

[Dai:Si] - A Modular Dataset Retrieval Framework with a Semantic Search for Biological Data

Fateme Shafiei¹, Felicitas Löffler¹, Sven Thiel¹, Kobkaew Opasjumruskit², Denis Grabiger¹, Pauline Rauh¹ and Birgitta König-Ries^{1,3,4}

¹Heinz Nixdorf Chair for Distributed Information Systems, Department of Mathematics and Computer Science, Friedrich Schiller University Jena, Jena, Germany

²Software Systems for Digitalization, Institute of Data Science, German Aerospace Center (DLR), Jena, Germany

³German Center for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Germany

⁴Michael-Stifel-Center for Data-Driven and Simulation Science, Jena, Germany

Abstract

Dataset search is receiving increasing attention in a scholar's daily research practice. In biodiversity research, dataset retrieval in particular is a challenging and time-consuming task as most search services in current data portals only offer a simple keyword-based search. In this work we introduce [Dai:Si], a modular framework for dataset retrieval with a semantic search for biological data. [Dai:Si] is based on a former semantic search service developed within the scope of the GFBio project. It allows the expansion of query keywords with related terms using GFBio's Terminology Service. This new version provides an enhanced user interface (UI) with explanations of related semantic terms upon demand. Due to its modular structure, [Dai:Si]'s semantic service can now be used independently of the user interface (UI).

Keywords

Dataset search, Dataset retrieval, Semantic search, Query expansion, Biodiversity informatics

1. Introduction

Dataset search is an increasingly important task in daily research practice. In particular, in biodiversity research, the search for datasets and their reuse has steadily increased over the last decade [1]. However, scholars report difficulties in finding relevant datasets [2, 3]. "Inadequate search tools" [3] constitute one obstacle. Most data portals only offer a keyword-based search along with a faceted search to look for scientific datasets, e.g., *DataOne*¹ or *Zenodo*². In these search systems, relevant datasets can only be found when a query keyword syntactically matches the content of a dataset. As biological terms are often fuzzy [4], further related semantic terms should be taken into account in dataset search. So far there are very few approaches that


S4BioDiv 2021: 3rd International Workshop on Semantics for Biodiversity, held at JWOW 2021: Episode VII The Bolzano Summer of Knowledge, September 11–18, 2021, Bolzano, Italy

✉ fateme.shafiei@uni-jena.de (F. Shafiei); felicitas.loeffler@uni-jena.de (F. Löffler); sven.thiel@uni-jena.de (S. Thiel); kobkaew.opasjumruskit@dlr.de (K. Opasjumruskit); denis.grabiger@uni-jena.de (D. Grabiger); pauline.rauh@uni-jena.de (P. Rauh); birgitta.koenig-ries@uni-jena.de (B. König-Ries)

ORCID: 0000-0001-9731-9496 (F. Shafiei); 0000-0001-6423-7427 (F. Löffler); 0000-0003-3093-5635 (S. Thiel); 0000-0002-9206-6896 (K. Opasjumruskit); 0000-0002-2382-9722 (B. König-Ries)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹DataOne, <https://www.dataone.org/>

²Zenodo, <https://zenodo.org/>

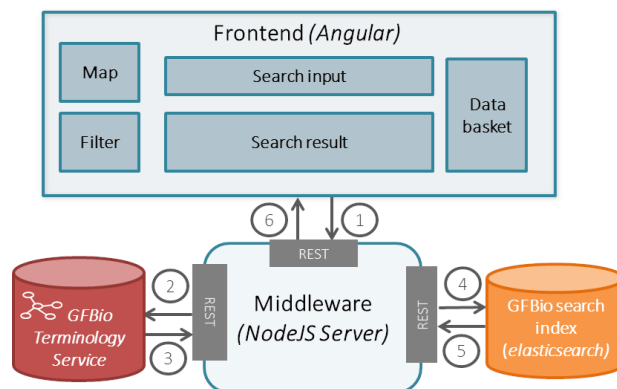


Figure 1: *[Dai:Si]*'s architecture and the overall flow of the semantic search.

additionally support scholars with semantic services. In the biomedical domain, *Datamed* [5]³ offers a search portal that expands query terms with related terms using the UMLS service⁴. In biodiversity research, within the scope of the *GFBio*⁵ project a semantic search based on query expansion has been introduced [6]. The knowledge base behind the system is *GFBio*'s own Terminology Service [7] offering tailored ontology services for biodiversity research.

In this work, we present a new version of this semantic search for biological datasets. It is part of a modular framework - *[Dai:Si]* (the name is an abbreviation of the phonetic spelling of 'dataset search') - that allows developers to use the semantic search independently of the user interface. In addition, explanations of the expanded terms are now available on demand, and the search can be expanded with narrower or broader terms. The code is publicly available in our GitHub repository: <https://github.com/fusion-jena/DaiSi>.

2. Architecture

The architecture is presented in Figure 1. The framework consists of a middleware, implemented with *NodeJS*⁶, and a front-end, implemented with *Angular*⁷. The *NodeJS* server communicates through a REST API with both back-end applications, the *GFBio* Terminology service⁸ and the *GFBio* search index.

Modularity is one of the main aims of the framework. Therefore, domain and business specific logic are separated from functional components. This allows an easy integration of additional search indexes. For each search index a new module is added to the middleware. The search index only has to provide some fields that need to be mapped to the underlying data model. More details are described on our GitHub page. The terminology service can be replaced by other services in a configuration file in the middleware. However, as no protocols

³Datamed, <https://datamed.org/>

⁴UMLS, <https://www.nlm.nih.gov/research/umls/>

⁵GFBio, <https://www.gfbio.org>

⁶NodeJS, <https://nodejs.org/en/>

⁷Angular, <https://angular.io/>

⁸GFBio TS, <https://terminologies.gfbio.org/>

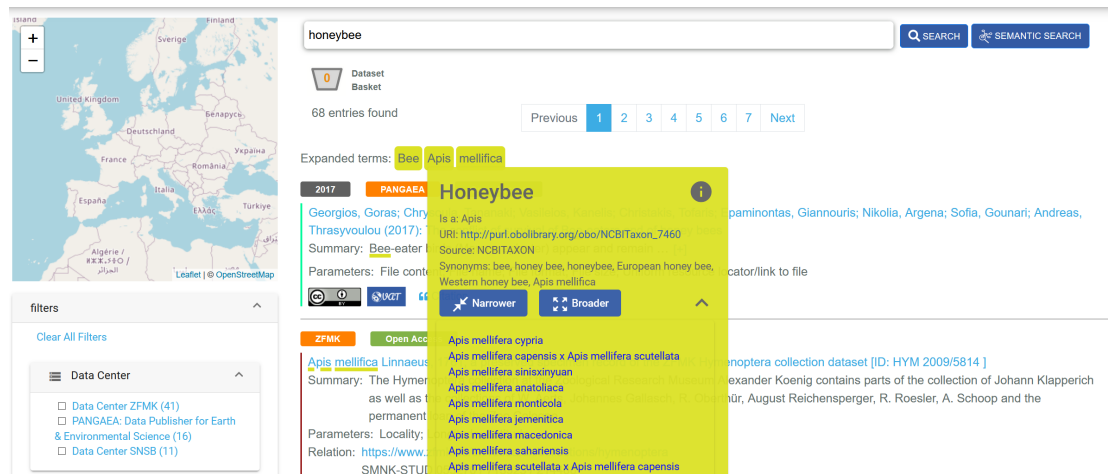


Figure 2: Screenshot of *[Dai:Si]*'s semantic search. A search for 'honeybee' will extend the query to related terms such as 'bee' or 'Apis mellifica'.

or standards for terminology services exist yet, changing the service might require adjustments to the middleware functions based on the API requests and responses.

[Dai:Si]'s UI consists of four main components: filters, map, search input and search result. For collecting documents during search, e.g. for later download, a data basket is provided. When users look for datasets in the semantic search (1), all query terms are sent to *GFBio*'s terminology service (2). Matching URIs are looked up per query term and are sent back to the middleware (3). In the current version, only synonyms including scientific and common names are considered. Afterwards, the expanded search terms are sent to the search index (4). The result (5) is forwarded to the frontend and the returned datasets are displayed (6). For now all search terms are combined with a logical OR. However, if the search index supports boosting of results containing all or most search terms, datasets with the largest match are presented on top. Figure 2 presents a screenshot of *[Dai:Si]*'s semantic search. The user can obtain more information about the expanded terms by hovering over them. An explanation dialog displays the URIs found, their ontologies and a description. This supports users in understanding the relation between the originally entered keyword and the expanded terms. Users can also query for further semantic relations such as child (narrower) or parent (broader) concepts on demand. These related terms can be added to the search input field with a double-click.

3. Demonstration

We provide a demonstration of *[Dai:Si]* with *GFBio*'s search index: <https://dev.gfbio.uni-jena.de/daisi>. Users can either search for datasets with the original search without query expansion, or they can try out the semantic search. All middleware services, including the semantic search, are also accessible separately: <https://dev.gfbio.uni-jena.de/daisi-api/api-docs/>.

4. Conclusion

In this work, we presented *[Dai:Si]* - a new modular dataset retrieval framework with a semantic search for biological data. We aim to enhance the semantic search to permit the usage of AND, OR, NOT and quotation marks in the search input field. We would also like to integrate further semantic services to highlight important biological entities, e.g., species, environmental terms or data parameters.

Acknowledgments

We would like to thank the following colleagues for their support in terms of the search index and terminology services: U. Schindler, A. Behnken, F. Becker and N. Karam.

References

- [1] GBIF, GBIF Science Review 2020, Technical Report, 2020. doi:10.35035/bezp-jj23.
- [2] A. Culina, T. W. Crowther, J. J. C. Ramakers, P. Gienapp, M. E. Visser, How to do meta-analysis of open datasets, *Nature Ecology & Evolution* 2 (2018) 1053–1056. doi:10.1038/s41559-018-0579-2.
- [3] K. Gregory, P. Groth, A. Scharnhorst, S. Wyatt, Lost or found? Discovering data needed for research, *Harvard Data Science Review* 2 (2020). doi:10.1162/99608f92.e38165eb.
- [4] A. E. Thessen, H. Cui, D. Mozzherin, Applications of natural language processing in biodiversity science, *Advances in Bioinformatics* (2012). doi:10.1155/2012/391574.
- [5] X. Chen, A. E. Gururaj, B. Ozyurt, R. Liu, E. Soysal, T. Cohen, F. Tiryaki, Y. Li, N. Zong, M. Jiang, D. Rogith, M. Salimi, H.-E. Kim, P. Rocca-Serra, A. Gonzalez-Beltran, C. Farcas, T. Johnson, R. Margolis, G. Alter, I. M. Fore, L. Ohno-Machado, J. S. Grethe, H. Xu, Datamed - an open source discovery index for finding biomedical datasets, *Journal of the American Medical Informatics Association* 25 (2018) 300–308. doi:10.1038/s41559-018-0579-2.
- [6] F. Löffler, K. Opasjumruskit, N. Karam, D. Fichtmüller, U. Schindler, F. Klan, C. Müller-Birn, M. Diepenbroek, Honey bee versus *Apis Mellifera*: A semantic search for biological data, Springer International Publishing, Cham, 2017, pp. 98–103. doi:10.1007/978-3-319-70407-4_19.
- [7] N. Karam, C. Müller-Birn, M. Gleisberg, D. Fichtmüller, R. Tolksdorf, A. Güntsch, A terminology service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data, *Datenbank-Spektrum* 16 (2016) 195–205. doi:10.1007/s13222-016-0231-8.