Graph based Visual Search Engine

Pawandeep Kaur Betz, German Aerospace Center (DLR), Institute of Software Technology, Software for Space Systems and Interactive Visualization

Tobias Hecking, German Aerospace Center (DLR), Institute of Software Technology, Intelligent and Distributed Systems

Date of submission: 31st March 2023

Abstract

One of the first and important principles of the FAIR data is that it should be 'Findable'. This principle is difficult to achieve when the data repositories are extremely large and contains datasets from numerous interconnected domains. Common search mechanisms seen within the NFDI4Earth data repositories (Pangea, DKRZ, EEA, ICOS, etc.) are simple grid-based, facet-based and keyword-based search. We believe these search techniques are not efficient enough for big data repositories. Additional search and exploration techniques are needed to exploit interconnected datasets and fully fulfil complex information needs. We propose to build a visualization-enabled search interface on top of a graph database to visualize the collection of datasets from different providers and to offer easy and multidimensional access to the datasets in one search portal. We planned to connect two to three repositories for this pilot proposal as proof of concept.

I. Introduction

As more and more datasets from variant data publishers are added to the consortium, simple grid based or keyword-based search engines to access these datasets are not efficient enough. The reasons are standard search engines firstly don't provide an overview of the large collection of datasets and secondly, connections between particular datasets are not explicit, and thus, they don't provide options to find related datasets based on their content. Due to the limited scope of the simple search environments, many of the important datasets remain unfindable and inaccessible, thus declining one of the goals of NFDI4Earth project. To answer these problems our objectives for this project are 1) to provide a common interface to access the data from different data providers; 2) interlink large collections of data entries based on metadata and extracted information in a data property graph and 3) to provide a search interface that assists in finding related datasets by using different search dimensions: spatial, temporal, network, hierarchy etc. To fulfill these objectives, we propose visualization enabled search environment on top of the data property graph that assists users in navigating through interconnected datasets.

To solve these problems other data domains, for e.g., digital humanities[1] are applying the visual interfaces in the search of their datasets. Very limited work has been seen in Earth System Sciences. German Federation for Biological Data (GFBIO) VAT[2] provides their spatial platform to search their datasets, however, their system lacks the contextual search of interrelated datasets. We still miss examples of visualization enable search engines that can connect multiple and disparate data providers and can provide multi-dimensional data overview and access. Our research, where fits these three missing blocks, will also contribute in finding different visual dimensions and visual encodings to dataset search.

II. Pilot description

To tackle the problem of limited findability of relevant research data and to make connections between datasets more explicit, we will analyse data entries in the downloaded repositories for relevant elements and connect them in a graph database.

The general workflow is depicted in Figure 1. In order to harmonize access to the data entries of different providers, we will develop specific importers for the selected data portals. For building the aforementioned graph, we will make use of our Corpus Analytics Graph framework (el Baff et al., 2023) for building property graphs from document corpora. It is built upon a “create and annotate” pattern. In the create phase, available metadata of NFDI4Earth data entries will be extracted and transformed into a graph structure stored in ArangoDB[3]. Nodes in the graph are apart from references to the dataset also keywords, data descriptions, authors, funding references, geo-coordinates, related artefacts, etc. In this way, different data entries will be connected based on shared metadata. In the annotate phase, the graph becomes enriched by specific annotator components that analyse dataset descriptions to extract further scientific concepts and their relations. This will be done based on the SPERT information extraction model (Eberts & Ulges, 2019). It is trained on the SciERC corpus (Luan et al., 2018) and extracts concepts in the categories “Method”, “Task”, “Evaluation Metric”, “Material” and “Generic Scientific Terms” and different types of relations among them from texts. In this way further connections between data entries will be added to the graph. Managing information about data entries as a property graph allows for complex search queries beyond simple keyword or coordinate search. Examples are finding the most common research topic for a region, datasets and corresponding geographic regions that connect given keywords or concepts, or the impact of funding measures.

On top of this backend we aim to develop visual exploration and search tools that assist users in navigating through the graph in an intuitive way. Different visualizations assist in finding information in different dimensions. For example: maps → spatial context, line charts → temporal context, network diagrams or graphs → interrelations, word clouds → thematic context, etc. We will connect different data dimensions from the graph database to different visualizations. These dimensions, for e.g are., dataset’s location → maps, measurement dates → temporal charts, a word cloud to show different

[1] https://uclab.fh-potsdam.de/ddb/netzwerke/

[2] https://www.gfbio.org/visualize

[3] https://www.arangodb.com/

context of the datasets and hierarchal diagrams to show datasets based on different taxonomical levels. These connected and coordinated visualizations will enhance a dataset search in multiple granularities. This will be done at the interface level with the declarative javascript framework and different visualization libraries.

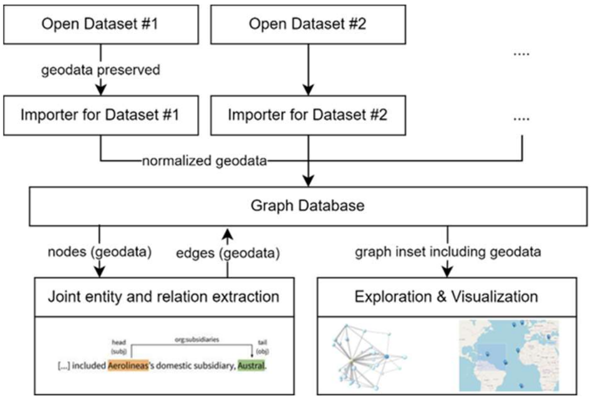

Figure 1: A general workflow of the project

III. Relevance for the NFDI4Earth

This proposal targets Task Area 2, M2.5 of the NFDI4 Earth framework. It proposes to provide an advanced data search tool for FAIR research data management. This framework and tool will be relevant to the environmental scientists who search for the datasets, and for them, it is an enhanced search application. The overview of the collection of the datasets with their metadata contents will assist data administrators in better decision-making. The NFDI4Earth external data infrastructure projects can deploy and use the tool for their private repositories with minimal overhead. Due to the limited timeline in this project, we will only be connecting the data from those repositories which provide an API to access the metadata from the server programmatically (for example, Pangea etc.).

IV. Deliverables

By end of the project we will deliver:

- A visual data search and exploration tool based on three data repositories.
- Open source and citable code repository
- Report on the conducted internal DLR workshop
- Project report
- Presentation and publication (at least on pre-print level).

V. Work Plan & Requested funding

The project is divided into three main work packages (WPs). WP1 and WP2 combine the scientific and research elements of this project. These work packages will be developed in an iterative manner aiming at a first prototype after 6 month that will be evaluated and refined in a second iteration during the second half of the project and integrated into the final system (See work plan in Figure 2). In WP3 we plan to conduct different dissemination activities including a user workshop with domain experts at the German Aerospace Center (DLR).

Data Property Graph Construction (WP1) – 6 months: As the backbone of our visual search application, in this work package we will first develop importers for different data providers. As stated above we will adapt the Corpus Analytics Graph framework (el Baff et al., 2023) for the task of creating a graph database in ArangoDB from a corpus of data entries. First references to data entries are then added to a graph database as nodes and linked to other nodes that are created from their associated metadata. In this way connections between datasets become more explicit and visualizable. The resulting graph becomes further enriched by applying concept and relation extraction to dataset descriptions, which results in additional links between datasets. We will further create a search index over the nodes in the graph. In this way data elements can be retrieved via keyword and raster queries as in standard search engines but in addition also more complex queries that utilize links to related datasets can be supported as well.

Visual Search Interface (WP2) – 6 months: In the second work package, we will explore different dimensions of the constructed graph from the previous work package. Based on that suitable visual dimensions, related visualizations will be explored. Accordingly, suitable visualization framework will be developed that will decide on needed technologies, interactive elements and the data workflow for the search interface.

Dissemination (WP3) – 3 months: In the second half of the project we will conduct an expert workshop with environmental scientists at the German Aerospace Center (DLR) to collect feedback on the developed prototype, which will then be incorporated into the development of the final system. This project work will be disseminated via documentation and project reports and publication. Code will be made open source and citable.

WPs	Jan - March	April - June	July - Sept	Oct – Dec
WP1 Knowledge Graph Construction	▨		▨	
WP2 Visual Search Interface		▨		▨
WP3 Dissemination		▨	▨	▨

Figure 2: Time line of the project

References

(Baff et al. 2023) Roxanne El Baff, Tobias Hecking, Andreas Hamm, Jasper W. Korte, & Sabine Bartsch (Year is required!). Corpus Annotation Graph Builder (CAG): An Architectural Framework to Create and Annotate a Multi-source Graph. In The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023): : System Demonstrations . Association for Computational Linguistics.

(Eberts & Ulges, 2019) Eberts, M., & Ulges, A. (2019). Span-based joint entity and relation extraction with transformer pre-training. arXiv preprint arXiv:1909.07755.

(Luan et al., 2018) Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. arXiv preprint arXiv:1808.09602.