A Scalable Framework for Dynamic Data Citation

Stefan Pröll\(^1\)\(^,\)\(^2\) and Andreas Rauber\(^2\)

Motivation
- Precise data set citation for reuse and verification of experiments
- Increasingly large data sets require subset mechanisms enabling precise identification
- Evolving and dynamic data needs to be referenceable

Current Approaches and their Shortcomings
- Redundant copies of data for each experiment do not scale
- Textual descriptions are neither precise nor machine actionable
- Versioned releases are not fine granular enough

Requirements for Data Citation
1) Arbitrary subsets must be citable
2) Dynamic data (inserts, updates and deletes) can be handled
3) Scalability for huge volumes of data
4) Transparent implementation
5) Machine actionability
6) Verifiable and auditable
7) Provenance metadata

The Proposed Solution
- Data store is time-stamped
- Data store is versioned (2)
- Unique (primary) key for all records
- Queries used for subset creation are (1,3)
  - uniquely sorted
  - time-stamped with the execution date
  - stored in a query store
- Hash keys are computed for verification purposes (6)
- PIDs get assigned that can be used in references
- PID resolution leads to landing page that re-executes query and returns identical subset of the version (4, 5,7)

Dynamic Data Citation for Relational Databases
1) Adapt database tables for storing time stamped data
2) Implement stored procedures or triggers for versioning
3) Ensure unique sorting of the data
4) Store queries which created the subset
5) Hash the result set
6) Store queries with metadata
7) Assign persistent identifier (e.g. DOI) to the queries
8) Retrieve subset again by querying historical data

Enabling Data Citation for a Diverse Range of Domains
- Many scientific disciplines still rely on simple data formats such as CSV, JSON, XML...
- Such formats are often only used for smaller scale data sets, but now also bigger sets need proper citation
- We currently develop a Web based prototype which manages uploaded CSV or native SQL data and allows to cite subsets.
- Trace creation process of data sets and store the provenance data
- Provide flexible citation mechanism including result set hashing
- Enables query based subsetting and data sharing

Acknowledgements

Supported by the projects APARSEN, TIMBUS and SCAPE, partially funded by the EU under the FP7 contracts 269977, 269940 and 270137.

Contacts
- www.sba-research.org/~sproell
- http://ifs.tuwien.ac.at/dp
- @stefanproell

A Query Centric Data Citation Model

A Query Centric Data Citation Model

Enabling Data Citation for a Diverse Range of Domains

- Many scientific disciplines still rely on simple data formats such as CSV, JSON, XML...
- Such formats are often only used for smaller scale data sets, but now also bigger sets need proper citation
- We currently develop a Web based prototype which manages uploaded CSV or native SQL data and allows to cite subsets.
- Trace creation process of data sets and store the provenance data
- Provide flexible citation mechanism including result set hashing
- Enables query based subsetting and data sharing

Acknowledgements

Supported by the projects APARSEN, TIMBUS and SCAPE, partially funded by the EU under the FP7 contracts 269977, 269940 and 270137.

Contacts
- www.sba-research.org/~sproell
- http://ifs.tuwien.ac.at/dp
- @stefanproell