Data Citation WG
Making Dynamic Data Citable

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

Andreas Rauber, Ari Asmi, Dieter van Uytvanck

research data sharing without barriers
rd-alliance.org

Agenda

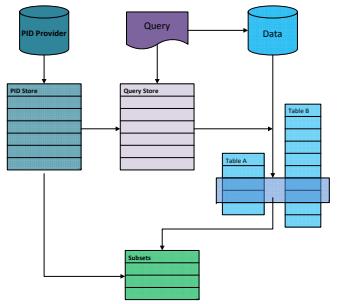
- Intro: Goals, approach and progress so far
- Reports from pilots
  - Results from the NERC workshop
  - CLARIN presentation
  - MSD presentation
  - XML data
  - National Supercomputing Facility Australia
  - Results from the VAMDC workshop
  - <other?>
- Open challenges
  - Time stamping and versioning
  - Hash Key calculation / Result set verification
  - Data Citation Zoo and request for examples
- Discussion & future plans

Accomplishments & Goals

- Working group endorsed in March 2014
- Goals
  - Concentrating on the problems of **dynamic (changing) datasets** (But: should work also for non-dynamic data)
  - Assigning PIDs flexibly to **arbitrary subsets of data**
    - Entire dataset, arbitrary subsets of rows/columns, sub-graphs, …
    - SQL, XML, RDF, LOD, CSV, …
  - Does NOT deal with metadata, landing page design, bibliometrics, …
  - Creating concepts, recommendations and pilots/demonstrators
  - Starting with conceptual evaluation of the approach, studying fitness, impact, scalability, changes required, …
  - Followed by actual pilot implementation

Data Citation – Making Dynamic Data Cite-able

- **Goal**:
  - Ensure cite-ability of data at arbitrary levels of granularity, particularly when data is large-volume and dynamic
  - Machine-actionable, variety of data types
- **Approach**:
  - Ensure data is time-stamped and versioned
  - Assign PID to time-stamped query/selection expression
- **Deliverables**:
  - Requirements: minimal set of functionality for data and service
  - Pilot implementations for selected data types
  - Reference architecture and guidelines for implementation
- **Pilots**
  - https://rd-alliance.org/groups/data-citation-wg/wiki/use-cases-pilots-precise-citation-data.html



Data Citation – Making Dynamic Data Cite-able

- **Pilots in progress (selected)**
  - LNEC: Portugese Civil Engineering Lab (SQL)
  - NERC: UK Natural Environment Research Council data centres (SQL)
  - CLARIN: XML
  - MSD: Million Song Database (CSV, SQL)
  - VAMDC: Virtual Atomic and Molecular Data Centre: Distributed SQL
- **Pilots under preparation**
  - XML prototype for scalablity analysis
  - LOD

Data Citation – Making Dynamic Data Cite-able

- **Open issues identified**
  - Which timestamp to assign?
    (query, last update, last change to result set)
  - Hashing for verification
  - Annotating compulsory deletion
  - How to handle distributed datasets
  - Migrating data and queries

Agenda

- Intro: Goals, approach and progress so far
- Reports from pilots
  - Results from the NERC workshop
  - CLARIN presentation
  - MSD presentation
  - XML data
  - National Supercomputing Facility Australia
  - Results from the VAMDC workshop
  - <others?>
- Open challenges
  - Time stamping and versioning
  - Hash Key calculation / Result set verification
  - Data Citation Zoo and request for examples
- Discussion & future plans

7 minutes each!

Pilots

Results from NERC Workshop

June 1-2 2014, London

John Watkins

Centre for Ecology and Hydrology

jww@ceh.ac.uk

Data Citation WG – July 2014 London workshop report

Following RDA Plenary 3 in Dublin –

- Plans made for a workshop looking at specific use cases of data citation
- WG members were invited to British Library in London to contribute use cases
- This workshop was arranged and mainly attended by UK Natural Environment Research Council data centres and hosted by British Library

Aims of workshop

- To present RDA WG conceptual model addressing citation of dynamic data to a group of data curation practitioners
- To assess goodness of fit of the model for the requirements of users, curators, publishers, authors
- To extend and/or improve the model to meet the widest range of data users
- To plan test implementations of the citation model with various dynamic data curated by the group
- To provide input to WG reporting at RDA P4

Workshop facilitation:

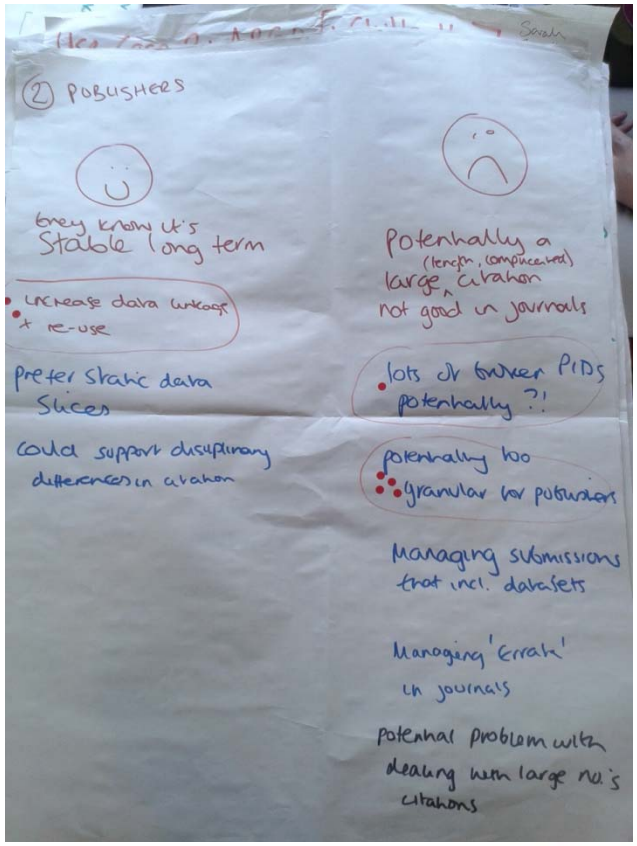
- We had a facilitator!
- We had a detailed plan for the 2 days
- We made sure everybody got out of their seats and contributed ideas
- We captured all the work done and tried to present a summary report

What we did:

- We looked at the current model from 4 perspectives
  - Data user
  - Data depositor
  - Data Centre / repository
  - Data publishers / journals
- We looked at the current model from 4 use cases
  - Butterfly monitoring
  - Ocean buoy network
  - Sociological Archive
  - National hydrological archive

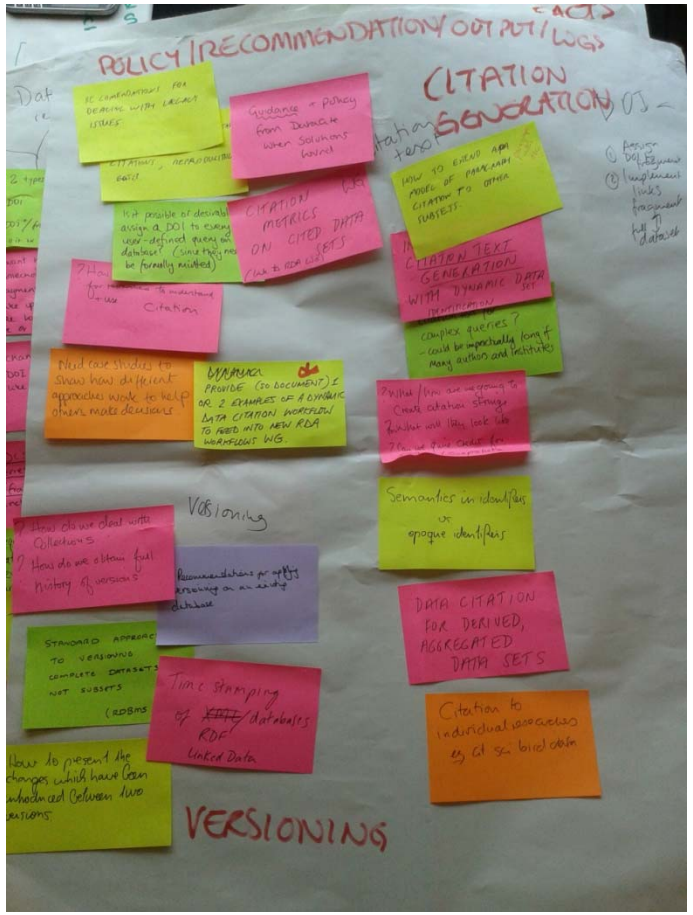
We worked hard on positives and negatives of the model:

- Data centres liked PIDs but didn't want thousands of them
- Publishers didn't want very fine grained PIDs into data sets
- Originators wanted recognistion for data sets produced
- We voted on what we thought were the most important aspects to think of improvements



We came up with ideas of general improvements and practical steps

- We looked at different issues in the model especially subsets and versioning in file repositories
- UKDA gave a clear demonstration of how to cite parts of documents
- We found need for board PIDs for collections and finer PIDs for versions and additions
- We looked at improvements to the working of the current use cases



Sometimes we had to stop for a rest but Andi kept us working:

- We documented the ideas recommendations in the report
- The 3 use cases worked on proposed improvements to citation
- The participant felt they have a much better understand of how they might deal with dynamic data citation and how other organisations like DataCite were working on it.

Outcome of workshop

- Stefan has generated a paper for IEEE detailing the approach and use cases
- The ARGO buoy network has a draft proposal for how to implement data citation of the SeaDataNet
- Other UK NERC data centres are continuing with their data citation developments
- The ESIP (Federation of Earth science Information Partners) may hold a workshop in early 2015 (Washington DC?) to go through case studies
- A working group maybe organized in China next year by Inst of Scientific and Technical Information of China (ISTIC)

Support for Dynamic Data Citation in CLARIN

Dieter van Uytvanck

Max Plank Institute

dieter.vanuytvanck@mpi.nl

Use case: field linguistics

- Field Linguistics:
  language archive: transcriptions
- Type: well-described XML files (stand-off annotations to stable video/audio files)
- Size: small (few 100 KBs)
- Dynamics: rather tens (or even <10) than hundreds of versions
- Citation practice: rather fragments than the whole file (illustration, examples, counter-examples)

Use case: field linguistics

- https://corpus1.mpi.nl/ds/imdi_browser/versioninginfo.jsp?nodeid=MPI2002357%23
- the only timestamps displayed are those at the moment a new version is created (hover over the date) - but this is rather a limit of the displaying application

Use case: field linguistics – handle record

- There is an MD5 checksum and timestamp however in the handle records:
  - http://hdl.handle.net/1839/00-0000-0000-001E-8DA4-1?noredirect
  - http://hdl.handle.net/1839/00-0000-0000-001E-8DB5-6?noredirect
- not standardized however to store this in handle record, see eg:
  - http://hdl.handle.net/10.4121/uuid:496d8244-5449-4036-b32f-3bf97f298689?noredirect

Use case: field linguistics - issues

- Permission system: by default older versions are not accessible for others than the resource owner

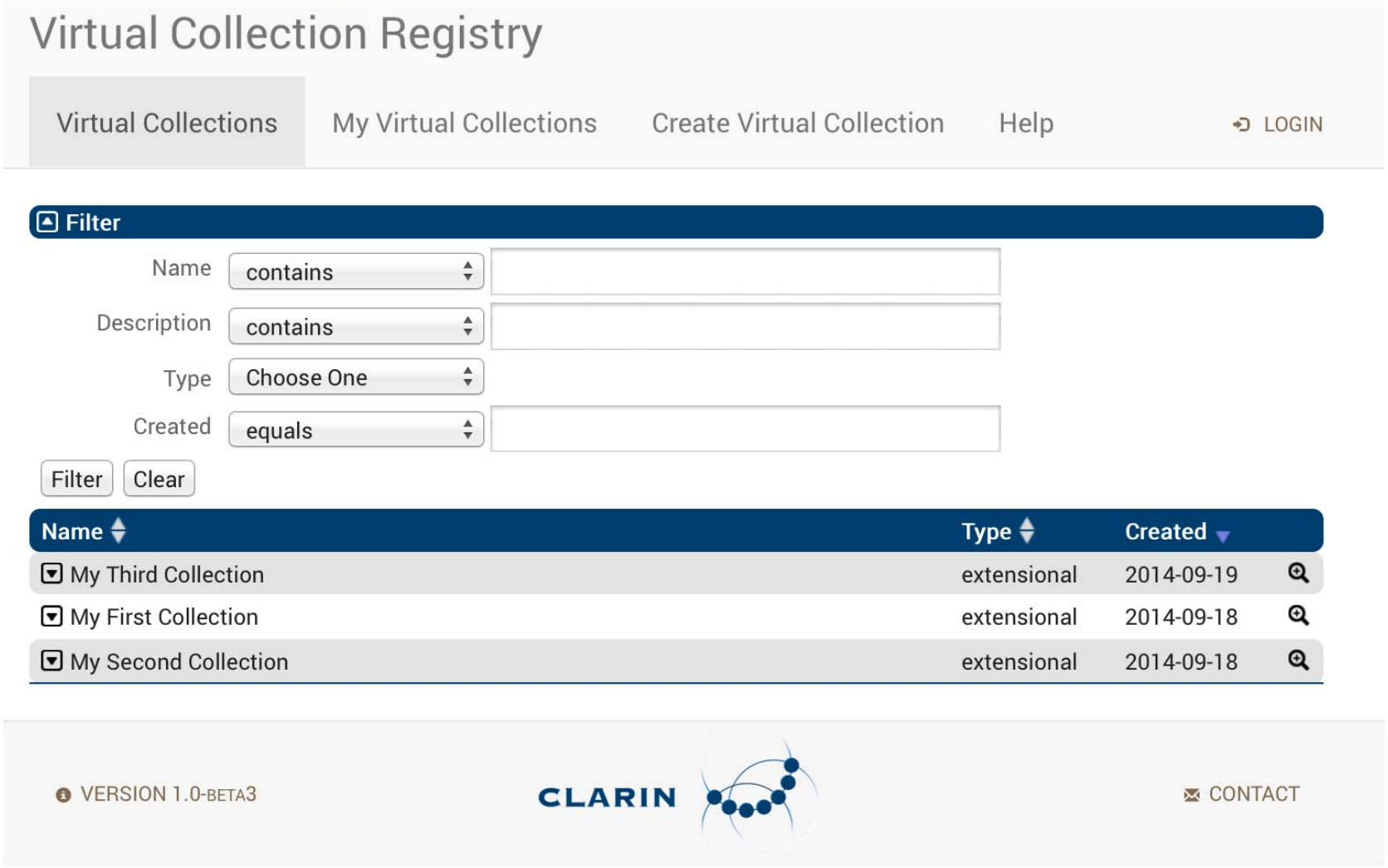
Virtual Collections

- Concept: set of references to data objects
- Both human and machine readable (HTTP content negotiation: e.g. HTML or XML)
- Accessible via REST API or OAI-PMH harvesting
- Possible references:
  - URLs
  - Persistent Identifiers (DOIs, handles)
  - Queries (= plain-text strings) > no execution mechanism although they can also be encoded as URL
- Indication of type of reference:
  - Intensional (query) vs extensional (list of links)
  - Stable vs instable
  - Intended use

Virtual Collections - Implementation

- **Beta** version available of the Virtual Collection Registry: http://clarin.eu/vcr
- Official release planned for early October
- Free software, GPL v3
- Comes with
  - Federated Identity
  - Persistent Identifiers
  - Metadata harvesting
- Connection to this WG: allows to publish links to versioned datasets in an easy way

Virtual Collections - Implementation



Virtual Collections - plans

- Improve reference checking on creation
- Integrate with DataCite:
  - Assign DOIs on publication
  - Integrate in metadata catalogue
- Integrate with ORCID API for author referencing

Dynamic Data Citation for CSV Data

Stefan Pröll

Secure Business Austria

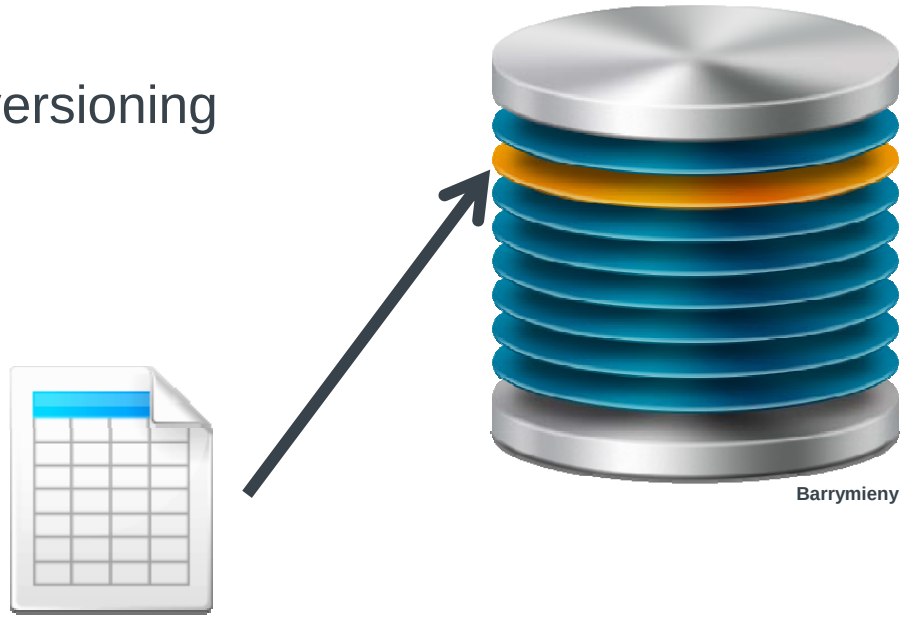
sproell@sba-research.org

Dynamic Data Citation for CSV Data

- **Goals**:
  - Ensure cite-ability of CSV data
  - Enable subset citation
  - Support particularly small and large volume data
  - Support dynamically changing data
- **Why CSV data?**
  - Well understood and widely spread
  - Simple and flexible

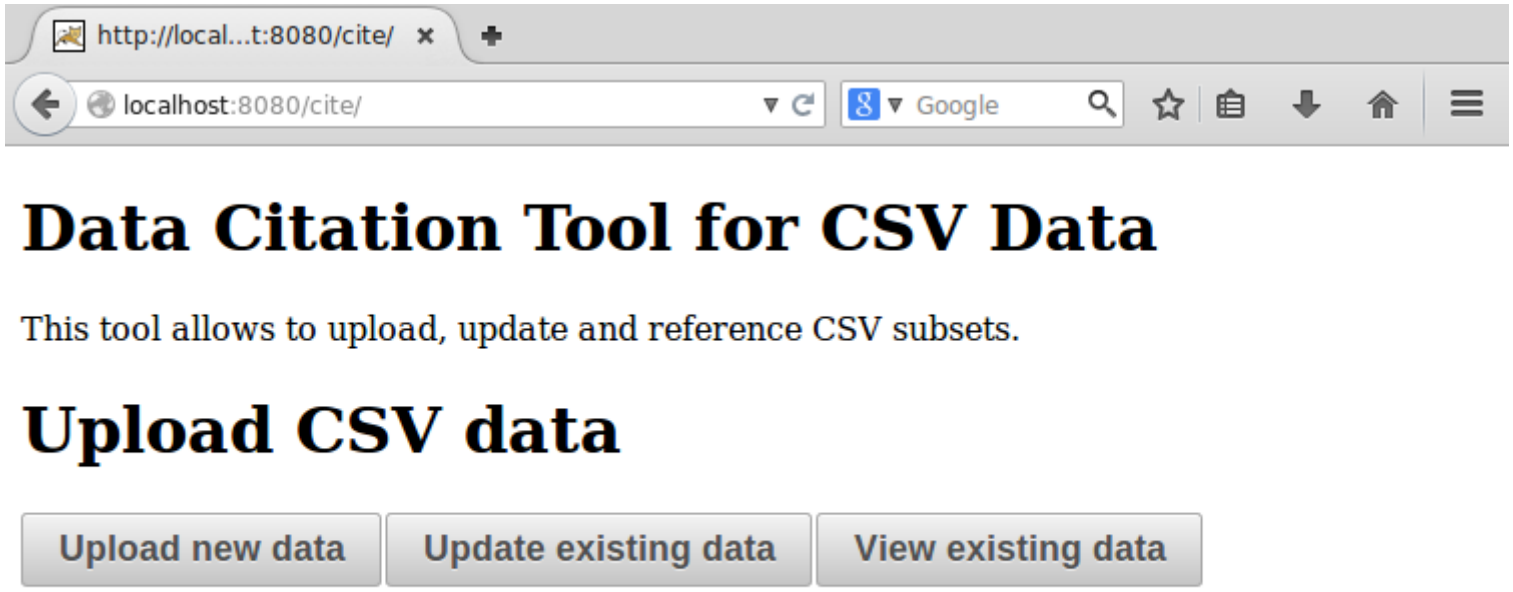
Basic Steps

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

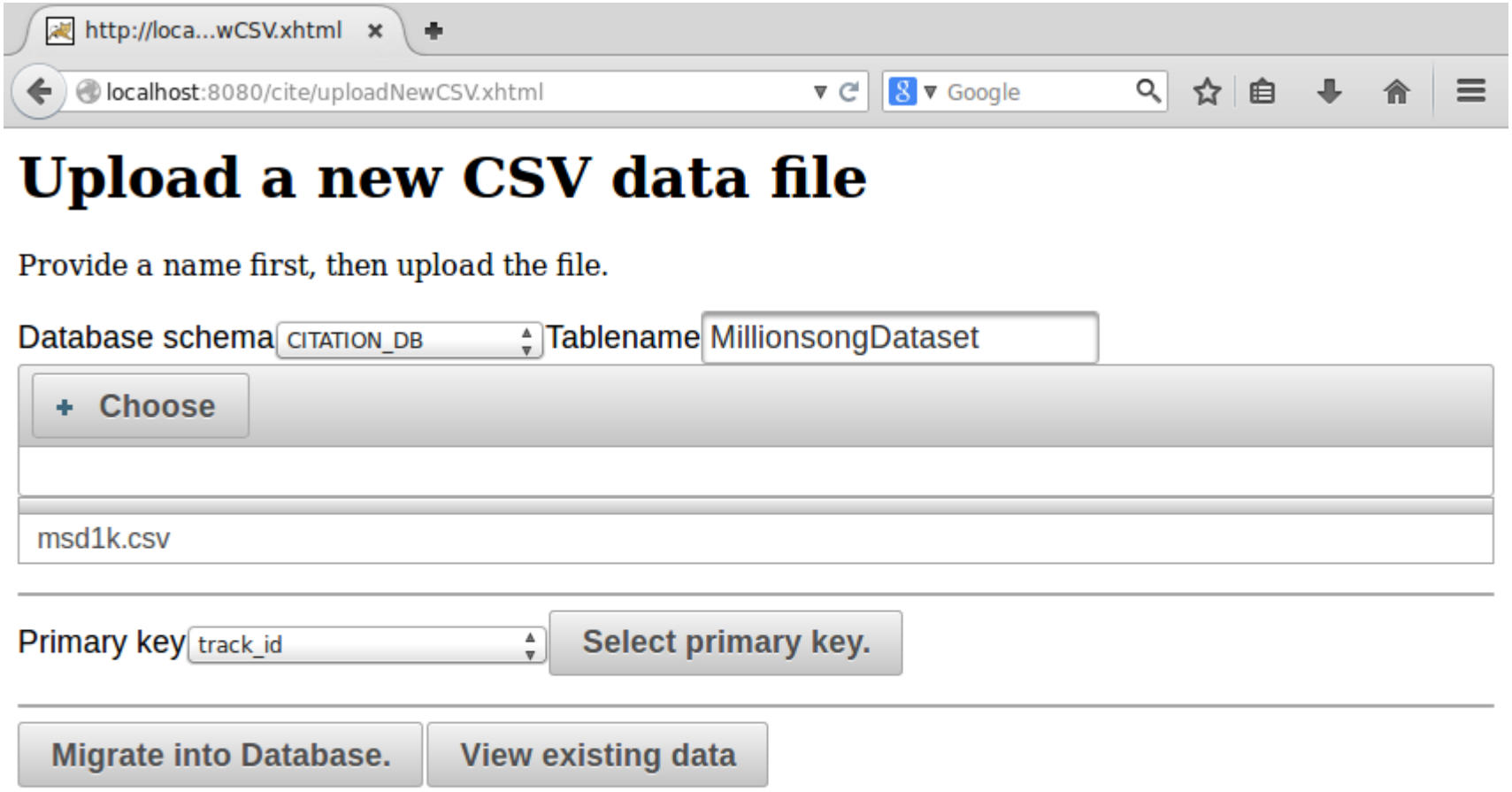


Barrymieny

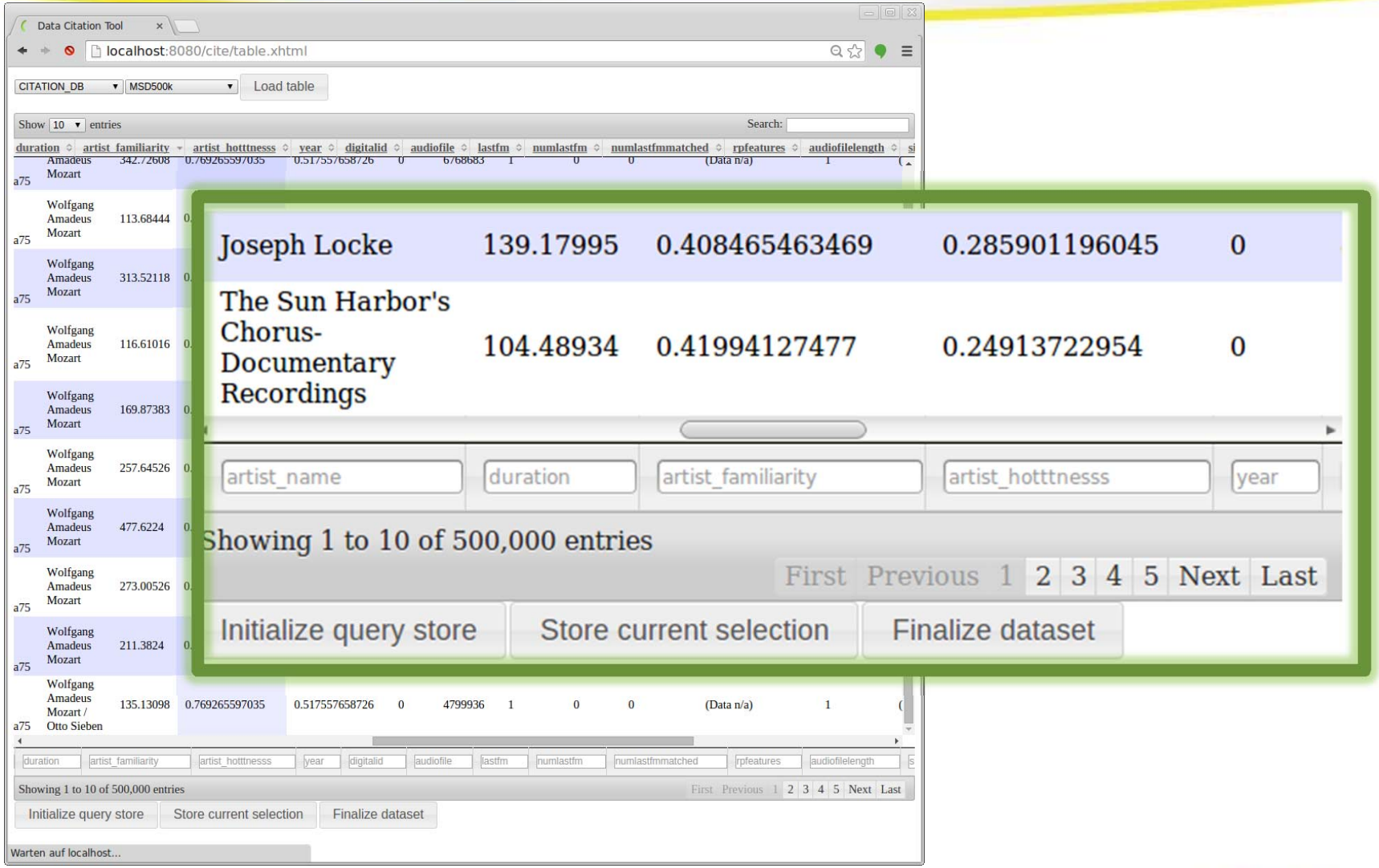
CSV Data Prototype



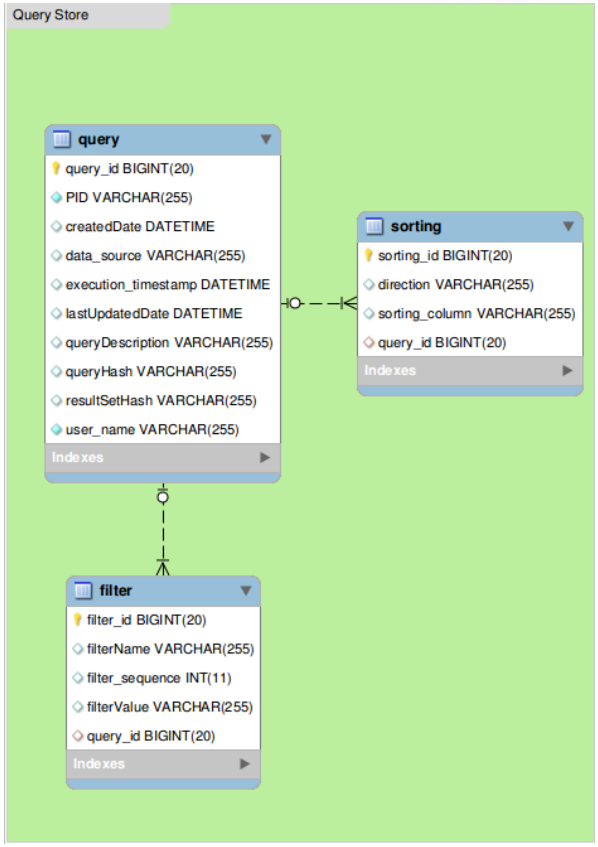
CSV Data Prototype

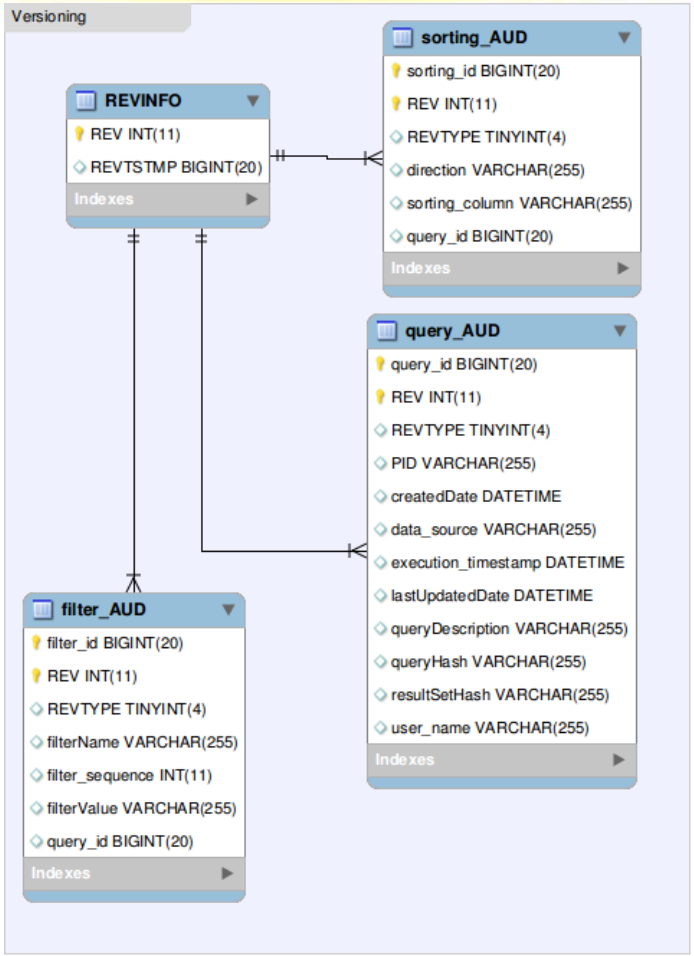


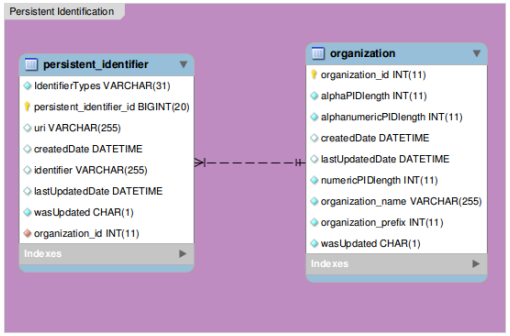
CSV Data Prototype



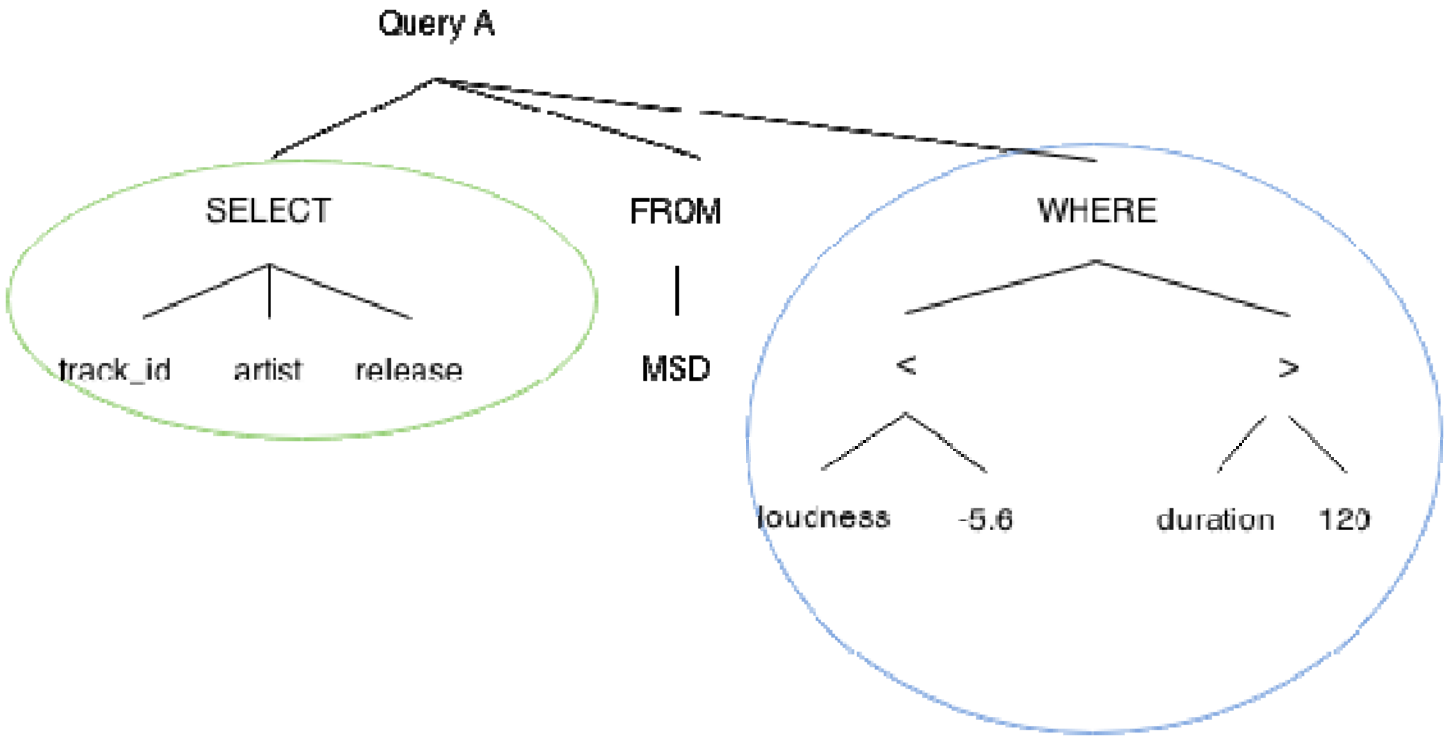
Query Store Database Scheme



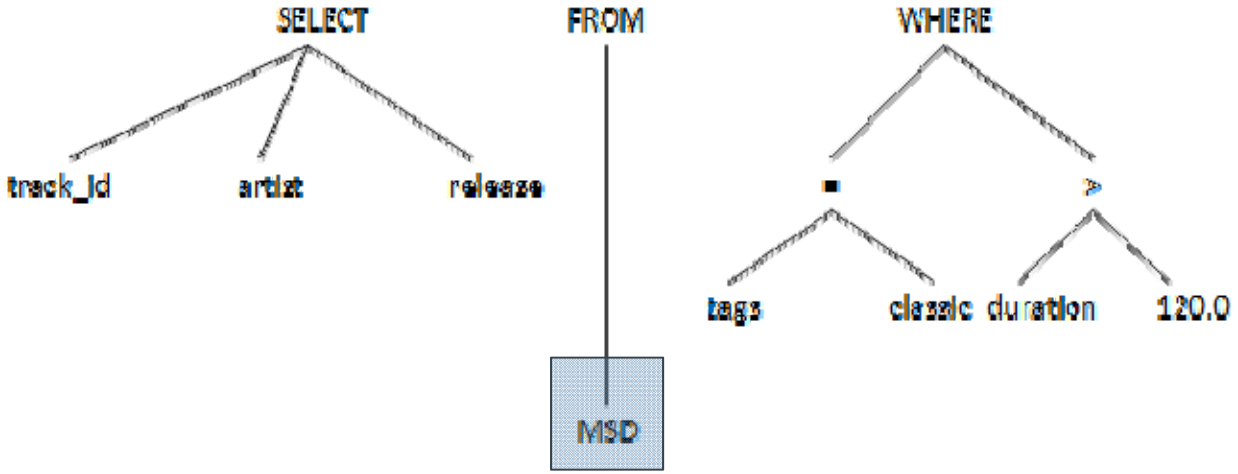




Storing Queries



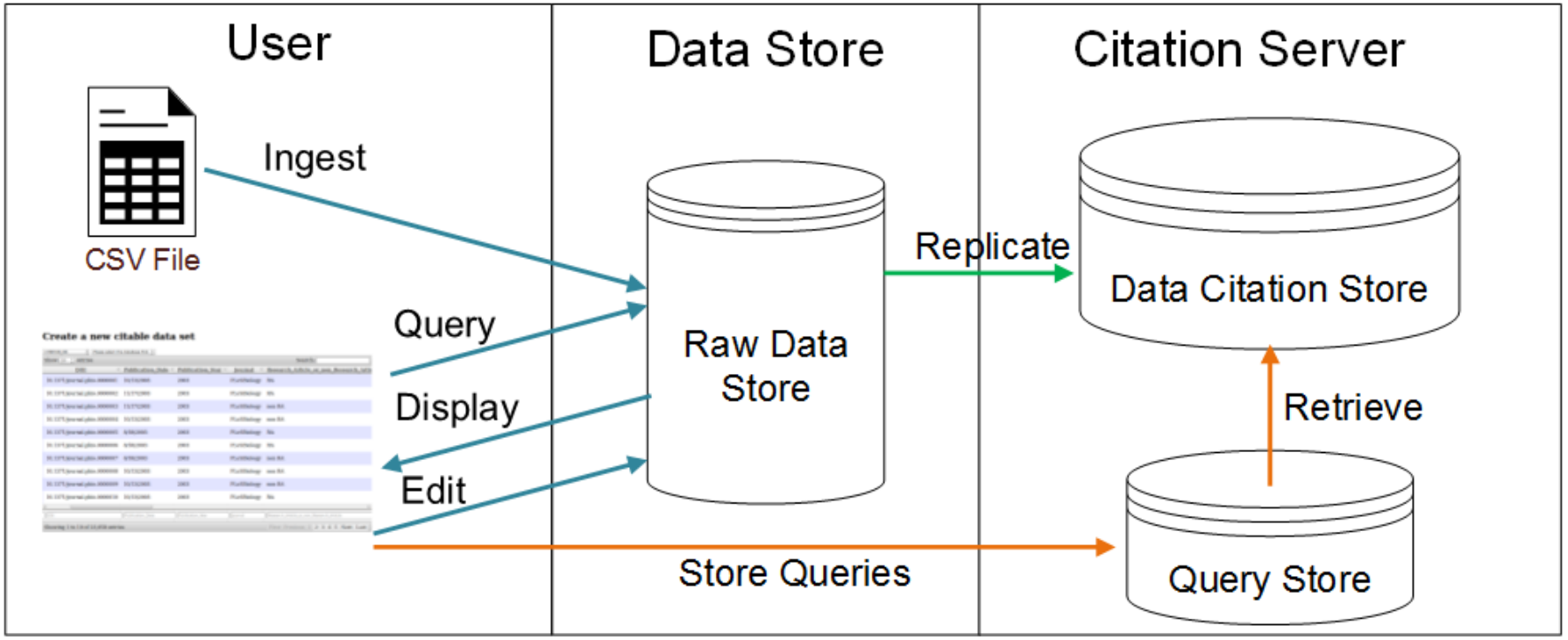
Retrieving Historical Resultsets




```
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;
```

Extension: Replicated Prototype for CSV Data



Modularized Approach

- Modules
  - CSV Tools
  - SQL Database Layer
  - Query Store
  - Web interface
  - ResultSet Verification
  - Persistent Identification

Near future

- Finalize module integration
- Landing page design
- Resolver service
- Apply real world use cases
  - Testing and evaluating

Dynamic Data Citation for XML Data

Stefan Pröll, Secure Business Austria

sproell@sba-research.org

Overview

- Introduction
- Goals
- Native XML Databases
- Next steps

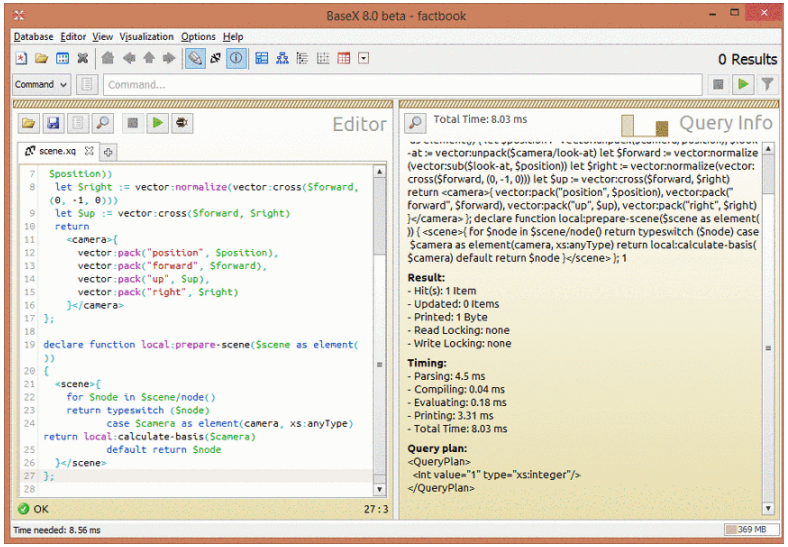
Dynamic Data Citation for XML Data

- **Goals**:
  - Cite arbitrary subsets of XML data
    - Subsets, nodes, attributes
  - Enable dynamic data
  - Utilize query available languages

- **Why XML data?**
  - Used in many different settings
  - Complex structures possible
  - Schema available

XML Data Approaches

- **Apply data citation framework**
  - Add metadata for versioning
  - Mark insert, update and delete operations
  - No actual deletes



- **Approaches**
  - Copy branches upon updates and deletes
    - Simple approach, but uses storage space
  - Introduce parent/child relationships and
    - Resolution more complex

Dynamic Data Citation for XML Data

- **XML Database: Base X**
  - Lightweight system
  - Client/Server architecture,
  - ACID safe transactions
  - XPath/XQuery 3.0 Processor
  - Scalable
- **E.g. used in UK Data Service Use Case**
  - Textual transcripts in XML format
  - Unique identification of (sub) sections

XML Database Implementation Concept

- **Adapt XQuery parser**
  - Rewrite operations for
    - Inserting
    - Deleting
    - Replacing
    - Renaming
- Reuse query parser for alternative implementations
  - E.g. eXistDB

XML Database Implementation Concept

Current Status and Future Work

- Student working on implementation as a thesis project
  - Copy branch approach implemented
- XQuery parser supports basic operations
  - Provides modules for XBase and eXistDB
- **Future work:**
  - Parent/child approach
  - Comfortable interface for access
  - Real world use cases/datasets

Pilots

National Supercomputer Facility (NCI), Australia

Adrian Burton, Australian National Data Service
adrian.burton@ands.org.au

Results from VAMDC Workshop

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

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

Changes adopted following the workshop

- An existing dataset will never be deleted nor modified
- If a correction and/or addition to an existing data node are/is needed, this will be associated with the creation of a new dataset
- Automatically maintain the genealogy for the families of datasets
- Users and data-providers will be able to know the creation date of a dataset, its ancestor and its descendants

VAMDC

- If new datasets enrich the VAMDC infrastructure, a user will always be able to obtain exactly the same results he/she obtained at a given date, by restricting his/her query on the datasets already existing at that date
- The data-set ID will be returned in each result file coming from the VAMDC infrastructure. A user will always know what are the datasets used for satisfying his/her query and may easily cite it
- The XSAMS format (the VAMDC standard for formatting the results) will be modified to natively include references to the datasets used for its composition

Agenda

- Intro: Goals, approach and progress so far
- Reports from pilots
  - Results from the NERC workshop
  - CLARIN presentation
  - MSD presentation
  - XML data
  - National Supercomputing Facility Australia
  - Results from the VAMDC workshop
  - <others?>
- Open challenges
  - Time stamping and versioning
  - Hash Key calculation / Result set verification
  - Data citation Zoo and request for examples
- Discussion & future plans

PID Assignment

- PID assigned to a query identifying a new dataset
- When to assign an existing/new PID to a query?
  - **Existing PID**: Identical query (semantics) with identical result set, i.e. no change to any element touched upon by query since first processing of the query
  - **New PID**: whenever query semantics is not absolutely identical (irrespective of result set being potentially identical!)
- Note:
  - Identical result set alone does not mean that the query semantics is identical
  - Will assign different PIDs to capture query semantics
  - Need to normalize query to allow comparison
    -> query re-writing

Query Re-Writing

- Query rewriting to ensure unique sort of result list
  - Most databases are set-based
  - Most subsequent processing is sequence-based
  - Need to re-write query to apply unique sort on any table prior to applying any user-defined sort for repeatability

- Hashing of result set to verify identity of result
  - Compute over entire result set: comprehensive, potentially slow
  - Computer over column headers and row IDs:
    - verifies correctness of attributes and data items selected
    - does not safeguard against unmonitored changes to attribute values
  - Other? Is this an issue?

Timestamping

- Which timestamp to assign to new query?
  - Timestamp of **query processing**
  - Timestamp of **last change** to DB (global)
  - Timestamp of **last change to result set** touched upon by query (including deletes)

    most complex approach in terms of query re-writing required to select with deletes, extract latest TS, then filter

Agenda

- Intro: Goals, approach and progress so far
- Reports from pilots
  - Results from the NERC workshop
  - CLARIN presentation
  - MSD presentation
  - XML data
  - Results from the VAMDC workshop
  - <others?>
- Open challenges
  - Time stamping and versioning
  - Hash Key calculation / Result set verification
  - Data citation Zoo and request for examples
- Discussion & future plans