1st International Conference on Research Data Infrastructures

Enabling RDM

https://doi.org/10.52825/CoRDI.v1i.360

Published: 07 Sept. 2023

On the Design and Implementation of Easy Access to External Spatiotemporal Datasets in NFDI

Christian Beilschmidt[1] [https://orcid.org/0009-0001-6297-0921]
Dominik Brandenstein[2] [https://orcid.org/0009-0002-1901-9935]
Johannes Drönner[1] [https://orcid.org/0009-0003-9629-2844]
Nikolaus Glombiewski[2] [https://orcid.org/0000-0003-2876-3918]
Michael Mattig[1] [https://orcid.org/0009-0006-1893-5391]
Bernhard Seeger[1 2] [https://orcid.org/0000-0002-9362-153X]

[1]Geo Engine GmbH, Am Kornacker 68, 35041 Marburg, Germany

[2]University of Marburg, Dept. of Mathematics and Computer Science, Hans-Meerwein-Str. 6, 35032 Marburg, Germany

Abstract: n.a.

Keywords: Spatiotemporal data access, Workflow platform, Geo Engine

Across many scientific domains, the ability to process large amounts of heterogeneous spatiotemporal data from various sources is crucial for solving challenging research questions. For example, researchers in NFDI4Biodiversity [1] must combine observational data with satellite images to correlate biodiversity loss with climate change variables.

In general, large datasets are not available on the system (called consumer) where the processing is performed, but first have to be retrieved from one or multiple external systems (called providers) that offer a corresponding service. Moreover, a consumer is often unaware of the datasets the providers offer. Ideally, a provider follows FAIR principles [2] and thus supports mechanisms to simplify data exchange. However, in practice, multiple providers with valuable datasets are not as FAIR as desired or lack spatiotemporal-specific support for data exchange. Instead of improving each potential provider at the source, we propose an intermediary spatiotemporal data exchange layer (SDExL) that helps simplify data exchange so that domain experts can easily access valuable data with little technical know-how.

Based on practical experience and guided by the FAIR principles, in the following, we postulate four requirements for building an SDExL (Figure 1). Then, we discuss two reference implementations within Geo Engine [3], a flexible analytical processing platform for spatiotemporal data used in projects like NFDI4Biodiversity and FAIR Data Spaces [4].

The following three-step process of an SDExL summarizes the general communication between consumers and providers.

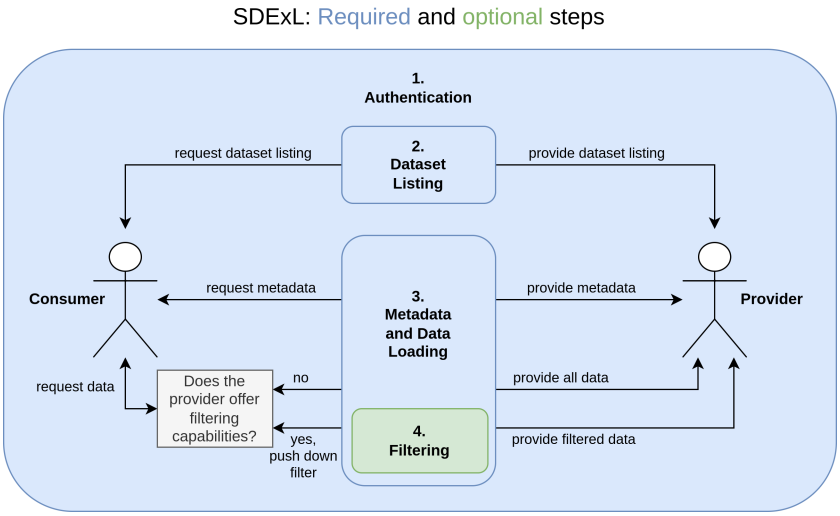

Figure 1. Overview of the spatiotemporal data exchange layer (SDExL).

1. First, users of a consumer have to authenticate themselves in the provider to gain access to datasets. Afterwards, SDExL abstracts away all access control mechanisms until the system's authorization policy requires user input. Furthermore, the provider needs to be uniquely identified in the consumer through an identifier.
2. Next, the users can request a list of all datasets they can access. SDExL translates a list of unique dataset IDs of the provider into a list of unique dataset IDs for the consumer.
3. The users select one of the datasets from their list in a two-step approach. First, SDExL requests any available metadata for the desired dataset. Second, based on the metadata, the actual dataset is requested and delivered to the consumer.

Each step in the process is a strict requirement for implementing an SDExL. The fourth optional requirement minimizes processing time and thus increases the overall usability.

4. At any point, SDExL leverages the filtering capabilities of the provider such that the consumer receives only the data required for the scientific task at hand.

Geo Engine implements multiple connectors for interacting with providers, e.g., standard services based on WCS [5] and STAC [6], following the SDExL process. We now discuss two specialized connectors in detail.

Aruna Object Storage (AOS) [7] is a FAIR cloud-based storage platform developed in NFDI4Biodiversity. As an object storage, AOS has limited spatiotemporal processing capabilities but offers scientific data objects to users. The four requirements are realized as follows.

1. Users of AOS create a key for authorization and pass it to Geo Engine. This key is attached to each request of Geo Engine.
2. Dataset listings are natively available through AOS.
3. AOS provides for each dataset a metaobject containing the loading information.
4. AOS offers labels for datasets, e.g., species names. Geo Engine leverages these labels as a filter to only return a subset of datasets in listings and subsequently to reduce loading of the actual data.

GBIF [8] offers open access to the largest observation database about life on Earth, with over 2 billion records. Many scientific applications use GBIF, enrich the data with domain-specific data, and offer the resulting data products in the open-source database system PostgreSQL [9] with its powerful spatial extension PostGIS [10]. For such a setting, our four requirements are realized as follows.

1. User credentials are required to connect to a PostgreSQL database and access its data records.
2. Dataset listings are retrieved through SQL queries processed by PostgreSQL. E.g., a list of species can be retrieved with a "SELECT UNIQUE species. . . " query, serving as identifiers for occurrences of a species. In order to impose restrictions on data visibility, it is possible to utilize PostgreSQL's role-based access control mechanisms.
3. The PostGIS extension of PostgreSQL offers the required metadata (e.g., spatial reference system) for loading data into Geo Engine. Then, Geo Engine uses this metadata as input for the PostgreSQL GDAL driver [11] to load the actual data.
4. The SQL query interface of PostgreSQL supports accessing datasets and many means of filtering data records. For example, when Geo Engine only requires data within a specific spatial bounding box, the GDAL driver leverages PostgreSQL's filtering capabilities by adding a filter condition to the SQL query.

Finally, Geo Engine can manage datasets internally, either uploaded by an administrator or by the user. These datasets can be combined with datasets provided by SDExL through spatiotemporal processing operators in Geo Engine. The processing steps are stored as a graph of operators, which serves as a provenance workflow with a unique ID. Since workflows in Geo Engine can be accessed like a dataset through standard OGC [12] protocols, Geo Engine is a provider that fulfills requirements 1-3. In addition, it offers spatiotemporal filtering capabilities (requirement 4) that consumers can leverage. Thus, it fulfills our four requirements. By acting as a facilitator for data exchange, Geo Engine is a service that helps researchers solve difficult research questions while ensuring interoperability within the overall spatiotemporal processing landscape.

For future work, we plan to connect Geo Engine to many more data spaces which adhere to our postulated requirements. For example, connecting the Data and Information Access Services [13] will make the satellite data from the Copernicus mission available. Moreover, we will implement SDExL for time series database systems in the project FAIR Data Spaces. Finally, Geo Engine will specify SDExL as a self-describing service in a federated GAIA-X [14] catalog making it available to a large user community, including the one of NFDI [15].

Data availability statement

This submission is not based on data.

Underlying and related material

The source code of Geo Engine is publicly available on GitHub: https://github.com/geo-engine/geoengine.

Author contributions

Nikolaus Glombiewski wrote the initial draft (CRediT ID: 43ebbd94-98b4-42f1-866b-c930ce- f228ca). All authors reviewed and edited it to create the final draft (CRediT ID: d3aead86-f2a2-47f7- bb99-79de6421164d).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially funded by the German Research Foundation DFG under the grant agreement number 442032008 (NFDI4Biodiversity). The project is part of NFDI, the National Research Data Infrastructure Programme in Germany. This work was partially funded by the BMBF project FAIR Data Spaces (FAIRDS10). This work was partially funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under grant numbers O3EUPHE069 and 50EE2303B.

References

[1] "NFDI4Biodiversity." (2023), [Online]. Available: https://www.nfdi4biodiversity.org/ (visited on 04/25/2023).

[2] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[3] C. Beilschmidt, J. Drönner, M. Mattig, P. Schweitzer, and B. Seeger, "Geo engine: Workflow-backed geo data portals," in *BTW 2023*, Bonn: Gesellschaft für Informatik e.V., 2023, pp. 837–849, ISBN: 978-3-88579-725-8.

[4] "FAIR Data Spaces." (2023), [Online]. Available: https://www.nfdi.de/fair-data-spaces/ (visited on 04/25/2023).

[5] "Web Coverage Service." (2023), [Online]. Available: https://www.ogc.org/standard/wcs/ (visited on 04/25/2023).

[6] "SpatioTemporal Asset Catalogs." (2023), [Online]. Available: https://stacspec.org/ (visited on 04/25/2023).

[7] "Aruna Object Storage." (2023), [Online]. Available: https://www.uni-giessen.de/de/fbz/fb08/Inst/bioinformatik/software/aruna (visited on 04/25/2023).

[8] "Global Biodiversity Information Facility." (2023), [Online]. Available: https://www.gbif.org/ (visited on 04/25/2023).

[9] "PostgreSQL." (2023), [Online]. Available: https://www.postgresql.org/ (visited on 04/25/2023).

[10] "PostGIS." (2023), [Online]. Available: https://postgis.net/ (visited on 04/25/2023).

[11] "GDAL PostgreSQL Driver." (2023), [Online]. Available: https://gdal.org/drivers/vector/pg.html (visited on 04/25/2023).

[12] "Open Geospatial Consortium." (2023), [Online]. Available: https://www.ogc.org/ (visited on 04/25/2023).

[13] "Data and Information Access Services." (2023), [Online]. Available: https://www.copernicus.eu/en/access-data/dias (visited on 04/25/2023).

[14] "The Gaia-X Ecosystem - A Sovereign Data Infrastructure for Europe." (2023), [Online]. Available: https://www.bmwk.de/Redaktion/EN/Dossier/gaia-x.html (visited on 04/25/2023).

[15] "NFDI - Nationale Forschungsdateninfrastruktur." (2023), [Online]. Available: https://www.nfdi.de/ (visited on 04/25/2023).