Scalable Dynamic Data Citation - Working Group

Working Group Chairs:
Andreas Rauber - Vienna University of Technology; Dieter Van Uytvanck- CLARIN; Ari Asmi- University of Helinski

Editors:
Herman Stehouwer - Max Planck Society
Yolanda Meleco - RDA/United States

Research Data Sharing Without Barriers

What is the Problem?

Digitally driven research is dependent on quickly evolving technology. As a result, many existing tools and collections of data were not developed with a focus on long term sustainability. Researchers strive for fast results and promotion of those results, but without a consistent and long term record of the validation of their data, evaluation and verification of research experiments and business processes is not possible.

To track data projects, data used in research needs to be precisely identified; however, researchers rarely use an entire dataset as it was provided to them. Instead, they select subsets of the entire dataset based on a specific time-range, a set of measurements, etc. Due to these factors, there is a strong need for data citation mechanisms that identify arbitrary subsets of large data sets with precision in a machine-actionable way, also known as scalable, dynamic data citation.

Researchers also need the ability to reliability cite data that is subject to change. Mechanisms should be put in place that allow researchers to cite data as it is used during a particular experiment. When data gets updated, modified or deleted, those changes should be reflected in the citation information, which include time-stamps and version history, and be recoverable by the citation system. These mechanisms need to be user-friendly, transparent, machine-actionable, scalable and applicable to various static and dynamic data types.

What was the Goal?

The goal of the Dynamic Data Citation Working Group is to ensure the identification of sub subsections of data at the time of addition, deletion or modification, regardless of their database management system (DMBS), to enable efficient processing of that data and linking from publications.

What is the Solution?

The solution to accomplish these goals include the following components:

- Version control for data collection, which ensures changes and deletions to data are assigned timestamps.
- Assignment of Persistent Identifier Definitions (PIDs) and timestamps to queries and expressions that identify a cited subset of data.
- Computation of hash keys for the selection result to ensure subsequent verification of identity.
- Consideration of potential problems that may occur as future process work on data sets and related sequences create barriers to retrieving the original cited data.

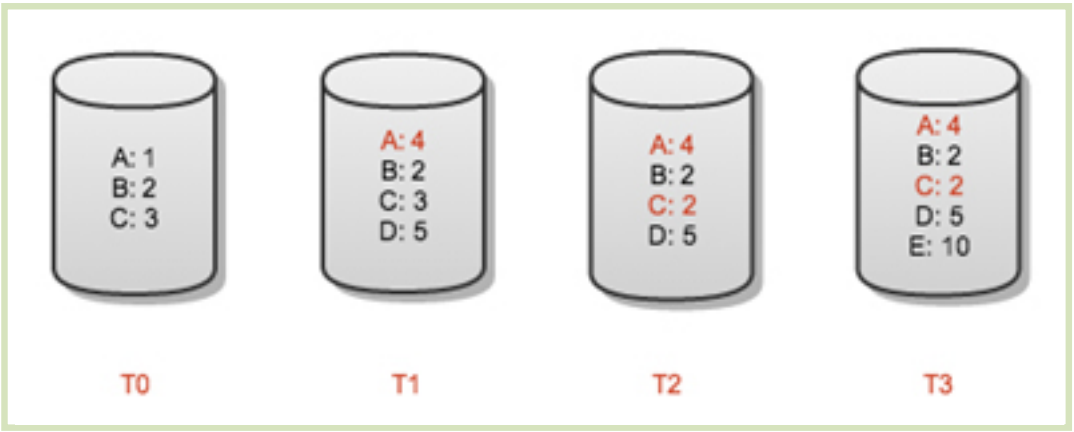

Figure 1: Shows change in simple database over time (updated values in red)

These components should be included across all DBMSs where there is a combination of data sources and operations that include subsets at specific points in time.

Although the exact technical implementation depends on existing local data structures and procedures, evaluations of numerous pilot projects involving various data types (SQL, CSV, XML) indicate the success of this solution.

What is the Impact?

The main impact of this solution is ensuring reproducibility of scientific research by allowing for a database to be dynamically updated when information is added, updated or deleted, while still enabling for the reproduction of time-specific data conditions. The approach detailed above has several advantages over current practices, including storing subsets as redundant data deposits. Additionally, by having the query/ expression as a basis for identifying the dataset, valuable semantic information is provided on the way the specific dataset was constructed, as opposed to merely having a data dump.

Furthermore, scalable dynamic data citation allows the user to re-execute the query with the original time stamp and retrieve the original data, while also obtain the current version of the data with all additions and changes. This enables them to compare the resulting differences. As data migrates to new representations, the queries can also be migrated, ensuring stability across changing technologies.

This approach works consistently for both small and large datasets and for static and highly dynamic data. By promoting a consistent approach, decision making and scientific research based on data will become more transparent and reproducible.

When Can We Use This?

As demonstrated by our first successful pilots, this approach can be applied right now.

What is the RDA?

The RDA is an international organization that was formed in 2013 through funding from the National Science Foundation, the European Commission and the Australian government, with a mission to reduce barriers to data sharing and exchange, and accelerate data driven innovation worldwide.

Contact Information:

For more information on the solutions detailed above or to learn more about the Dynamic Data Citations Working Group, please visit https://rd-alliance.org/groups/data-citation-wg.html.