Data Citation of Evolving Data

Recommendations of the Working Group on Data Citation (WGDC)

Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll

Draft – Request for Comments

Revision of May 7th 2015

I. Making Data Citable

These WGDC recommendations enable researchers and data centers to identify and cite data used in experiments and studies. Instead of providing static data exports or textual descriptions of data subsets, we support a dynamic, query centric view of data sets. The proposed solution enables precise identification of the very set and version of data used, supporting reproducibility of processes, sharing and reuse of data.

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends to solve this challenge by:

- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

II. WG Recommendations

To realise the goal of rendering arbitrary data sets citeable, from single values to entire DBs in settings that range from static data to highly dynamic data streams, the WG recommends the following steps:

A. *Preparing the Data and the Query Store*

- **R1 – Data Versioning**: For retrieving earlier states of data sets the data needs to be versioned.
- **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 used to select data and associated metadata.

B. *Persistently Identify Specific Data sets*

When a data set should be persisted, the following steps need to be applied:

- **R4 – Query Uniqueness**: Re-write the query to a normalised 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 a 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 either 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.
- **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 normalised query, query & result set checksum, timestamp, superset PID, data set description and other) in the query store.
- **R10 – Citation Text**: Provide a recommended citation text and the PID to the user.

C. *Upon Request of a PID*

- **R11 – Landing Page**: PIDs should resolve to a human readable landing page of the data set, which provides metadata including a link to the superset (PID of the data source) and citation text snippet.
- **R12 – Machine Actionability**: the landing page should be machine-actionable and allow retrieving the data set by re-executing the timestamped query.

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), the queries and associated checksums need to be migrated.
- **R14 – Migration Verification**: Successful query migration should be verified by ensuring that queries can be re-executed correctly.

III. Benefits

The proposed solution has several benefits compared to current approaches relying on individual data exports for each data set or ambiguous natural language descriptions of data set characteristics.

- It allows identifying, retrieving and citing the precise data set with minimal storage overhead by only storing the versioned data and the queries used for creating the data set. In many environments data versioning is considered a best practice. Data sets can be re-created on demand.
- It allows retrieving the data both as it existed at a given point in time as well as the current view on it, by re-executing the same query with the stored or current timestamp, thus benefiting from all corrections made since the query was originally issued. This allows tracing changes of data sets over the time and comparing the effects on the result set.
- The query stored as a basis for identifying the data set provides valuable provenance information on the way the specific data set was constructed, thus being semantically more explicit than a mere data export.
- Metadata such as checksums support the verification of the correctness and authenticity of data sets retrieved.
- The recommendations are applicable across different types of data representation and data characteristics (big or small data; static or highly dynamic; identifying single values or the entire data set). If data is migrated to new representations, the queries can also be migrated, ensuring stability across changing technologies. Distributed data sources can be managed by relying on the local timestamps at each node, avoiding the need for expensive synchronization in loosely coupled systems.

IV. Frequently Asked Questions

- May data be deleted? Yes, of course, given appropriate policies. Queries may then not be re-executable anymore against the original timestamp anymore (but still against the current timestamp), specifically as the landing pages should persist. .
- Does the system need to store every query? No, only data sets that should be persisted for citation and later re-use need to be stored. Persisting queries can be decided individually or policy-based in an automated fashion.
- Can I obtain only the most recent data set? Queries can be re-executed with the original timestamp or with the current timestamp or any other timestamp desired. This allows retrieving the semantically identical data set but incorporating all changes, corrections or updates applied before the given timestamp.
- Which PID system should be used? Any PID system can, in principle, be applied according to the institutional policy. More information on PIDs can be found at the PID Interest Group[1].
- How are the queries created? Queries can either be created manually via an interface/workbench or applications create the proper queries automatically. Both methods require the adaption of the query by adding metadata and timestamps.
- How can I share parts of my database? The query centric view allows selecting any particular view or data subset of the data from the complete data set.
- How does this support giving credit and attribution? Attribution and giving credit is supported by providing a provenance chain from a subset/view of a data to the data set it was derived from, allowing to document intellectual contributions on the way. Note, however, that analysis and recommendations on how to aggregate bibliometrics and credits is not addressed in the context of this WG.

V. Next Steps

The set of recommendations is undergoing continuous evaluation in a series of pilots in different domains. It is also open for discussion and we encourage interested community members to participate and provide improvements, comments, suggestions and general feedback via the working space of the WG[2]. We are very interested in further real world use cases to act as evaluation pilots.

VI. Get Involved

You can find additional information RDA Working Group Page[3]. Please register on the mailing list to stay informed. The community feedback will be collected in the Wiki page[4].

[1] https://rd-alliance.org/groups/pid-interest-group.html
[2] https://rd-alliance.org/group/data-citation-wg/wiki/collaboration-environments.html
[3] www.rd-alliance.org/group/data-citation-wg.html
[4] https://rd-alliance.org/group/data-citation-wg/wiki/wgdc-dynamic-data-citation-recommendations.html