Honey Bee Versus Apis Mellifera: A Semantic Search for Biological Data

Conference Paper · May 2017

DOI: 10.1007/978-3-319-70407-4_19

8 authors, including:

Felicitas Löffler
Friedrich Schiller University Jena
8 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Kobkaew Opasjumruskit
Friedrich Schiller University Jena
6 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Naouel Karam
Freie Universität Berlin
20 PUBLICATIONS **56** CITATIONS
SEE PROFILE

David Fichtmueller
Freie Universität Berlin
11 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Honey Bee Versus Apis Mellifera: A Semantic Search for Biological Data

Felicitas Löffler[1(✉)], Kobkaew Opasjumruskit[1], Naouel Karam[2], David Fichtmüller[3], Uwe Schindler[4], Friederike Klan[1], Claudia Müller-Birn[2], and Michael Diepenbroek[4]

[1] Heinz-Nixdorf Endowed Chair for Distributed Information Systems, Friedrich Schiller University Jena, Jena, Germany
{felicitas.loeffler,kobkaew.opasjumruskit,friederike.klan}@uni-jena.de
[2] Institute of Computer Science, Freie Universität Berlin, Berlin, Germany
naouel.karam@fu-berlin.de, clmb@inf.fu-berlin.de
[3] Botanic Garden and Botanical Museum (BGBM), Freie Universität Berlin, Berlin, Germany
d.fichtmueller@bgbm.org
[4] MARUM, University of Bremen, Bremen, Germany
{uschindler,mdiepenbroek}@pangaea.de

Abstract. While literature portals in the biomedical domain already enhance their search applications with ontological concepts, data portals offering biological primary data still use a classical keyword search. Similar to publications, biological primary data are described along meta information such as author, title, location and time which is stored in a separate file in XML format. Here, we introduce a semantic search for biological data based on metadata files. The search is running over 4.6 million datasets from GFBio - The German Federation for Biological Data (GFBio, https://www.gfbio.org), a national infrastructure for long-term preservation of biological data. The semantic search method used is query expansion. Instead of looking for originally entered keywords the search terms are expanded with related concepts from different biological vocabularies. Hosting our own Terminology Service with vocabularies that are tailored to the datasets, we demonstrate how ontological concepts are integrated into the search and how it improves the search result.

Keywords: Semantic search · Query expansion · Biological data · Life sciences · Biodiversity

1 Introduction

Scholars in life sciences are faced with an increasing amount of biological data. One example is the biodiversity domain dealing with the variety and variability of species, habitats and their relationships on earth. In this research area, different types and formats of data such as observational data, images or genome

E. Blomqvist et al. (Eds.): ESWC 2017 Satellite Events, LNCS 10577, pp. 98–103, 2017.
https://doi.org/10.1007/978-3-319-70407-4_19

sequences need to be collected and analyzed. Enabling researchers to find relevant data for their information need requires effective filtering and search techniques that allow data retrieval across scientific domains. Even though a large number of ontologies exist in life sciences, only data portals offering biomedical literature such as MEDLINE articles enhance their search systems with semantic concepts and ontological filtering [4,5]. Data portals offering a search over biological datasets still rely on classical keyword-based techniques. Based on metadata files containing information about author, title, location and parameters of collected data, search applications in data portals such as GBIF[1], Data One[2] or Dryad[3] present datasets containing only user entered search terms. This does not allow cross-domain retrieval where keywords are used for searching and dataset descriptions refer to various terminologies. This hampers data retrieval, in particular, in the biodiversity domain which is inherently interdisciplinary.

Expanding search queries with semantically related concepts is a common technique for enhancing search engines. This idea is used in GFBio's [3] data search that currently contains around 4.6 million datasets from data centers specialized on nucleotide and environmental data, e.g., PANGAEA[4] and data centers focused on natural science collections, e.g., BGBM[5]. When entering a search term, the system calls web services from GFBio's open access Terminology Service [7]. Providing only vocabularies and ontologies that are tailored to biological datasets, the Terminology Service returns related terms that are added to the originally entered keywords. According to our previous findings [8] we only expand queries with synonyms, scientific and common names. For example, when looking for datasets about *Apis mellifera*, which is the scientific name of honey bee, the system also retrieves datasets with the common name *Honey bee* and synonyms such as *Bee*. This paper presents the underlying architecture of our semantic search feature, describes its components and use cases we will showcase to demonstrate how end users benefit from the system. A life demonstration can be found at https://www.gfbio.org/semantic-search.

2 Architecture

Figure 1 depicts the main components of GFBio's semantic search and the information flow between. Before calling the search engine, the system invokes web services from Terminology Service to get synonyms of the entered search terms. All terms, the originally entered and expanded ones, are finally sent to GFBio's standard search application (Sect. 2.4). In the following section, we describe all components individually.

[1] GBIF, http://www.gbif.org/.
[2] Data One, https://dataone.org/.
[3] Dryad, https://datadryad.org/.
[4] PANGAEA, https://www.pangaea.de/.
[5] BGBM, https://www.bgbm.org/.

2.1 Query Expansion with Synonyms

For a given search term t, we denote $\mathbf{S}_t = \{s_1, s_2, \ldots, s_n\}$ as a set of synonyms for t. From a linguistic point of view, synonyms are two different words with the same meaning including different spellings and different languages. In contrast to most biological literature, we also subsume common and scientific names under this term. All s in $\mathbf{S}_t$ are appended to t with a logical OR (Fig. 1). By default, several search terms are connected with a logical AND. If there is no result for AND, an OR search is processed.

Our experiments in a previous study [8] point out that scholars are experts in their research domain, however, they are not familiar with all taxonomic terms. Given a user's search query, datasets with broader (superclass label) or narrower terms (subclass label) in the result list were not considered as relevant. For instance, when looking for data about butterflies, datasets containing scientific names of butterfly species such as *Vanessa atalanta* (Red admiral) got low relevance ratings. Interviews afterwards revealed that those datasets were not irrelevant per se, however, the scholars marked them as 'not relevant' since they were not aware of the taxonomic relationship between the entered keywords and the expanded terms. In contrast, datasets with synonyms got high relevance ratings. Therefore, in a first version, we only expand the search terms with synonyms.

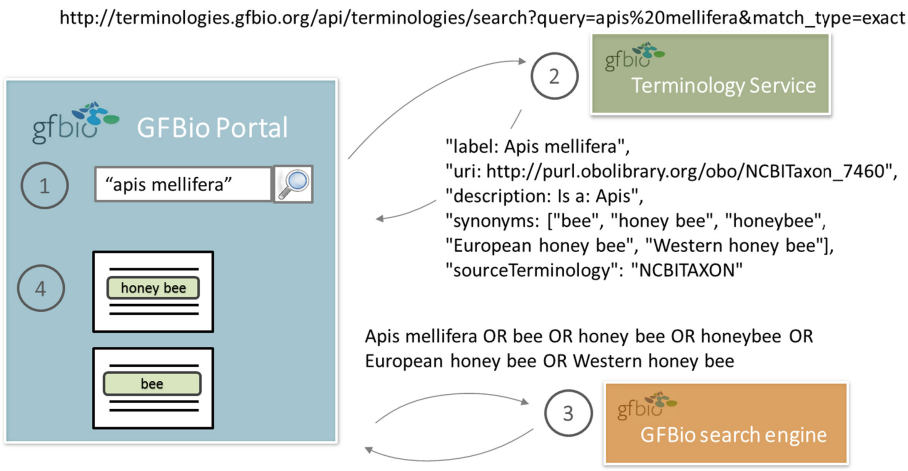

Fig. 1. (1) For all search terms, concepts from Terminology Service are looked up. (2) In a successful match, synonyms such as scientific and common names are extracted. (3) The expanded terms are sent to the search engine and (4) lead to a higher number of results.

2.2 User Interface

The user interface (Fig. 2) consists of a search field and two buttons for search, one for the ordinary keyword search and one for the semantic search. When entering a search term, auto-complete functionality suggests existing terms and phrases in the data repository. Quotes can be used to keep search terms together in the search result. All matched terms are highlighted in the result list.

2.3 Terminology Service

The GFBio Terminology Service (GFBio TS) [7] is the core semantic component of GFBio's infrastructure. It enables access to terminological knowledge necessary for annotation, semantic search, and integration of the increasingly heterogeneous project related datasets. Unlike existing terminology repositories like *Bioportal* or *Ontology Lookup Service* [2,9], the primary focus of GFBio TS is to provide tailored terminologies and services for the GFBio community. The GFBio community drives the selection and integration of terminologies, which can be either well-established ontologies like ENVO [1] or ontologies provided by the GFBio community like the KINGDOM ontology, describing a GFBio agreed list of species kingdoms. Terminologies are either internally hosted in a Semantic Web repository (Virtuoso) or externally accessed via their web services. Access to the GFBio TS is provided via a RESTful API, available terminologies can be accessed in a uniform way regardless of their degree of complexity and whether they are internally stored or externally accessed. An adapter component enables to harmonize the different internal and external schemas into a common Semantic Web compliant format. The service endpoints are grouped into four categories: metadata services, search services, information services and hierarchy-oriented services. The API documentation is available at: http://terminologies.gfbio.org/developer_section/api.html.

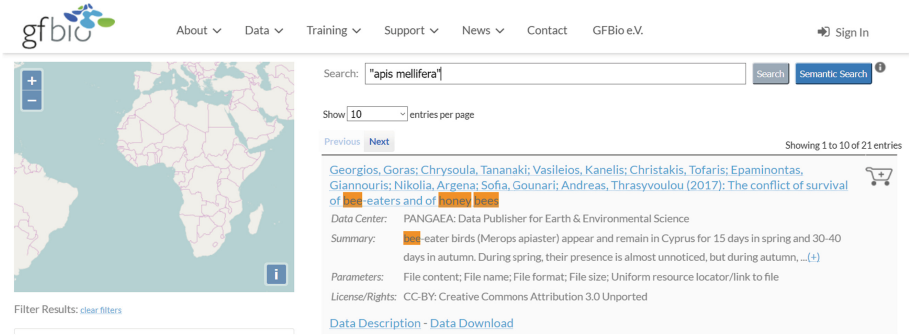

Fig. 2. User interface of GFBio's semantic search. The user's entered keywords are expanded in the background with related terms from GFBio's Terminology Service.

2.4 Search Engine

GFBio's metadata files are stored in *pansimple* format[6] which was primarily developed for PANGAEA. It is mainly based on Dublin Core metadata standard but contains additional fields such as parameters from primary data. An excerpt from an example file is presented in the listing below. GFBio uses *elasticsearch*[7] as search engine and TF-IDF weights as ranking function. Using elasticsearch's query-time boosting, originally entered keywords are higher ranked than expanded terms.

Listing 1.1. Excerpt from a biodiversity metadata file in pansimple format [6].

```
<dataset>
<dc:title>Wild bee monitoring in six agriculturally dominated landscapes of
    Saxony-Anhalt (Germany) in 2014</dc:title>
<dc:creator>Frenzel, Mark</dc:creator>
<dc:creator>[...]</dc:creator>
<dc:source>Helmholtz Centre for Environmental Research - UFZ</dc:source>
<dc:publisher>PANGAEA</dc:publisher>
<dataCenter>PANGAEA: [...]</dataCenter>
<dc:date>2016-09-29</dc:date>
<dc:type>Dataset</dc:type>
<dc:format>text/tab-separated-values, 47557 data points</dc:format>
<dc:identifier>doi:10.1594/PANGAEA.865100</dc:identifier>
<parentIdentifier>doi:10.1594/PANGAEA.864908</parentIdentifier>
<dc:relation>Papanikolaou, Alexandra D; Kuehn, Ingolf; Frenzel, Mark;
    Schweiger, Oliver (2016): Semi-natural habitats mitigate the effects of
    temperature rise on wild bees. Journal of Applied Ecology, doi:10
    .1111/1365-2664.12763</dc:relation>
[...]
</dataset>
```

3 Demonstration

In this demo[8], visitors will be able to use the semantic search and to compare search results of the standard and the semantic search. We will provide users with example queries, but they are also welcome to explore on their own. If interested, users can also directly access the Terminology Service.

Acknowledgements. This work was funded by the Deutsche Forschungsgemeinschaft (DFG) within the scope of the GFBio project.

References

1. Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C., Lewis, S., The ENVO Consortium: The environment ontology: contextualising biological and biomedical entities. J. Biomed. Semant. **4**, 43 (2013)

[6] Pansimple, https://ws.pangaea.de/schemas/pansimple/pansimple.xsd.
[7] Elasticsearch, https://www.elastic.co.
[8] https://www.gfbio.org/semantic-search.

2. Côté, R.G., Jones, P., Apweiler, R., Hermjakob, H.: The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinform. **7**, 97 (2006)
3. Diepenbroek, M., Glöckner, F., Grobe, P., Güntsch, A., Huber, R., König-Ries, B., Kostadinov, I., Nieschulze, J., Seeger, B., Tolksdorf, R., Triebel, D.: Towards an integrated biodiversity and ecological research data management and archiving platform: GFBio. In: Informatik (2014)
4. Dietze, H., Schroeder, M.: Goweb: a semantic search engine for the life science web. BMC Bioinform. **10**(S-10), 7 (2009)
5. Faessler, E., Hahn, U.: Semedico: a comprehensive semantic search engine for the life sciences. In: ACL 2017 - Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Vancouver, Canada, July 30–August 4 2017
6. Frenzel, M., Dussl, F., Höhne, R., Nickels, V., Creutzburg, F.: Wild bee monitoring in six agriculturally dominated landscapes of Saxony-Anhalt (Germany) (2014). doi:10.1594/PANGAEA.865100. In: Frenzel, M., Preiser, C., Dussl, F., Höhne, R., Nickels, V., Creutzburg, F.: (2016): TERENO (Terrestrial Environmental Observatories) wild bee monitoring in six agriculturally dominated landscapes of Saxony-Anhalt (Germany). Helmholtz Centre for Environmental Research - UFZ. doi:10.1594/PANGAEA.864908
7. Karam, N., Müller-Birn, C., Gleisberg, M., Fichtmüller, D., Tolksdorf, R., Güntsch, A.: A terminology service supporting semantic annotation, integration, discovery and analysis of interdisciplinary research data. Datenbank-Spektrum **16**(3), 195–205 (2016)
8. Löffler, F., Klan, F.: Does term expansion matter for the retrieval of biodiversity data? In: Martin, M., Cuquet, M., Folmer, E. (eds.) Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems (SEMANTiCS 2016). CEUR Workshop Proceedings (2016)
9. Noy, N., Shah, N., Whetzel, P., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D., Storey, M., Chute, C., Musen, M.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. **37**(Web-Server-Issue), 170–173 (2009)