Advancing caching and automation with FDO

Amirpasha Mozaffari‡, Niklas Selke‡, Martin Schultz‡

‡ Jülich Supercomputing Centre (JSC), Jülich, Germany

Corresponding author: Amirpasha Mozaffari (a.mozaffari@fz-juelich.de)

Abstract

Introduction: Geosciences are utilising big data that is constantly updated, modified, and changed with an ever-growing stream of new measured, modelled accumulated data (Reichstein et al. 2019). Many of these data reside in databases and are frequently revised and recalculated with new data, corrections, or recalibrations. Thus, versioning the data is a known challenge for the earth sciences system (ESS) community. To ensure reusability of the data and traceability of associated information, it is crucial to document this stream of changes in an efficient, human-readable, and machine-actionable manner. As previous studies (Iump et al. 2021, Schultes et al. 2022) and community driven-efforts have shown, we believe the FAIR Digital Object (FDO) (De Smedt et al. 2020) concept could provide a neat solution by encapsulating the data, metadata, data version, and associated information and identifying it with a persistent identifier (PID) (Philipson 2019). In addition, FDO could be a path to avoid rerunning expensive, energy-demanding computations and data duplication as PID and metadata could enable a detailed search and cataloguing of available statistical aggregations and other products. In this conceptual work, we want to explore the FDO capability for data versioning combined with a state-of-the-art caching system for relational databases to provide reusable and mutable data products. Such an FDO-enabled caching system would enable us to identify recurring access patterns to data and store them as FDOs. Moreover, we believe such a concept could be integrated into an automated workflow where highly anticipated computation or user requests that require intensive computation are generated and submitted to High-performance Computing (HPC)'s.

Case study: TOAR database: The Tropospheric Ozone Assessment Report (TOAR) database (Schultz 2017) consists of an extensive collection of global air quality measurements focusing on ground-level ozone. We use a PostgreSQL (PostgresSQL 2022) database, a widely used database for relational data, and provide the database schema and all related code via a free and open-source git repository (Jülich Supercomputing Centre 2022). A vital asset of the TOAR database is the associated REST API which has been implemented with Python/fast API and includes a module for statistical analysis of TOAR data. Users can request one or several of over 30 different statistical aggregates and define one of 5 target temporal resolutions from daily to annual, and the API will trigger online calculations based on the original data, which are stored at hourly time resolution.

As there is an apparent demand from the scientific community to expand these analysis capabilities and allow for multi-variable and multi-station analysis (for example, to evaluate numerical model simulations), it will be necessary to design new parallel workflows to enable such calculations in a reasonable time.

Two specific challenges to overcome in the design of automated workflow with an FDO-enabled caching system are ensuring that the query stays connected to the correct data and establishing a schedule for pre-calculating the most frequently used statistical aggregates. In the following, we discuss these two challenges in more detail.

Caching system: There are other data providers in the field, but they commonly focus mainly on archiving the measurement data. We want an analysis tool with the fastest possible response times for the users. Furthermore, we want to look into FDOs to ensure the preservation of queries and make them reusable and traceable. For the caching itself, it is very important that the cache key created for a query allows for verification that the data used in computing the query the first time did not change when trying to reuse the cached result. In our conceptual work, we want to develop a concept and a demonstrator of an atmospheric data analysis cache. This includes choices for the underlying technical solution (e.g. PostgreSQL, MongoDB, Redis...), the definition of data structures and hash codes, design of a mechanism for the triggering of re-calculations, definition of a schedule for automated cache updates, and various aspects related to query documentation and reproducibility of results. Technical obstacles caused by the expected size of up to 0.8 Terabytes for the TOAR and complicated scalability issues that can arise should be considered for a possible solution. Ideally, the caching system should be agnostic of the underlying database/server choice to enhance portability.

Automated workflow: The second challenge to address here is to combine the envisioned caching system with a flexible workflow scheme. Such a workflow setup enables preparing pre-compile and calculating the most frequently used statistical aggregate ahead of user demand. Queries can either be triggered by a user (demand-driven) or by an automatic system which will compute commonly used queries without a user having to trigger it (provider-driven) to have as many query results as possible ready to go for users, so they do not have to wait for the results after they have sent a request. User requests are categorised according to the availability of the statistical products and the required computation effort. Some might have already been calculated and stored as FDO can be quickly reloaded and processed further. While some queries might be new, but still possible to be calculated on the fly, and the responses could be delivered on near real-time basis. In contrast, some more intensive statistical aggregations require HPC. We believe an automated FDO-enabled caching system will utilise the metadata and FDO to provide on-demand data requests and reduce repetitive computation. It paves the way for intelligent computation that can be scheduled at different times of the day based on the priority and availability of resources, and reduction of the energy consumption and carbon footprint of computing.

Outlook and next steps: In our conceptual work and demonstrator, we aim not to use the FDO only to ensure long-term preservation but rather to create a use case where FDO is

used for practical reusability in the daily operation of the database. An automated FDO-enabled caching system requires multiple components to work synchronously. In the following months, we will focus on creating a demonstrator for the caching system, adopting an FDO typing that could fit the best with the planned tasks, and creating a workflow management system that could support such a dynamic system with interfaces to API-enabled web-services, cloud computing resources and conventional HPC resources.

Keywords

FAIR Digital Object (FDO), database, caching system, automated workflow

Presenting author

Amirpasha Mozaffari

Presented at

First International Conference on FAIR Digital Objects, presentation

Hosting institution

Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich, Germany

Conflicts of interest

References

- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8: 21. https://doi.org/10.3390/publications8020021
- Jülich Supercomputing Centre (2022) https://gitlab.jsc.fz-juelich.de/esde/toar-data/toardb_fastapi
- lump J, Wyborn L, Wu M, Martin J, Downs RR, Asmi A (2021) Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles. Data Science Journal 20: 12: 1-13. https://doi.org/10.5334/dsj-2021-012
- Philipson J (2019) Identifying PIDs playing FAIR. Data Science 5 (2): 229-244. https://doi.org/10.3233/DS-190024
- PostgresSQL (2022) https://www.postgresql.org/
- Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566 (7743): 195-204. https://doi.org/10.1038/s41586-019-0912-1

- Schultes E, Roos M, Silva Santos LO, Guizzardi G, Bouwman J, Hankemeier T, A B, Mons B (2022) FAIR Digital Twins for Data-Intensive Research. Front. Big Data 5 (883341). https://doi.org/10.3389/fdata.2022.883341
- Schultz M, et al. (2017) Tropospheric Ozone Assessment Report: Database and metrics data of global surface ozone observations,Elementa. Science of the Anthropocene 5: 58. https://doi.org/10.1525/elementa.244