id;
  
```



Reference Implementation for CSV Data

Stefan Pröll
sproell@sba-research.org

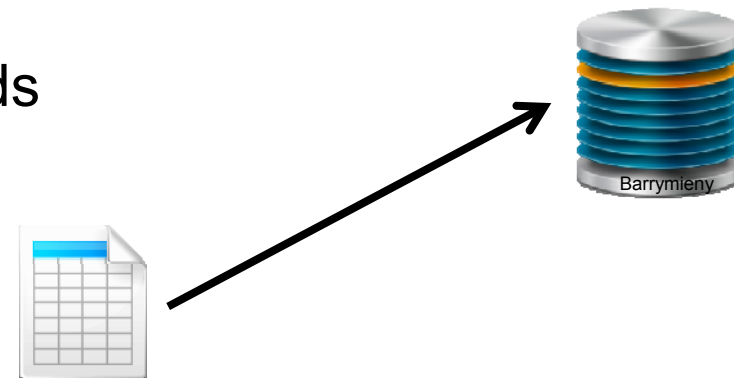
research data sharing without barriers
rd-alliance.org

Dynamic Data Citation for CSV Data

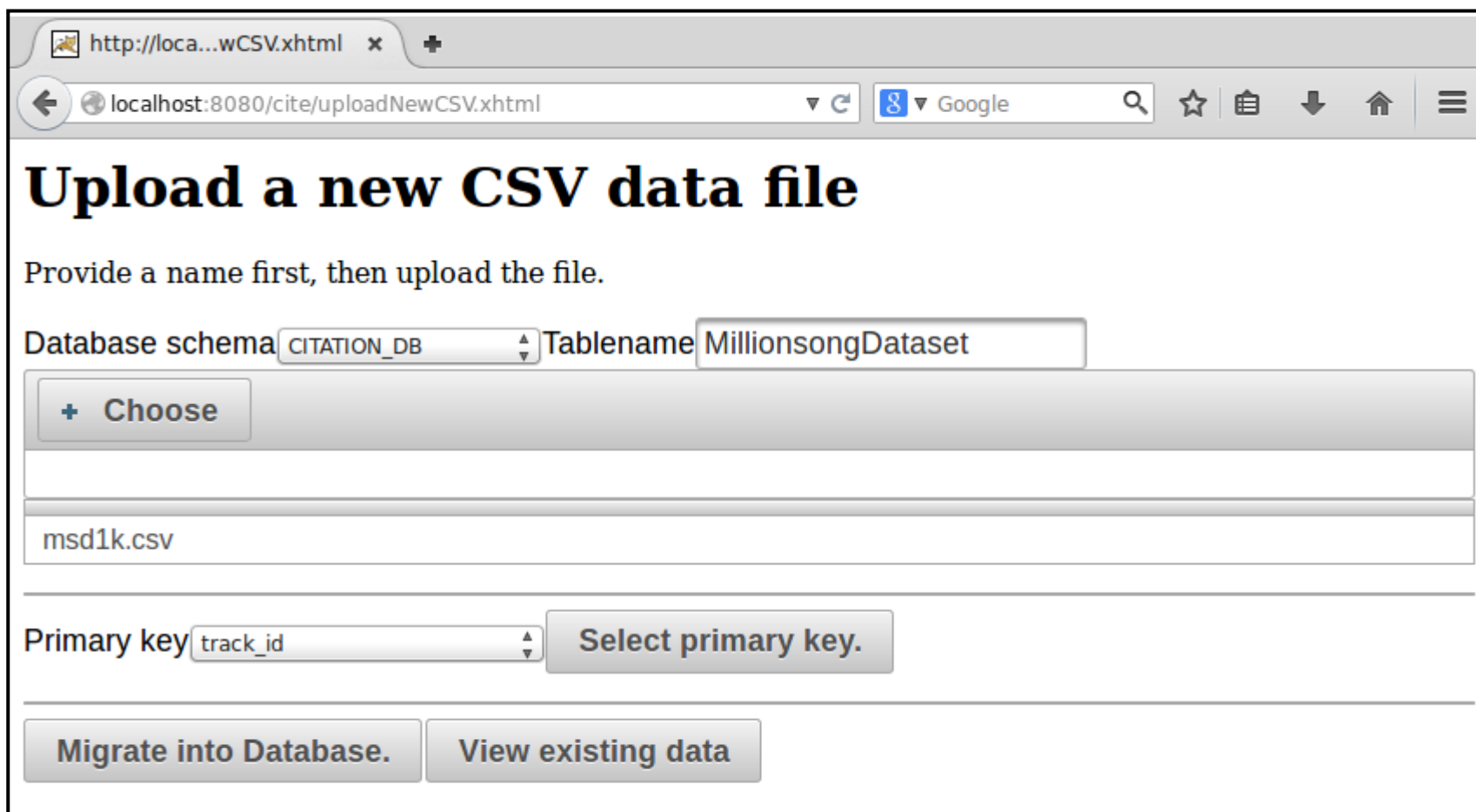
- Why CSV data? (not large, not very dynamic...)
 - Well understood and widely used
 - Simple and flexible
 - Most frequently requested during initial RDA meetings
- Goals:
 - Ensure cite-ability of CSV data
 - Enable subset citation
 - Support particularly small and large volume data
 - Support dynamically changing data
- 2 Options:
 - Versioning system (subversion/svn, git, ...)
 - Migration to RDBMS

CSV Prototype: Basic Steps

- Upload interface for CSV files
- 2 approaches:
 - Migrate CSV file into RDBMS
 - Generate table structure, identify primary key
 - Add metadata columns for versioning, indices
 - Use GIT for data and separate branch for queries
- Dynamic data
 - Update / delete existing records
 - Append new data
- Access interface
 - Track subset creation
 - Store queries



CSV Data Prototype



http://loca...wCSV.xhtml x +

localhost:8080/cite/uploadNewCSV.xhtml

Google

Upload a new CSV data file

Provide a name first, then upload the file.

Database schema CITATION_DB Tablename MillionsongDataset

+ Choose

msd1k.csv

Primary key track_id Select primary key.

Migrate into Database. View existing data

CSV Data Prototype

Data Citation Tool

localhost:8080/cite/table.xhtml

CITATION_DB ▼ MSD500k ▼ Load table

Show 10 entries Search:

duration	artist_familiarity	artist_hottness	year	digitalid	audiofile	lastfm	numlastfm	numlastfmmatched	rpfeatures	audiofilelength	si
342.72608	0.769265597035	0.517557658726	0	6768683	1	0	0	(Data n/a)	1	(
Amadeus Mozart											
113.68444	0.769265597035	0.517557658726	0	9032098	1	0	0	(Data n/a)	1	(
Wolfgang Amadeus Mozart											
313.52118	0.7692										
Wolfgang Amadeus Mozart											
116.61016	0.7692										
Wolfgang Amadeus Mozart											
169.87383	0.7692										
Wolfgang Amadeus Mozart											
257.64526	0.7692										
Wolfgang Amadeus Mozart											
477.6224	0.7692										
Wolfgang Amadeus Mozart											
273.00526	0.7692										
Wolfgang Amadeus Mozart											
211.3824	0.7692										
Wolfgang Amadeus Mozart											
135.13098	0.769265597035	0.517557658726	0	4799936	1	0	0	(Data n/a)	1	(
Amadeus Mozart / Otto Sieben											

Showing 1 to 10 of 500,000 entries

First Previous 1 2 3 4 5 Next Last

Initialize query store Store current selection Finalize dataset

Showing 1 to 10 of 500,000 entries

First Previous 1 2 3 4 5 Next Last

Initialize query store Store current selection Finalize dataset

Warten auf localhost...

CSV Data Prototype

Suggested citation text:

Stefan Pröll (2015) "jj test" created at 2015-02-19 11:33:54.0, PID [ark:12345/5l86eH4qMX].
Subset of Stefan Pröll: "Adresses", PID [ark:12345/Ojfl4gUmFo]

Download area

Download CSV Subset

↓ Download

Download the CSV data of this subset at the execution time of the query

Download Latest Subset

↓ Download

Download the CSV data of this subset at its current state

Download Full DB

↓ Download

Download the full database as CSV file

Download Diff CSV file

↓ Download

Download the differences as CSV between the subset at its original execution time and now.

CSV Data Prototype

SQL string

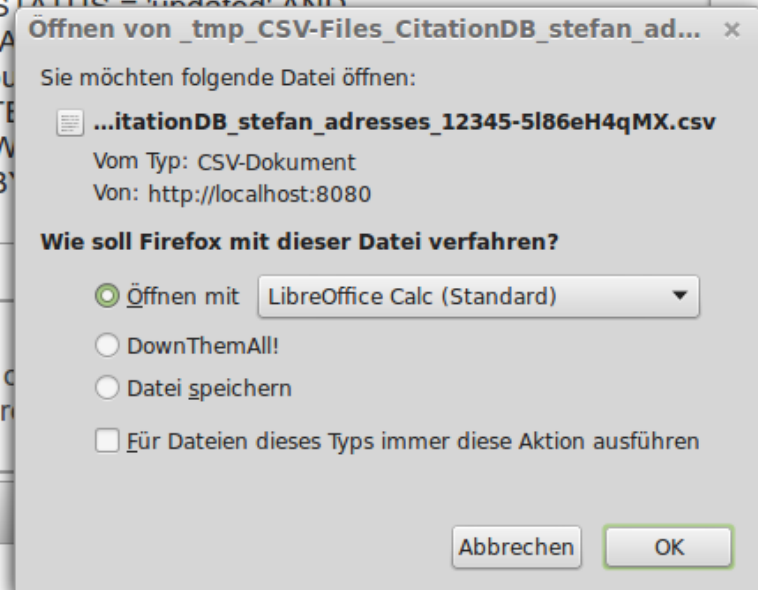
```
(innerSELECT.RECORD_STATUS = 'inserted' OR  
innerSELECT.RECORD_STATUS = 'updated' AND  
innerSELECT.LAST_UPDATE  
LAST_UPDATE) innerGroup  
innerGroup.LAST_UPDATE  
innerGroup.mostRecent W  
UPPER('%jj%') ORDER B
```

Suggested citation text:

Stefan Pröll (2015) "jj test" c
Subset of Stefan Pröll: "Adre
eH4qMX].

Download area

Download CSV Subset	↓ Download	Download the CSV data of this subset at the execution time of the query
Download Latest Subset	↓ Download	Download the CSV data of this subset at its current state
Download Full DB	↓ Download	Download the full database as CSV file
Download Diff CSV file	↓ Download	Download the differences as CSV between the subset at its original execution time and now.



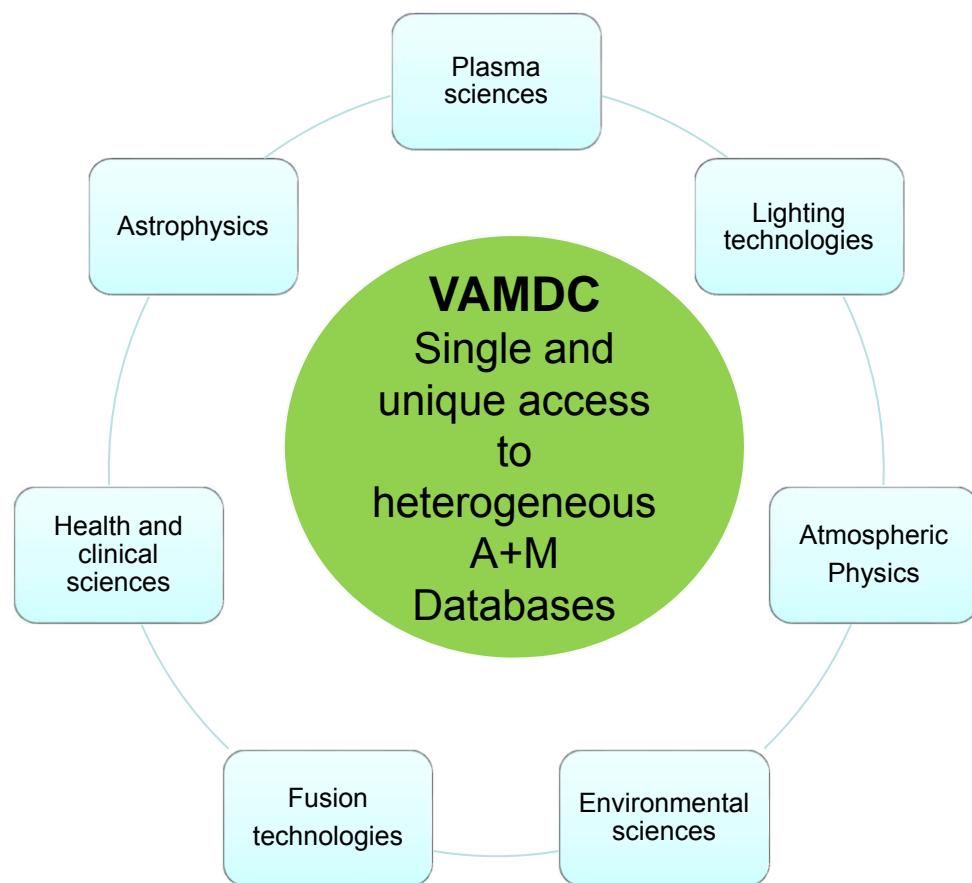


Progress on Data Citation within VAMDC

C.M. Zwölf and VAMDC Consortium
carlo-maria.zwolf@obspm.fr

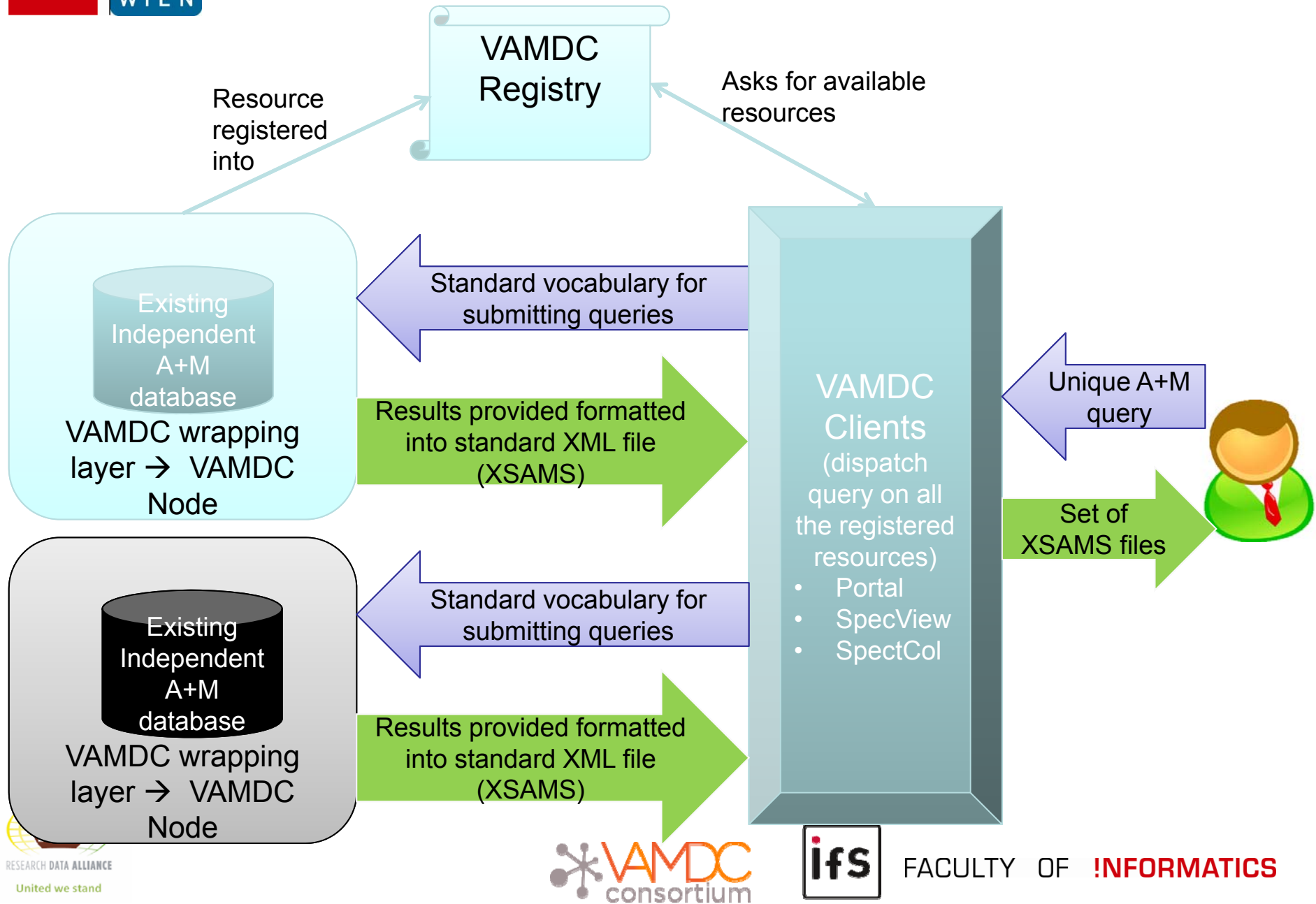
research data sharing without barriers
rd-alliance.org

Virtual Atomic and Molecular Data Centre

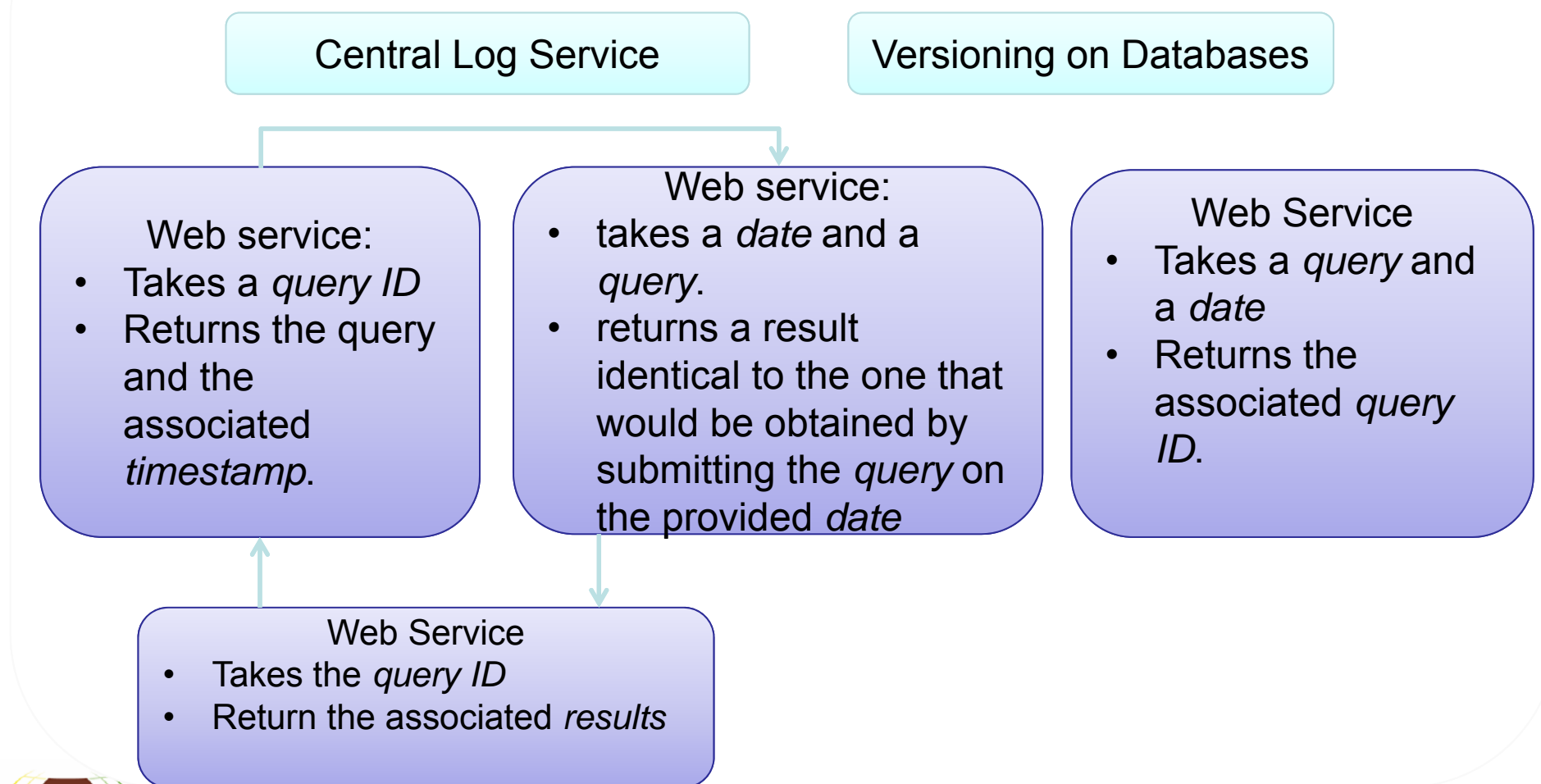


- Federates 28 heterogeneous databases
<http://portal.vamdc.org/>
- Distributed infrastructure with no central management system
- The “V” of VAMDC stands for Virtual in the sense that the e-infrastructure does not contain data. The infrastructure is a wrapping for exposing in a unified way a set of heterogeneous databases.
- Relies on a strong and sustainable technical and political organisation.

VAMDC Infrastructure



Architecture of the query store



VAMDC - Implementation

Query Store

Tagging Datasets with Ids
(Relational Database case)

- On January 2016 we submitted a paper for a special issue of the *Journal of Molecular Spectroscopy* dedicated to *New visions in Spectroscopic Databases*
 - *New paradigm for datasets citation and extraction reproducibility in VAMDC, C.M. Zwölf, N. Moreau and M.-L. Dubernet.*
- We describe in this paper
 - How the internal VAMDC standards evolve for achieving data-tagging and versioning
 - How a query store will be integrated within the existing VAMDC infrastructure.

VAMDC - Implementation

Query Store

Tagging Datasets with Ids
(Relational Database case)

- The design of the query store is almost finished.
- The development will start on spring 2016, in joint collaboration with the RDA Europe project.
- We have started collaborations with editors for the integration of the query store into their digital platforms and journal repositories.



WG Data Citation Pilot CBMI @ WUSTL

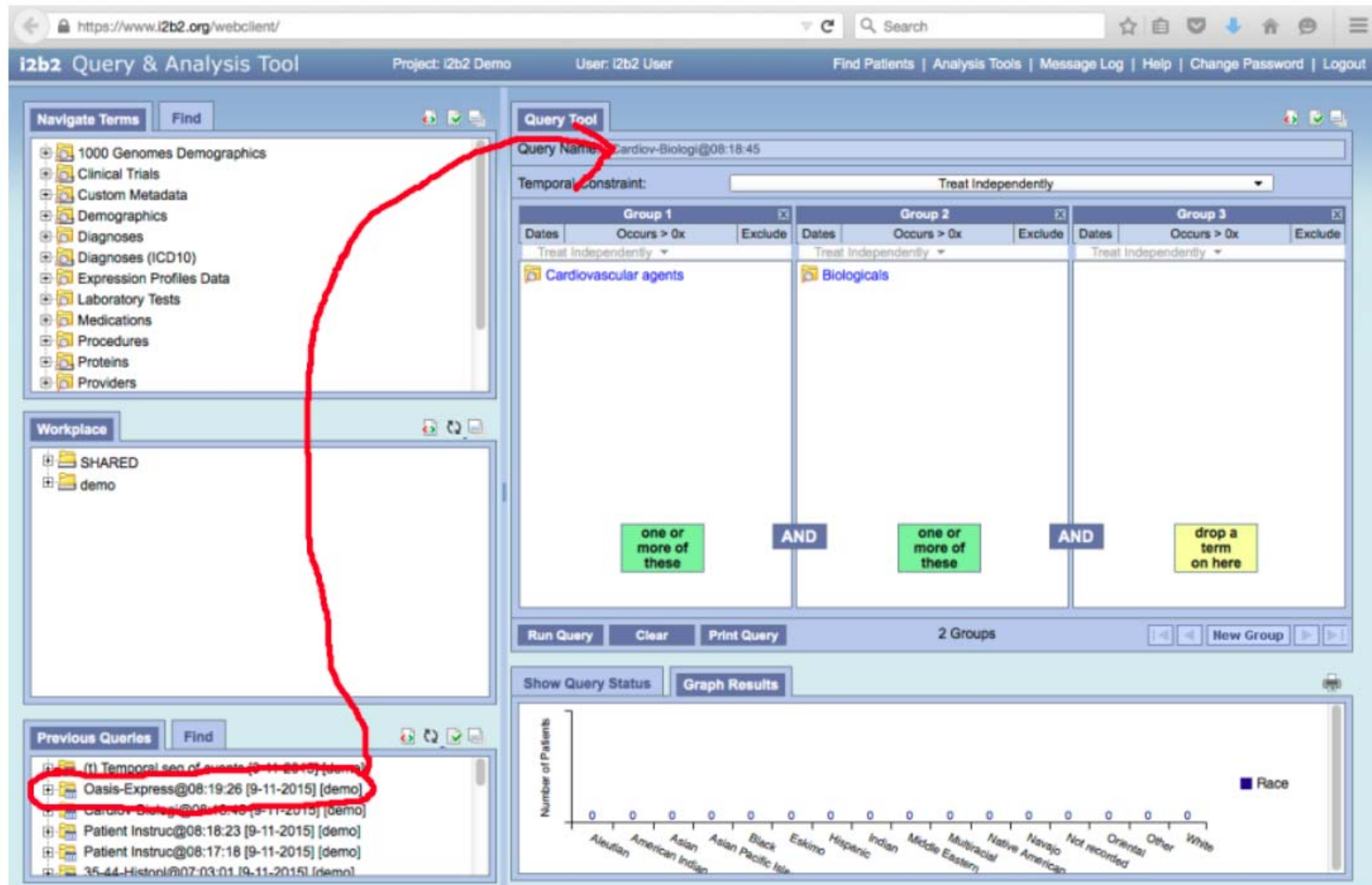
**Leslie McIntosh, Cynthia Hudson Vitale,
Snehil Gupta
Washington University in St.Luis**

research data sharing without barriers
rd-alliance.org



CBMI @ WUSTL

- Center for Biomedical Informatics, Washington University in St. Louis
- Electronic medical health record aggregator **i2b2** (Informatics for Integrating Biology and the Bedside) NIH-funded Health Care System, OS SW
- Electronic patient medical records (EMR)
- i2b2 instance with de-identified data from local hospitals and outpatient clinics
 - Overall approx. 2 billion records
 - 4 mio patients, 48 mio encounters, 82 mio medications, 674 mio lab results, 385 mio vital sign data, ...
- Obtained funding to implement WGDC recommendations
- Timeframe: 9 months



The screenshot displays the i2b2 Query & Analysis Tool interface. The left sidebar contains a 'Navigate Terms' section with a tree view of data categories. A red line highlights the 'Diagnoses' category. The main area is titled 'Query Tool' and shows a query named 'Cardio-Biologi@08:18:45'. It features three groups for building a query: Group 1 (Cardiovascular agents), Group 2 (Biologicals), and Group 3 (empty). The query is built using 'AND' operators and 'one or more of these' constraints. The 'Temporal constraint' is set to 'Treat Independently'. At the bottom, there is a 'Show Query Status' section with a 'Graph Results' tab showing a bar chart of 'Number of Patients' by 'Race'.

Goals of this project are to:

- 1. Integrate the RDA recommendations for Data Citation of Evolving Data into the WU CBMI instance of i2b2** by modifying the open source code to account for the recommendations at the query or project level.
- 2. Contribute all source code back to the i2b2 open source community via GitHub** for wider adoption and implementation among all i2b2 users (currently 50 translational science centers, 34 academic health centers, 2 HMO's, 20 international organizations, 4 companies).
- 3. Gather feedback about RDA WGDC-compliant i2b2 code from established i2b2 installations;** obtain projected timelines of compliant code integration, barriers to implementation



Data Citation for ENVRIplus

Ari Asmi
ari.asmi@helsinki.fi

research data sharing without barriers
rd-alliance.org

ENVRI Plus – ICOS Data Citation

- Part of ENVRI PLUS data citation Workpackage
- ICOS – Integrated Carbon Observation System (infrastructure)

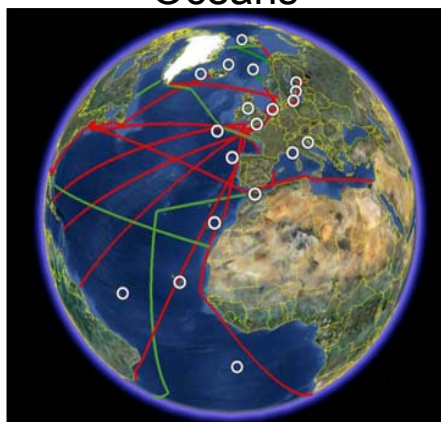
Atmosphere



Ecosystems



Oceans

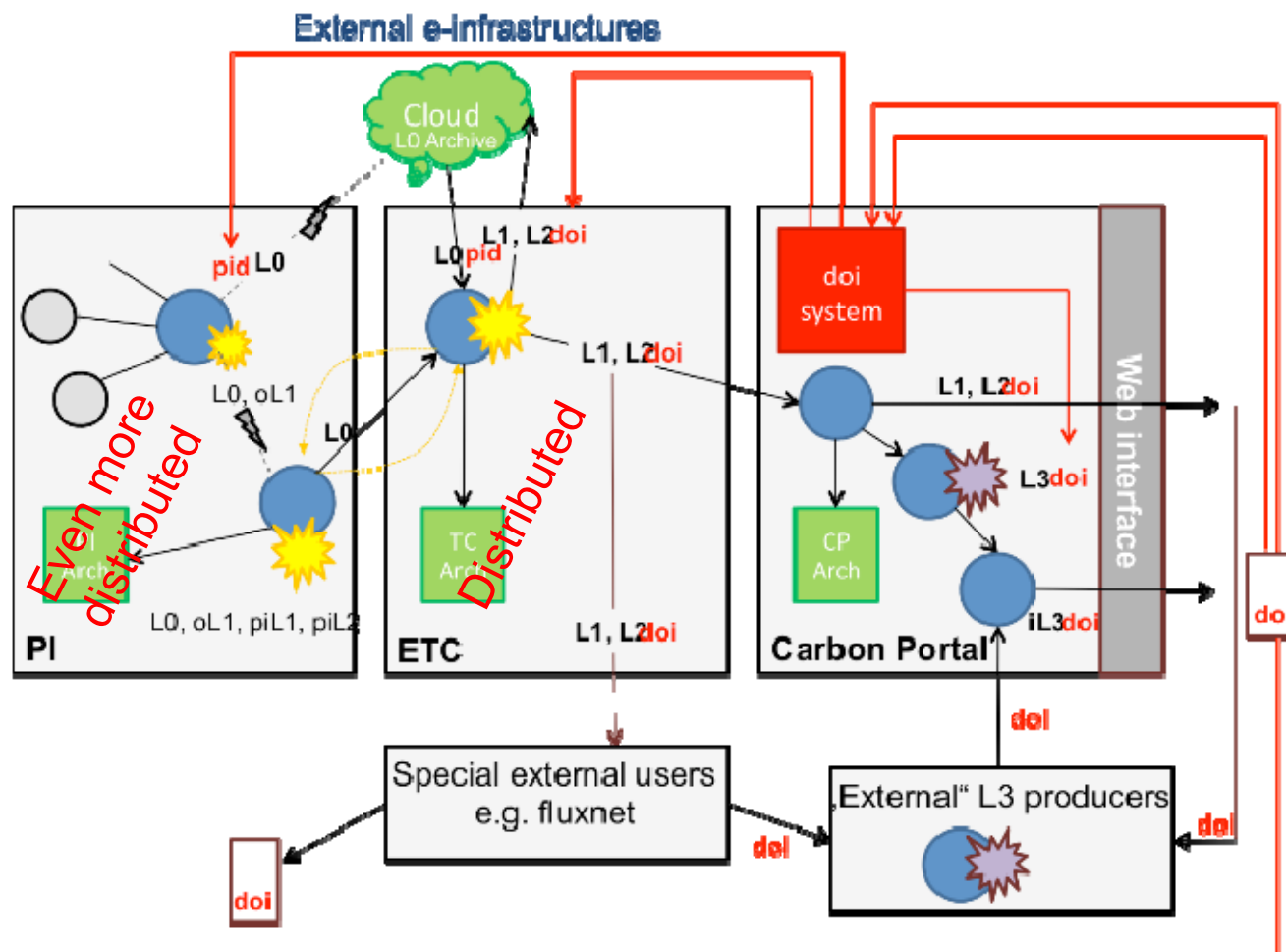


- Distributed data production
- Distributed data storage
 - Centralized “high level” data sets
- Updated daily
- Usage wide in Carbon observation science
 - Some NRT
 - Some “high level” data storage

ENVRIplus ICOS Implementation (in progress)

Versioning DB – OK!

Distributed Data delivery – Potential issue (if users bypass web interface)





Data Citation for ARGO (ODIP II Project)

Helen Glaves

hmg@bgs.ac.uk

research data sharing without barriers
rd-alliance.org

Argo data use case

- Aims & objectives
 - Resolve the ambiguity in the syntax for citation of dynamic data
 - Agree and ratify a common syntax for dynamic data citation
 - Publish results in authoritative documentation e.g. DataCite metadata schema
 - Implement dynamic data citation for Argo data



Application scenario

- Argo data held by several international data centres
 - IFREMER
 - NCEI (formerly NOAA National Climatic Data Center, the National Geophysical Data Center, and the National Oceanographic Data Center)
 - BODC
- Validation of method using a real world exemplar
- Results reported to RDA via DCWG and MDH IG
- Feed into related activities in ODIP, ENVRIplus, EUDAT etc.



Adoption of Data Citation Outcomes by BCO-DMO, R2R

Cynthia Chandler, Adam Shepherd

research data sharing without barriers
rd-alliance.org

- BCO-DMO
 - Biological and Chemical Oceanography Data Management Office (WHOI)
 - Curation of marine ecosystem system data contributed by NSF funded investigators
- R2R
 - Rolling Deck to Repository
 - Curation of routine, underway data from US academic fleet, and authoritative expedition catalog
- Members of Marine Data Harmonization IG

- **Evaluation**
 - Evaluate recommendations
 - Try implementation in existing systems
- **Trial**
 - BCO-DMO: R1-11 fit well with current architecture; R12 doable; test as part of DataONE node membership
 - R2R: curation of original field data and selected subset of post-field products (ship track); so no evolving data

RDA Recommendations for Data Citation of Evolving Data

R1 Data Versioning
R2 Time Stamping
R3 Query Store
R4 Query Verification
R5 Stable Sorting
R6 Result Set Verification
R7 Query Time Stamping
R8 Query PID
R9 Citation Text
R10 Landing Page
R11 Machine Actionability
R12 Technology Migration
R13 Migration Verification

CARIACO zooplankton data subset, since 2000

/OCB/CARIACO/Zooplankton.html?Date>20000101,
Cruise_ID,lon,lat,Date,zoop_DW_200,zoop_ash_200,
zoop_DW_500,zoop_ash_500

[Directory](#)[Documentation](#)[Download & Other Operations](#)[Level 0](#)[Next Level](#)[Flat Listing](#)

```
# version 08 July 2009
#
# PI: Jaimie Rojas (EDIMAR/FLASA)
#
# CARIACO Zooplankton biomass data
#
```

```
=====
Cruise_ID  lon    lat    Date      zoop_DW_200  zoop_ash_200  zoop_DW_500  zoop_ash_500
-----
93HG_071   -64.67  10.5   20011009   26.83        6.84          15.95         3.31
93HG_072   -64.67  10.5   20011106   7.43         0.64          2.34          0.07
93HG_073   -64.67  10.5   20011211   5.60         0.14          0.71          0.18
93HG_074   -64.67  10.5   20020110   10.24        0.51          4.93          0.18
93HG_075   -64.67  10.5   20020214   12.92        2.56          nd            nd
93HG_076   -64.67  10.5   20020312   10.55        1.78          5.47          0.77
93HG_077   -64.67  10.5   20020402   8.71         1.82          8.31          2.15
93HG_078   -64.67  10.5   20020507   4.50         0.43          1.31          0.09
93HG_079   -64.67  10.5   20020612   10.63        1.07          4.54          0.35
```

BCO-DMO - New capabilities

- Preserve the data subset
- Request a DOI
- Store data subset, query, and create new landing page for data subset DOI

WG Pilots

Name	Data	Type	Status	Notes
Timbus	RDBMS	research	finished	Sensor data, pilot
XML-Reference	XML	research	finished	eXist-DB
DEXHELPP	CSV/RDBMS	research	running	Social security data
CSV-Reference	CSV/RDBMS	reference	running - β	Reference implem.
GIT-Reference	<ASCII>	reference	running - α	Reference implem.
VAMDC	SQL/NoSQL/ ASCII -> XML	deployment	running	Distributed data center
CBMI@wustl	RDBMS	deployment	starting	integration into i2b2
CCCA	NetCDF	deployment	starting	climate data
ENVRIplus		deployment	starting	ICOS: Carbon Obs.Infr.
ARGO	NetCDF	deployment	starting	ODIP-II, RDA-Europe
BCO-DMO	CSV	deployment	starting	RDA-US
VMC (Vermont)	VMC data cat.	deployment	starting	Forest Research Data
<a few others>	CSV, RDBMS	deployment	planned	Conceptual evaluation, seeking funding

-
- Recap: Challenges addressed by the WG
 - Recommendation of the RDA Working Group
 - Pilots and Adoption
 - Summary
-

Benefits

- Retrieval of precise subset with low storage overhead
- Subset as cited or as it is now (including e.g. corrections)
- Query provides provenance information
- Query store supports analysis of data usage
- Checksums support verification
- Same principles applicable across all settings
 - Small and large data
 - Static and dynamic data
 - Different data representations (RDBMS, CSV, XML, LOD, ...)
- Would work also for more sophisticated/general transformations on data beyond select/project

Adoption Activities

- 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, ...
- How could we organize this?
 - New roles to join the WG coordinators
 - Ambassadors
 - Domain-specific contact points
 - Others? Links to supporting infrastructure (PIDs, ...)
 - Collection of on-going pilots
 - Structuring: contact, descriptions, results, lessons learned
 - Outcomes: reports, slides, publications, code, discussions

Adoption Activities

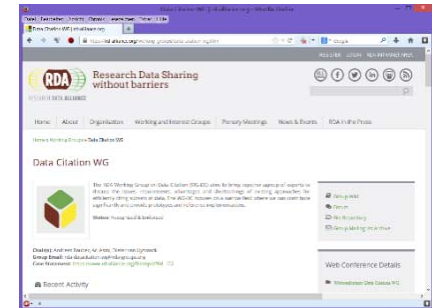
- Joint projects:
 - Individual funding per data center
 - RDA call for Collaboration Projects (EU, US)
 - H2020 call: *EINFRA-21-2017: Platform driven e-infrastructure innovation*” (and particularly the item *1-Universal discoverability of data objects and provenance*)
 - Others?

- Other needs & means of support?
 - No new WG at the moment, adoption phase, lessons to learn
 - Any crucial open topics?
 - Citation texts, attribution, ...
 - New WG if new issues arise leading to revision of recommendations or addressing new topics

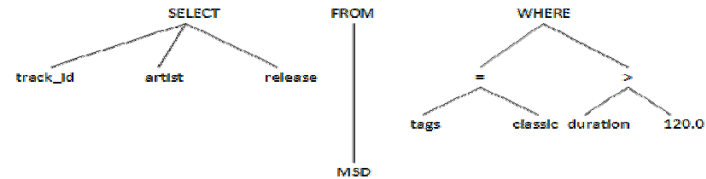
Join RDA and Working Group WGDC

If you are interested in joining the discussion, contributing a pilot, wish to establish a data citation solution, ...

- Register for the RDA WG on Data Citation:
 - Website:
<https://rd-alliance.org/working-groups/data-citation-wg.html>
 - Mailinglist:
<https://rd-alliance.org/node/141/archive-post-mailinglist>
- Contact us if you plan to implement the recommendations
- Let us know your feedback, concerns, issues identified, ...



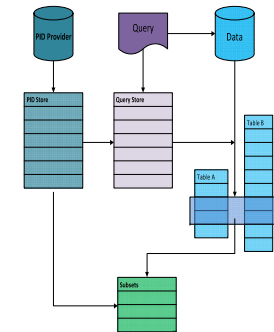
Thank you!



```

SELECT results.track id, results.artist, results.release
FROM MSD AS results JOIN (
  SELECT track id, max(timestamp) AS latestTimestamp
FROM MSD
WHERE timestamp <= (SELECT @queryExecutionTimestamp)
AND (track id NOT IN
  (SELECT track id FROM MSD AS deletedRecords
  WHERE deletedRecords.status mark = 'deleted'
  AND (deletedRecords.timestamp < @queryExecutionTimestamp))
) AS version ON results.track id = version.track id AND results.timestamp = version.latestTimestamp
WHERE
  results.tags = 'classic' AND results.duration > 120
ORDER BY results.track_id;

```



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

