# Does Term Expansion Matter for the Retrieval of Biodiversity Data?

Felicitas Löffler, Friederike Klan
Friedrich-Schiller-University Jena
Department of Mathematics and Computer Science
Ernst-Abbe-Platz 2-4, 07743 Jena, Germany
{felicitas.loeffler,friederike.klan}@uni-jena.de

## ABSTRACT

While term expansion techniques are well investigated for many domains, semantic enrichment of keyword queries for the retrieval of scientific datasets is still paid little attention to. In particular, a systematic analysis of which kind of semantically related concepts lead to the most relevant results is missing. Based on query expansion techniques, we semantically enriched search queries provided by biodiversity researchers to answer specific research questions. We applied them to a system indexing over 92,856 biological metadata files harvested from GFBio - the German Federation for Biological Data. We compared the outcome with the original keyword-based query. The result reveals that enriched keywords deliver a larger number of relevant datasets and that datasets retrieved based on keywords and their synonyms were judged more relevant. Query expansion with other related concepts returned a mixed picture.

## CCS Concepts

•**Information systems** → **Evaluation of retrieval results;** *Data management systems; Information retrieval query processing; Retrieval models and ranking;*

## Keywords

semantic search, biodiversity data, life sciences, evaluation

## 1. INTRODUCTION

Researchers in the life sciences are faced with an increasing abundance of data in scientific repositories. Thus, handling and filtering becomes a more and more time-consuming and challenging task [4]. Dataset retrieval is mainly based on metadata, i.e. describing information of the primary data, that differ in size, format and quality for the individual disciplines. This heterogeneity hampers scholars in finding the information they are looking for, since most scholarly data portals are still based on conventional keyword-based retrieval models (e.g., GFBio[1] or DataOne[2]). Furthermore, searching over biological datasets requires to consider two different perspectives: the view of the scholar providing data and the view of an information seeker looking for data. Given various research backgrounds and the implicit knowledge that is associated with each discipline, it often happens that scholars who offer data describe it using other terms than a person looking for that kind of data. For instance, a marine biologist who has conducted research on oxygen uptake rates in the Atlantic Ocean might describe his data mentioning the words *oxygen uptake* and *Atlantic Ocean*. An ecologist who is entering the keywords *respiration* and *marine area* into the search field would not be able to find this dataset although it is relevant. Disagreements or different spellings in species names and taxonomy are further obstacles in search applications that make it hard to find relevant results with a plain keyword-based search. Semantic query expansion is a common technique to enhance the quality of search results. However, there is no single strategy on *how* to expand the search terms based on underlying ontologies. In this paper, we will therefore investigate whether, compared to a classical keyword-based search, employing semantic query expansion techniques for retrieving biodiversity datasets will lead to a larger portion of relevant datasets, and which type of hierarchical relations are most effective. We developed a prototype of a semantic search based on query expansion operating on a data store with biodiversity metadata files from GFBio and conducted a user study with domain experts.

[1]GFBio, http://www.gfbio.org
[2]DataOne, http://www.dataone.org


*SEMANTICS 2016: Posters and Demos Track*
*September 13-14, 2016, Leipzig, Germany*

## 2. LITERATURE REVIEW

According to the definition by the semantic search workshops held in conjunction with *ESWC* and *WWW* from 2008-2011, semantic search systems can be roughly divided into two groups: (1) *Semantic Data Search* that focuses on the retrieval and ranking of data in triple stores [8] and (2) *Semantic Driven Information Retrieval* that enriches conventional search techniques with semantic data [5, 1]. Since 2011 a third group is arising aiming at transforming search queries in natural language into structured data so that Semantic Web techniques can be applied [9]. Driven by the annual *Question Answering over Linked Data (QALD)*[3] evaluation campaign, different aspects of linked data retrieval such as hybrid approaches with structured and unstructured data and different domains, e.g., the retrieval of bio-medical datasets, have already been addressed. However, little attention has been paid to the integration of hierarchical terms into the retrieval process.

An automatic query expansion idea for the bio-medical domain has been suggested by Haslhofer et al. [6]. They indexed PubMed articles with *Apache Lucene* and used the SKOS-based *MeSH*[4] ontology for annotating entities with URIs. In order to expand search queries, all kinds of label properties are considered, other hierarchical relations are left out. The first approach considering hierarchy in ranking is the filtering algorithm proposed by Maidel et al [10]. Developed for a personalized news-portal, they represent documents and user information as concepts of ontologies. The similarity measure distinguishes between a perfect, close or weak

[3]QALD, http://qald.sebastianwalter.org/
[4]Medical Subject Headings,https://www.nlm.nih.gov/mesh/

match depending on the hierarchical distance of the concepts. It turned out that including hierarchical terms led to a higher number of relevant results in any case. The best results were returned when integrating expansion terms from one level higher of the item profile. Including grandchildren or grandparent nodes seemed to be a not effective expansion method.

Common standards and methods as they exist in Information Retrieval (IR) with the annual TREC[5] competition are still missing for the evaluation of semantic search systems [12]. First steps towards common methodologies and metrics have been made in the SEALs project [13] that examines the user experience of using the search interface. It favors a two-phase approach comprising a fully automated assessment and a user-study. Motivated by the high demand for a test collection for semantic search engines and the associated high costs for setting up a new test corpus with relevance judgments, Blanco et al. [2] proposed an evaluation framework based on adapted TREC collections and crowdsourcing, which turned out to be a reliable and cost-effective way to evaluate a search application. Current evaluation of semantic search applications lack of analyzing the dependency individual system components may have on the produced search results. Layered evaluation strategies such as suggested by Paramythis et al. [11] for interactive adaptive applications, mitigate that problem. The core idea of their proposed evaluation framework is to analyze a system's behavior in layers, starting from the collection of input data over the used adaption strategy to the final personalized system. Each layer is evaluated with different methods that range from user tasks, discussion in expert groups to heuristic evaluation.

# 3. EXPANDING BIODIVERSITY TERMS

One issue we encounter when dealing with biodiversity data is that metadata descriptions often contain just the scientific name of a species a dataset refers to (e.g. a certain butterfly). This hampers dataset retrieval when users are interested in broader groups of plants or animals, since additional taxonomic information is missing in the metadata. Moreover, researchers might use the common name instead of the scientific name to search for data related to a certain species (e.g. butterfly instead of 'lepidoptera'). Hence, search techniques based on query expansion modeling the missing implicit knowledge are promising candidates to overcome these issues.

*Experimental Setup*: We randomly harvested 92,856 biodiversity metadata files from GFBio and indexed them with the search engine *GATE Mímir* [3]. The internal retrieval model is based on the classical TF-IDF approach. Afterwards, we requested 6 experienced biodiversity researchers (3 post-docs in marine biology, 2 post-docs in ecology, 1 scientific researcher in ecology) from 4 different organizations to provide five research questions related to their field of expertise each and also asked them to give proper search terms they would enter to find relevant data. For all provided search terms, we looked for matching concepts on two ontology collections (exact match to a concept label), namely the Terminology Server hosted by GFBio (GFBio TS) [7] and *Bioportal*[6]. A preference was given to the GFBio TS since its ontologies are tailored to the datasets that are available via the GFBio portal. In case of a successful match, both, the original set of keywords and the expanded version were sent to the search engine for dataset retrieval. This led to two different result sets displayed to the study participants side by side in a portal-based user interface. The overall flow is presented in Fig. 1 (left). We assume, that for a given search term all its synonyms as well as terms referring to narrower, i.e., more specific concepts, can lead to relevant results, too. For instance, a user interested in *Lepidoptera* would probably like to obtain more specific results, such as *Cameraria* (certain group of butterflies). Thus, for each concept with a label that matches one of the given search terms, we fetched all synonyms and direct sub-concepts and added them as expansion terms. If no narrower concepts were available, we selected the next broader term and all sibling nodes. For species names, the genus was regarded as the most specific concept, since scientific names typically already contain the genus. Thus, adding labels of sub-concepts, e.g., species names, would not lead to a higher recall. All expansion terms derived from an original term were connected with a logical OR, all original terms were combined among each other with a logical AND. If no corresponding concept was found for a given search term, just the term itself was included into the expanded query.

*Evaluation Method*: Following the layered approach suggested in [11], we investigated each step of the retrieval procedure individually. For evaluating the expansion strategy, we conducted post-interviews where all search terms used for query expansion were presented to the users. We asked the subjects to assess the relevance of these expanded terms on a binary scale and discussed the expansion strategy informally which delivered supplementary qualitative feedback. Finally, we evaluated whether the enriched query actually led to the retrieval of additional and relevant datasets. For that purpose, we compared the retrieval results based on the user-provided keywords with those resulting from the expanded query. For each query result, the users evaluated the relevance of the Top25 datasets. The relevance was assessed on a 7-point-Likert-scale from 0 (irrelevant) to 6 (highly relevant). Supplementary material to this evaluation study can be found on our web page[7].

# 4. RESULTS

In total, we got 581 human relevance judgments[8] for the retrieved datasets resulting from 38 queries (19 queries each as given and expanded). 34 out of the 36 provided search terms led to matching concepts in the ontologies. The number of extracted terms per keyword varied between 10 and 209,000. In particular for insects the number of expansion terms was very high which forced us to cut the number down. We only picked those expansions where metadata mentioning that term were available in the corpus.

## 4.1 Expansion Strategies

At first, we wanted to find out which term expansion strategies lead to additional relevant keywords, independently of the fact whether these keywords actually lead to a larger number of relevant datasets. We considered four kinds of expansion terms: synonyms, sub-classes (narrower concepts), sibling classes and super classes (broader concepts). For each provided search query, we presented a list of all potential expansion terms that could be derived from the considered ontologies (see supplementary material[7]) and asked the users to indicate if these terms are relevant with respect to their query or not. The result is shown in Fig. 1 (right). Depicted is the portion of terms that were marked as relevant and not relevant respectively, grouped by term type and averaged over all queries. It becomes evident, that among all types of expansion terms there is a substantial portion of terms that are relevant. Thus, all investigated types of terms are worth to be considered for term expansion. Surprisingly, the synonyms of a provided input keyword were considered as relevant in just two thirds of the cases (65%). Post-interviews afterwards revealed that the supposed synonyms

[5]TREC, http://trec.nist.gov/

[6]Bioportal, http://bioportal.bioontology.org/

[7]http://fusion.cs.uni-jena.de/fusion/semantics-2016/

[8]a researcher just judged the results for the queries he/she posed

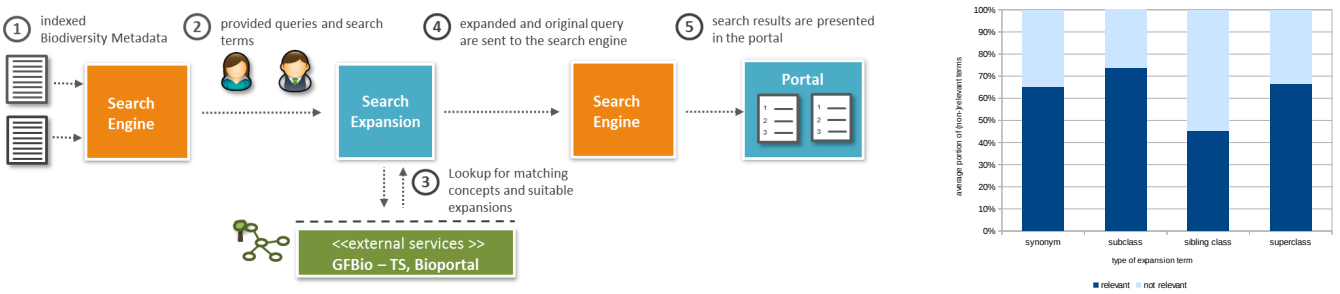


Figure 1: Experimental setup (left) and relevance of expansion term types (right)

turned out to be broader terms actually (modeling errors in the underlying ontologies). Expansion terms referring to sub-concepts are deemed relevant in nearly three quarters of the cases (74%). Terms pointing to more general concepts are also often (67%) pertinent. Sibling concepts are as often relevant (45%) as not.

## 4.2 Quality of the Search Results

Automatic query expansion is meaningless, if the extended keyword set does not lead to an improved set of search results. Thus, we compared the outcome of a classical keyword-based search with the result obtained from a search based on semantic query expansion. In particular, we investigated whether (a) term expansion permits to retrieve a higher portion of relevant datasets and (b) whether the retrieved datasets were considered more relevant by the users.

*Portion of Relevant Datasets:* To verify hypothesis (a), we measured precision and recall over the Top 25 and Top 10 datasets that have been retrieved using the original keywords provided by a user and on the expanded keyword set, respectively. Due to the large number of datasets in our repository and missing ground truth information about the relevance of these datasets, the total number of relevant datasets is unknown. Hence, we decided to determine the relative recall and precision of the keyword-based retrieval mechanism (KS) compared to the search with expanded search terms (ES). Relative precision (w.r.t. the Top 25) was calculated as

$$(1+\frac{\#\text{ relevant datasets KS}}{\min(\#\text{ datasets KS}, 25)})/(1+\frac{\#\text{ relevant datasets ES}}{\min(\#\text{ datasets ES}, 25)}).$$

If no datasets were retrieved, the absolute precision was set to 0 (just happened for the keyword search). Relative recall is defined as

$$(1+\ \#\text{ relevant datasets KS})/(1+\#\text{ relevant datasets ES}).$$

Relative precision and recall over the Top 10 are calculated analogously. These measures allow us to compare both retrieval mechanisms with respect to precision and recall. A value of 1 means equal precision/recall, a value lower/higher than 1 indicates that precision/recall of the semantic search is higher/lower than that of the pure keyword search.

Fig. 3 shows relative precision and recall averaged over all queries using synonyms (Fig. 3 left) and all queries using sub-concept labels (Figs. 3 middle and right). We only considered those queries where the added keywords were judged relevant by the users[9]. For the Figs. 3 left and middle, precision and recall were determined with respect to the Top 25 results, Fig. 3 right refers to the Top 10 results. In order to distinguish relevant from irrelevant datasets, we set a threshold for the ratings. Datasets with a higher rating were marked as relevant, those with a lower rating as not relevant. Fig. 3 presents the relative precision and recall for all possible choices of such a threshold (relevance value $> 0$, $> 1$, etc.).

From all 3 charts depicted in Fig. 3, it is evident that term expansion did not have much effect on the precision of the retrieval process (relative precision at around 1). Synonyms are an exception. Here, we observed a slight increase of the precision for the semantic search (relative precision lower than 1). A much stronger effect can be observed for the recall where the values are increased for the expanded search (relative recall lower than 1). The retrieval based on the term expansion strategy returns a larger fraction of relevant datasets including even datasets that have not been found by the original keyword-based search. The effect is very strong for synonyms and sub-classes but minor for super classes and sibling labels (not depicted, see supplementary material[7]). Finally, when comparing the Figs. 3 middle and right, it can be realized that the increase in recall for the semantically enriched search is larger when looking at the Top 25 results than when restricting it to the Top 10 results. This indicates that the ranking of the results is not optimal (see supplementary material[7] for other retrieval metrics). The observed deficiencies are not surprising since the ranking model of GATE Mímir is agnostic to the source of a search term.

*Relevance:* Apart from figuring out whether term expansion permits to retrieve a larger number of relevant datasets, we also wanted to analyze whether the retrieved datasets were actually more pertinent to a user's query. Therefore, we measured the fraction of the Top 25 results that were judged with a given relevance and averaged it over all queries that used synonyms (Fig. 2 left), all queries that used sub-concept labels (Fig. 2 middle) and all queries that contained labels of super classes and sibling concepts of the input terms (Fig. 2 right). Again, we considered only those queries where the added keywords were judged relevant by the users. In case that the retrieved datasets are actually more relevant, we would observe a decrease of the portions for low relevance values (relevance values $0-2$) and an increase of the portions for high relevance values (relevance values $4-6$). Fig. 2 left depicts the results for synonyms. In fact, we observe less results of low relevance (values 0 and 2) and more highly relevant results (values of 5 and 6), which means expanding keywords with their synonyms increases the portion of highly relevant results. Adding subclass labels produces a larger fraction of irrelevant datasets but increases the amount of relevant datasets (relevance values $3-5$). Adding super classes and sibling labels does not seem to have a positive effect on the relevance of the retrieved datasets. It leads to more results of low relevance (values 1 and 2) and less results of high relevance (value 5).

In post-interviews we asked the users why they gave low ratings to datasets obtained from hierarchical terms. It turned out that they did not know if the presented narrower or broader concepts are somehow related to their original search terms. Obviously, even domain experts are not familiar with all taxonomic terms, they need

[9]Since we wanted to evaluate if additional relevant keywords actually lead to relevant datasets, we thereby eliminated irrelevant datasets that would have resulted from irrelevant keywords.

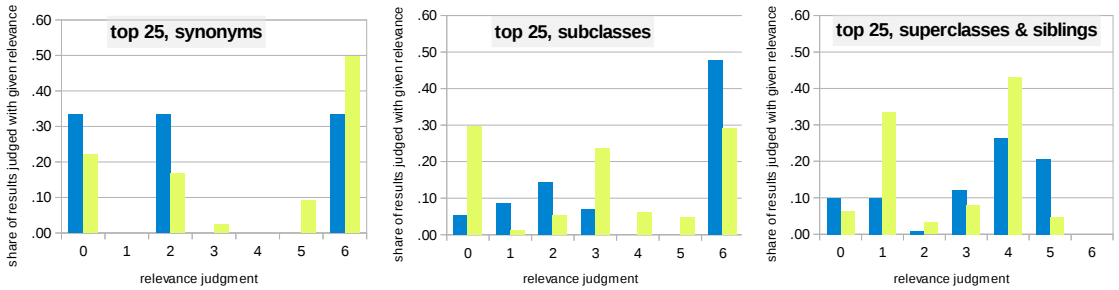


Figure 2: Distribution of average relevances for keyword-based (blue bars) and semantic search (green bars)

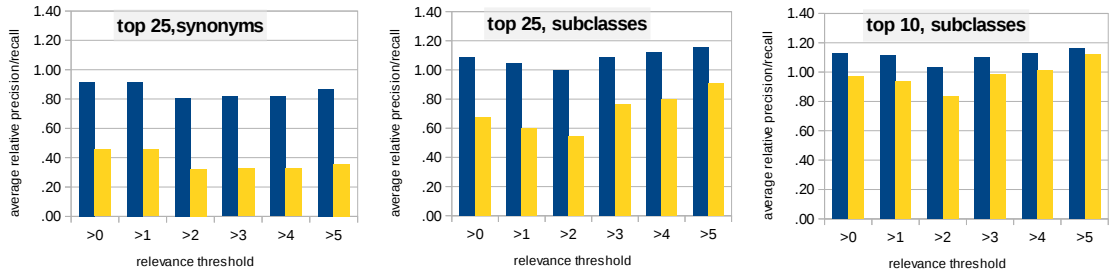


Figure 3: Average relative precision (blue) and recall (yellow)

explanations why expanded terms appear in the result set.

## 5. CONCLUSION

We developed a prototype of a semantic search based on query expansion and evaluated our system with a layered approach to account for possible dependencies between its individual components. The outcome reveals that enriching user queries with synonyms and subclass labels is an effective way for retrieving a larger number of relevant results and for finding datasets that are highly relevant. Other relations might be useful as well but require explanations how the expansions are related to the original query.

## 6. ACKNOWLEDGMENTS

We thank all users for their time and their valuable feedback. The work has been (partly) funded by the *Deutsche Forschungsgemeinschaft* (DFG) as part of GFBio and CRC 1076 AquaDiva.

## 7. REFERENCES

[1] A. Bakhtin, Y. Ustinovskiy, and P. Serdyukov. Predicting the Impact of Expansion Terms Using Semantic and User Interaction Features. CIKM '13, USA, 2013.

[2] R. Blanco, H. Halpin, D. M. Herzig, P. Mika, J. Pound, H. S. Thompson, and T. Tran. Repeatable and Reliable Semantic Search Evaluation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 21(0), 2013.

[3] H. Cunningham, V. Tablan, I. Roberts, M. Greenwood, and N. Aswani. Information Extraction and Semantic Annotation for Multi-Paradigm Information Management. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29. Springer, 2011.

[4] M. Diepenbroek, F. Glöckner, P. Grobe, A. Güntsch, R. Huber, B. König-Ries, I. Kostadinov, J. Nieschulze, B. Seeger, R. Tolksdorf, and D. Triebel. Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: GFBio. In *Informatik 2014*, 2014.

[5] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. IJCAI'07, USA, 2007.

[6] B. Haslhofer, F. Martins, and J. a. Magalhães. Using SKOS Vocabularies for Improving Web Search. WWW '13 Companion, USA, 2013.

[7] N. Karam, D. Fichtmueller, M. Gleisberg, V. Bohlen, R. Tolksdorf, and A. Güntsch. The Terminology Server of GFBio. http://terminologies.gfbio.org/.

[8] Y. Lei, V. Uren, and E. Motta. SemSearch: A Search Engine for the Semantic Web. In *Proc. of EKAW 2006*, pages 238–245, Czech Republic, 2006. Springer-Verlag.

[9] V. Lopez, M. Fernández, E. Motta, and N. Stieler. PowerAqua: Supporting Users in Querying and Exploring the Semantic Web Content. In *Semantik Web Journal*. IOS Press, 2011.

[10] V. Maidel, P. Shoval, B. Shapira, and M. Taieb-Maimon. Ontological Content-Based Filtering for Personalised Newspapers: A Method and its Evaluation. *Online Information Review*, 34(5), 2010.

[11] A. Paramythis, S. Weibelzahl, and J. Masthoff. Layered Evaluation of Interactive Adaptive Systems: Framework and Formative Methods. *User Modeling and User-Adapted Interaction*, 20(5), Dec. 2010.

[12] V. Uren, M. Sabou, E. Motta, M. Fernandez, V. Lopez, and Y. Lei. Reflections on Five Years of Evaluating Semantic Search Systems. *Int. J. Metadata Semant. Ontologies*, 5(2), 2010.

[13] S. N. Wrigley, K. Elbedweihy, D. Reinhard, A. Bernstein, and F. Ciravegna. Evaluating Semantic Search Tools Using the SEALS Platform. In *Proc. of ISWC 2010*, 2010.