Approaches to Making Dynamic Data Citeable

Recommendations of the RDA Working Group

Andreas Rauber

Vienna University of Technology
rauber@ifs.tuwien.ac.at
http://www.ifs.tuwien.ac.at/~andi

RESEARCH DATA ALLIANCE
United we stand

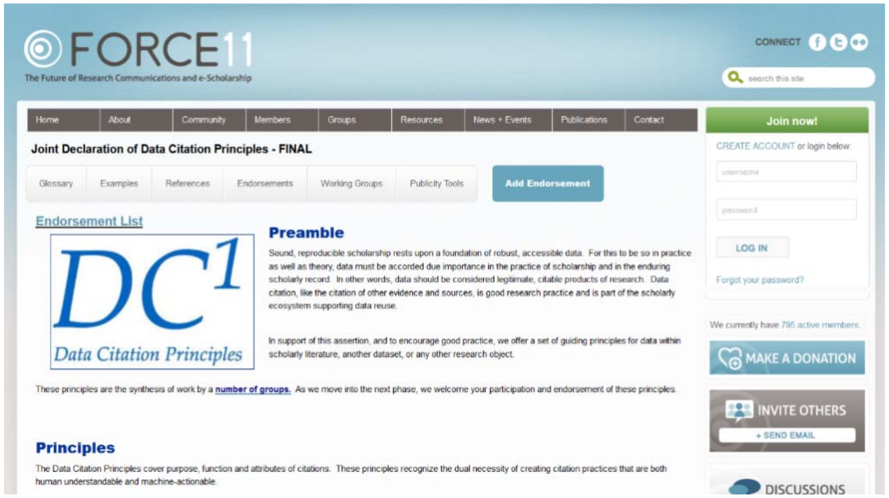
FACULTY OF !NFORMATICS

Outline

- Joint Declaration of Data Citation Principles
- Challenges in non-trivial settings
- Recommendation of the RDA Working Group
- Pilots
- Summary

Joint Declaration of Data Citation Principles

- 8 Principles created by the Data Citation Synthesis Group
- https://www.force11.org/datacitation
- The Data Citation Principles cover purpose, function and attributes of citations
- Goal: Encourage communities to develop practices and tools that embody uniform data citation principles



Joint Declaration of Data Citation Principles - FINAL

Preamble

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

In support of this assertion, and to encourage good practice, we offer a set of guiding principles for data within scholarly literature, another dataset, or any other research object.

These principles are the synthesis of work by a **number of groups.** As we move into the next phase, we welcome your participation and endorsement of these principles.

We currently have 795 active members.

Principles

The Data Citation Principles cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation practices that are both human understandable and machine-actionable.

RDA RESEARCH DATA ALLIANCE
United we stand

Joint Declaration of Data Citation Principles (cont‘d)

1) **Importance**

   Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance as publications.

2) **Credit and Attribution**

   Data citations should facilitate giving credit and normative and legal attribution to all contributors to the data.

Joint Declaration of Data Citation Principles (cont‘d)

3) **Evidence**

Whenever and wherever a claim relies upon data, the corresponding data should be cited.

4) **Unique Identification**

A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

Joint Declaration of Data Citation Principles (cont‘d)

5) **Access**

Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

6) **Persistence**

Unique identifiers, and metadata describing the data, and its disposition, should persist - even beyond the lifespan of the data they describe.

Joint Declaration of Data Citation Principles (cont‘d)

7) **Specificity and Verifiability**

Data citations should facilitate identification of, access to, and verfication of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verfiying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

Joint Declaration of Data Citation Principles (cont'd)

8) **Interoperability and flexibility**

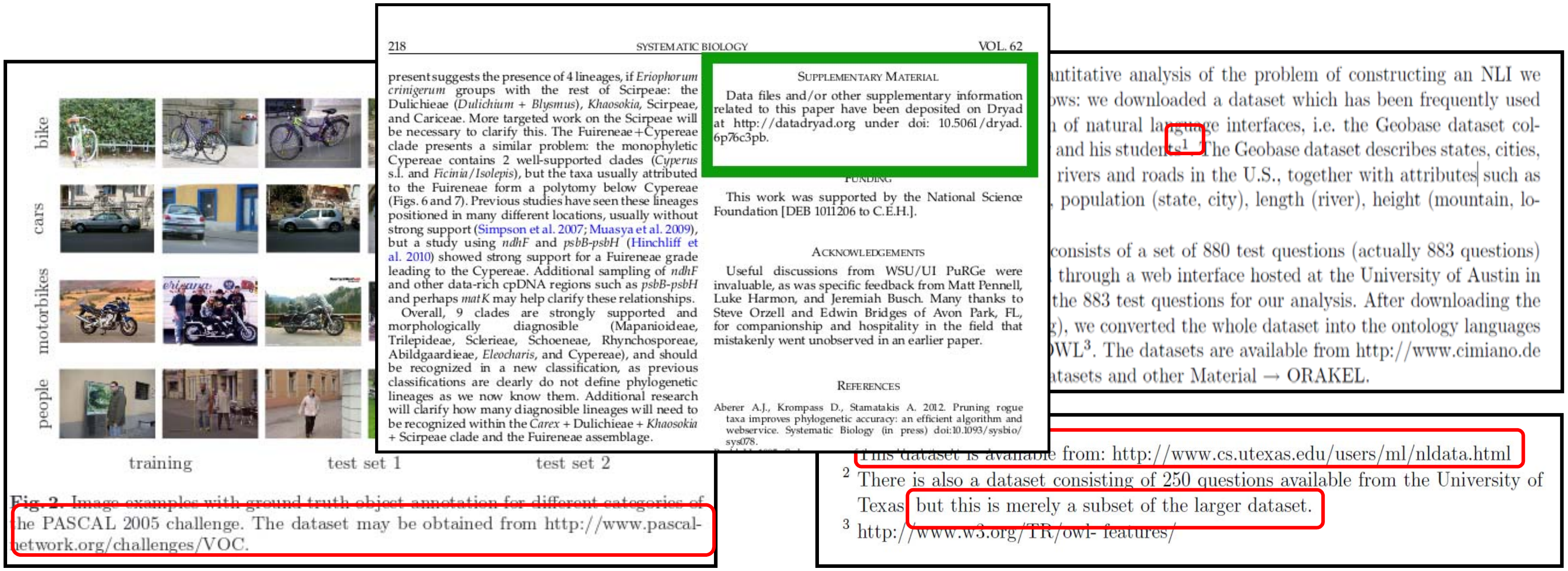
Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

Outline

- Joint Declaration of Data Citation Principles
- Challenges in non-trivial settings
- Recommendation of the RDA Working Group
- Pilots
- Summary

Data Citation

- 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?



bike
cars
motorbikes
people
training
test set 1
test set 2

~~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.

218 SYSTEMATIC BIOLOGY VOL. 62

present suggests the presence of 4 lineages, if *Eriophorum crinigerum* groups with the rest of Scirpeae: the Dulichieae (*Dulichium* + *Blysmus*), *Khaosokia,* Scirpeae, and Cariceae. More targeted work on the Scirpeae will be necessary to clarify this. The Fuireneae+Cypereae clade presents a similar problem: the monophyletic Cypereae contains 2 well-supported clades (*Cyperus* s.l. and *Ficinia/Isolepis*), but the taxa usually attributed to the Fuireneae form a polytomy below Cypereae (Figs. 6 and 7). Previous studies have seen these lineages positioned in many different locations, usually without strong support (Simpson et al. 2007; Muasya et al. 2009), but a study using *ndhF* and *psbB-psbH* (Hinchliff et al. 2010) showed strong support for a Fuireneae grade leading to the Cypereae. Additional sampling of *ndhF* and other data-rich cpDNA regions such as *psbB-psbH* and perhaps *matK* may help clarify these relationships.

Overall, 9 clades are strongly supported and morphologically diagnosible (Mapanioideae, Trilepideae, Sclerieae, Schoeneae, Rhynchosporeae, Abildgaardieae, *Eleocharis*, and Cypereae), and should be recognized in a new classification, as previous classifications are clearly do not define phylogenetic lineages as we now know them. Additional research will clarify how many diagnosible lineages will need to be recognized within the *Carex* + Dulichieae + *Khaosokia* + Scirpeae clade and the Fuireneae assemblage.

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.

ntitative analysis of the problem of constructing an NLI we ows: we downloaded a dataset which has been frequently used n of natural language interfaces, i.e. the Geobase dataset col- and his students[1]. 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 OWL[3]. The datasets are available from http://www.cimiano.de atasets and other Material → ORAKEL.

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/

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

Granularity of Data Citation

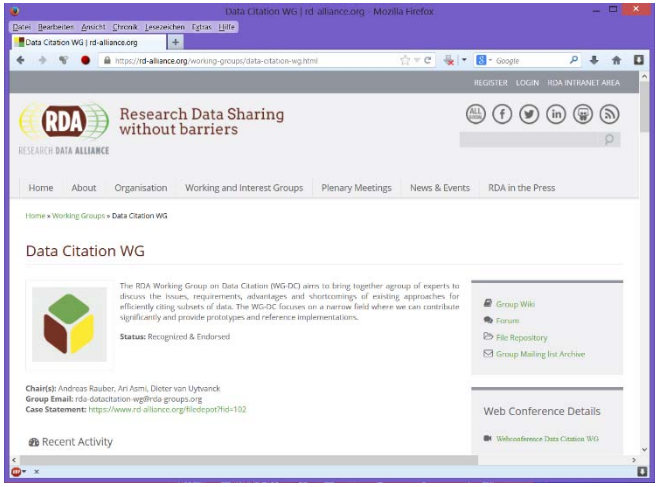
- What about the **granularity** of data to be cited?
  - Databases collect enormous amounts of data over time
  - Researchers use specific subsets of data
  - Need to identify precisely the subset used
- Current approaches
  - Storing a copy of subset as used in study -> scalability
  - Citing entire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

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

RDA WG Data Citation

- Research Data Alliance
- WG on **Data Citation:**
  Making Dynamic Data Citeable
- WG officially endorsed in March 2014
  - Concentrating on the problems of **large, dynamic (changing) datasets**
  - Focus!
    Not: PID systems, metadata, citation string, attribution, …
  - Liaise with other WGs and initiatives on data citation (CODATA, DataCite, Force11, …)

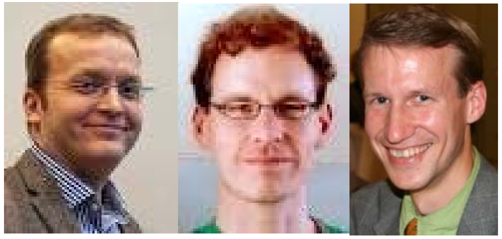

Data Citation WG | rd-alliance.org - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
Data Citation WG | rd-alliance.org
https://rd-alliance.org/working-groups/data-citation-wg.html
Google
REGISTER LOGIN RDA INTRANET AREA
Research Data Sharing without barriers
Home About Organisation Working and Interest Groups Plenary Meetings News & Events RDA in the Press
Home » Working Groups » Data Citation WG
Group Wiki
Forum
File Repository
Group Mailing list Archive
Web Conference Details
Webconference Data Citation WG
Recent Activity

Data Citation WG

The RDA Working Group on Data Citation (WG-DC) aims to bring together agroup of experts to discuss the issues, requirements, advantages and shortcomings of existing approaches for efficiently citing subsets of data. The WG-DC focuses on a narrow field where we can contribute significantly and provide prototypes and reference implementations.

Status: Recognized & Endorsed

Chair(s): Andreas Rauber, Ari Asmi, Dieter van Uytvanck
Group Email: rda-datacitation-wg@rda-groups.org
Case Statement: https://www.rd-alliance.org/filedepot?fid=102



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

Outline

- Joint Declaration of Data Citation Principles
- Challenges in non-trivial settings
- Recommendation of the RDA Working Group
- Pilots
- Summary

Making Dynamic Data Citeable

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

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

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

- ubset of data
- Upon executing selection („download") user gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citati n text (e.g. bTeX)
- PID resolves to land g page
  - Provides detailed metadata, link 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

RDA RESEARCH DATA ALLIANCE United we stand

ifs FACULTY OF !NFORMATICS

Data Citat

Note:

- 1 & 2 are already pretty much standard in many (RDBMS-) research databases
- Different ways to implement
- A bit more challenging for some data types (XML, LOD, …)

System set-up to supp dynamic data:

1. **Ensure data is time-stamped**
   i.e. any additions, deletions are marked with a timestamp
   (*optional, if data is dynamic*)

2. **Ensure data is versioned**
   i.e. updates not implemented as overwriting an earlier value, but as *marked-as-deleted* and *re-inserted* with new value, both time-stamped
   (*optional, if data is dynamic and access to previous versions is desired*)

3. **Create a query store for queries and metadata**

Data Citation – Recommendations *(Draft, 2/4)*

When a specific subset of data needs to be persistently identified (i.e. not necessarily for all subsets!):

1. **Re-write the query to a normalized form** (*optional*)
2. **Specifically: re-write the query to ensure unique sort of the result set** (*optional*)
3. **Compute a hash key** of the normalized query to identify identical queries (*optional*)
4. **Assign a time-stamp to the query**
   Execution time ***or:*** last update to the entire database ***or:*** last update to the subset of data affected by the query
5. **Compute a hash key of the result set** (*optional*)
6. **Assign PID to the query** *(if query/result set is new)*
7. **Store query and metadata** in query store

Data Citation – Recommendations *(Draft, 3/4)*

Upon request of a specific subset:

1. **PID resolves to landing page** of the subset, provides metadata including link to the super-set (PID of the DB)
2. **Landing page allows** (transparently, in a machine-actionable manner) to **retrieve the subset by re-executing the query**

- Query can be re-executed with the **original time stamp** or with the **current timestamp**, retrieving the semantically identical data set but incorporating all changes/corrections/updates applied since.
- Storing the query string provides comprehensive **provenance information** (description of criteria that the subset satisfies)

Data Citation – Recommendations *(Draft, 4/4)*

Upon modifications to the data management system:

1. **When data is migrated** to a new representation (new DBMS system, new schema), the **queries need to be migrated**
2. **Hash keys** for the query strings may need to be **re-computed**
3. **Hash input function** for result set may need to be **adapted** to ensure that the result sets are presented in the same form to the hash function
4. **Successful re-writing should be verified** by ensuring that queries can be re-executed resulting in the correct result set hash key

Initial Pilots

- Devised concept
- Identified challenges (unique sorting, hash-key computation, distribution, different data types, …)
- Evaluated conceptually in different settings
  - How to apply versioning/time-stamping efficiently?
  - How to perform query re-writing?
  - How easy to adopt / changes required within RI?
- Started implementing pilots
  - SQL: LNEC, MSD
  - CSV: MSD, open source prototype
  - XML: xBase

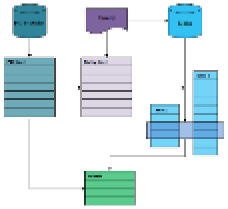
Outline

- Joint Declaration of Data Citation Principles
- Challenges in non-trivial settings
- Recommendation of the RDA Working Group
- Pilots
- Summary

WG Pilots

- Pilot workshops and implementations by
  - Various EU projects (TIMBUS, SCAPE,…)
  - NERC (UK Natural Environment Research Council Data Centres)
  - ESIP (Earth Science Information Partners)
  - CLARIN (XML, Field Linguistics Transcriptions)
  - Virtual Atomic and Molecular Data Centre

- Prototype solutions for
  - SQL, CSV, XML (partially)
  - LOD/RDF, triple-store DBs in the queue
  - Distributed data



Dynamic Data Citation - Pilots

Dynamic Data Citation for SQL Data

LNEC, MSD Implementation

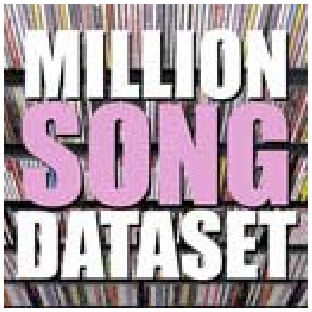
FACULTY OF !NFORMATICS

SQL Prototype Implementation

- 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

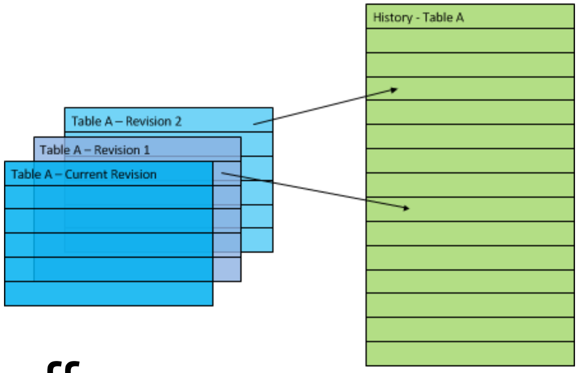
SQL Prototype Implementation



- 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

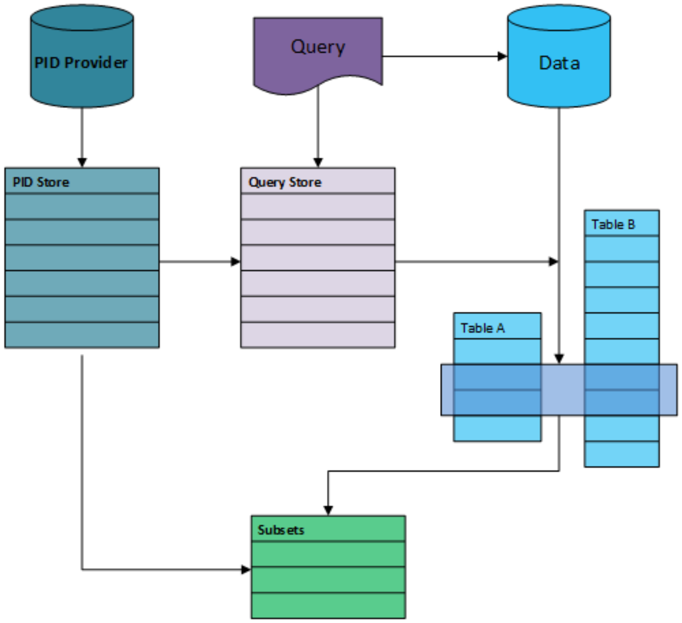
- Integrated
  - Extend original tables by temporal metadata
  - Expand primary key by record-version column
- Hybrid
  - Utilize history table for deleted record versions with metadata
  - Original table reflects latest version only
- Separated
  - Utilizes full history table
  - Also inserts reflected in history table



- Solution to be adopted depends on trade-off
  - Storage Demand
  - Query Complexity
  - Software adaption

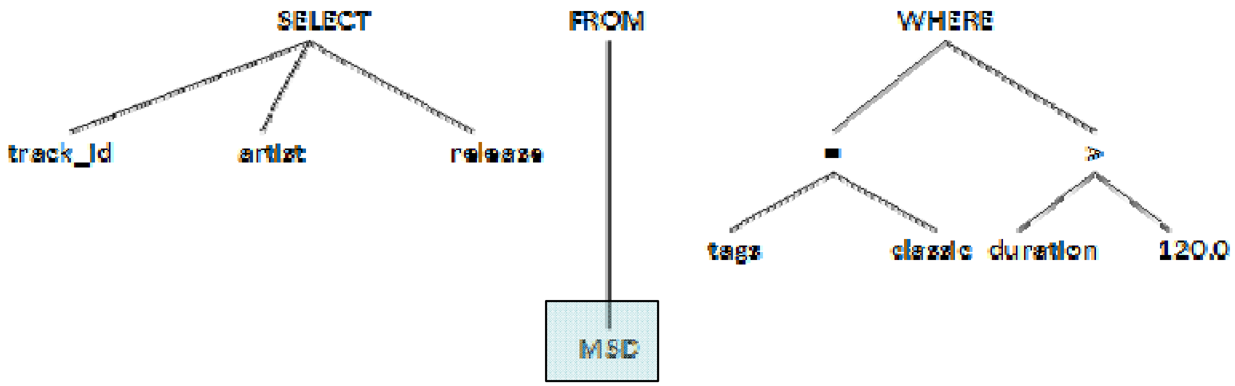
SQL: Storing Queries

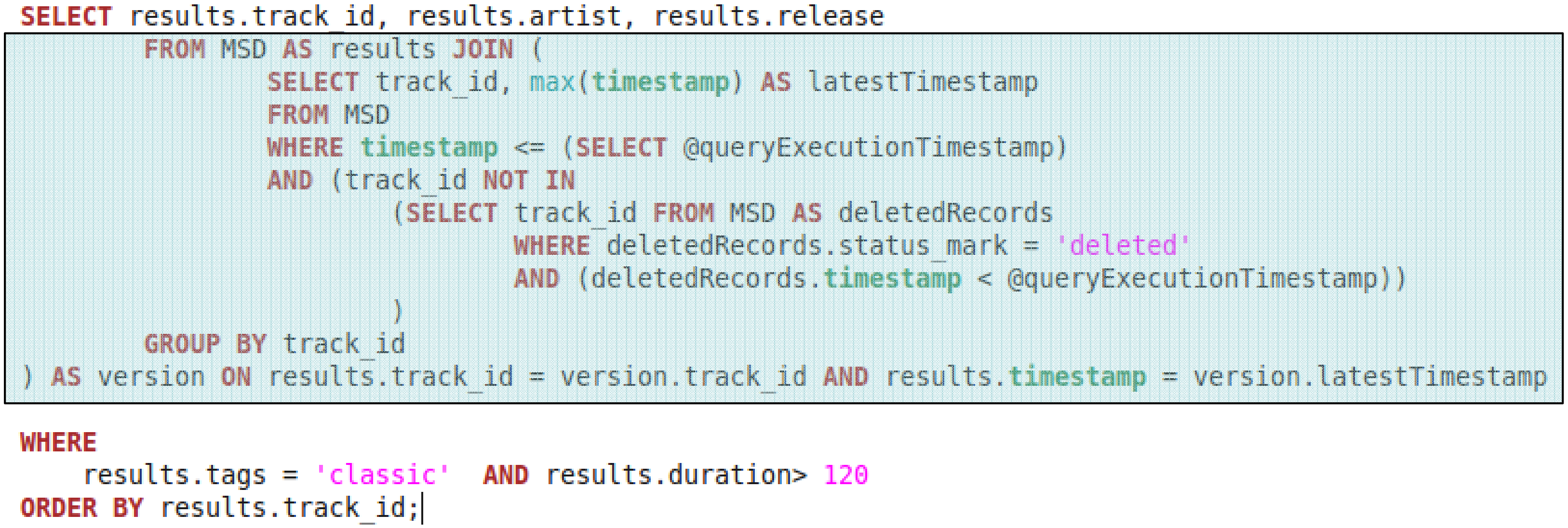
- Add query store containing
  - PID of the query
  - Original query
  - Re-written query + query string hash
  - Timestamp
    (as used in re-written query)
  - Hash-key of query result
  - Metadata useful for citation / landing page
    (creator, institution, rights, …)
  - PID of parent dataset
 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

Dynamic Data Citation for CSV Data

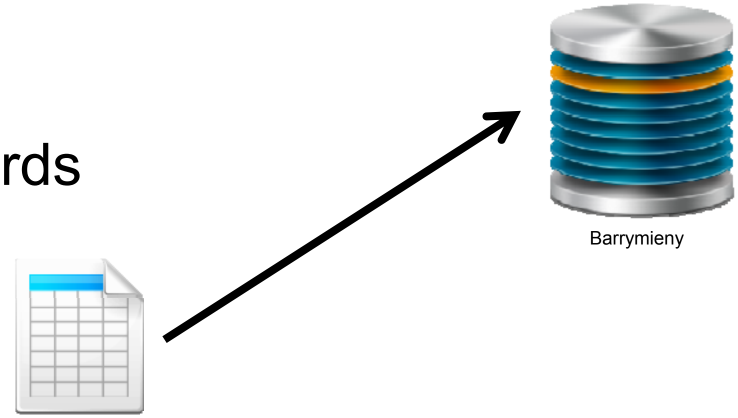
Open Source Reference Implementation

Dynamic Data Citation for CSV Data

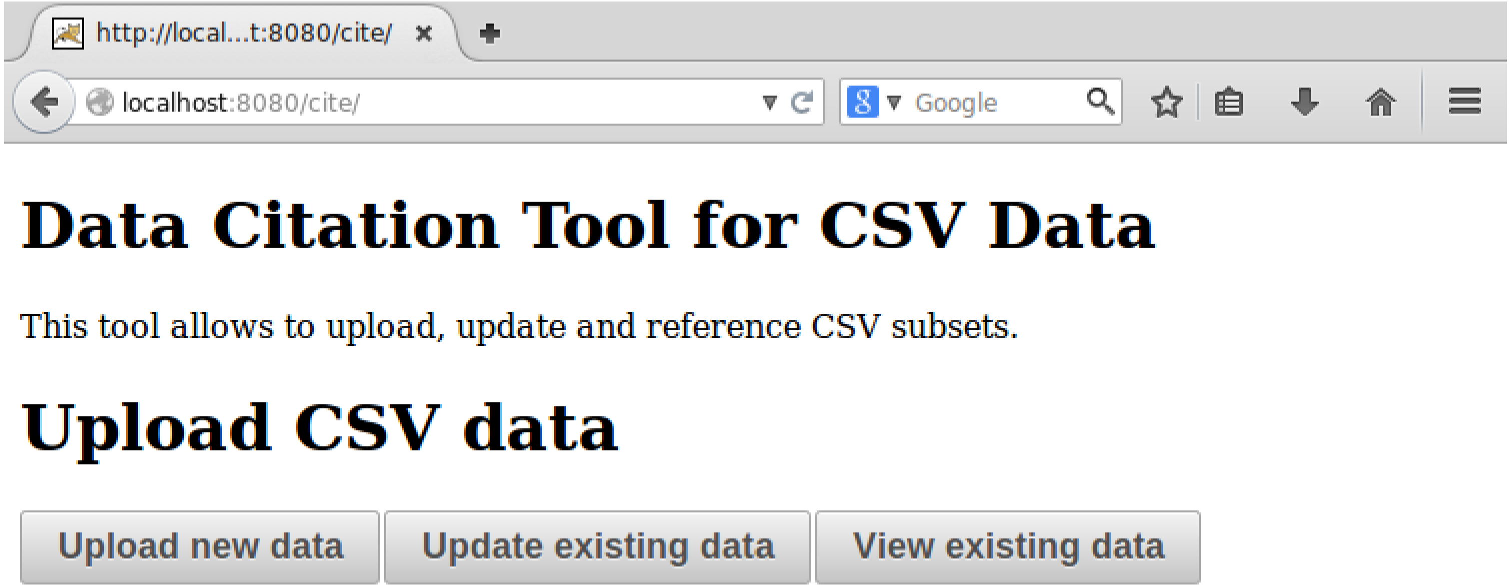
- Why CSV data? (not large, not very dynamic…)
  - Well understood and widely used
  - Simple and flexible
  - Most frequently requested during initial RDA meetings
- Goals:
  - Ensure cite-ability of CSV data
  - Enable subset citation
  - Support particularly small and large volume data
  - Support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

CSV Prototype: Basic Steps

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



CSV Data Prototype

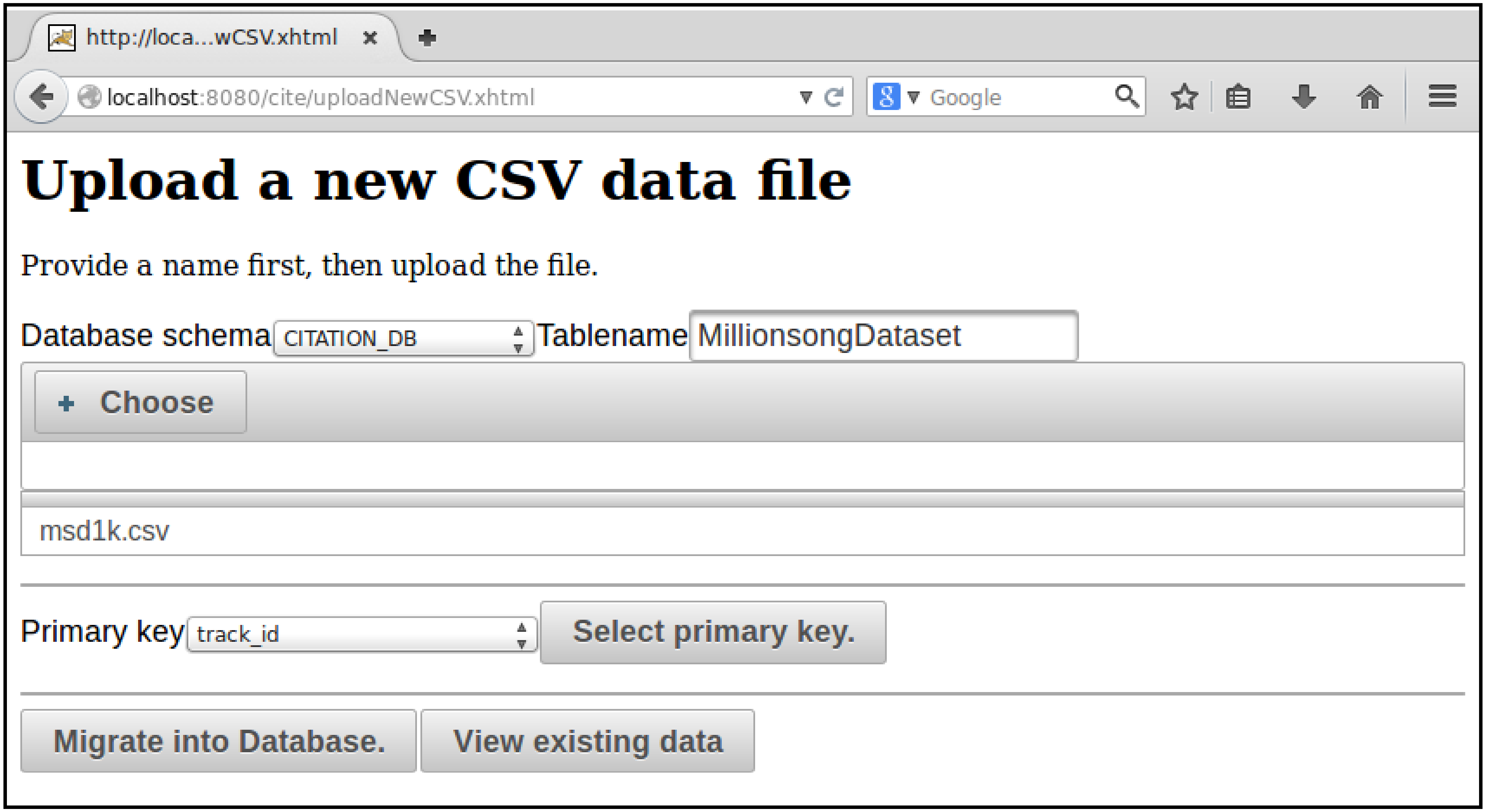


Data Citation Tool for CSV Data

This tool allows to upload, update and reference CSV subsets.

Upload CSV data

CSV Data Prototype



Upload a new CSV data file

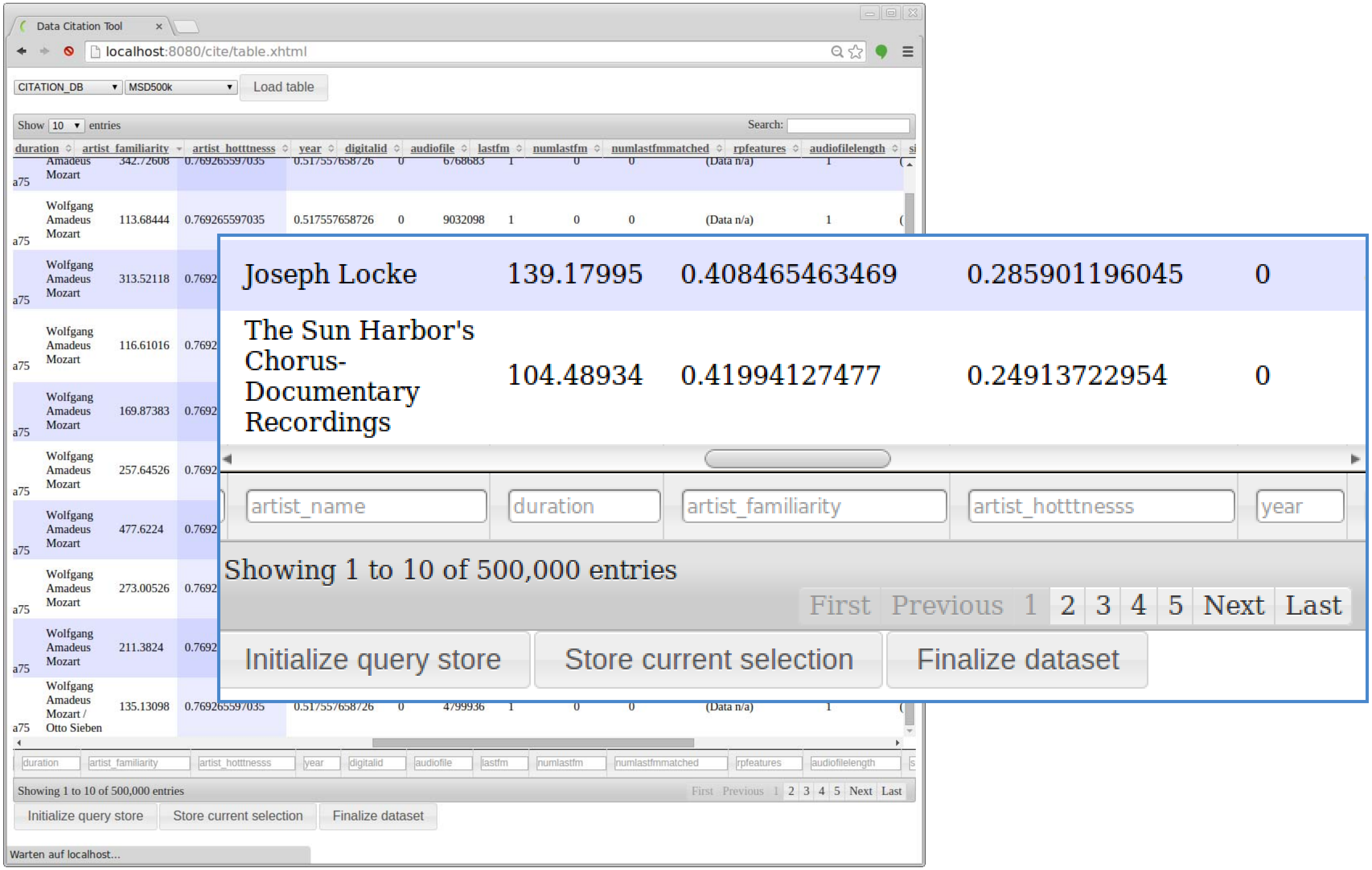
Provide a name first, then upload the file.

Database schema CITATION_DB Tablename MillionsongDataset

msd1k.csv

Primary key track_id

CSV Data Prototype



	duration	artist_familiarity	artist_hotttnesss	year	digitalid	audiofile	lastfm	numlastfm	numlastfmmatched	rpfeatures	audiofilelength	si
a75	Amadeus Mozart	342.72608	0.769265597035	0.517557658726	0	6768683	1	0	0	(Data n/a)	1	(
a75	Wolfgang Amadeus Mozart	113.68444	0.769265597035	0.517557658726	0	9032098	1	0	0	(Data n/a)	1	(
a75	Wolfgang Amadeus Mozart	313.52118	0.7692									
a75	Wolfgang Amadeus Mozart	116.61016	0.7692									
a75	Wolfgang Amadeus Mozart	169.87383	0.7692									
a75	Wolfgang Amadeus Mozart	257.64526	0.7692									
a75	Wolfgang Amadeus Mozart	477.6224	0.7692									
a75	Wolfgang Amadeus Mozart	273.00526	0.7692									
a75	Wolfgang Amadeus Mozart	211.3824	0.7692									
a75	Wolfgang Amadeus Mozart / Otto Sieben	135.13098	0.769265597035	0.517557658726	0	4799936	1	0	0	(Data n/a)	1	(

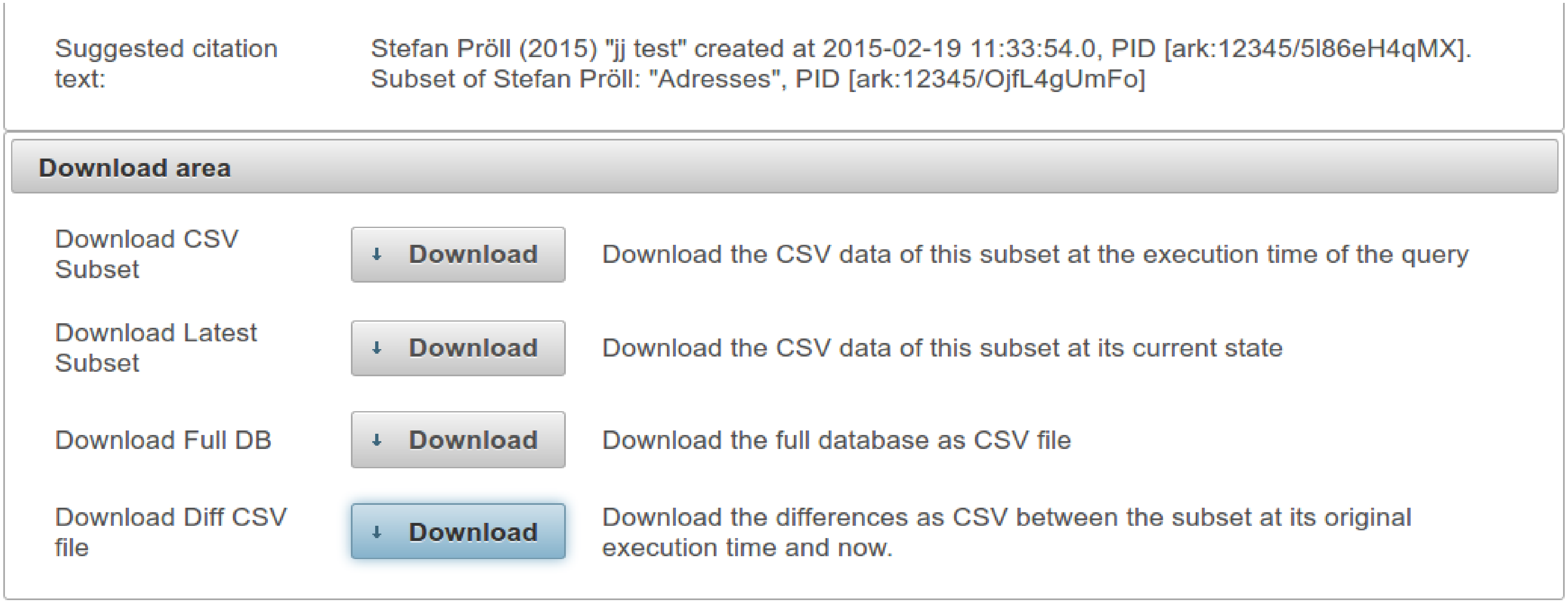
Showing 1 to 10 of 500,000 entries

Joseph Locke	139.17995	0.408465463469	0.285901196045	0
The Sun Harbor's Chorus-Documentary Recordings	104.48934	0.41994127477	0.24913722954	0

Showing 1 to 10 of 500,000 entries

FACULTY OF !NFORMATICS

CSV Data Prototype

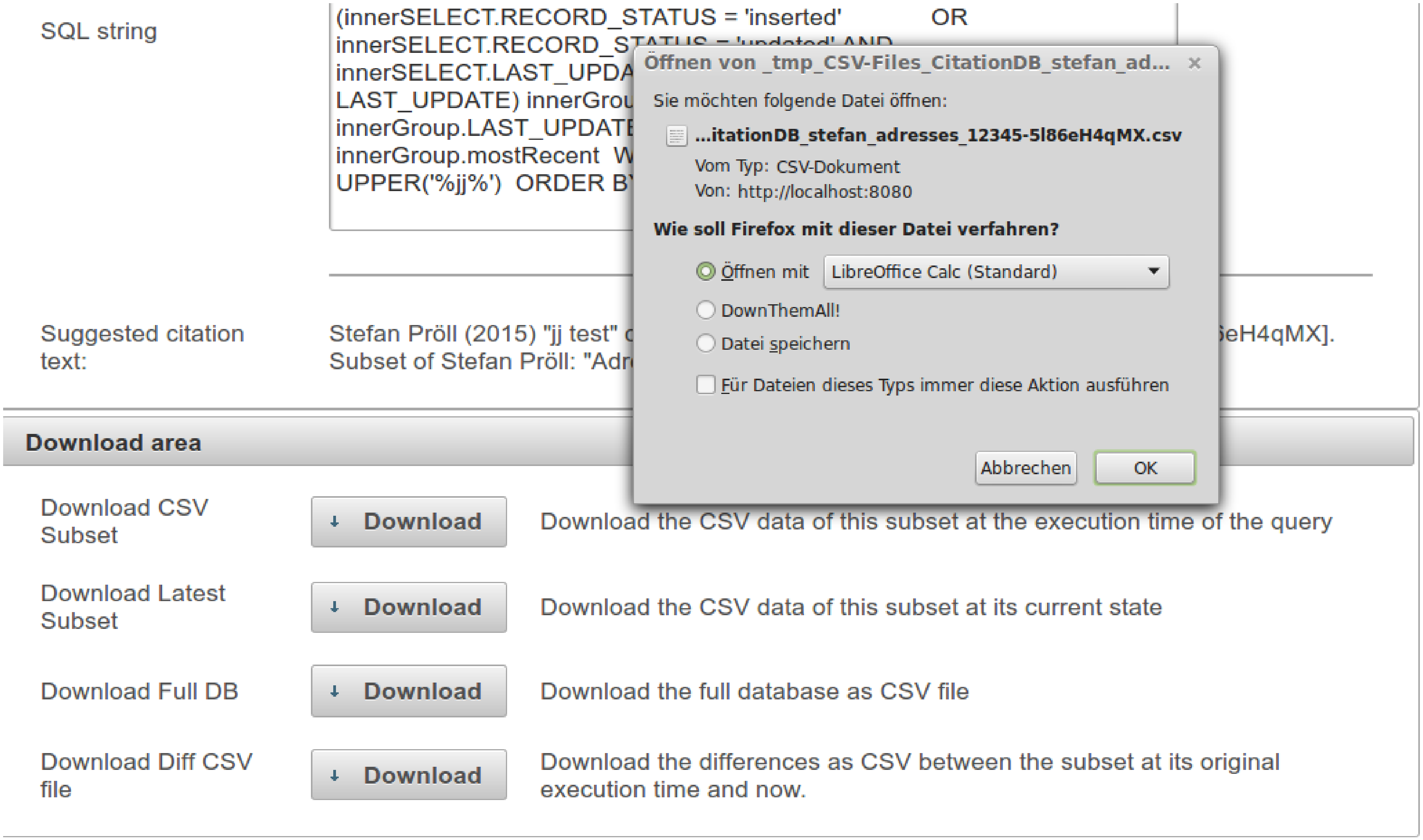


Suggested citation text:	Stefan Pröll (2015) "jj test" created at 2015-02-19 11:33:54.0, PID [ark:12345/5l86eH4qMX]. Subset of Stefan Pröll: "Adresses", PID [ark:12345/OjfL4gUmFo]	

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_UPDA
LAST_UPDATE) innerGrou
innerGroup.LAST_UPDATE
innerGroup.mostRecent  W
UPPER('%jj%')  ORDER B
```

Suggested citation text: Stefan Pröll (2015) "jj test" c eH4qMX]. Subset of Stefan Pröll: "Adr

Öffnen von _tmp_CSV-Files_CitationDB_stefan_ad... ×

Sie möchten folgende Datei öffnen:

...itationDB_stefan_adresses_12345-5l86eH4qMX.csv

Vom Typ: CSV-Dokument

Von: http://localhost:8080

Wie soll Firefox mit dieser Datei verfahren?

● Öffnen mit LibreOffice Calc (Standard)

○ DownThemAll!

○ Datei speichern

□ Für Dateien dieses Typs immer diese Aktion ausführen

Abbrechen | OK

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 update from VAMDC Distributed Data Centre

Carlo Maria Zwölf

Virtual Atomic and Molecular Data Centre
carlo-maria.zwolf@obspm.fr

VAMDC

- Virtual Atomic and Molecular Data Centre
- Worldwide e-infrastructure federating 41 heterogeneous and interoperable Atomic and Molecular databases
- Nodes decide independently about growing rate, ingest system, corrections to apply to already stored data
- Data-node may use different technology for storing data (SQL, No-sql, ASCII files),
- All implement VAMDC access/query protocols
- Return results in standardized XML format (XSAMS)
- Access directly node-by-node or via VAMDC portal, which relays the user request to each node

VAMDC

Workshop prior to RDA P4

Issues identified

- Each data node could modify/delete/add data without tracing
- No support for reproducibility of past data extraction

Proposed Data Citation WG Solution:

- Considering the distributed architecture of the federated VAMDC infrastructure, it seemed very complex to apply the “Query Store” strategy
  - Should we need a QS on each node?
  - Should we need an additional QS on the central portal?
  - Since the portal acts as a relay between the user and the existing nodes, how can we coordinate the generation of PID for queries in this distributed context?

VAMDC

Status / Progress since RDA P4

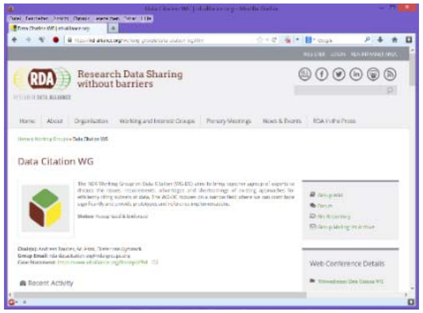
- Versioning adopted prior to P4
- Central service registering user interactions with data
- At each client SW notifies tracing service that a given **user** is using, at a given **time**, that specific **software** for submitting a given **query**
- Will assign single identifier for each unique query centrally
- Query store initially private (confidentiality issues)

Further Pilots

- NERC: UK Natural Environment Research Council
  - ARGO buoy network: SeaDataNet
  - Butterfly monitoring, Ocean buoy network, National hydrological archive, …
- ESIP: BCO-DMO
- XML Data in Field Linguistics (CLARIN, XBase)
- Further Pilots on XML, LOD, …
- Workshops:
  - NERC Workshop, London, July 1/2 2014
  - ESIP Mtg in Washington, Jan 8 2015: Earth Science Data
  - Data Citation Workshop, Riva di Garda, April 20/21
  - Bilateral meetings with data centers

Join RDA and Working Group

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
  - Web Conferences: https://rd-alliance.org/webconference-data-citation-wg.html
  - List of pilots: https://rd-alliance.org/groups/data-citation-wg/wiki/collaboration-environments.html

Summary

- Trustworthy and efficient e-Science based on data
- Data as “1st-class citizen”
- Support for identifying arbitrary subsets of dynamic data
  - Time-stamping and versioning of data
  - Storing (and citing) time-stamped queries
- Allows retrieving exact view on data set as used
- No need for artificial “versioning”, delaying release of new data, or redundant storage of data subset dumps
- Helps tracing provenance (semantics) of data selection
- Future work: distributed datasets, data & query migration

TU WIEN

Thank you!

NERC SCIENCE OF THE ENVIRONMENT

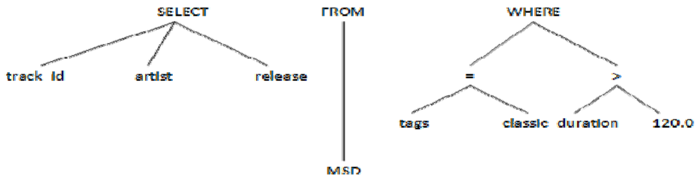
CLARIN

ESIP

VAMDC consortium

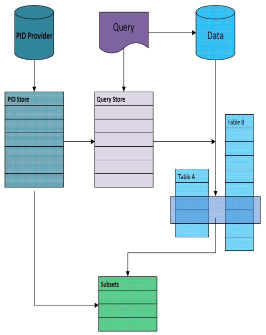
MILLION SONG DATASET

BASE X

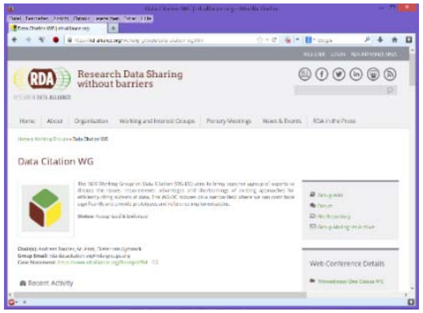


```
SELECT results.track_id, results.artist, results.release
        FROM MSD AS results JOIN (
                SELECT track_id, max(timestamp) AS latestTimestamp
                FROM MSD
                WHERE timestamp <= (SELECT @queryExecutionTimestamp)
                AND (track_id NOT IN
                        (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;
```



http://www.ifs.tuwien.ac.at/imp


Research Data Sharing without barriers

Home About Organisation Working and Interest Groups Plenary Meetings News & Events RDA in the Press

Data Citation WG

Web Conference Details

Recent Activity

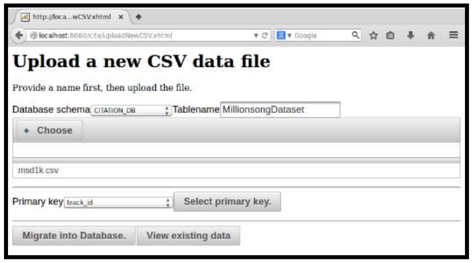
DC1

Data Citation Principles

RDA

RESEARCH DATA ALLIANCE

United we stand


http://loca...wCSV.xhtml

localhost:8080/cite/uploadNewCSV.xhtml

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

RDA

RESEARCH DATA ALLIANCE

United we stand

ifs FACULTY OF !NFORMATICS