Linking to and Citing Data in non-trivial Setings

Andreas Rauber
rauber@ifs.tuwien.ac.at

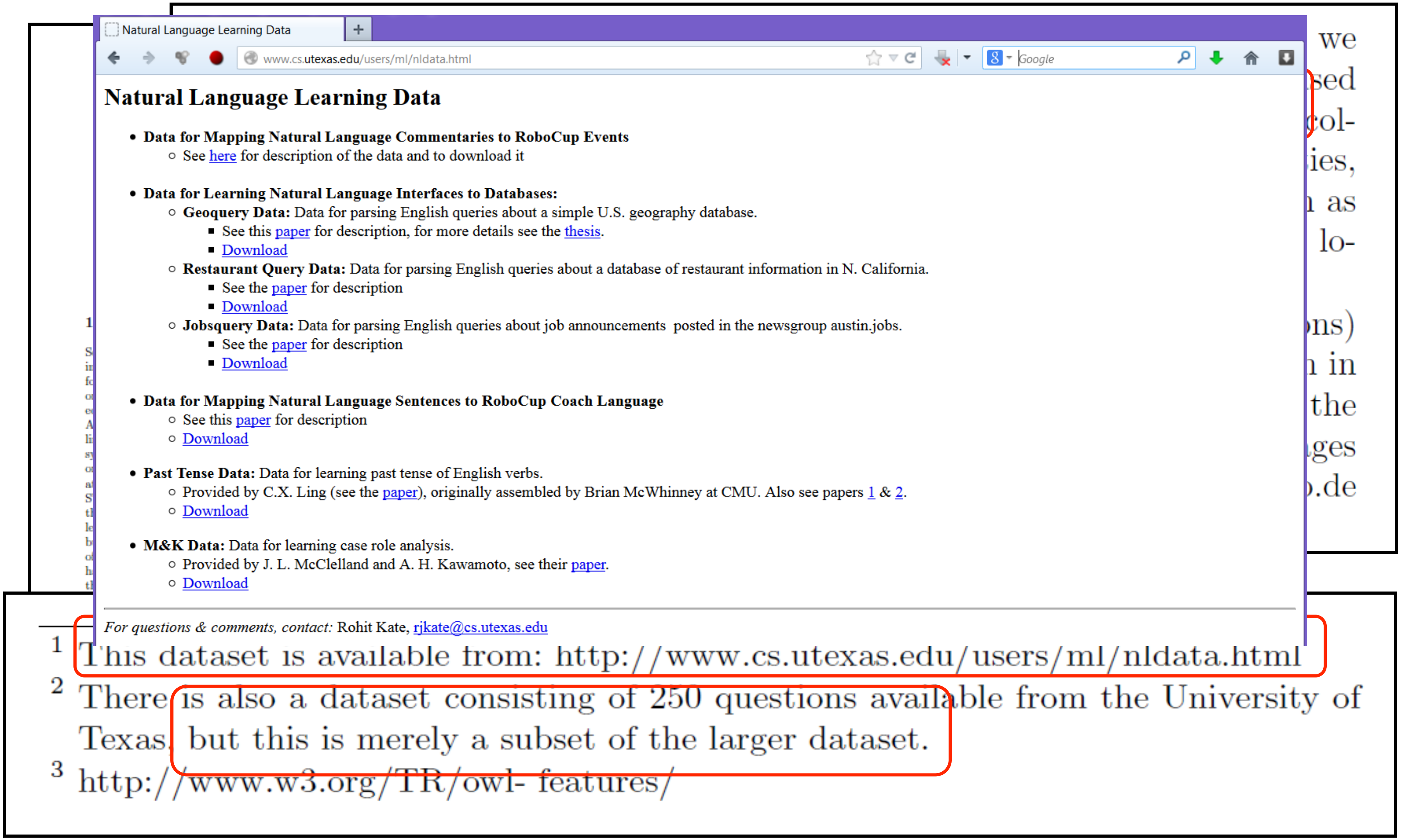
Outline

- What are the challenges in citing dynamic data?
  - Data Citation: the status quo and requirements
- How can we enable precise citation of dynamic data?
  - Making Dynamic Data Citeable
- A prototype solution for SQL
  - Solution and challenges
- Does this work for all data?
  - Next steps, open issues, and the RDA working Group

Dynamic Data Citation

- So far citable datasets have to be static
  - Fixed set of data, no changes:
    no corrections to errors, no new data being added
  - Sometimes solved by assigning version numbers or releases
- But: research data is dynamic
  - Correcting errors, adding new data, enhancing data quality, …
  - Changes sometimes highly dynamic, at irregular intervals
- Granularity?
  - Researchers use specific subsets of data
  - Storing a copy of every subset does not scale
  - Assigning PIDs to every record does not scale
  - How to create specific subsets?
  - How to reference subsets in a dynamic environment?

How should we cite data?



Natural Language Learning Data
www.cs.utexas.edu/users/ml/nldata.html
Google

Natural Language Learning Data

- **Data for Mapping Natural Language Commentaries to RoboCup Events**
  - See here for description of the data and to download it

- **Data for Learning Natural Language Interfaces to Databases:**
  - **Geoquery Data:** Data for parsing English queries about a simple U.S. geography database.
    - See this paper for description, for more details see the thesis.
    - Download
  - **Restaurant Query Data:** Data for parsing English queries about a database of restaurant information in N. California.
    - See the paper for description
    - Download
  - **Jobsquery Data:** Data for parsing English queries about job announcements posted in the newsgroup austin.jobs.
    - See the paper for description
    - Download

- **Data for Mapping Natural Language Sentences to RoboCup Coach Language**
  - See this paper for description
  - Download

- **Past Tense Data:** Data for learning past tense of English verbs.
  - Provided by C.X. Ling (see the paper), originally assembled by Brian McWhinney at CMU. Also see papers 1 & 2.
  - Download

- **M&K Data:** Data for learning case role analysis.
  - Provided by J. L. McClelland and A. H. Kawamoto, see their paper.
  - Download

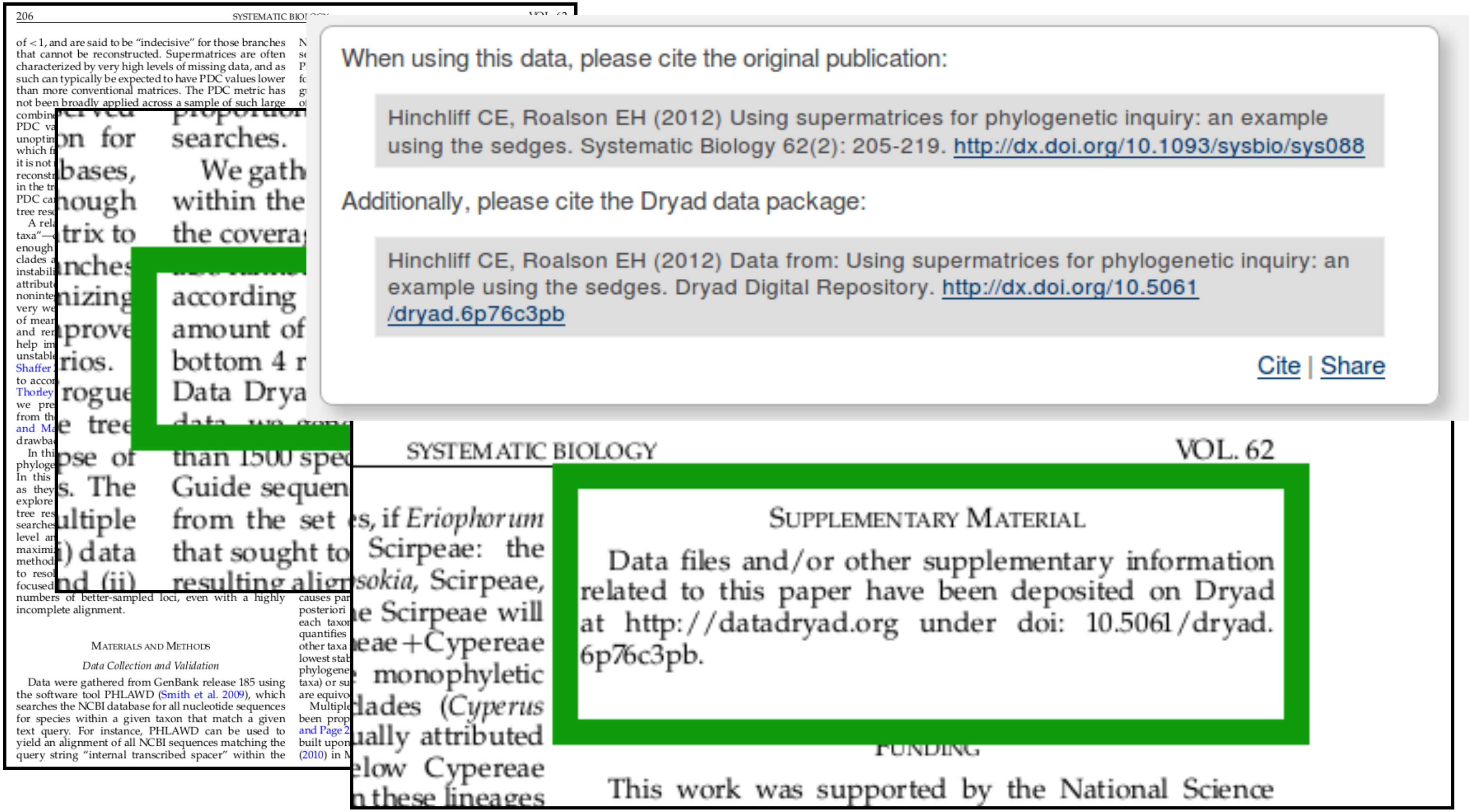
For questions & comments, contact: Rohit Kate, rjkate@cs.utexas.edu

[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/

How should we cite data?



206 SYSTEMATIC BIOLOGY

of < 1, and are said to be "indecisive" for those branches that cannot be reconstructed. Supermatrices are often characterized by very high levels of missing data, and as such can typically be expected to have PDC values lower than more conventional matrices. The PDC metric has not been broadly applied across a sample of such large

numbers of better-sampled loci, even with a highly incomplete alignment.

MATERIALS AND METHODS

Data Collection and Validation

Data were gathered from GenBank release 185 using the software tool PHLAWD (Smith et al. 2009), which searches the NCBI database for all nucleotide sequences for species within a given taxon that match a given text query. For instance, PHLAWD can be used to yield an alignment of all NCBI sequences matching the query string "internal transcribed spacer" within the

When using this data, please cite the original publication:

Hinchliff CE, Roalson EH (2012) Using supermatrices for phylogenetic inquiry: an example using the sedges. Systematic Biology 62(2): 205-219. http://dx.doi.org/10.1093/sysbio/sys088

Additionally, please cite the Dryad data package:

Hinchliff CE, Roalson EH (2012) Data from: Using supermatrices for phylogenetic inquiry: an example using the sedges. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.6p76c3pb

Cite | Share

SYSTEMATIC BIOLOGY VOL. 62

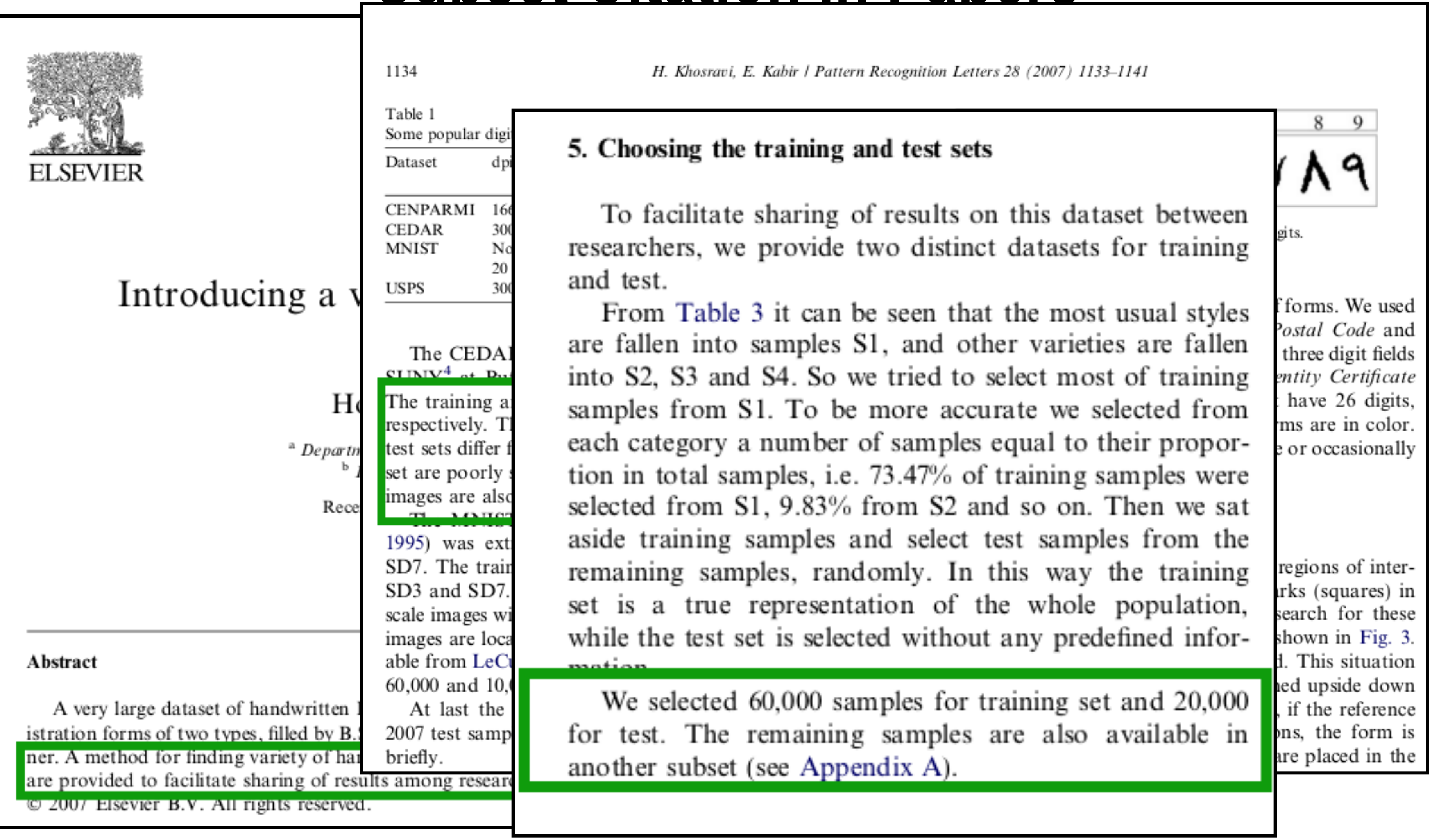
SUPPLEMENTARY MATERIAL

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

Subset Citation in Papers



ELSEVIER

Introducing a v

H

a Departm

b

Rece

Abstract

A very large dataset of handwritten
istration forms of two types, filled by B.
ner. A method for finding variety of ha
are provided to facilitate sharing of results among resear
© 2007 Elsevier B.V. All rights reserved.

Table 1
Some popular digi

Dataset	dpi
CENPARMI	166
CEDAR	300
MNIST	No 20
USPS	300

The CEDA
SUNY4 at Bu
The training a
respectively. T
test sets differ f
set are poorly
images are also
The MNIST
1995) was ext
SD7. The train
SD3 and SD7.
scale images wi
images are loca
able from LeCu
60,000 and 10,0
At last the
2007 test samp
briefly.

8 9

gits.

f forms. We used
Postal Code and
three digit fields
entity Certificate
have 26 digits,
rms are in color.
or occasionally

regions of inter-
rks (squares) in
search for these
shown in Fig. 3.
l. This situation
ed upside down
, if the reference
ons, the form is
are placed in the

5. Choosing the training and test sets

To facilitate sharing of results on this dataset between researchers, we provide two distinct datasets for training and test.

From Table 3 it can be seen that the most usual styles are fallen into samples S1, and other varieties are fallen into S2, S3 and S4. So we tried to select most of training samples from S1. To be more accurate we selected from each category a number of samples equal to their proportion in total samples, i.e. 73.47% of training samples were selected from S1, 9.83% from S2 and so on. Then we sat aside training samples and select test samples from the remaining samples, randomly. In this way the training set is a true representation of the whole population, while the test set is selected without any predefined information.

We selected 60,000 samples for training set and 20,000 for test. The remaining samples are also available in another subset (see Appendix A).

Khosravi, Hossein, and Ehsanollah Kabir. "Introducing a very large dataset of handwritten Farsi digits and a study on their varieties." Pattern Recognition Letters 28.10 (2007): 1133-1141.

Data Citation
Current Approaches

- Persistent Identifier (PID) e.g. DOI, URI, ARK, … currently provided for
  - entire data sets, copies of subsets
  - static data, sometimes release of versions
  - cited in their entirety with textual description of subsets
- This is insufficient in many settings
  - imprecise
  - not machine-actionable
  - not scalable for large data sets
  - insufficient support for data that changes
  - insufficient support for arbitrary subsets (rows/columns)

Data Citation – Requirements for Citing

- Arbitrary subsets of data
  - rows/columns, time sequences, …
  - from single number to virtually the entire set
- Changing data
  - corrections, additions, …
- 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

Outline

- What are the challenges in citing dynamic data?
  - Data Citation: the status quo and requirements
- How can we enable precise citation of dynamic data?
  - Making Dynamic Data Citeable
- A prototype solution for SQL
  - Solution and challenges
- Does this work for all data?
  - Next steps, open issues, and the RDA working Group

Making Dynamic Data Citeable

Data Citation: Data + Means-of-access

- Data → time-stamped & versioned

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, timestamping
  - **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

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 re-writing needed to
  - **Standardization/Normalization** of query to help with identifying semantically identical queries
  - Re-write to **adapt to versioning approach** chosen (versioning in operational tables, separate history table, …)
  - **Add timestamp** to any select statement in query
  - Potentially re-write to **identify last change to result set touched** upon (i.e. select including elements marked deleted, check most recent timestamp, to determine correct PID assignment)
  - **Apply unique sort** to any table touched upon in query prior to query to ensure unique sort

Query Re-Writing

- **Normalization of query string**
  - Upper / lower case spelling
  - Sorting of filtering criteria
    (order does not influence result semantics)
  - Compute hash-key over query string to identify whether identical query has been issued already
  - If identical query found, re-run and check for changes in result set based on time-stamps of data records added/deleted
  - If different, assign new PID, otherwise existing PID

Query Re-Writing

- 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

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

Making Dynamic Data Citeable

- Building blocks of supporting dynamic data citation:
  - Uniquely identifiable data records
  - Versioned data, marking changes as insertion/deletion
  - Time stamps of data insertion / deletions
  - “Query language” for constructing subsets
- Add modules:
  - Persistent query store: queries and the timestamp (either: <when issued> or <of last change to data>)
  - Query rewriting module
  - PID assignment for queries that enables access
- Stable across data source migrations (e.g. diff. DBMS), scalable, machine-actionable

Data Citation – Deployment

- Researcher uses workbench to identify subset of data
- Upon executing selection („download“) user gets
  - Data (pac
  - PID (e.g.
  - Hash valu
  - Recommended citation text (e.g. bTeX)

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

- PID resolves to landing page
  - Provides detailed metadata, link t 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

Outline

- What are the challenges in citing dynamic data?
  - Data Citation: the status quo and requirements
- How can we enable precise citation of dynamic data?
  - Making Dynamic Data Citeable
- A prototype solution for SQL
  - Solution and challenges
- Does this work for all data?
  - Next steps, open issues, and the RDA working Group

Prototype Implementation

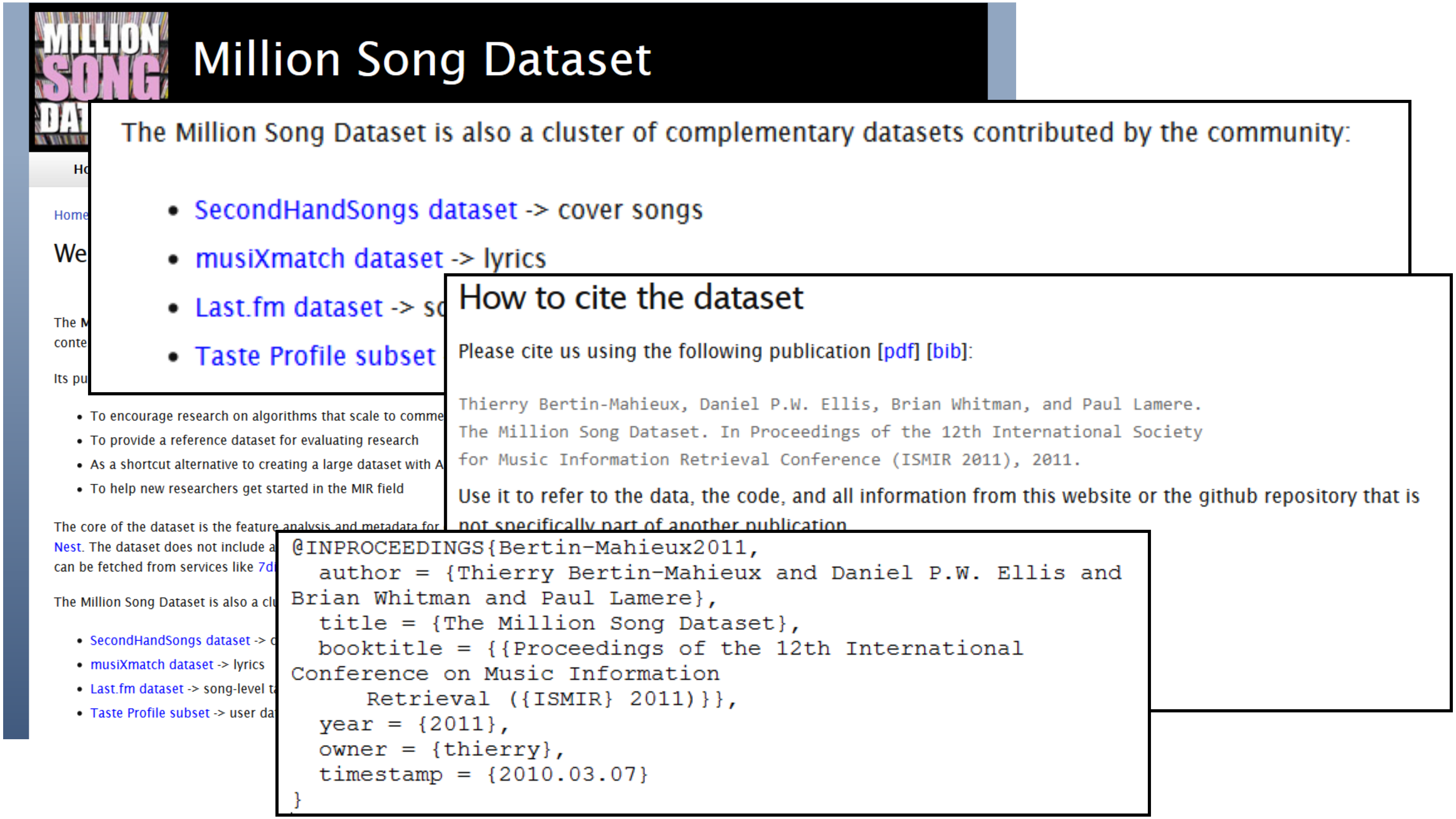
- LNEC Laboratory of Civil Engineering, Portugal
- Monitoring dams and bridges
- 31 manual sensor instruments
- 25 automatic sensor instruments
- Web portal
  - Select sensor data
  - Define timespans
- Report generation
  - Analysis processes, produces
  - Latex, produces
  - PDF report

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

Prototye Implementation

- Million Song Dataset
  http://labrosa.ee.columbia.edu/millionsong/
- Larges benchmark collection in Music Retrieval
- Original set provided by Echonest
- No audio, only set of features
- 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,…

Prototype Implementation



MILLION SONG DAT

Million Song Dataset

Ho

Home

We

The M
conte

Its pu

- To encourage research on algorithms that scale to comme
- To provide a reference dataset for evaluating research
- As a shortcut alternative to creating a large dataset with A
- To help new researchers get started in the MIR field

The core of the dataset is the feature analysis and metadata for
Nest. The dataset does not include a
can be fetched from services like 7d

The Million Song Dataset is also a clu

- SecondHandSongs dataset -> c
- musiXmatch dataset -> lyrics
- Last.fm dataset -> song-level t
- Taste Profile subset -> user da

The Million Song Dataset is also a cluster of complementary datasets contributed by the community:

- SecondHandSongs dataset -> cover songs
- musiXmatch dataset -> lyrics
- Last.fm dataset -> so
- Taste Profile subset

How to cite the dataset

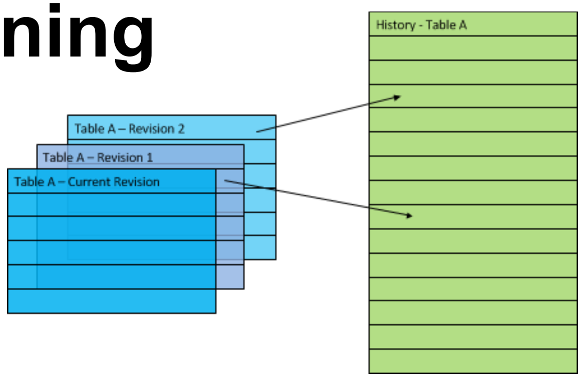
Please cite us using the following publication [pdf] [bib]:

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

Use it to refer to the data, the code, and all information from this website or the github repository that is not specifically part of another publication

```
@INPROCEEDINGS{Bertin-Mahieux2011,
  author = {Thierry Bertin-Mahieux and Daniel P.W. Ellis and
Brian Whitman and Paul Lamere},
  title = {The Million Song Dataset},
  booktitle = {{Proceedings of the 12th International
Conference on Music Information
      Retrieval ({ISMIR} 2011)}},
  year = {2011},
  owner = {thierry},
  timestamp = {2010.03.07}
}
```

Time-Stamping and Versioning

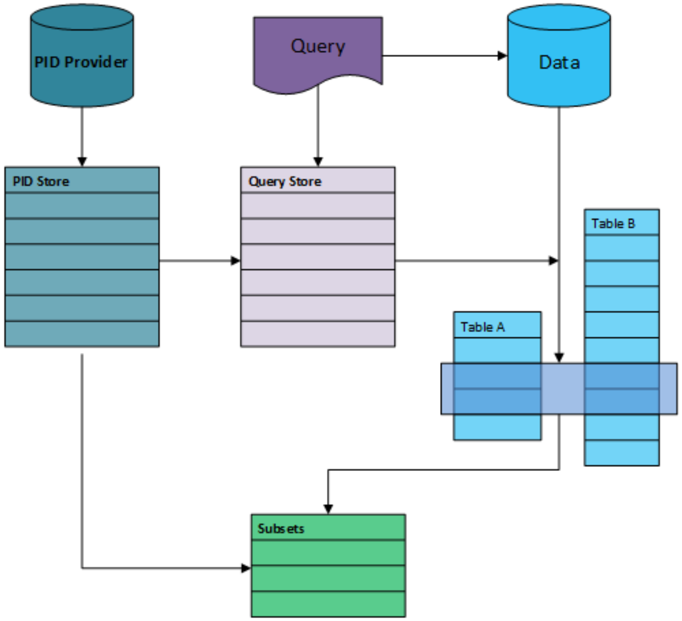


- Integrated
  - Extend original tables by temporal metadata
  - Expand primary key by version column
- Hybrid
  - Utilize history table for deleted record versions with metadata
  - Original table reflects latest version only
- Separated
  - Utilizes full history table
  - Also inserts reflected in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

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

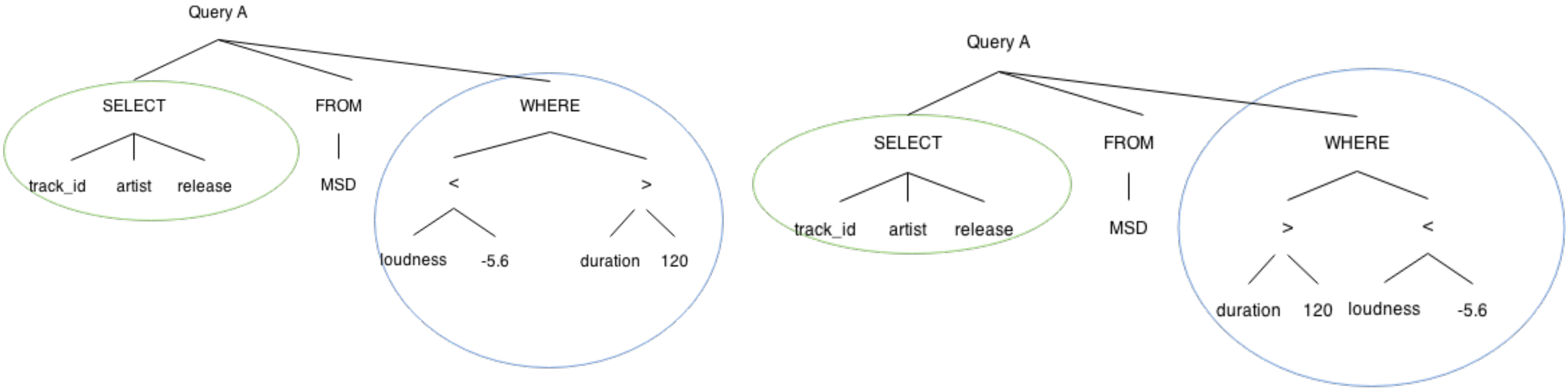
Storing Queries

- Add query store containing
  - PID of the query
  - Original query
  - Re-written query + query string hash
  - Timestamp (as included 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)



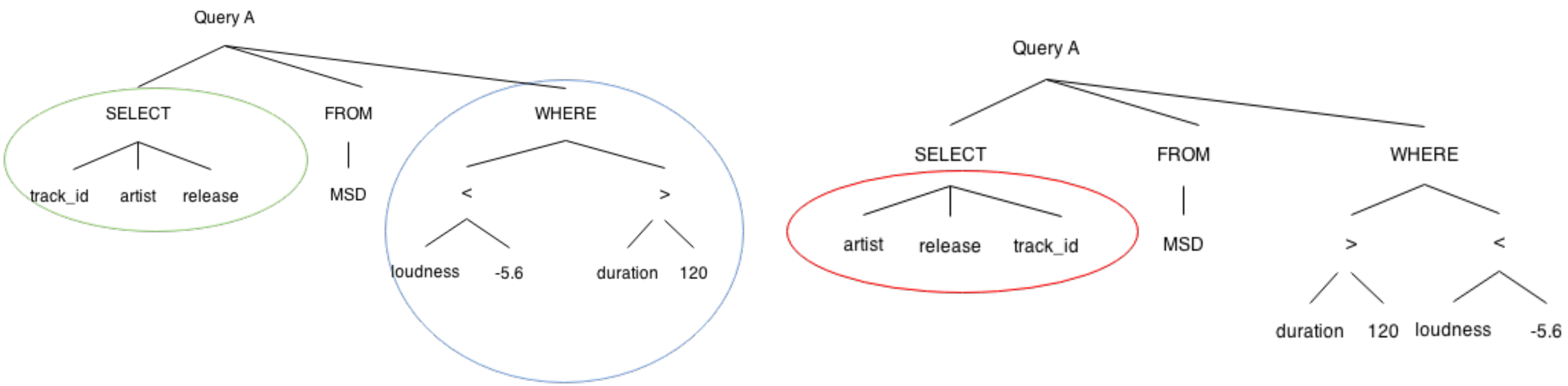
Query Re-Writing

- Normalizing queries to detect identical queries
  - WHERE clause sorted
  - Calculate query string hash
  - Identify semantically identical queries

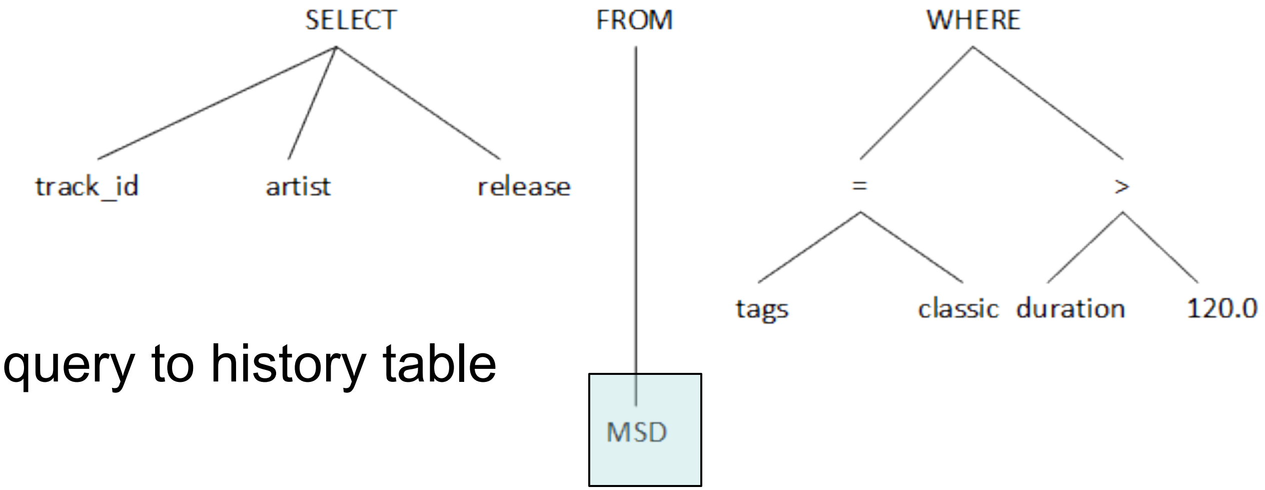


Query Re-Writing

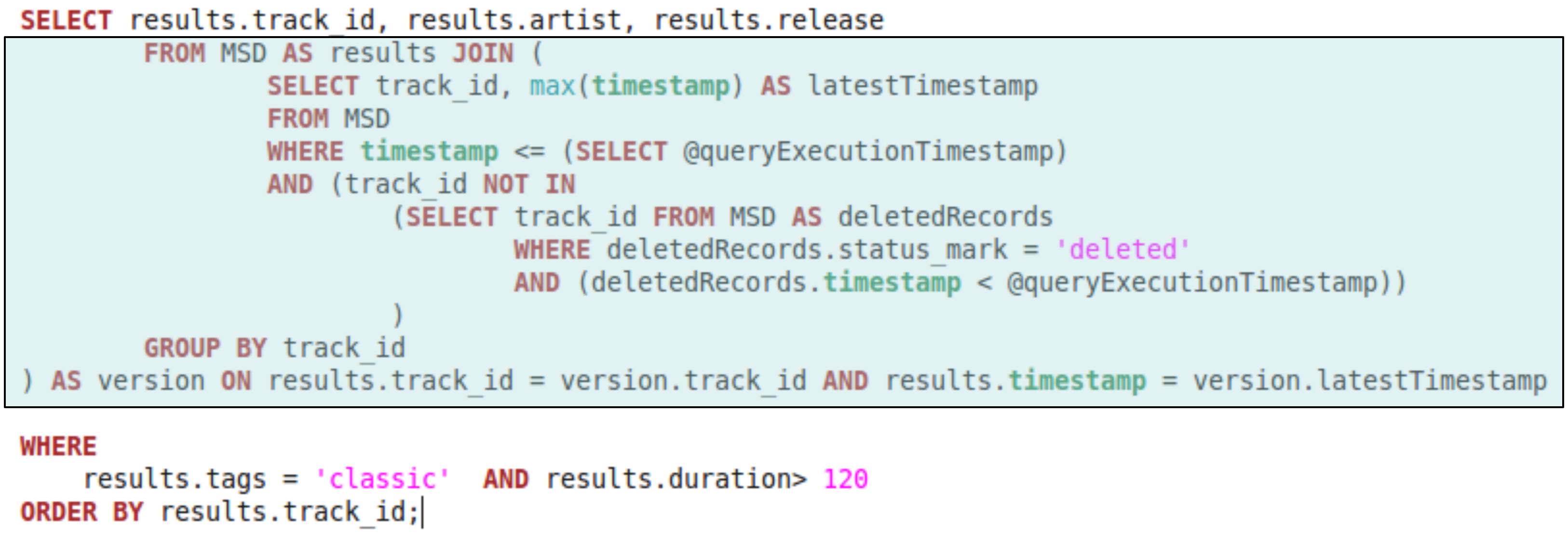
- Normalizing queries to detect identical queries
  - WHERE clause sorted
  - Calculate query string hash
  - Identify semantically identical queries
  - → non-identical queries: columns in different order



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;
```

Outline

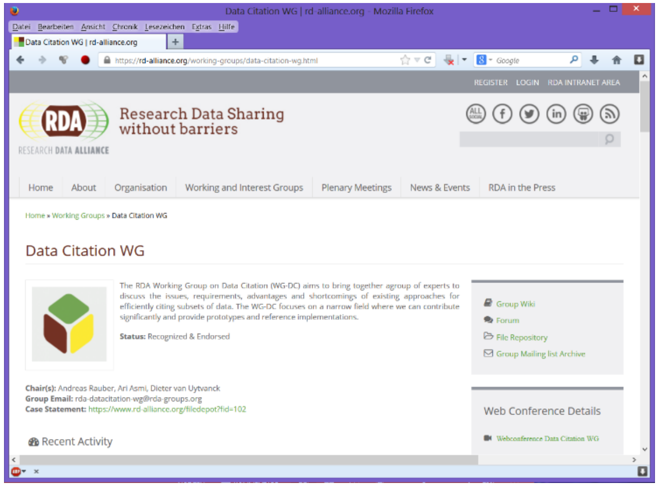
- What are the challenges in citing dynamic data?
  - Data Citation: the status quo and requirements
- How can we enable precise citation of dynamic data?
  - Making Dynamic Data Citeable
- A prototype solution for SQL
  - Solution and challenges
- Does this work for all data?
  - Next steps, open issues, and the RDA working Group

Data Citation: Next steps

- Solution devised for SQL -> expand to other data types
  - SQL: LNEC, MSD
  - Pilot for CSV: MSD
  - Analyze how to make XML and RDF time-stamped, versioned
- Verify pilots conceptually
  - Does it work?
  - Impact on data center (size, operations, APIs, …) specifically: how to realize versioning
  - How to integrate in workbenches?
- Implement several pilots and verify
- Test stability under migrations of data management systems

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 **dynamic (changing) datasets**
  - Focus!
  - Liaise with other WGs on attribution, metadata, …
  - Liaise with other 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
RDA
RESEARCH DATA ALLIANCE
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.

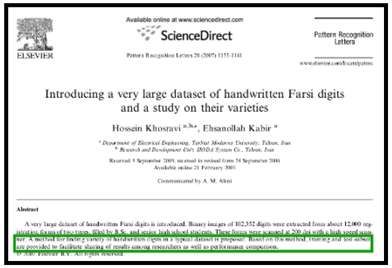
Status: Recognized & Endorsed

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

Join RDA and Working Group

If you are interested in joining the discussion, 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



Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters 28 (2007) 1133–1141

Pattern Recognition Letters

www.elsevier.com/locate/patrec

ELSEVIER

Introducing a very large dataset of handwritten Farsi digits and a study on their varieties

Hossein Khosravi [a,b,*], Ehsanollah Kabir [a]

[a] Department of Electrical Engineering, Tarbiat Modarres University, Tehran, Iran
[b] Research and Development Unit, HODA System Co., Tehran, Iran

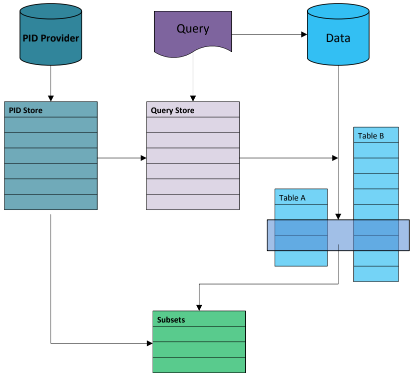
Received 5 September 2005; received in revised form 24 September 2006
Available online 21 February 2007

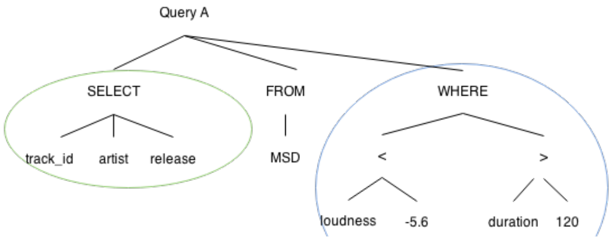
Communicated by A. M. Alimi

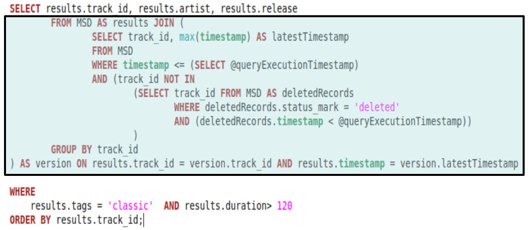
Abstract

A very large dataset of handwritten Farsi digits is introduced. Binary images of 102,352 digits were extracted from about 12,000 registration forms of two types, filled by B.Sc. and senior high school students. These forms were scanned at 200 dpi with a high speed scanner. A method for finding variety of handwritten digits in a typical dataset is proposed. Based on this method, training and test subset are provided to facilitate sharing of results among researchers as well as performance comparison.

© 2007 Elsevier B.V. All rights reserved.



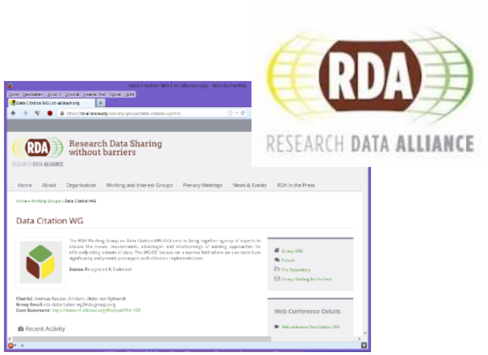




```
SELECT results.track_id, results.artist, results.release
        FROM MSD AS results JOIN (
                SELECT track_id, max(timestamp) AS latestTimestamp
                FROM MSD
                WHERE timestamp <= (SELECT @queryExecutionTimestamp)
                AND (track_id NOT IN
      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

Thank you for your attention.



Data Citation WG | rd-alliance.org - Mozilla Firefox

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

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

Data Citation WG

The RDA Working Group on Data Citation (WG-DC) aims to bring together a group 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

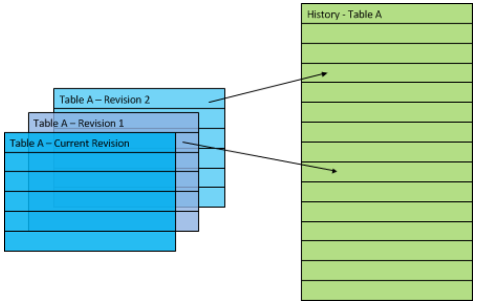
Chair(s): Andreas Rauber, Ari Asmi, Dieter van Uytvanck
Group Email: rda-datacitation-wg@rda-groups.org
Case Statement: https://www.rd-alliance.org/filedepot?fid=102

Group Wiki
Forum
File Repository
Group Mailing list Archive

Web Conference Details

Webconference Data Citation WG

Recent Activity



Literature and Links

- http://www.dlib.org/dlib/march07/altman/03altman.html
- http://www.dcc.ac.uk/resources/how-guides/cite-datasets
- http://www.dlib.org/dlib/january11/starr/01starr.html
- http://dx.doi.org/10.1109%2F2.901164
- http://www.doi.org/factsheets/DOIKeyFacts.html
- http://www.datacite.org
- http://www.handle.net
- http://www.mpi.nl/DAM-LR/meeting5/Persistent%20Identifiers.pdf
- https://wiki.ucop.edu/display/Curation/ARK
- http://www.doi.org/factsheets/DOIHandle.html
- http://n2t.net/ezid/home/understanding
- http://sagecite.knowledgeblog.org/2011/07/28/why-do-we-need-datacitation

Further Pointers

- http://www.ariadne.ac.uk/issue56/tonkin
- http://ands.org.au/guides/persistent-identifiers-working.html
- http://hdl.handle.net/
- http://dx.doi.org/
- http://www.dcc.ac.uk/resources/how-guides/appraise-select-data