

# 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 September 24<sup>th</sup> 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 subset 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 solving 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*

Prepare existing data sources and provide the required infrastructure, which is needed for implementing the query based approach.

- **R1 – Data Versioning:** Apply versioning to ensure earlier states of data sets can be retrieved.
- **R2 – Timestamping:** Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- **R3 – Query Store Facilities:** Provide means for storing queries and the associated metadata in order to re-execute them in the future.

#### 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 that the sorting of the records in the data set is unambiguous and reproducible
- **R6 – Result Set Verification:** Compute fixity information (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 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 the time a user issued a query.
- **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 normalized query, query & result set checksum, timestamp, superset PID, data set description, and other) in the query store.
- **R10 – Automated Citation Texts:** Generate citation texts in the format prevalent in the designated community for lowering the barrier for citing the data. Include the PID into the citation text snippet.

### C. Resolving PIDs and Retrieving the Data

- **R11 – Landing Page:** Make the PIDs resolve to a human readable landing page that provides the data (via query re-execution) and metadata, including a link to the superset (PID of the data source) and citation text snippet.
- **R12 – Machine Actionability:** Provide an API / machine actionable landing page to access metadata and data via query re-execution.

### 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), migrate also the queries and associated fixity information.
- **R14 – Migration Verification:** Verify successful data and query migration, 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 subsets 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.
- The query store offers a valuable, central basis for analyzing data usage.
- Metadata such as checksums support the verification of 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 rely on local timestamps at each node, avoiding the need for expensive synchronization in loosely coupled systems.

## IV. FREQUENTLY ASKED QUESTIONS

- May data be deleted? Yes, given appropriate policies. Queries may then not be re-executable against the original timestamp anymore (but still against the current timestamp). Landing pages should persist.
- Does the system need to store every query? No. Only data sets that should be persisted for citation / 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 be applied according to the institutional policy.
- 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 via a provenance chain from a subset/view of data to the data set it was derived from, allowing to document intellectual contributions on the way. 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 evaluation in a series of pilots in different domains. We encourage interested community members to participate and provide improvements, comments, suggestions and general feedback via the working space of the WG<sup>1</sup>. We are very interested in further real world use cases to act as pilots.

## VI. GET INVOLVED

You can find additional information RDA Working Group Page<sup>2</sup>. Please register on the mailing list to stay informed. The community feedback will be collected in the Wiki page<sup>3</sup>.

<sup>1</sup> <https://rd-alliance.org/group/data-citation-wg/wiki/collaboration-environments.html>

<sup>2</sup> [www.rd-alliance.org/group/data-citation-wg.html](http://www.rd-alliance.org/group/data-citation-wg.html)

<sup>3</sup> <https://rd-alliance.org/group/data-citation-wg/wiki/wgdc-dynamic-data-citation-recommendations.html>