FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

User-Centered Semantic Dataset Retrieval

Dissertation
zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik
der Friedrich-Schiller-Universität Jena

von Felicitas Maria Löffler, geb. Ritz
geboren am 31.05.1982 in Torgau, Germany

Gutachter

1. Prof. Dr. Birgitta König-Ries
   Friedrich-Schiller-Universität Jena, Institut für Informatik, Heinz-Nixdorf-Professur für verteilte Informationssysteme, Jena

2. Dr. habil. Clement Jonquet
   L'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE/ MISTEA) und Université de Montpellier (LIRMM), France

3. Prof. Dr. Harald Sack
   FIZ Karlsruhe, Leibniz-Institut für Informationsinfrastruktur und Karlsruher Institut für Technologie (KIT), Karlsruhe

Tag der öffentlichen Verteidigung: 07. Juni 2023

Ehrenwörtliche Erklärung

Hiermit erkläre ich,

- dass mir die Promotionsordnung der Fakultät bekannt ist,
- dass ich die Dissertation selbst angefertigt habe, keine Textabschnitte oder Ergebnisse eines Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen in meiner Arbeit angegeben habe,
- dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts haben mich folgende Personen unterstützt:

- Prof. Dr. Birgitta König-Ries
- Dr. Friederike Klan

Ich habe die gleiche, eine in wesentlichen Teilen ähnliche bzw. eine andere Abhandlung bereits bei einer anderen Hochschule als Dissertation eingereicht: Ja / Nein.

Jena, den 27. November 2022

[Felicitas Löffler]

To my parents and grandparents who were not allowed to study or finish their studies.

Acknowledgments

This thesis would have not been completed without the support and guidance of several people. At first, I would like to thank my main supervisor Prof. Dr. Birgitta König-Ries. In particular, I would like to thank for her patience and guidance through the past ten years, for an inspiring and pleasant working environment, for financial support and the opportunity for several research stays in Montréal, Canada. I'm also very thankful for all the fruitful discussions, hints and advice of my former colleague Dr. Friederike Klan. Thank you for all your contribution to finish this thesis!

My thanks also go to Bahar Sateli and René Witte from the Semantic Software Lab in Montréal, Canada for making my stays in Canada as productive and pleasant as possible. Under their supervision I have learned the foundations about text mining and information extraction. My research benefited greatly from their support.

I would also like to thank my colleagues Kobkaew Opasjumruskit and Sven Thiel for our time together in the GFBio project. It was always a pleasure and motivation to work with you. In addition, I'm also very thankful for the support of Fateme Shafiei. She made a significant contribution to the redesign and implementation of the search frontend. Many thanks also go to colleagues and partners in various biodiversity projects, e.g., for giving feedback in several workshops, for taking part in my requirement studies and evaluations. I also appreciated a collegial working atmosphere in the Fusion group. It was great to work with you all!

I also like to thank the external reviewers of this thesis, Dr. Clement Jonquet and Prof. Dr. Harald Sack for their time and effort to review this work.

I would also like to thank my family, my in-laws and friends for good conversations and happy evenings during the PhD time. Finally, my thanks also go to my husband Martin and my sons Richard and Constantin for believing in me and giving support during the long days of writing this work. Thank you for being in my life.

Abstract

The importance of finding relevant research data in daily research practice is increasing. Scholars search for pertinent data to compare and evaluate their scientific outcomes with other available data, to integrate data from various sources for developing novel research ideas or to reuse results instead of repeating experiments. One research domain with a high number of such data discovery tasks is biodiversity research, a research field dealing with the variety of species, genetic diversity and ecological diversity.

In various studies, scholars in the Life Sciences report on several difficulties they have in retrieving relevant data. Dataset search tasks are time consuming, users retrieve only partial pertinent data and important information can not be displayed in the user interface, because the information is missing in the metadata. In addition, most data portals use classical keyword based retrieval models in their search applications enabling only syntactic matches of query terms and content in datasets. Moreover, user studies are missing to examine the current obstacles in dataset search in more detail and to explore usability issues in search interfaces.

Overcoming these problems has motivated a number of research efforts in computer science such as text mining (extracting important information from text) and semantic search (a search with results going beyond the entered keywords). In particular, the emergence of the Semantic Web opens a variety of novel research perspectives, as the Life Sciences are a very active application domain providing around one third of all available terminologies and datasets in the Linked Open Data Cloud.

Motivated by these challenges, the overall aim of this work is to analyze the current obstacles in dataset search and to propose and develop a novel semantic dataset search for biodiversity research. We address the above mentioned obstacles in three main parts: (1) We evaluate the current situation in dataset search in a user study, and we compare a first version of a semantic search with a classical keyword search to explore the suitability of semantic web technologies for dataset search. (2) We generate a question corpus and develop an information model to figure out on what scientific topics scholars in biodiversity research are interested in. Moreover, we also analyze the gap between current metadata and scholarly search interests, and we explore whether metadata and user interests match. (3) We propose and develop an improved dataset search based on three components: (A) a text mining pipeline, enriching metadata and queries with semantic categories (entity

types) and URIs, (B) a retrieval component with a semantic index over categories and URIs (and various entity expansion strategies and entity-based retrieval models) and (C) a user interface that enables a search within categories and a search including further hierarchical relations. Following user centered design principles, we ensure user involvement in various user studies during the development process.

Our results show that scholars only obtain moderate relevant results in a data portal offering data for biodiversity research and that semantic technologies help to retrieve more relevant data. The analysis on the question corpus and metadata reveal that scholarly search interests in biodiversity research can be grouped into a limited set of semantic categories. These categories are less reflected in current metadata, as most data portals utilize general metadata standards without domain specific metadata fields. Another problem are inconsistent keywords and vocabularies in metadata fields. However, we show that text mining approaches can support the discovery of hidden but relevant information in descriptive metadata fields such as abstract and title. Finally, our improved semantic dataset search presents results going beyond the entered keywords and offers a search within domain specific categories. The system is independent from specific metadata formats and provides a user friendly interface with a keyword input. Now, users are able to retrieve more specific terms when looking with keywords from a higher semantic hierarchy level. Explanations on used terminologies and URIs are provided on demand. We evaluate the system on different levels: The novel text mining pipeline BiodivTagger, enabling the annotation of metadata with entity types and URIs, achieves moderate to good results with respect to the identification of relevant domain categories. The evaluation of various entity based retrieval models show that the expansion of the original query concept with descendant nodes of domain ontologies results in the highest accuracy. Therefore, the final system considers descendant nodes in the result set automatically. The usability evaluation with 20 scholars reveal that for unknown search tasks users prefer the classical single input field. However, the results for a second user interface, offering a form-based search with the possibility to search within domain categories, are very close to the results for the single input field. Concerning the provided explanations, users appreciate the highlightings of search terms and their related terms as well as the highlightings of additional biological terms.

Zusammenfassung

Relevante Daten zu finden ist eine zunehmend wichtige Aufgabe in der täglichen Forschungspraxis. Wissenschaftler suchen zum Beispiel nach Daten, um eigene Ergebnisse mit verfügbaren Daten zu vergleichen oder Ergebnisse wieder zu verwenden anstatt sie zu wiederholen. Eine Forschungsdomäne mit einer hohen Anzahl solcher Suchaufgaben ist die Biodiversitätsforschung, eine Wissenschaft, die sich mit der Artenvielfalt, der genetischen Vielfalt und der ökologischen Vielfalt beschäftigt. In mehreren Studien berichten Wissenschaftler in den Lebenswissenschaften über die Schwierigkeiten, die sie haben relevante Daten zu finden. Die Suche nach passenden Datensätzen ist Zeit aufwändig, Anwender finden nur teilweise relevante Daten und wichtige Informationen können in der Benutzeroberfläche nicht angezeigt werden aufgrund fehlender Informationen in den Metadaten. Zudem verwenden die meisten Datenportale klassische, schlagwortbasierte Suchalgorithmen, die nur syntaktische Übereinstimmungen von Anfrage und Inhalt im Datensatz abgleichen können. Des Weiteren fehlen Benutzerstudien, um die aktuellen Schwierigkeiten in der Datensatzsuche weiter zu untersuchen und Usability-Probleme aufzudecken. Diesen Problemen mit Methoden der Informatik zu begegnen ist Gegenstand verschiedener Forschungsrichtungen, z.B. im Text Mining, einem Forschungsgebiet, das darauf abzielt, wichtige Informationen aus Texten zu extrahieren und in der semantischen Suche, einer Suche, die Ergebnisse anzeigt, die über die eingegebenen Schlagworte hinaus gehen. Insbesondere das Aufkommen des Semantic Web hat eine Vielzahl an neuen Forschungsperspektiven eröffnet, da die Lebenswissenschaften eine sehr aktive Anwendungsdomäne sind, die ungefähr ein Drittel aller verfügbaren Terminologien und Datensätze in der Linked Open Data Cloud ausmachen.

Motiviert durch diese Herausforderungen, ist das Ziel dieser Arbeit, die aktuellen Probleme in der Datensatzsuche der Biodiversitätsforschung zu untersuchen und eine neuartige semantische Datensatzsuche für diese Domäne zu entwickeln. Wir adressieren die beschriebene Problematik in drei Teilen: (1) In einem ersten Teil untersuchen wir die aktuellen Probleme in der Datensatzsuche in einer Benutzerstudie. Wir vergleichen außerdem eine erste Version einer semantischen Suche mit einer klassischen, schlagwortbasierten Suche, um herauszufinden, ob semantische Technologien für den Einsatz in der Datensatzsuche grundsätzlich geeignet sind. (2) Wir erstellen einen Korpus mit Suchanfragen und entwickeln ein Informationsmodell, um zu verstehen, für welche Fachthemen sich

Wissenschaftler in der Biodiversitätsforschung interessieren. (3) Wir stellen eine verbesserte Datensatzsuche auf Basis von drei Komponenten vor: (A) eine Text Mining-Pipeline, die Metadaten mit semantischen Kategorien (Entitätstypen) und eindeutigen Identifikatoren (URIs) verknüpft, (B) eine Suchkomponente mit einem semantischen Index über Kategorien und Entitäten (und verschiedenen Entitäts-Expansionsstrategien sowie entitätsbasierten Suchmodellen) und (C) einer Benutzeroberfläche, die eine Suche innerhalb von Kategorien ermöglicht und weitere hierarchische Beziehungen in die Suche einbindet. Außerdem stellen wir sicher, dass gemäß benutzerzentrierter Gestaltungsprinzipien verschiedene Benutzerstudien während der Entwicklung durchgeführt werden.

Unsere Ergebnisse zeigen, dass Forschende nur moderate Suchergebnisse in einem Datenportal für die Biodiversitätswissenschaften erhalten und dass semantische Technologien helfen mehr relevante Daten zu finden. Die Analyse des Fragekorpus ergibt, dass sich wissenschaftliche Informationsbedürfnisse in der Biodiversitätswissenschaft durchaus in eine begrenzte Anzahl von semantischen Kategorien gruppieren lassen. Diese Kategorien sind bisher wenig in vorhanden Metadaten reflektiert, da die meisten Datenportale allgemeine Metadatenstandards ohne fachspezifische Felder verwenden. Ein weiteres Problem sind inkonsistent genutzte Begriffe in Metadatenfeldern. Wir demonstrieren, dass Text Mining-Ansätze das Erkennen von versteckten aber relevanten Informationen in beschreibenden Metadatenfeldern, wie z.B. 'Abstract' unterstützen. Im letzten Teil stellen wir unsere verbesserte semantische Datensatzsuche vor, die Ergebnisse anzeigt, die über die eingegeben Schlagworte hinaus gehen und eine Suche innerhalb von Kategorien ermöglicht. Das System ist unabhängig von spezifischen Metadatenformaten und stellt eine benutzerfreundliche Oberfläche mit einer schlagwortbasierten Eingabe bereit. Die Anwender erhalten spezifischere Ergebnisse, wenn sie mit Schlagworten suchen, die auf einem höheren semantischen Hierarchielevel liegen. Erklärungen über verwendete Terminologien und URIs werden nach Bedarf angezeigt. Wir evaluieren das System auf verschiedenen Ebenen: Die neuartige Text Mining-Pipeline *BiodivTagger* erreicht moderate bis gute Ergebnisse in Bezug auf die Identifikation von relevanten Fachkategorien. Die Evaluation der verschiedenen entitätsbasierten Suchmodelle zeigt, dass die Expansion der Entitäten aus den Eingabeschlagworten mit Nachkommen aus der Terminologie die höchsten Werte in Bezug auf die Genauigkeit aufweist. Daher enthält das finale System im Suchergebnis automatisch auch Nachfolgerkonzepte aus der Terminologie. Die Usability Evaluation mit 20 Wissenschaftlern zeigt, dass Anwender für unbekannte Suchaufgaben das klassische, Eine-Sucheingabefeld bevorzugen. Die Ergebnisse für die zweite Benutzeroberfläche mit einer formbasierten Suche, in der die Forschenden innerhalb von Kategorien suchen können, sind jedoch sehr nah an den Ergebnissen für das klassische Sucheingabefeld. Hinsichtlich der Erklärungen schätzen die Anwender sowohl die Hervorhebungen der eingegeben Suchbegriffe und ihrer verwandten Terme in den Suchergebnissen als auch die Hervorhebungen für zusätzliche biologische Begriffe.

Wissenschaftliche Publikationen und Vorträge der Promovendin

Publikationen

1. Löffler, Felicitas/ Shafiei, Fateme/ Witte, René/ König-Ries, Birgitta/ Klan, Friederike (2023): *Semantic Search for Biological Datasets: A Usability Study on Modes of Querying and Explaining Search Results* in König-Ries, B., Scherzinger, S., Lehner, W., Vossen, G. (Eds.), BTW 2023, Gesellschaft für Informatik e.V., https://doi.org/10.18420/BTW2023-56

2. Abdelmageed, Nora / Löffler, Felicitas / König-Ries, Birgitta (2023): *BiodivBERT: a Pre-Trained Language Model for the Biodiversity Domain* in Yamaguchi, A., Splendiani, A., Marshall, M. S., Baker, C., Bolleman, J., Burger, A., Castro, L. J., Eigenbrod, O., Österle, S., Romacker, M., Waagmeester, A. (Eds.): 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (SWAT4HCLS 2023), https://ceur-ws.org/Vol-3415/paper-7.pdf

3. Abdelmageed, Nora/ Löffler, Felicitas/ Feddoul, Leila/ Algergawy, Alsayed/ Samuel, Sheeba/ Gaikwad, Jitendra/ Kazem, Anahita/ König-Ries, Birgitta (2022): *BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain*. Biodiversity Data Journal 10: e89481. https://doi.org/10.3897/BDJ.10.e89481

4. Shafiei, Fateme/ Löffler, Felicitas/ Thiel, Sven/ Opasjumruskit, Kobkaew/ Grabiger, Denis/ Rauh, Pauline/ König-Ries, Birgitta (2021): *[Dai:Si] - A Modular Dataset Retrieval Framework with a Semantic Search for Biological Data* in Sanfilippo, E. M., Kutz, O., Troquard, N., Hahmann, T., Masolo, C., Hoehndorf, R., Vita, R., Algergawy, A., Karam, N., Klan, F., Michel, F., Rosati, I. (Eds.): S4BioDiv 2021: 3rd International Workshop on Semantics for Biodiversity, held at JOWO 2021: Episode VII The Bolzano Summer of Knowledge, September 11–18, 2021, Bolzano, Italy, http://ceur-ws.org/Vol-2969/paper4-s4biodiv.pdf

5. Löffler, Felicitas/ Schuldt, Andreas/ König-Ries, Birgitta/ Bruelheide, Helge/ Klan, Friederike (2021): *A Test Collection for Dataset Retrieval in Biodiversity Research*, Research Ideas and Outcomes , Vol. 7, Pensoft Publishers, https://doi.org/10.3897/rio.7.e67887

6. Löffler, Felicitas/ Wesp, Valentin/ König-Ries, Birgitta/ Klan, Friederike (2021): *Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs?*, PLOS ONE , Vol. 16, No. 3, Public Library of Science, p. 1-36, https://doi.org/10.1371/journal.pone.0246099

7. Löffler, Felicitas/ Wesp, Valentin/ Babalou, Samira/ Kahn, Philipp/ Lachmann, René/ Sateli, Bahar/ Witte, René/ König-Ries, B. (2020): *ScholarLensViz: A Visualization Framework for Transparency in Semantic User Profiles* in Taylor, K., Gonçalves, R., Lecue, F., Yan, J. (Eds.): Proceedings of the ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 19th International Semantic Web Conference (ISWC 2020), Globally online, November 1-6, 2020., http://ceur-ws.org/Vol-2721/paper485.pdf

8. Löffler, Felicitas/ Abdelmageed, Nora/ Babalou, Samira/ Kaur, Pawandeep/ König-Ries, Birgitta (2020): *Tag Me If You Can! Semantic Annotation of Biodiversity Metadata with the QEMP Corpus and the BiodivTagger*, Proceedings of The 12th Language Resources and Evaluation Conference (LREC), European Language Resources Association: Marseille, France, p. 4557-4564, https://aclanthology.org/2020.lrec-1.560/

9. Sateli, Bahar/ Löffler, Felicitas/ König-Ries, Birgitta/ Witte, René (2017): *ScholarLens: extracting competences from research publications for the automatic generation of semantic user profiles*, 2017-07, PeerJ Computer Science , Vol. 3, https://doi.org/10.7717/peerj-cs.121

10. Löffler, Felicitas/ Pfaff, Claas-Thido/ Karam, Naouel/ Fichtmüller, David/ Klan, Friederike (2017): *What do Biodiversity Scholars Search for? Identifying High-Level Entities for Biological Metadata* in Algergawy, A., Karam, N., Klan, F., Jonquet, C. (Eds.): Proceedings of the 2nd Semantics for Biodiversity Workshop held in conjunction with ISWC2017, Vienna, Austria, http://ceur-ws.org/Vol-1933/poster-paper-10.pdf

11. Löffler, Felicitas/ Opasjumruskit, Kobkaew/ Karam, Naouel/ Fichtmüller, David/ Schindler, Uwe/ Klan, Friederike/ Müller-Birn, Claudia/ Diepenbroek, Michael (2017): *Honey Bee Versus Apis Mellifera: A Semantic Search for Biological Data* in Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., Hartig, O. (Eds.): The Semantic Web: ESWC 2017 Satellite Events: Portorož, Slovenia, Springer International Publishing, p. 98-103, https://link.springer.com/chapter/10.1007/978-3-319-70407-4_19

12. Löffler, Felicitas/ Klan, Friederike (2016): *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 - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16), co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), CEUR Workshop Proceedings, http://ceur-ws.org/Vol-1695/paper2.pdf

13. Sateli, Bahar/ Löffler, Felicitas/ König-Ries, Birgitta/ Witte, René (2016): *Semantic User Profiles: Learning Scholars' Competences by Analyzing their Publications* in Semantics, Analytics, Visualisation: Enhancing Scholarly Data Workshop co-located with the 25th International World Wide Web Conference 2016, Montréal, Canada, https://link.springer.com/chapter/10.1007/978-3-319-53637-8_12

14. Beckstein, Clemens/ Böcker, Sebastian/ Bogdan, Martin/ Bruelheide, Helge/ Bücker, Martin/ Denzler, Joachim/ Dittrich, Peter/ Grosse, Ivo/ Hinneburg, Alexander/ König-Ries, Birgitta/ Löffler, Felicitas/ Marz, Manja/ Müller-Hannemann, Matthias/ Winter, Marten/ Zimmermann, Wolf (2014): *Explorative Analysis of Heterogeneous, Unstructured, and Uncertain Data: A Computer Science Perspective on Biodiversity Research* in Helfert, M./ Holzinger, A./ Belo, O./ Francalanci, C. (Eds.): Proceedings of the 3rd International Conference on Data Management Technologies and Applications, DATA 2014, Vienna, Austria, August 29–31, 2014, SCITEPRESS, p. 251-257, https://doi.org/10.5220/0005098402510257

15. Löffler, Felicitas/ Sateli, Bahar/ König-Ries, Birgitta/ Witte, René (2014): *Towards Semantic Recommendation of Biodiversity Datasets based on Linked Open Data* in G. Specht, H. Gamper, F. Klan (eds.): Proceedings of the 26th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 21.10.2014 - 24.10.2014, Bozen, Italy, http://ceur-ws.org/Vol-1313/paper_12.pdf

16. Löffler, Felicitas/ Sateli, Bahar/ König-Ries, Birgitta/ Witte, René (2013): *Semantic Content Processing in Web Portals* in Witte, R./Baker, C. J. O./ Butler, G./ Dumontier, M. (Eds.) Proceedings of the 4th Canadian Semantic Web Symposium part of the Semantic Trilogy 2013, Montreal, QC, Canada, July 10th, 2013 , Vol. 1054, http://ceur-ws.org/Vol-1313/paper_12.pdf

17. Rößner, Susanne/ Löffler, Felicitas/ Engelien, Heike (2006): *Konzeption einer IT-Kooperationsplattform für den Export von Dienstleistungen im Rahmen des Forschungsprojektes IDEE* in Meißner, K. / Engelien, M. (Eds.) Gemeinschaften in Neuen Medien (GeNeMe) 2006, Dresden, Germany, 2006. Tagungsband Technische Universität Dresden, p. 385-397

Vorträge

1. Löffler, Felicitas (2020): *What are main search interests in biodiversity research?*, RDA Data Discovery Paradigms IG Session, RDA 16th Plenary Meeting - Costa Rica (Virtual), November 2020

2. Löffler, Felicitas/ König-Ries, Birgitta (2019): *What's in this Collection Dataset? Semantic Annotation with GATE*. Biodiversity Information Science and Standards 3: e37184. https://doi.org/10.3897/biss.3.37184

3. Grygorova, Eelena/ Müller, Friedrich/ Hohmuth, Martin/ Schöne, David/ Löffler, Felicitas / König-Ries, Birgitta (2019): *How to increase dataset findability in Google? Enriching BEXIS 2 data with schema.org entities*, idiv Conference 2019

4. Löffler, Felicitas/ Astor, Tina/ Müller-Birn, Claudia (2018): *Unfolding Existing Data Publication Practice in Research Data Workflows in the Biological and Environmental Sciences – First Results from a Survey*, 10th International Conference on Ecological Informatics (ICEI2018). `https://doi.org/10.22032/dbt.37803`

5. Löffler, Felicitas/ Klan, Friederike/ König-Ries, Birgitta (2018): *How to Search for Biological Data? A Comparison of User Interfaces in a Semantic Search*, 10th International Conference on Ecological Informatics (ICEI2018). `https://doi.org/10.22032/dbt.37854`

6. König-Ries, Birgitta/ Triebel, Dagmar/ Huber, Robert/ Glöckler, Falko/ Güntsch, Anton/ Felden, Janine/ Löffler, Felicitas/ Hoffmann, Jana (2017): *Setting up an Interdisciplinary Data Infrastructure: Why Cooperation between Domain Experts and Computer Scientists Matters - An Experience Report from the GFBio Project*, Biodiversity Information Science and Standards, Vol. 1, Pensoft Publishers, `https://doi.org/10.3897/tdwgproceedings.1.20198`

7. Löffler, Felicitas (2015): *GFBio – The German Federation for Biological Data – yet another data infrastructure?*, iDiv conference 2015

Contents

Chapter 1

Introduction

> *"In the same sense that we need archives for journal publications, we need archives for the data "*
>
> - Jim Gray in "Fourth Paradigm", *Computer scientist and winner of the Turing Award*

In 2007, Jim Gray predicted a "Fourth Paradigm" [Hey et al., 2009], a data-intensive research practice across all disciplines being dependent on computer science and algorithms. Fiveteen years later, this vision is a reality. In particular, integrating scientific data to gain new insights is a key challenge that requires large computing capabilities and sophisticated algorithms.

In order to integrate research data, scholars need to find relevant data. Search systems - or the underlying research field information retrieval - are a well investigated and discussed research field in computer science. The result are a variety of retrieval algorithms available in search engines such as Elasticsearch[1] or Apache Solr[2]. Research data repositories utilize these search engines for dataset search. However, finding pertinent data is challenging and time consuming for scholars. These challenges have not been sufficiently addressed so far. The introduction of the Semantic Web [Berners-Lee and Hendler, 2001] and Linked Data [Heath and Bizer, 2011] broadens the possibility to enrich data sources and populate additional knowledge to the search process. The value of additional information in the form of ontologies in search have already been shown in semantic search approaches for document retrieval in the Life Sciences. However, semantic search systems focusing on scientific data and their special needs are lacking.

Understanding the user experience is essential to improve a dataset search. Studying the context of use allows the creation of requirements for enhanced data retrieval. Therefore, users need to be observed, interviewed and involved into the development process to ensure a product with a high user satisfaction. This becomes challenging when semantic technologies are utilized. For instance, it is still unclear whether Linked Data improves

[1] Elasticsearch, `https://www.elastic.co`

[2] Apache Solr, `https://solr.apache.org/`

dataset search and what additional information from knowledge bases are benefical for scholars. User centric semantic dataset retrieval needs more attention and is the main scope of this thesis.

1.1 Motivation

The following sections elaborately discuss the motivation for this thesis. Starting with a subsection about data-intensive research 1.1.1 explaining why computing capabilities are required in today's research practice, Subsection 1.1.2 introduces the new research field dataset retrieval and current challenges followed by a Subsection 1.1.3 about semantic search with its opportunities and challenges for dataset search. Subsection 1.1.4 presents our motivation for exploring the user experience in dataset search applications.

1.1.1 Data-Intensive Research

Automatic monitoring systems, computer simulations or the digitization of legacy data produce a vast amount of research data that are not manually manageable. For instance, in molecular biology, there is a massive demand for large computing capabilities, as molecular sequencing produces very large files. For example, a couple of years ago, the EMBL-EBI[3], a global molecular data repository, announced on its website that it contains 8.7 petabytes of data. Remote sensing for Earth observations is another example where high end computing capabilities for data scientists are required to handle the large number and size of images.

Apart from the exponentially increasing amount of data, progressively, scientific data management practices are changing. In many scientific fields, it is still common to store collected data only on a hard disk. This is difficult, as data formats and software change over time. If data stored on hard disk are not updated, in the worst case, it can be lost forever. Among many reasons, as [Güntsch et al., 2012] indicated, staff change and retirement resulted in tremendous loss of data. Nowadays, the awareness that data sharing and good data management are important key factors to improve data analysis and gain scientific insights is rising. Driven by stakeholders from academia, scientific publishers and funding agencies, scholars are expected to describe, publish and share research data. This development was additionally promoted by the publication of the FAIR principles [Wilkinson et al., 2016]. The adherence of these four principles ensures that data is findable, accessible, interoperable and reusable. Thus, in an increasing number of scientific publications the underlying data are being made openly accessible. This "Long-tail of Data" holds a great potential of unexplored data that are small in volume and lots in

[3]EMBL-EBI, `https://www.ebi.ac.uk/`

number. In order to find these data, new algorithms and techniques for efficient dataset retrieval and analysis are needed.

A discipline with a large number of unexplored data is biodiversity research, a field dealing with the diversity of life on earth - the variety of species, genetic diversity, diversity of functions, interactions and ecosystems [idiv, 2019, Loreau, 2010]. Collected and stored in different data formats, the datasets often contain or link to spatial, temporal and environmental data [Walls et al., 2014]. Many important research questions cannot be answered by working with individual datasets or data collected by one group, but require meta-analysis integrating a wide range of topics [Culina et al., 2018].

As application domain, this work uses scientific datasets from biodiversity research. In particular, we conducted our study with datasets and users from long-term biodiversity projects such as the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig[4], AquaDiva[5] and The German Federation for Biological Data (GFBio)[6]. Further information on the projects can be found in Chapter 2.4.

1.1.2 Dataset Retrieval

Datasets are defined as "a collection of related observations organized and formatted for a particular purpose" [Chapman et al., 2019]. In contrast to classical information retrieval, where unstructured, textual resources are the main source, dataset retrieval is mainly based on metadata. Metadata are a supplementary or supporting descriptions of primary data. They are mainly provided in half-structured formats such as XML[7] or JSON[8]. A survey among 98 data repositories showed that data archives either use all available metadata (58%) or partial metadata (52%) in search [Khalsa et al., 2018]. Only in some research fields, primary data can be directly used for indexing. For instance, taxonomic data in GBIF (Global Biodiversity Information Facility)[9] are provided in XML based standards such as ABCD[10] or Darwin Core[11].

Retrieving pertinent data from data repositories is becoming increasingly important. For instance, the annual GBIF report [GBIF, 2018] registers a growing number of GBIF data being reused. Scholars reuse data for various purposes. For some research tasks it is not feasible to collect all data in the field, or scholars want to compare their data to data collected at a different time or place. Hence, retrieving relevant datasets is nowadays an essential part in daily research practice. However, studies

[4] idiv, `https://www.idiv.de/`
[5] AquaDiva, `https://www.aquadiva.uni-jena.de/`
[6] GFBio, `https://gfbio.org`
[7] XML, `https://www.w3.org/standards/xml/`
[8] JSON, `https://www.json.org/`
[9] GBIF, `https://www.gbif.org`
[10] ABCD, `https://abcd.tdwg.org/`
[11] Darwin Core, `https://dwc.tdwg.org/`

are rising that describe obstacles and report on challenges in dataset search. For instance, a large study by [Kacprzak et al., 2018] reveals that users had to send direct requests to data portals, because they could not find relevant data. And a recent study by [Gregory et al., 2020] reports that scholars searched for data in publications or utilized text search engines such as Google[12]. Obviously, even the launch of the Google Dataset Search [Benjelloun et al., 2020] with almost 30M datasets from various domains such as Social Sciences 26.2 %, Geosciences 19.0% and Biology 15.2% has not improved dataset discovery yet, and scholars prefer to search in the text search rather than in the dataset search.

Only very limited evaluation studies are publicly available studying these issues. Therefore, this thesis aims to explore current obstacles in dataset retrieval.

1.1.3 Semantic Search

Semantic search is a very broad field with no exact definition what it comprises. The only common agreement exists on the term *semantic*. *Semantic* stands for "meaningful" or just "meaning". [Bast et al., 2016] define a semantic search as a "search with meaning".

A semantic search goes beyond the classical keyword-search, which only returns documents syntactly matching a user's entered query. This broader search supports the discovery of relevant documents containing semantically related terms. For instance, a search for a profession could also return famous people, who have practiced this profession, or a search for a city could also return highlights for sightseeing. This related information comes from structured knowledge bases, graph-based storages providing data in triples and formal languages such as the Resource Description Format (RDF)[13] or the Web Ontology Language (OWL)[14]. In order to retrieve data from these knowledge bases, novel query languages such as SPARQL[15] have been developed. A brief introduction into common Semantic Web standards and techniques are provided in Chapter 2.3.

Over the last decade, numerous knowledge bases were developed for a variety of research domains. A very active domain are the Life Sciences (see also Chapter 2.3.4). Various taxonomies (hierarchies with no semantic relationship between the individual concepts, only `is-a` relations) and ontologies (structured knowledge with concepts and more complex relations such as `part_of` or `develops_from`) have been launched from this community. For instance, UMLS[16] and Bioportal[17] provide large numbers of on-

[12]Google, `https://www.google.com`
[13]RDF, `https://www.w3.org/RDF/`
[14]OWL, `https://www.w3.org/OWL/`
[15]SPARQL, `https://www.w3.org/TR/rdf-sparql-query/`
[16]UMLS, `https://www.nlm.nih.gov`
[17]Bioportal, `https://bioportal.bioontology.org/`

tologies and services for the biomedical domain. The BioASQ[18] or TREC Live QA[19] challenges are annual competitions in the fields of question answering. Furthermore, workshops in conjunction with Computer Linguistic or Semantic Web conferences have been organized such as the BioNLP[20], BioCreative[21] and S4biodiv[22].

Driven by these efforts, multiple semantic search systems in the Life Sciences have already been introduced (see Chapter 3.1.2). Most of them are based on unstructured text such as scientific articles, or they work solely over Linked Data. Approaches that attempt to include ontological knowledge into textual resources are rare. Moreover, due to this lack of approaches for combined data sources, there are also few studies exploring the suitability of the multitude of available semantic relations for dataset search. It has not been studied so far whether semantic techniques are benefical for dataset discovery. Therefore, we address this topic and examine different semantic strategies and approaches for dataset retrieval. A second open issue in semantic search applications are appropriate user interfaces. A key challenge is to hide the complexity of SPARQL statements but to enable the power of this new technology for end users. We also address this challenge and propose a user interface for a semantic dataset search.

1.1.4 User Experience

Users' experiences with search applications are often influenced by the user interface. A proper structure, a clear navigation and consistency in the design elements enhance the user acceptance, the usability (the characteristic to what extent users can complete their tasks, in what time and in which context, see also Chapter 2.1) and hence trust in a novel system. On the contrary, too many colors with no contrast, buttons in different shapes and colors, hidden navigations and dead or unhighlighted links distract users and even prevent users from fulfilling their tasks. In addition to the user interface, the underlying search engines often provide different functionalities. For instance, a search for "Apis mellifera" (the scientific name for 'honey bee') in PANGAEA[23] leads to the same number of results, no matter whether a user enters the query with quotes or without quotes. In contrast, a search in Zenodo[24] leads to different results when searching with or without quotes. Further issues can occur when users search with plural forms instead of singular or make mistakes in spelling.

For dataset search and semantic search, very few user studies are publicly available. User studies are time consuming and personnel intensive, which are very likely the main

[18] BioASQ, `http://bioasq.org/`
[19] TREC Live QA, `https://trec.nist.gov/data/qamain.html`
[20] BioNLP,`https://bionlp.sourceforge.net/index.shtml`
[21] BioCreative, `https://www.biocreative.org/`
[22] S4biodiv, `https://fusion.cs.uni-jena.de/s4biodiv2021/`
[23] PANGAEA, `https://pangaea.de`
[24] Zenodo, `https://zenodo.org/`

reason for the lack of studies. Study planning, recruiting participants and organizing the evaluation setup as well as the analysis afterwards take a lot of time and require additional expertise. However, usability issues can only be identified and eliminated by means of user studies. A user interface with a good user experience saves money for service and support, as it prevents users from making mistakes. This finally enhances the acceptance of and trust in digital systems. Therefore, the process of analyzing current obstacles in dataset search as well as the development of a novel semantic dataset search should be accompanied by various user studies.

1.1.5 Goal

The aim of this thesis is to identify obstacles in a dataset search application and propose an improved semantic dataset search for biodiversity research using a user-centric design process. The identification of search interests and obstacles in dataset search applications form the basis before a concept and an enhanced semantic search system is introduced and evaluated. Various user studies are integrated into the requirement and development process.

1.2 Use Cases

To emphasize the necessity for improved retrieval approaches with semantic technologies, we introduce two use cases in biodiversity research, which so far are not sufficiently supported by dataset search applications. The uses cases were originally described by [Thessen et al., 2015]. For each use case, we provide the research question that should be answered by datasets and summarize the current workflow scientists have to go through.

- **Use Case 1:** "Which traits are common to or vary the most in beetles collected from deserts?" [Thessen et al., 2015]

  In a first step, scholars have to think about specific traits to search with, e.g., "cuticular texture, hairiness, and color" [Thessen et al., 2015]. Afterwards, they can start searching for beetle datasets with the determined traits. However, not all data descriptions provide information on phenotypes or do not provide it in a format or with a terminology that is searchable. Therefore, another workaround is to remember descriptions from older literature in Latin. Finally, if datasets have been found, scholars need to verify the datasets found to ensure that the retrieved species actually lived in deserts at the time of collection in the field. To ensure reliability and authenticity, scholars refer to additional data sources such as the Global Biodiversity Information Facility (GBIF)[25]. Sometimes scholars have to visit biological

[25] GBIF, `https://www.gbif.org/`

collections to verify the taxonomic, spatial and temporal information from the labels located on the specimens.

- **Use Case 2:** "Do species that have independently reduced or lost their eyes share common environments now or in the past?" [Thessen et al., 2015]

  To answer this research questions, scholars visit and examine several hundred species of freshwater fishes in natural history museums and biological collections. They classify fish species into three categories: absent eyes, reduced eyes and fully developed eyes. After further scientific evaluations, they conclude that "eye reduction and loss has occurred independently several times in this clade" [Thessen et al., 2015]. This results in a new hypothesis: fish species adapt their eyes to their habitat. In order to test this hypothesis, scholars go to museums again and look closer into the data descriptions, specifically into the element "verbatimLocality" of the Darwin Core standard[26] (a data standard for biodiversity data, see also Chapter 4.3) describing the habitat. They notice that the descriptions vary greatly, but there are also terms that could be synonyms. After further research, they find that "species with reduced or absent eyes are all from subterranean environments" [Thessen et al., 2015]. Now, they can continue research on other environmental factors that might have an influence on that habitat.

1.3 Problem Statement

The use cases of current practices in biodiversity research reveal, that research data are scattered across various repositories and institutions and scholars need to search at different places and portals to obtain relevant data. Once suitable datasets are found, they are in different formats, or descriptions are incomplete.

Searching for datasets is an increasingly important task in daily research work, but scientists are not satisfied with the search results returned by dataset search applications. A study by [Kacprzak et al., 2018] confirms that 40% of the users of an open data portal said that they could not find the data they were interested in and thus directly requested the data from the repository manager. In different publications, ecologists mentioned obstacles when looking for datasets, e.g., missing primary data, missing information on how the studies were conducted [Parker et al., 2016] and scattered data across various data portals [Culina et al., 2018, Ramakers et al., 2018]. The participants in a study by [Gregory et al., 2020] reported on dataset search as a challenging (73%) or even difficult (19%) task. From the user perspective, the most challeging issues are "inadequate search

[26] verbatimLocality, `http://rs.tdwg.org/dwc/terms/verbatimLocality`

tools, a lack of skill in searching for data, or the fact that their needed data are not digital" [Gregory et al., 2020]. As a consequence, scholars look for datasets in publications mentioning research data, or they utilize general search engines such as Google[27].

So far, **we do not know why current search applications in data portals do not work properly for dataset search**. Only very few user studies explored obstacles in dataset search. For instance, user studies are missing on existing retrieval techniques and their appropriateness for dataset search. Search applications usually consist of a given corpus, a query and a retrieval model to return a ranked list of documents that match a query (see also Chapter 2.2.2). In dataset retrieval, the underlying documents are metadata files, descriptive information about the primary data. Most data portals utilize conventional retrieval models such as the Vector-Space-Model [Khalsa et al., 2018], a keyword based approach returning data that match the user's query exactly. It is successfully used in a wide range of applications and performs well when queries are very specific and users want to retrieve the documents that exactly contain their search terms. Yet, little research has been done whether keyword based retrieval leads to relevant results in dataset search. In particular, there is a lack of investigations in applied domains. Hence, further efforts and studies are necessary to analyze how scholars search for scientific data and to tackle the root causes for the current problems.

The Semantic Web with its vision of a web of data opens a variety of new research opportunities in computer science. Due to the large amount of available terminologies and datasets in the Linked Open Data (LOD) cloud with respect to the Life Sciences, numerous research activities in computer science have emerged to address the above mentioned obstacles. One example for such a research field is semantic search. In particular, a variety of approaches exist for the biomedical domain that link search terms to entries in taxonomies or ontologies. These enriched queries are then sent to search engines and return not only exact matches, but the result set also contains documents with synonyms or alternate labels for the originally entered query terms, e.g., [Thomas et al., 2012, Wei et al., 2019, Ernst et al., 2019]. However, it is still unclear whether query expansion is beneficial for dataset retrieval. We still do not know **how to properly integrate additional semantic information into the retrieval process for dataset search**. For instance, are any specific semantic relations relevant for dataset discovery? And how to further enhance semantic search for dataset retrieval? In particular, it is still a technical challenge to design and develop a search over metadata and knowledge bases in one search process. Utilizing entities (concepts) in the retrieval process instead of keywords would allow a more powerful search - an intelligent search that can expand the scope not only on synonyms and alternate labels, but also enables an expansion on available semantic relations. A second issue in semantic search systems are **missing appropriate user interfaces**. Quering knowledge bases requires SPARQL knowledge, which is too complex for end

[27] Google, `https://www.google.com`

users. Open questions are for instance how to design a user interface that hides the complexity of ontologies and vocabularies, but uses the full potential of SPARQL queries. And what concepts or additional information are necessary in the presentation of search results to foster comprehensibility and reliability?

1.4 Hypotheses

Based on the described problem statement (Section 1.3), we define the main hypothesis as follows:

It is possible to analyze, describe and enrich scientific datasets and scholarly information needs so that scholars are able to retrieve, understand, and reuse scientific datasets with an improved retrieval system.

Due to our association with various biodiversity research related projects, the predominantly studied domain is biodiversity science. The main hypothesis can be divided into the following sub-hypotheses:

Hypothesis H_1:
Current dataset search systems do not support (biodiversity) scholars sufficiently to find relevant datasets.

Hypothesis H_2:
A major problem are insufficiently described metadata. Metadata fields do not entirely reflect scholarly search interests. Data providers' search applications are mainly based on these insufficient metadata.

Hypothesis H_3:
Descriptive parts in metadata such as titles or abstracts implicitly contain useful information reflecting information needs of the respective domain. This hidden data can be extracted and semantically enriched with text mining techniques.

Hypothesis H_4:
A concept-based retrieval model that considers semantic relations in ranking returns more relevant results than a concept-based retrieval model using no additional semantic relations in ranking.

Hypothesis H_5:
Users are more efficient in a semantic dataset search that allows a search over semantic categories in comparison to a semantic dataset search offering a single input field as search

input. Users find explanations including utilized semantic technologies and terminologies more useful than a dataset search with less semantic information.

1.5 Objectives

The main goal of this thesis is to analyze current obstacles in dataset search and to propose and develop an improved dataset search for biodiversity research based on user-driven design principles. In order to verify our hypotheses, we define two sub-objectives:

1. O1: Identify obstacles in current dataset search systems for the green Life Sciences considering all three building blocks in dataset retrieval: the data source, information needs (user queries) and the retrieval process (addressing H_1 - H_2).

2. O2: Design and develop an improved dataset search that takes scholarly information needs into account, describes scientific data appropriately and returns relevant results that are comprehensible and reusable (addressing H_3 - H_5).

1.6 Overview of the Proposed Solution

An overview of the thesis is illustrated in Figure 1.1. The proposed solution is designed in accordance with the three parts of a retrieval system (Chapter 2): user input (queries), retrieval process and data source. Each row in the table displays the outcomes for the identification of obstacles (O1) and a solution (O2). User-centered design principles and evaluations are considered in all steps. Each cell in the table contains novel developments, gold standards or test collections denoted with light icons and/or evaluations. We distinguish between user requirement studies and formative evaluations (gear icon) and summative evaluations (checklist icon). The user icon denotes evaluations with user involvement.

At first, we carried out an evaluation of an existing dataset search (Hypothesis H_1, Chapter 3) for biodiversity research to analyze if users are satisfied with the given retrieval system and if they are able to find relevant datasets. A user survey was conducted afterwards to obtain further insights on current obstacles. Based on these findings, we wanted to know if semantic enrichment actually improves dataset retrieval systems. Therefore, we developed a first prototype of a semantic search, and we conducted tests with scholars to determine if a semantically enhanced search outperforms a keyword based search (Hypothesis H_1, Chapter 3). Subsequent interviews complemented this formative study.

The results pointed us back to the user input and led to a large study on scholarly information needs in biodiversity research (Hypothesis H_2, Chapter 4.2). The aim of this study was to understand what search interests and domain categories are important for biodiversity research, and therefore should be considered in a retrieval system. The

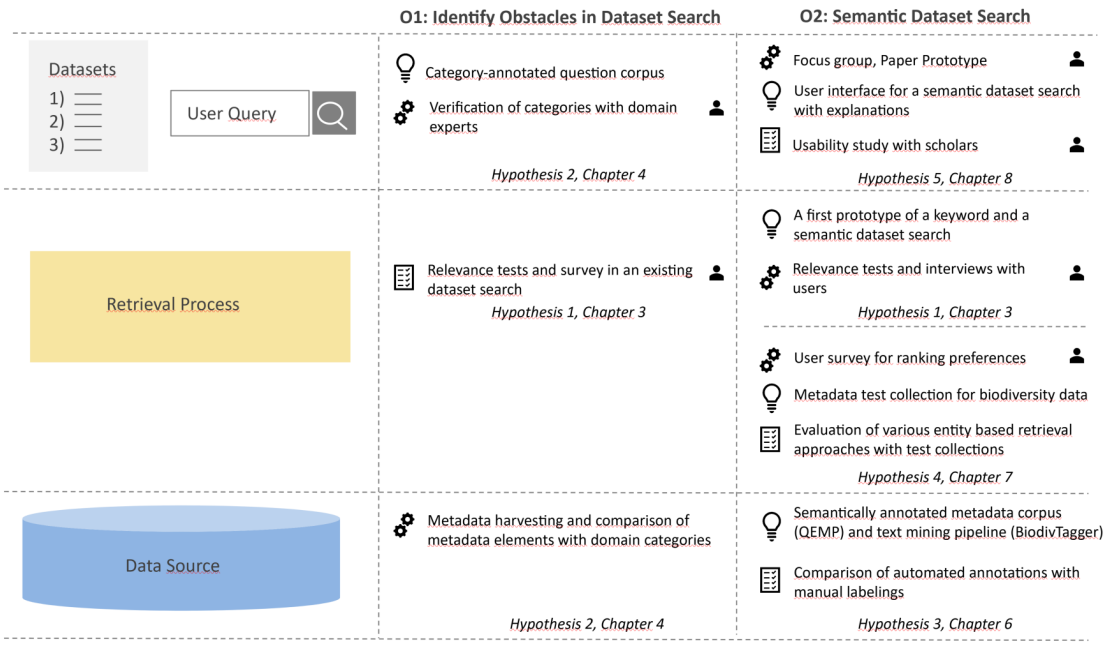


Figure 1.1: Overview of the thesis. The light bulb denotes reusable outcomes such as gold standards, test collections and development parts. The checklist icon indicates requirement studies and evaluation types (formative and summative). Entries with a user icon highlight evaluations with user involvement.

result is a question corpus containing both search questions (questions, scholars have in mind when searching for datasets) and research questions (questions that are not necessarily tailored to a search but require datasets implicitly) to analyze the research context a scholar is involved. We annotated the questions' noun entities and their nested components manually with domain categories (entity types). In a subsequent evaluation with domain experts, we explored the comprehensiveness and representativeness of the proposed categories. Afterwards, we analyzed the occurence of the identified categories in common metadata standards and whether the categories are reflected in existing metadata in five large data repositories (Hypothesis H_2, Section 4.3 and Section 4.4). This allowed us to quantify the gap between existing metadata and scholarly information needs in biodiversity research.

Due to the findings of these studies, we were able to list and describe obstacles in dataset retrieval systems. Hence, we propose an improved retrieval system for existing research data using semantic enrichment in all three building blocks (Chapter 5). The first component is a text mining pipeline linking metadata with concepts from selected domain ontologies and assigning an entity type (Hypothesis H_3, Chapter 6). We evaluated the pipeline with a novel gold standard for metadata. The second part describes the retrieval component (Hypothesis H_4, Chapter 7). Based on the semantic annotations created in the previous step, a concept based retrieval is now possible. In order to get a better understanding of scholars' ranking preferences and what kind of semantic relations support dataset retrieval, we conducted a user survey. The outcome gave insights on the selection

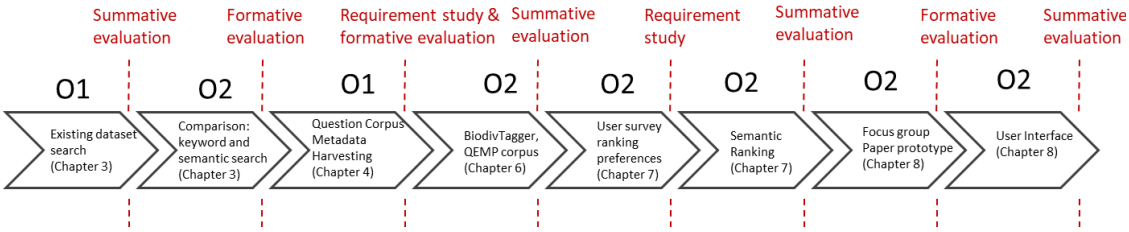


Figure 1.2: Research methodology of the thesis: According to an agile development, O1 and O2 were processed in iterations. Following a user-centered design, we collected user requirements and conducted formative and summative evaluations iteratively.

of entity based retrieval models and semantic relations to be considered. We evaluated the retrieval approach with two test collections from the Life Sciences. One of this ground thruth information was created in cooperation with a domain expert. We developed a user interface (UI) framework (Hypothesis H_5, Chapter 8) to provide convenient access to the novel semantic dataset search. We propose two user interfaces based on discussions and decisions in a focus group. A final evaluation with scholars took place to assess the user experience and usability.

1.7 Research Methodology

The research methodology in this thesis oriented on an agile development process (Figure 1.2) consisting of several iterations. The outcome of one increment served as input for the following increment [Sommerville, 2021].

User feedback was considered throughout the entire thesis process. We collected user requirements, conducted formative evaluations to assess a solution in an early stage to make improvements or to obtain further insights for the development process, and we carried out summative evaluations to support the assessment of the overall experience or accuracy of a proposed solution 2.

At first, we conducted a relevance evaluation on a current dataset search in a summative evaluation (Chapter 3). Afterwards, we developed a prototype of a semantic search and evaluated the prototyp in a comparative, formative study at the very beginning of the thesis to estimate whether users actually benefit from semantic technologies in a dataset search. This outcome mainly inspired and influenced the following iterations. Subsequently, we conducted a more thorough analysis of user' information needs and available data before we developed a final concept for a semantic dataset search. In a formative evaluation, we verified the proposed domain categories (Chapter 4.2). The domain experts sorted provided terms and phrases into pre-defined categories (card sorting). In an analysis on existing metadata standards and metadata in data repositories, we collected quantitative data to determine the biggest obstacles in metadata descriptions. These findings gave further insights on the subsequent development of an improved dataset search.

In order to get feedback on ranking preferences in a dataset search, we presented search scenarios in a survey to domain experts (Section 7.3). The results of this user requirement study led to the entity expansion strategies and retrieval approaches introduced in Section 7.5 and Section 7.6. We evaluated the retrieval part with two test collections. One of those was created in the scope of this work (Section 7.4). The discussions and testings of the suggested user interfaces (Chapter 7) in a focus group (Section 8.2.3) facilitated the decision on the final user interfaces and the usability methods for the subsequent user study. In a final summative evaluation with 20 scholars, we assessed the usability of two main search components (query input and explanations) in two proposed user interfaces for dataset search (Chapter 8).

In addition to evaluations, comprehensive literature studies for the different involved research fields complemented the requirement analysis for each iteration. Due to the heterogeneity of the research fields, each iteration required separate literature studies. It was necessary to gain comprehensive knowledge in user-centered design and user evaluations. Moreover, familiarity with information retrieval and Natural Language Processing was required as well as knowledge in semantic web technologies. Finally, in order to be able to work and understand the data sources, basic foundations had to be gained in biodiversity research and biodiversity informatics. We provide basic knowledge for each research field in Chapter 2. We also decided to provide related work at the beginning of each chapter as we think, this guides readers better through the thesis and presents relevant literature where it is needed.

Further additional insights for the proposed solutions were gained during the participation in the GFBio[28] research project. GFBio is a national infrastructure and contact point for issues concerning research data management and standardization of biological and environmental data. The project was established in 2013 with the main goal to support scholars in all steps of the data life cycle, from acquisition to archiving and data publication. The team at the Friedrich Schiller University Jena (FSU Jena) was responsible for the data portal, in particular for the data search, user management and integration of tools and services developed by GFBio partners. Being part of the project from 2013 to 2021, the involvement in the design, implementation and maintenance process of GFBio's dataset search strongly influenced this thesis. In particular, several workshops, telephone conferences and task groups gave important hints towards a concept for an improved search. First results of the thesis were also incorporated into the GFBio search. GFBio's semantic search [Löffler et al., 2017] was developed based on the findings of the first semantic prototype of this thesis. In this novel semantic search, query terms were expanded with services from the Terminology Service [Karam et al., 2016] and extended search results on related terms such as synonyms, common and scientific names. The outcomes of the first research goal (O1) also led to improvements of GFBio's faceted search.

[28] GFBio, `https://www.gfbio.org`

Chapter	Contribution	Data and Code Availability	Publication
Chapter 3, H_1	A first prototyp of a semantic search based on query expansion was developed to evaluate whether semantic techniques are benefical for dataset search. For this methodology, we provide a JAVA based command line tool to obtain synonyms, alternate labels, descendant and ancestor nodes from two terminology services. A demonstration of this approach (limited to synonym expansion) is integrated into the GFBio search (`https://search.gfbio.org/`)	`https://doi.org/10.5281/zenodo.7374539`,	[Löffler and Klan, 2016], [Löffler et al., 2017]
Chapter 4, H_2	We created a manually annotated and evaluated question corpus for biodiversity research representing main search interests (information categories) in biodiversity research. In addition, scripts and survey templates are provided for creating own surveys for information modelling and analyis. As a second contribution, we developed a methodology to harvest and analyze metadata from data repositories.	`https://doi.org/10.5281/zenodo.7385600`	[Löffler et al., 2021]
Chapter 6, H_3	We manually annotated a metadata corpus (QEMP) with some of the identified information categories (entity types) from H_2. As a second contribution, we developed a text mining pipeline to automatically extract these entity types. In addition, key terms are linked to entities of selected ontologies.	`https://doi.org/10.5281/zenodo.7385638`	[Löffler et al., 2020]

Chapter 7, H_4	For the evaluation of different entity based expansion strategies and entity based retrieval models, we created a metadata ground truth in cooperation with a biodiversity expert. We also implemented two entity-based retrieval models being available as GATE Mímir (`https://gate.ac.uk/mimir/`) plugins.	`https://doi.org/10.5281/zenodo.4638089`
		
		`https://doi.org/10.5281/zenodo.7396799`
			[Löffler et al., 2021]

Chapter 8, H_5	The developed user interface framework for a semantic dataset search provides two interfaces (a structured, category based input and a classical one input field). We also setup a jupyter notebook to analyze the user results, and we implemented scripts to compile several survey files into one large file. As a supplement outcome, we upgraded the Semantic Assistants framework [Witte and Gitzinger, 2008] to publish text mining pipelines for Named Entity Recognition tasks.	`https://github.com/fusion-jena/semantic-assistants-2.0`
		
		`https://doi.org/10.5281/zenodo.7394890`
		
		`https://doi.org/10.5281/zenodo.7391991`
			[Shafiei et al., 2021], [Löffler et al., 2023]

Table 1.1: Research contributions with source code and data availability and respective publications.

All GFBio services are now part of the *Nationale Forschungsdateninfrastruktur für die Biodiversität* (NFDI4Biodiversity)[29]. The former search UI was revised and is now based on the first version of the user interface framework being developed in the scope of this thesis (Chapter 8) [Shafiei et al., 2021].

We developed the functional and non-functional requirements on the proposed system after the domain specific requirements and conditions were identified (Chapter 4). The functional and non-functional requirements are introduced in Chapter 5.

[29]NFDI4Biodiversity, `https://www.nfdi4biodiversity.org/de/`

1.8 Research Contributions

Table 1.1 lists the research contributions of our work. We focused on reusable artifacts such as gold standards, test collections, frameworks and further reusable code. We provide a brief overview about each contribution with its availability and respective publications.

1.9 Structure of the thesis

At first, we present the research areas our work deals with in a background chapter (Chapter 2). The following chapters are mainly organized along the hypotheses. We conduct different evaluations on the retrieval process, search interests and data sources to determine requirements for an improved dataset search (addressing H_1 and H_2) in Chapter 3 and Chapter 4. Chapter 5 summarizes the findings from first two chapters, provides ideas for improvements and introduces the proposed solution. The focus of the last three chapters is on the improved dataset search. In separate chapters, we present each of the three components addressing H_3 - H_5. Chapter 6 introduces the pre-processing component with the novel text mining pipeline. In Chapter 7, we present the evaluation on different entity expansion strategies and entity-based retrieval models. Chapter 8 describes the presentation component and the user study on the usability of specific components in a dataset search. We summarizes all achievements in Chapter 9, and we give an outlook on future work in Chapter 10.

Chapter 2

Background

> *"If I had asked people what they wanted, they would have said faster horses."*
>
> - attributed to Henry Ford, *Founder of Ford Motor Company*

Searching for information on the web is an interactive and highly complex process. Users have different backgrounds, attitudes and expectations. And systems are based on different technologies, algorithms and data. Asking users what should be improved in a search application is therefore difficult. Hence, solid research from various research areas is required to explore the user experience and to design and develop improved retrieval systems.

This work deals with topics from three research areas: User experience and user-centered design, information retrieval and the Semantic Web. Basics about user-centered design and user evaluations are provided in Section 2.1. These foundations are necessary to design proper user studies, to collect user requirements and to develop systems in close contact with the target group. Information retrieval (IR) is the research field that has leveraged the development and success of search engines. Main algorithms and evaluation metrics in information retrieval are introduced in Section 2.2. As this works aims to explore the integration of semantic resources into the retrieval process, Section 2.3 gives an overview about basic Semantic Web formats. The application domain to which this work relates to is biodiversity research. Therefore, Section 2.4 defines what biodiversity research is, what it comprises and introduces the projects and data used in this work.

2.1 User-Centered Design

User-centered (or human-centered) design aims to develop systems that are usable, comprehensible and focused on user needs. This result is often also called *usability*. It describes to what extent users can complete their tasks, in what time and in which context [ISO 9241-11, 2018]. User-centered design is the process to reach usability.

Usability is characterized by five attributes [Nielsen, 1993]: *Lernability*, *Efficiency*, *Memorability*, *Errors* and *Satisfaction*. *Learnability* describes how easy it is to work with the system. The aim is to design systems that require no or less training to get started and to achieve the desired tasks. *Efficiency* comprises the amount of work and tasks users can achieve in a certain time once they are familiar with the system. *Memorability* characterizes how easy it is to remember the system with its navigation and functions after some period of time. Users should be able to continue their work without any additional training. *Errors* describe the amount of errors occuring while working with the system. Usually, the aim is to develop systems with low error rates. *Satisfaction* comprises the subjective estimation of users how satisified they are with the system and whether they like the system or not.

In order to achieve usability, multiple standards for user-centered design have been developed. A central entry point is the [ISO 9241-11, 2018] standard published by the International Organization for Standardization (ISO). Developed in the early nineties and revised over the years, ISO 9241 contains definitions, advices, guidelines and principles to develop and evaluate user-focused applications. One important sub-standard is [ISO 9241-210, 2019] (Human-Centred Design For Interactive Systems) comprising a definition and activities to ensure human-centered design (Definition 1).

Definition 1. *"Human-centered design is an approach to systems design and development that aims to make interactive systems more usable by focusing on the use of the system and applying human factors/ ergonomics and usability knowledge and techniques. This approach enhances effectiveness and efficiency, improves human well-being, user satisfaction, accessibility and sustainability; and counteracts possible adverse effects of use on human health, safety and performance."*

To ensure a user-centered design (Figure 2.1), four main steps should take place during the development process of an application: At first, the target group needs to be identified and studied, e.g., by observing daily tasks or visiting the working environments. It also needs to be examined for what purpose the target group will use the product (context of use). Afterwards, user requirements can be collected, sorted and prioritized. Based on these first insights, several prototypes are developed and tested. These steps are iterated over and over again until the final result is achieved. Several user experience (UX) methodologies have been developed for each of these phases. We introduce the UX methdologies used in this work in the following Subsection 2.1.1.

2.1.1 Usability Testing

Usability testing comprises various evaluation methodologies to explore whether a system is easy to use or not. Researchers observe or interview users while they are interacting

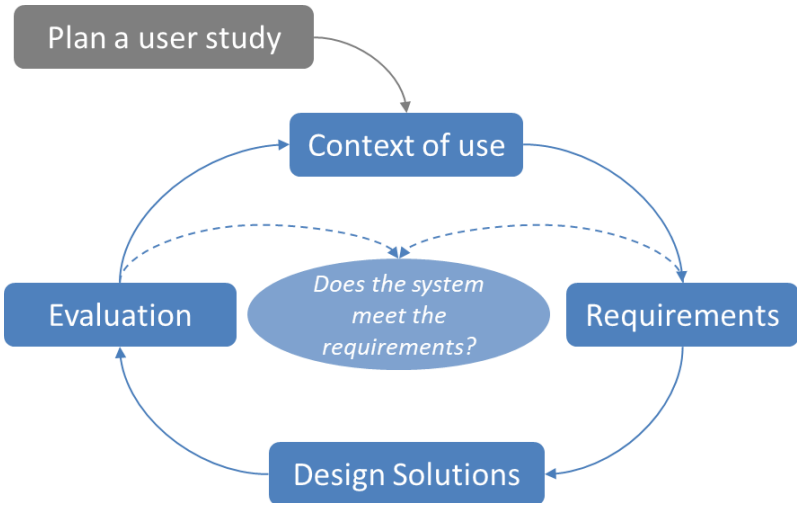


Figure 2.1: Four steps ensure a user-centered design process.

with the system. Usually, the main focus is on the user interface. The more users report on difficulties they have with a certain navigation, a function or the comprehensibility of a particular content, the more likely it is that this problem is a usability issue.

There are two main types of evaluations [Nielsen, 1993, Tullis and Albert, 2013a]: *formative* and *summative* evaluations. *Formative* evaluations are conducted to analyze a particular feature or design in an early stage of the development with the aim to improve the system before it is released. In contrast, *summative* evaluations are utilized to measure whether the system meets the defined requirements. For these final evaluations, which are usually not that frequently conducted than formative evaluations, often benchmarks are utilized or a larger number of users is recruited to examine the usability on a larger scale.

The data collected in usability studies distinguish between *quantitative* and *qualitative* data [Nielsen and Norman, 2022]. *Qualitative* data provide direct feedback on the usability of a system. When scholars observe users working or testing the system they can directly obtain information how easy the design or system is to use or what difficulties occur. In such live sessions, researchers can ask users why they encounter a problem or what they think about a particular design. It also give researchers the opportunity to change the methodology if needed (e.g., to ask more questions or discuss another solution) [Nielsen and Norman, 2022]. *Quantitative* data result in indirect findings on the usability of a design but deliver numbers on quantitiative measures such as error rate, time of task completion or success rate. However, scholars often need qualitiative data in addition to be able to interprete the numbers. For instance, if the task completion is only about 50%, it does not tell anything about the problems users had performing the tasks. Due to this disadvantage, usability studies that aim to gather quantitative data often also consider methodologies collecting qualitative data [Nielsen and Norman, 2022].

Table 2.1 provides an overview of the usability testing methods used in this work. For each methodology, we provide a short description and the information what data are

produced and for which evaluation type it is suited. Further usability testing methods are introduced in Nielsen's book on the foundations of usability testing [Nielsen, 1993].

Qualitative data can be collected in formative and summative evaluations and return a researcher's impressions and interpretations of studies with few participants. Quantitative data are mostly gathered in summative evaluations with a well-defined and controlled methodology and a larger number of participants. The result of quantitative data can be statistically analyzed. This encourages further research, which might aim to replicate or extend a study.

2.1.2 Usability Metrics

Usability metrics comprise all kinds of measurements to determine to what extent certain usability attributes have been addressed in a system. Based on a specific usability goal (e.g., *Efficiency of Use*), a model needs to be established to visualize which goal is addressed by which usability testing method and measured by which metric [Nielsen, 1993]. [Tullis and Albert, 2013a] distinguish between various types of usability studies (e.g., completing a transaction, comparing products or alternative designs) and provide suggestions for relevant usability testing methods and metrics. In the following, we introduce the usability metrics used in this work. Other metrics and more details for each metric are provided in [Tullis and Albert, 2013b, Tullis and Albert, 2013c, Tullis and Albert, 2013d]

Performance based metrics [Tullis and Albert, 2013b]: Performance based metrics describe all kinds of metrics observing or recording the user behaviour. They are the major metrics to measure the efficiency and effectiveness of a system. Important metrics being used in multiple usability studies are *task success*, *time-on-task*, *errors* and *efficiency*. All metrics can be either collected manually or automatically (e.g., by logging mechanisms).

- The *task success* metric marks the success or failure of pre-defined user tasks and can be either measured from the task or user perspective. Usually, success is measured on a scale with different levels. From the task perspective, success could be determined with three levels: complete success (all tasks achieved), partial success (only partial tasks achieved) and failure (no tasks achieved). In contrast, a scale based on the user perspective could distinguish between 'no problem', 'minor problems', 'major problems' and failure. Independent of which perspective is chosen, a proper definition and a quantitative measure of 'success' must be given, e.g., numbers of incomplete tasks or numbers of problems occurred.

- The *time-on-task* metric measures the time users need to reach a specific goal or task. The start and end time of an action is either manually or automatically recorded. Another opportunity besides start and end time are pre-defined thresholds. In

Method	Description	Data	Evaluation Type
Interview	Researcher conduct formal interviews with one or more users.	qualitative	formative
Field studies (question collection)	Researchers visit target group users at the environment that is relevant for the application to be developed.	qualitative	formative
Concept testing/ Competitive testing	An important function is introduced to the target group in an early stage of the developement. This can be combined with competitive testing in which several different concepts are discussed.	qualitative	formative
Focus group	Representative users of a target group are selected to be involved in the design and development process.	qualitative	formative
User stories	Daily tasks within the context of use are collected from several users of the target group.	qualitative	formative
Card sorting	Users classify information entities into categories.	qualitative, quantitative	formative
Clickable paper prototype	In an early stage of the developement online or offline paper prototypes support the identification of obvious usability issues, for instance in the navigation or comprehensibility of content.	qualitative	formative
Survey (questionnaire)	Surveys aim to measure users' satisfaction and subjective estimations on the use with the system.	quantitative	formative, summative
Moderated (remote) usability studies	A moderator guides users through different user tasks in a live (remote) session.	qualitative, quantitative	summative
Logging actual use	The system records the user behaviour automatically, for instance clicks on links and buttons, keywords entered in a search input field and occurring errors.	quantitative	summative

Table 2.1: Usability testing methods used in this work according to Nielsen [Nielsen, 1993, Nielsen and Norman, 2022]

this case, users need to fulfill a task within a certain time frame. In the later analysis, the threshold-based measurement allows a simple visualization, such as the amount of users being below the threshold for a specific task.

- *Errors* can be the result of usability problems and may lead to task failure. Errors can occur on a user's or on a system's side. Users may provide incorrect input data, select the wrong button or do the required actions in an incorrect sequence. System errors can occur if sub-systems are not available, the login is not successful or the maximum numbers of characters are reached in an input field. Again, similar to the definition of 'success', it is crucial to describe in advance what is an error and what is a required system or user behaviour.

- In supplement to *task success* and *time-on-task*, there are further metrics to measure *efficiency*. A common approach is to determine the actions that are needed to achieve a goal. Based on pre-defined start and end actions, the number of user actions (e.g., user clicks) are counted until the final state is reached. Other approaches combine task completion and time per task. The Common Format for Usability Test Reports [ISO, 2018] defines the "core measure of efficiency" as the "ratio of the task completion rate to the mean time per task" [Tullis and Albert, 2013b]. A variation of this ratio proposed by [Tullis and Albert, 2013b] divides the number of completed tasks by the "total time spent by the participant on all tasks (successful and unsuccessful)".

Issue-based metrics [Tullis and Albert, 2013c]: Issue-based metrics aim to collect information on usability issues. Usability issues occur while using a system and influence a user's behaviour. The range of issues can be very broad and may prevent users of completing a task or may lead to a wrong action while navigation through a page. A usability issue can also be the misinterpretation of a particular content or a user's unsatisfaction with a system. Issue-based metrics can be either collected in on-site studies with a thinking-aloud method 2.1.1 or can be determined with comment fields in automated studies (e.g., questionnaires). Not all usability issues have the same importance. Therefore, severity ratings facilitate to concentrate on the main issue only. In most cases, a three-level rating is sufficient. Finally, the most important usability issues can be either grouped in categories, or the issue classification can be based on the tasks or participants.

Self-reported metrics [Tullis and Albert, 2013d]: Self-reported metrics comprise all methodologies to collect direct feedback from users. The aim is to get information about users' perception about the system. Participants can be asked orally in on-site or remote sessions (interviews), but users can also fill a paper or an online questionnaire. The latter has the advantage that it is more likely that the participants give an unbiased feedback.

A study conducted by [Dillman et al., 2009] reveals that people tend to give more positive answers in interviews (in-person or on the phone). The reason for this is that people hestitate to give critical feedback to moderators or interviewers in direct interview situations. Hence, it is recommended to provide anonymous surveys, which the moderator will only see once the participant has left.

Likert scales are the most common rating scale used to collect self-reported metrics. Developed in 1932 by [Likert, 1932], this rating scale focuses on the level of agreement. Usually, a five- or seven-point Likert scale is utilized with agreement levels such as *stronly disagree*, *disagree*, *neither agree nor disagree*, *agree*, *strongly agree*. Another rating scheme are the semantic differential scales [Osgood et al., 1975] that provide oppositive adjectives such as weak/strong or beautiful/ugly. Similar to the Likert scale, the levels between these opposite terms are rated on a five- or seven-point scale.

Self-reported metrics are mostly collected after a task (post-task) and after a complete session (post-session). In particular for post-sessions, multiple questionnaires have been developed to gather anonymous user feedback such as the System Usability Scale (SUS) [Brooke, 1996], the Questionnaire for User Interface Satisfaction (QUIS) [Chin et al., 1988] or the Computer System Usability Questionnaire (CSUQ) [Lewis, 1995]. In 2004, [Tullis and Stetson, 2004] compared several post-session questionnaires and found out that the results from the SUS questionnaire, which is relatively short with 10 questions only, resulted in higher percentages of correct conclusions with respect to the percentages of correct conclusions achieved with the other questionnaires. They assume that the SUS questionnaire in particular is more consistant in its result as it contains five positive and five negative statements, which better keeps a user's attention. Another positive influence is the assessment of a system or page as a whole instead of splitting statements into smaller sections like in the other questionnaires.

2.2 Information Extraction and Information Retrieval

Information retrieval comprises research on how to find and obtain relevant information. Leveraged by the introduction of the world wide web in the nineties, one major outcome of this research area are search engines, which are utilized in numerous daily tasks for multiple purposes in various domains. In Subsection 2.2.2, we describe the main functionality of a retrieval process followed by the introduction of multiple evaluation metrics 2.2.3. Information retrieval is closely related to Natural Language Processing. Before indexing, documents need to pass a syntactic analysis and an optional semantic analysis, both being introduced in Subsection 2.2.1. The differences between classical document retrieval and dataset retrieval are presented in Subsection 2.2.4.

2.2.1 Natural Language Processing and Information Extraction

Natural Language Processing (NLP) is a research field that aims to analyze and to "understand" human written text with machines [Maynard et al., 2017]. One sub-field in this large research domain is *information extraction (IE)*. It comprises the processing of unstructured text to identify and extract structured information [Gaizauskas and Wilks, 1998]. For this purpose, three main steps have to be carried out subsequently [Maynard et al., 2017]: *Pre-processing* (to recognize basic elements in text), *named entity recognition (NER)* (to identify important entity types in text) and relation or event extraction (to determine relations or events).

Information extraction tasks: In a pre-processing step, the system splits the text into individual tokens such as words, punctuations or spaces. Afterwards, it identifies sentences, followed by part-of-speech tagging (POS), a syntactic analysis to classify tokens into categories, e.g., nouns, verbs and adjectives [Jurafsky and Martin, 2008]. Most text mining pipelines also provide a morphological analysis (an extended form of stemming or lemmatization) to determine the root form of terms. As a last step, parsing and chunking aim to identify further grammatical structures or phrases. Subsequent to this pre-processing step, named entity recognition can take place. Extracting important terms or phrases, such as names, organizations, geographic locations and classifying them into proper categories are the major tasks in this step [Maynard et al., 2017]. Once important words are identified, a further task can be added - entity linking. In this step, the extracted terms or phrases are linked to an entry in a knowledge base [Balog, 2018]. The distinction between named entities and simple entities is crucial, as named entities refer to concrete people (e.g., Jane Goddall, Barack Obama) or geographic locations (e.g., The Botanical Garden Jena, Mount Cook), whereas entities refer to classes or types, e.g., female scientist, president, garden, mountain [Maynard et al., 2017, Balog, 2018].

Two main approaches can be distinguished in linguistic processing, namely *knowledge based* approaches and approaches based on *machine learning* [Maynard et al., 2017]. Knowledge based approaches are the more conventional approaches based on rules written by NLP developers. Rules have the advantage that they are easy to understand [Maynard et al., 2017] and no large training data corpora are needed. However, changing or extending rules can get time-consuming, in particular for complex tasks. In contrast, machine learning approaches require only very little or no expertise in NLP. A variety of ready-made models exist for major entity extraction tasks. They are easy to setup in case sufficent training data exist [Maynard et al., 2017]. Supervised approaches based on labeled training corpora usually provide more precise results than unsupervised approaches. However, system or requirement changes can cause re-labeling, which can get time-consuming, too.

Evaluation: In order to evaluate information extraction pipelines, manual labelings are required. Usually, human judges manually annotate a text corpus based on given annotation guidelines. This manual assessment is called "gold standard" [Jurafsky and Martin, 2008]. In an evaluation, gold standards allow a comparision of automatically produced annotations and manually created labelings. Precision and recall metrics can be applied to compute the accuracy (Chapter 2.2.3).

2.2.2 The Retrieval Process

A retrieval system consists of a collection of documents (a *corpus*) and a user's information needs that are described with a few keywords (*query*). The retrieval process aims at returning a ranked list of documents that match a user's query. The architecture of a retrieval system is depicted in Figure 2.2: If the document corpus is not given, an optional *crawling process* has to be run beforehand to retrieve and collect documents [Baeza-Yates and Ribeiro-Neto, 2008]. The *indexing process* comprises pre-processing steps, such as stopword removal, stemming, and spell checks. They are important to clean documents from unnecessary information and to analyze only terms that represent the content of a document. Afterwards, the system counts word frequencies within a document and across all documents. The result is an inverted index. Similar to a book index, this is a list of terms with the number of occurrences of each term in each document and across all documents. These statistics, generated regularly in background processes, form the basis for a fast access to the documents at search time. The actual search takes place in the *retrieval and ranking process* whenever a user sends a query to the system and results in a ranked result set being returned to the user.

Based on the underlying *retrieval model*, different ranking functions have been developed to produce a score for the documents with respect to a query. Top-scored documents are returned first. In larger corpora, paging functions allow a subsequent retrieval of further documents. Classical retrieval models are for instance: the *Boolean Model* [Manning et al., 2008], where only documents are returned that exactly match a query. It is named *Boolean Model*, because a query may only contain operators from Boolean algebra such as AND, OR and NOT. The result is either TRUE or FALSE [Croft et al., 2009]. An example query looks like this:

butterfly **AND** grassland

The result set only contains documents were both terms are contained. Any further ranking does not take place, all documents in the retrieved set are equally relevant. Therefore, the *Boolean Model* is often used in search engines in combination with further retrieval models such as the *Vector Space Model* [Manning et al., 2008]. Here, documents and the query are represented by vectors that consist of term weights. These term

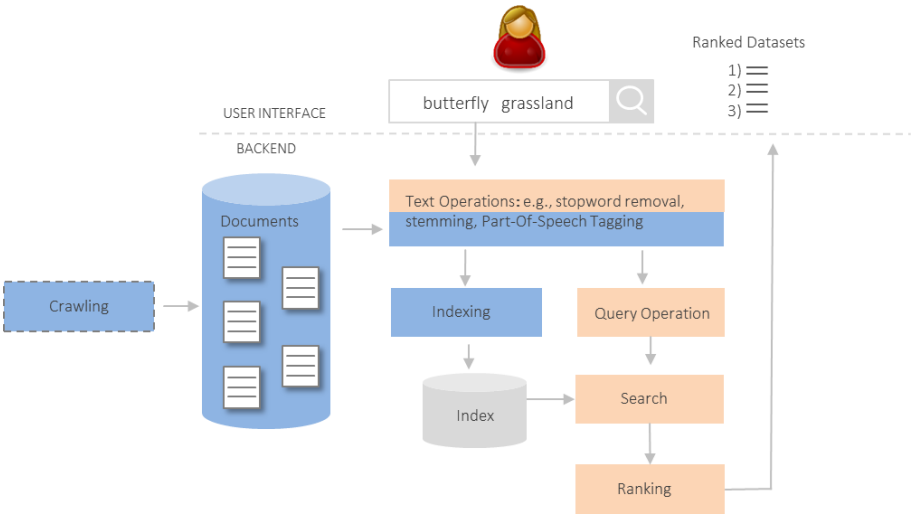


Figure 2.2: The architecture of an information retrieval system based on [Baeza-Yates and Ribeiro-Neto, 2008]: An optional *crawling process* (blue, dashed line) gathers documents. In the *indexing process* (blue) the documents are pre-processed before an index can be established. The *retrieval process* (orange) comprises the transformation of a user query into a format the search engine understands before the actual search and ranking takes place. Finally, users receive a ranked list of documents that match their query.

weights are computed based on statistical information from the documents. The term frequency (*TF*) counts how often a term appears in a document, whereas the document frequency (*DF*) denotes how often the term appears over all documents in the corpus. Both statistics are usually normalized in order to avoid giving preference to longer documents. A variety of normalizations have been explored in research [Manning et al., 2008, Croft et al., 2009] for *TF* values. A common modification is to use the logarithm for normalization as presented in Equation 2.1.

$$wf_t = \begin{cases} 1 + \log f_t & if \quad f_t > 0 \\ 0 & otherwise \end{cases} \tag{2.1}$$

For a term t, the weighted term frequency wf_t is computed by the frequency f_t, which denotes how often the term t occurs in the document. It is normalized with the logarithm. We add 1 because in case of f_t being 1, $\log 1$ would be 0. The second line of Equation 2.1 considers the case of f_t being 0, because $\log 0$ would be undefined.

$$idf_t = \log \frac{N}{df_t} \tag{2.2}$$

The document frequency is normalized as presented in Equation 2.2 where idf_t is the inverse document frequency of term t, N is the total number of documents in the corpus and df_t denotes the document frequency (the number of documents in which term t appears).

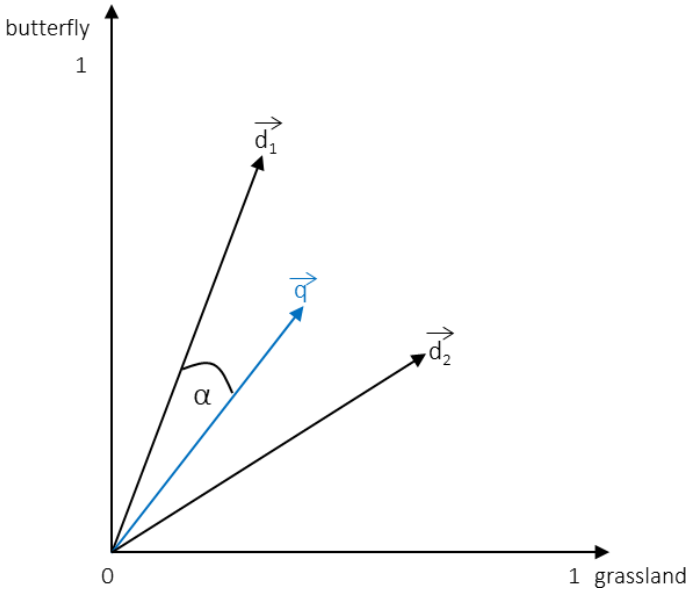


Figure 2.3: In the Vector Space Model, documents and queries are represented as vectors. The similarity between documents (and documents and the query) is computed: sim($\vec{d}$, $\vec{q}$) = cos α

$$w_t = TF - IDF = (1 + \log f_t) * \log \frac{N}{df_t} \tag{2.3}$$

Both together, term frequency and document frequency values form a weight for each term in a document. The final computation of term weights looks like as stated in Equation 2.3. In order to compute a final rank for each document, a subsequent scoring has to take place. A simple scoring measure is the *overlap score measure* [Manning et al., 2008] presented in 2.4.

$$score(q,d) = \sum_{t \in q} w_t \tag{2.4}$$

The score of a document d for a query q is the sum over all term weights, all $TF - IDF$ values of document d. Another scoring approach that considers all term weights in a document as vectors is the *Vector Space Model*. In Figure 2.3, each document d and the query q are represented as vectors in a vector space and one axis illustrates one term. Finally, the similarity of document and query vectors results from the distance between the document and query vectors (Figure 2.3). The cosine similarity (Equation 2.5) is the preferred similarity measure, as it considers normalization of the document lengths.

$$sim(d,q) = \cos \alpha = \frac{\vec{d} * \vec{q}}{\left|\vec{d}\right| * |\vec{q}|} \tag{2.5}$$

The numerator in Equation 2.5 is the dot product of the vectors $\vec{d}$ and $\vec{q}$, and the denominator represents the product of the Eucledian distance of both vectors. The results of the cosine similarity are values between -1 and 1. Negative values point to documents

containing contracting or opposing contents. Usually, values vary between 0 and 1, where values close to 1 denote documents with very similar content.

Apart from the *Vector Space Model*, which is the most common model, a variety of other retrieval models have been developed. *Probabilistic Models* [Manning et al., 2008] are based on computations of the probability of a document belonging to the relevant set. For languages where there are no word boundaries, e.g., in Eastern Asian Languages, *Language Models* [Jurafsky and Martin, 2008] are a solution to get a mathematical representation of the documents. The system analyzes the text documents by means of character-based sliding windows (*n-grams*) to determine word boundaries and to compute statistics.

All introduced retrieval models share one characteristic: They are keyword-based, and retrieval systems using these conventional models only return documents that exactly match a user's query terms.

2.2.3 Evaluation of Retrieval Systems

When setting up a retrieval system, various design decisions influencing different parts of the system have to be made. Examples of such decisions are whether to stem terms in the pre-processing phase or which terms to include in the stopword list. Numerous evaluation measures have been developed to determine the *effectiveness* of the systems, i.e., the accuracy of a result returned by a given retrieval algorithm. For this purpose, a test collection is required that consists of three things [Manning et al., 2008]: (1) a corpus of documents, (2) representative information needs expressed as queries, (3) a set of relevance judgments that are provided by human judges and that contain assessments of the relevance of a document for given queries. If judgments are available for the entire corpus, they serve as baseline ("gold standard") and can be used to determine how many relevant documents a search system finds for a specific query.

User queries should be representative for the target domain. Queries are either obtained from query logs of a similar application or domain users are asked to provide example queries [Croft et al., 2009]. The number of example queries influences the evaluation result. TREC (Text REtrieval Conference) is a long-running, very influential annual information retrieval competition that considers different retrieval issues in a number of *Tracks*, e.g., Genomics Track or Medical Track[1]. Various TREC experiments have shown that the number of queries used for the evaluation matters more than the number of documents judged per query [Croft et al., 2009]. Therefore, TREC experiments usually consist of around 150 queries (or so-called "topics") per track. Common evaluation metrics with respect to retrieval effectiveness are *Precision and Recall (PR)*. They are based on a confusion matrix presented in Table 2.2. The row with the POSITIVE entries denote the system's result set. The returned entries can be classified according to their relevance.

[1] TREC, `https://trec.nist.gov/`

POSITIVE	true positive (tp)	false positive (fp)
NEGATIVE	false negative (fn)	true negative (tn)

Table 2.2: Confusion matrix in a retrieval system. The POSITIVE row marks the result returned by the system, and the NEGATIVE row contains the sets that were not returned.

When the system is right, we say these entries are 'true positive'. When the system is wrong, we denote the entries as 'false positive'. The second row contains the missing entries in the result set ('false negative') and the entries that were correctly not returned ('true negative').

With these statistics, the performance metrics are computed as stated in Equation 2.8. The Precision denotes which fraction of the documents in the result set is relevant for a query, whereas the Recall describes which fraction of relevant documents was successfully retrieved. Both metrics are based on binary judgments, i.e., raters can only determine whether a document is relevant or non-relevant.

$$\begin{aligned} \text{Precision} &= \frac{tp}{tp+fp} \\ \text{Recall} &= \frac{tp}{tp+fn} \end{aligned} \tag{2.6}$$

The *F-Measure* (Equation 2.8) is the harmonic mean of Precision and Recall. If Precision and Recall values are very close, the F-measure is approximately the average of the two.

$$\text{F-Measure} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.7}$$

Precision and Recall can only be used, when a gold standard is provided containing the total number of documents in a corpus that are relevant for a query. However, in applied domains, where corpora are established specifically for a particular research field, gold standards are usually not available. Therefore, with the Recall being unknown, the *Mean Average Precision (MAP, Mean AP)* [Manning et al., 2008] is an alternative. The *MAP* only requires to get ratings for the TopN-ranked documents to compute an average precision. It is based on the assumption that users are only interested in the first entries of a search result and usually do not navigate to the last page. The top-ranked documents get higher scores than the lower ranked ones [Croft et al., 2009].

The MAP represents a measure for computing the quality of a system across several search queries. Hence, it requires to compute other intermediate metrics, first, such as the *Average Precision* per query and the *Precision@k*. The Precision@k is the Precision at a given cut-off rank, considering only the top-*k* documents returned by the system. In Equation 2.8,

$$\text{Precision@k} = \frac{1}{k} \cdot \sum_{d=1}^{k} rel(d), \tag{2.8}$$

k is the rank of the document that is considered, and $rel(d)$ denotes the binary rating for the iterating position d. For all relevant documents $d \in D$ per query q, the Average Precision (AP) of a query q is computed as follows 2.9:

$$\text{AP}(q) = \frac{1}{D_{rel_k}} \sum_{d=1}^{k} \text{Precision@k} \cdot rel(d), \tag{2.9}$$

where $rel(d)$ is 1 if the document d is relevant and 0 if it is not relevant. D_{rel_k} is the set of all relevant documents up to a certain cut-off rank k. Finally, the MAP is the sum of all Average Precision values for every query $q \in Q$, divided by the total number of queries as presented in Equation (2.10):

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \text{AP}(q). \tag{2.10}$$

Another metric proposed by Järvelin and Kekäläinen [Järvelin and Kekäläinen, 2002] is the *Discounted Cumulated Gain (DCG)*, a metric that uses a Likert-scale [Likert, 1932] as rating scheme and allows non-binary ratings. Given a list of documents, $rel(d)$ is the rating of each document d. When using Likert-scales, ratings can have arbitrary values that are equally distributed. For instance, a common scale contains five values such as 0 (not relevant), 1 (slightly relevant), 2 (less relevant), 3 (relevant) and 4 (highly relevant). In accordance to the Precision, the DCG supposes that higher ranked documents are more relevant than lower ranked. Therefore, the Logarithm reduces the influence of the lower ranked documents on the final DCG value. Per rank, so-called *cumulated gains* are computed as stated in Equation 2.11:

$$\text{DCG}_q = rel_1 + \sum_{i=2}^{|D|} \frac{rel(d)}{\log_2 i}. \tag{2.11}$$

Cumulated gains denote the sum of all relevance scores in a search result list. For a set of queries Q, let $rel(d)$ be the relevance score of document $d \in D$ for query $q \in Q$. Then, the DCG for every query as defined in [Croft et al., 2009] is the sum over all $|D|$ document gains. In order to consider different document length', the DCG values are normalized across all queries. Therefore, the documents are arranged in an ordered list called *Ideal DCG* (IDCG). This IDCG is utilized to determine the *normalized DCG* (nDCG) for a query q as follows (2.12):

$$\text{nDCG}_q = \frac{\text{DCG}_q}{\text{IDCG}_q}. \tag{2.12}$$

DCG does not penalize for wrong results, but only increases the scores of top-ranked documents.

All metrics introduced so far are based on the assumption that the provided relevance judgments are complete. For large test collections with millions of documents, manual inspection and judgment are not feasable. In the annual TREC competions, they utilize pooling strategies in combination with top k documents to overcome this obstacle. However, the above mentioned measures do not consider this incompletness appropriately [Buckley and Voorhees, 2004]. To address this problem, [Yilmaz et al., 2008] proposed an inferred measure *infAP* with randomly selected cut-off ranks. As also variance in the values of infAP can occur, they employed confidence intervals. Concerning an enhanced efficiency, the authors further improved the measure to also incorporate relevance judgments of non-random samples. Finally, [Yilmaz et al., 2008] showed that this stragegy can be applied to other metrics, such as nDCG, which they call *infNDCG*.

As discussed so far, numerous evaluation metrics focus on computing the relevance of documents. There are also other metrics concentrating on the *efficiency* (e.g., the time, memory and disk space required by the algorithm to produce the ranking [Croft et al., 2009]), *user satisfaction* on the provided result set and the presentation and *visualization* [Hearst, 2011] of search results.

Evaluation of User Interfaces in Retrieval Systems

A user-friendly and comprehensive user interface is as essential as a correct ranking. A common and classical search input provides one long input field with at least 27 characters covering the longest sequence of possible search terms [Nielsen, 1999]. Users enter keywords representing their information need. This so-called ad-hoc search returns a search engine result page (SERP) with a list of documents containing the entered keywords. SERP entries typically provide a snippet of the full content per document, the title and a link to the document source. Since a user's information needs are often imprecise, as only a few keywords can be entered into the input field, the interface has to be designed in a way that motivates and supports users in formulating their queries. It also has to present the results clearly and comprehensively and should provide the opportunity to select among the available resources. Common techniques in supporting users in search interfaces are auto-complete functions, spell correction, faceted filtering to enhance navigation and highlighting of search terms [Manning et al., 2008]. We introduce search interfaces of data portals in the Life Sciences and biodiversity research in the next Section 2.2.4.

All evaluation approaches of search applications with user involvement are comprised as *Interactive Information Retrieval (IIR)*. [Kelly, 2009] distinguishes between different types of interactive retrieval systems. They range from system-focused approaches in

which users assess the relevance of the results, studies where system or interface features are evaluated to evaluations that are solely focused on the search behaviour of users. Main contributions towards standard techniques for IIR were driven by the TREC Interactive Search [Dumais, 2005] tracks, a series of evaluations concentrating on both, system and interface features. Starting from TREC-7, users have to perform the search tasks on two systems (A/B testing), a system proposed by the TREC participants and a self-chosen information retrieval control system, which also meets the evaluation goals. Users need to save the documents that contain the relevant answers on the search task, and they have to fill out several questionnaires before (pre-)and after (post-search) the search and after each system (post-session). With TREC-9[2] the search time was limited to five minutes to minimize the cognitive effort for the users.

Even if a standard evaluation framework could not be established [Kelly, 2009], the TREC Interactive Search tracks supported the introduction of some standard techniques in IIR such as the logging of user data or the reduction of time per search task.

2.2.4 Dataset Retrieval

Scientific data in the Life Sciences comprise all components and information collected and generated during the research process. They can differ in format and size; examples are for instance: spread sheets, audio or video files, field books, specimens, questionnaires, models or software code. As it is difficult to search for these heterogenous data formats, scientific *primary data* are described by *metadata*.

Metadata are usually provided in semi-structured formats such as XML[3] or JSON[4] and contain descriptive information along the W-questions (What, When, Where and Why). Important metadata fields comprise information on the author and/or collector of the dataset, a title, an abstract, parameters that have been measured, information about the geographic location (including spatial information), collection time and data format. Data reuse and data citation also play an important role. Metadata provide several elements for digital identifiers, license information and citation, as it is demanded by the Joint Declaration of Data Citation Principles [Data Citation Synthesis Group, 2014] and practical guidelines for data repositories [Fenner et al., 2019]. Multiple metadata standards have been developed in different domains and for different purposes. We introduce different metadata standards being relevant for the Life Sciences in Chapter 4.3. An example metadata file is presented in the next Section 3.2.

Search engines need homogenous, text-based data for indexing. Therefore, data repositories mainly use metadata in their search indexes [Khalsa et al., 2018]. Metadata can be generated in different ways. In most cases, it is a manual or half-automated process.

[2]TREC-9, `https://trec.nist.gov/data/t9i/t9i.html`

[3]XML, `https://www.w3.org/standards/xml/`

[4]JSON, `https://www.ecma-international.org/publications-and-standards/standards/ecma-404/`

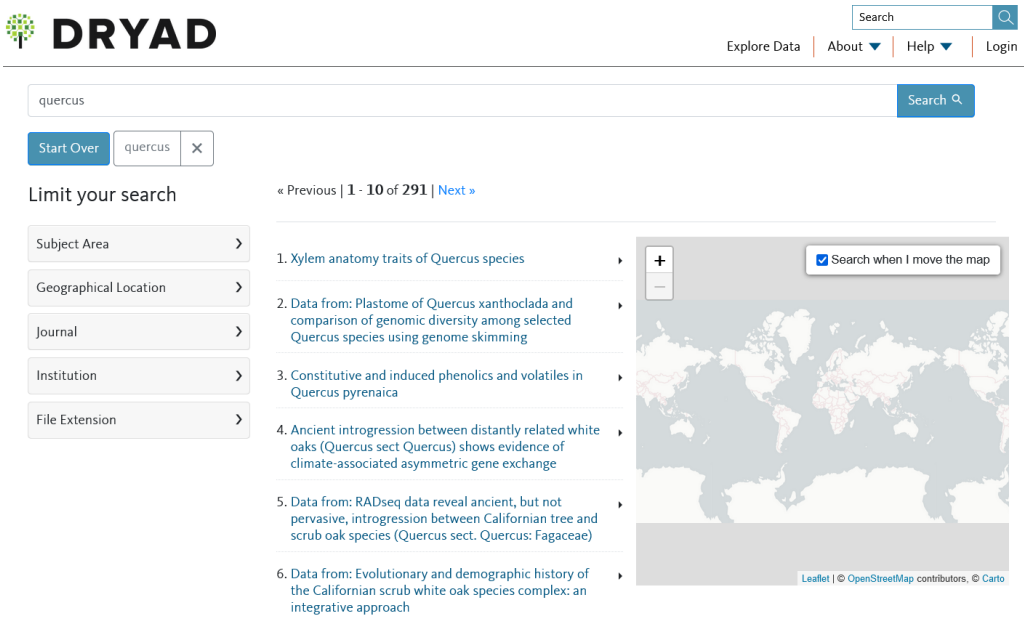


Figure 2.4: Screenshot of Dryad's search interface in March 2022.

When scholars submit research data, data centers either offer forms to ask for basic information about the dataset, or they guide scholars through a longer curation process and interact with the data submitter. Large publishers, such as Nature[5] recommend to submit research data to domain specific data repositories [Nature, 2018], e.g., PANGAEA[6] (environmental data) or EBI/ENA[7] (nucleic acid sequence data), as they are specialized on a particular domain and formats. In contrast, generalist repositories are able to archive different types of data, e.g., Dryad[8], Zenodo[9] or Figshare[10]. Federated approaches harvest data from different data sources and provide one entry point for data access. Examples are for instance DataONE[11], GFBio or the Google Dataset Search[12].

A typical *user interface* for dataset search consists of one input field, filters for search adjustment and a list of datasets utilizing different information from metadata fields. Figure 2.4 presents a screenshot from Dryad. In this example each entry in the result list only provides a title per dataset. Other information from metadata are omitted. This makes it difficult for users to decide at one glance whether a dataset is relevant to a query or not. In contrast, PANGAEA (Figure 2.5) presents a snippet per dataset with information on author, title, publication year and identifier per dataset. The RDA Data Discovery Interest group goes even further and recommends to display information on data access, data li-

[5]Nature, `https://www.nature.com/`
[6]PANGAEA, `https://pangaea.de/`
[7]EBI/ENA, `https://www.ebi.ac.uk/`
[8]Dryad, `https://datadryad.org`
[9]Zenodo, `https://zenodo.org/`
[10]Figshare, `https://figshare.com/`
[11]DataONE, `https://www.dataone.org/`
[12]Google Dataset Search, `https://datasetsearch.research.google.com`

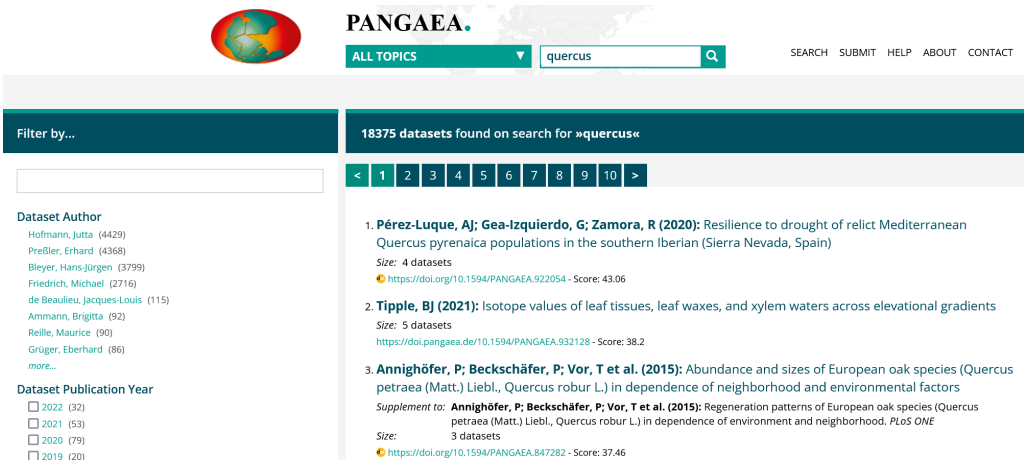


Figure 2.5: Screenshot of PANGAEA's search interface in March 2022.

cense as well as a preview or statistics of the primary data [Wu et al., 2019]. In addition, provenance information should be given, such as who collected the data, who is the data owner and which methods or devices were used to generate the data. Only these full metadata allow users to easily decide whether a dataset is worth to click and further to inspect or not. When selecting a dataset, users are usually directed to a separate page, a landing page, which displays more, and sometimes complete, metadata information. Landing pages usually use persistent identifiers in their address (Uniform Resource Locator [Berners-Lee et al., 2005]), e.g., `https://doi.pangaea.de/10.1594/PANGAEA.922054` [Pérez-Luque et al., 2020]. These unique identifiers foster the availability and accessibility of datasets in the web and are a prerequisite for Linked Data and semantic interoperability (see also Section 2.3).

Enhancements in dataset search such as auto-completion in the search field or faceted search facilities are nowadays state-of-the-art. Facets represent relevant categories of a domain [Hearst, 2006] and allow users to influence the result set, e.g., to narrow its scope and to search for datasets in a specific spatial coverage. Facets are based on metadata fields, grouped keywords or they can be created manuelly. Various approaches in computer science explore automatic facet creation [Hildebrand et al., 2006, Dakka and Ipeirotis, 2008, Xu and Zhuge, 2014]. There are also very first approaches built on semantic technologies and knowledge graphs [Moreno-Vega and Hogan, 2018, Feddoul et al., 2019].

Most data providers use one of the widely spread search engines such as *Apache Solr* [13], *Apache Lucene*[14] or *elasticsearch*[15] [Khalsa et al., 2018]. These search engines are keyword based and provide classical retrieval models, e.g., *TF-IDF* or *BM25* [Manning et al., 2008, Croft et al., 2009].

[13] Apache Solr, `http://lucene.apache.org/solr/`
[14] Apache Lucene, `http://lucene.apache.org/`
[15] elasticsearch, `https://www.elastic.co/`

2.3 Semantic Web

The Semantic Web is closely connected to the development of the World Wide Web (WWW)[16]. The WWW was originally designed to provide a machine-supported way of communication between users [Berners-Lee and Hendler, 2001] based on a client-server architecture. Web pages are placed on a server in (X)HTML[17] format. A client consumes the page with a browser that interprets the (X)HTML code and displays it in a user-friendly and readable form. The idea of the Semantic Web goes beyond this simple information consumption of humans. Instead of displaying information, the Semantic Web aims to enable machines to 'understand' web content [Berners-Lee and Hendler, 2001].

For this understanding, machines need logic structures. The development of the Semantic Web has the goal to provide these new formal languages to describe the real world in a formal and mathematical representation. It belongs to the research field of symbolic artifical intelligence (AI) and can be considered as an implementation of symbolic AI approaches. Symbolic AI has its origin in formalism and logic. It was the dominant research area in artifical intelligence until the mid-1990s [Russel and Norvig, 2021].

We briefly introduce the formal languages RDF (Subsection 2.3.1) and OWL (Subsection 2.3.2), and we describe how to query semantic data with SPARQL (Subsection 2.3.3). We also present the principles of Linked Open Data (LOD) and vocabularies being relevant for the Life Sciences (Subsection 2.3.4), followed by some examples of semantic metadata standards for datasets (Subsection 2.3.5).

2.3.1 RDF/RDFS

The Resource Description Framework (RDF)[18] is considered as "basic representation format for developing the Semantic Web" [Hitzler et al., 2009]. Instead of displaying information, it aims at processing information to provide a structure that is human and machine readable. In an RDF document, information is described as a directed graph that represents subject - predicate - object relations. Figure 2.6 presents an example. The book "The Lord of the Rings" was written by J.R.R. Tolkien. The book (subject) and the author (object) nodes are resources with unique identifiers. Theses identifiers evolved over time from Uniform Resource Identifiers (URIs) [Berners-Lee et al., 2005] based on ASCII characters to IRIs (Internationalized Resource Identifiers) [Duerst and Suignard, 2005], an extended version permitting further characters based on the Universal Character Set (Unicode/ISO 10646). They describe unique resources in the web, e.g., a web page, a mail address or a file. In our work, we only focus on URIs, as not all utilized terminologies and systems supported the extended character set. The arrow (predicate or property)

[16] WWW, `https://www.w3.org/`
[17] (X)HTML, `https://www.w3.org/standards/webdesign/htmlcss`
[18] RDF Primer, `https://www.w3.org/TR/rdf11-primer/`

Figure 2.6: An RDF graph visualizing the relationship between a book and an author.

"author" denotes the relation between book and author. It is also represented with a URI.

Various notation formats have emerged such as RDF/XML, notation 3/n3, Turtle, N-Triples, N-Quads and JSON-LD. The Listing below contains the graph of Figure 2.6 in Turtle notation.

Listing 2.1: RDF graph in Turtle notation.

```
@base <http://example.org/>
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>

<The_Lord_of_the_Rings>
        <hasAuthor> <JRR_Tolkien> ;
        <publicationDateStart "1954-07-29"^^xsd:date> ;
        <publicationDateEnd "1955-10-20"^^xsd:date> ;
```

Nodes can be either described as resources or literals. The hierarchy is represented by the nesting of the elements, e.g., `<The_Lord_of_the_Rings> <hasAuthor> <JRR_Tolkien>`). Literals are used when values are strings, numbers or dates. For instance, in Listing 2.1, publication date is described by the properties `<publicationDateStart>` and `<publicationDateEnd>`, as the book was published in three volumes over one year. The attribute `xsd:date` denotes that the following value is a date. In a similar way, numbers and strings can be expressed. The namespace information at the beginning of the document (`xsd`) aims at grouping identifiers and ensures that each occurring element in the group is unique. RDF graphs are stored in non-relational database structures, in so-called *triple stores* (e.g., Virtuoso[19], Apache Jena[20] or GraphDB[21]), which return triple statements as subjects, predicates and objects.

RDF data mainly describe instances, concrete subjects or objects (also called *A-Box*). In contrast, RDF Schema (RDF/S)[22] aims at modeling classes and their relationships (also called *T-Box*). The book instance "The Lord of the Rings" is a book (or belongs to the class "book") (Listing 2.2). Hierarchical IS-A relations can be expressed with the element `rdfs:subClassOf`. In the following Listing, the class `Book` is a subclass of class `Publication`.

Listing 2.2: RDF graph in Turtle notation.

```
@base <http://example.org/>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

[19] Virtuoso, `https://virtuoso.openlinksw.com/`
[20] Apache Jena, `http://jena.apache.org/`
[21] GraphDB, `http://graphdb.ontotext.com/`
[22] RDFS, `https://www.w3.org/TR/rdf-schema/`

```
<Lord_of_the_Rings> a <Book> .
<Book>
  a rdfs:Class   ;
  rdfs:label   "book"@en ;
  rdfs:subClassOf  [  rdfs:resource   "Publication"  ]  .
```

2.3.2 OWL

RDF/S allows the modeling of simple domain knowledge. However, modeling the real world is a difficult task as real life does not only consist of hierarchical structures.

In order to overcome these limitations and to provide more powerful elements, the Web Ontology Language (OWL)[23] has been developed. OWL reuses elements from RDF/S such as classes and properties. Properties are also called *roles*, and instances are named *individuals*. Class extensions that are not available in RDF/S are for instance `disjointClasses` to express that two classes must not have any overlap, individuals can only belong to one of these classes. OWL also provides elements to describe closed classes. In order to express that only specific individuals belong to a class, the class restriction `oneOf` can be used. Equivalent to RDF/S, OWL offers two role constructors: abstract roles (`owl:ObjectProperty`) and concrete roles (`owl:DatatypeProperty`). Abstract roles define relationships between classes, whereas concrete roles describe relationships between classes and data values, e.g., literals.

OWL is a W3C[24] Recommendation since 2004. Since 2009, OWL2 (second edition since 2012) is the officially used version for the modeling of ontologies. There are three sub-languages, namely OWL Full, OWL Lite and OWL DL. Depending on the purpose of the ontology to be developed and the degree of expressivity, one of the three sub-languages is used [Hitzler et al., 2009]. In order to model taxonomies, a less complex language as RDF/S has been evolved. The Simple Knowledge Organization System (SKOS)[25] allows the description of simple hierarchical structures with the elements`skos:broader` and `skos:narrower`. It also contain RDF/S elements to describe class/type associations. More descriptive language constructs are not provided.

Further detailed information on modeling languages and semantic web technologies can be found in [Hitzler et al., 2009, Antoniou et al., 2012]. Nowadays, a variety of vocabularies are available for multiple domains. We introduce main vocabularies for the Life Sciences in Subsection 2.3.4.

[23] OWL, `https://www.w3.org/TR/owl2-overview/`
[24] w3c, `https://w3c.org`
[25] SKOS, `https://www.w3.org/TR/skos-primer/`

2.3.3 SPARQL

A thorough selection of a suitable format and careful decisions about the used language constructs facilitates the query of a semantic web graph. The SPARQL Protocol And RDF Query Language[26] is a W3C Recommendation since March 2008, the latest version (SPARQL 1.1) was published in March 2013. SPARQL queries return all sub-graphs that match the graph described by the triples in the query. Queries are sent to SPARQL endpoints, and results are returned via HTTP(S) in various formats such XML, JSON, RDF and HTML. Listing 2.3 presents an example query to retrieve all book instances in a graph.

Listing 2.3: SPARQL example query: return all book instances that belong to class 'book'

```
# prefix declarations
PREFIX foo: <http://example.com/resources/>...
# dataset definition
FROM ...
# result clause
SELECT ?book
# query pattern
WHERE {
?book rdf:type ex:Book .
}
# query modifiers
ORDER BY ...
```

The PREFIX declarations provide the abbreviations of the URIs used. The following dataset definitions indicate which graphs are being required in the query, and the SELECT clause identifies the variables to appear in the query results. The WHERE clause provides the basic graph pattern to match against the data graph, and the ORDER BY information allows query modification, e.g., ordering of query results. The WHERE clause contains a graph patterns that needs to be successfully matched to be further processed. Patterns consists of triple patterns (subject, predicate, object). One or more parts of the pattern can be replaced with variables that are marked with a question mark, e.g., `?book`. Various language constructs are available to add constraints (FILTER), matching alternatives (UNION), exclude results (MINUS) or grouping (GROUP BY). Further information can be obtained from the SPARQL W3C website.

2.3.4 Linked Open Data and Vocabularies in the Life Sciences

All introduced semantic technologies aim at presenting or querying information in the web. However, the Semantic Web is not just about providing more meaningful informa-

[26] SPARQL, `https://www.w3.org/TR/sparql11-overview/`

tion, but has a strong focus on linking data, too. While the WWW is a network of web pages, the Web of Linked Open Data aims at linking any kind of resources that have unique identifiers (data). If every resource has its own URI, e.g., people, publications or products, we are able to create a network of data.

In 2006, Tim Berners-Lee published four rules that need to be fulfilled in order to call a web resource "fully semantic". Nowadays, his revised version from 2010, is known worldwide and comprises five principles[27].

- *one star*: A web resource gets one star once it is available in the web (it has a URI).
- *two stars*: If a web resource is available in a machine-readable and structured format (e.g., xlsx), it owns two stars.
- *three stars*: Three stars are assigned when the resource fulfills the two stars but in addition uses a non-proprietary format, e.g., csv.
- *four stars*: A web resource gets four stars when open W3C standards are used such as RDF and SPARQL to present and query the information.
- *five stars*: Five stars are granted if a web resource in addition to four provide links to other web sources.

These principles served as basis for further discussions on improving linked datasets, such as adding provenance information [Lebo et al., 2013] and reusing vocabularies [Janowicz et al., 2014]. In particular the latter is important for semantic interoperability. Not only the linking of web resources is important, but resources in datasets should refer to classes and properties in vocabularies to be interpretable and reusable. Over the past decade, numerous vocabularies for a variety of applications and domains have been developed and linked among each other. The current (but not complete) state of linked datasets and vocabularies are presented in the Linked Open Data cloud [McCrae et al., 2020], a visualization initiated by [Semantic Web Education and Outreach Interest Group, 2009]. In May 2020, the Linked Open Data cloud contains 1,301 datasets with 16,283 links. Figure 2.7 presents the current state. What stands out is that one third of the displayed vocabularies comes from the Life Sciences. Most of them have been developed in the bio-medical context such as the Gene Ontology [GO, 2021] or the Chemical Entities of Biological Interests (ChEBI) developed for describing molecular entities with chemical compounds [Hastings et al., 2016]. However, numerous other datasets describing various biological fields have emerged, e.g., the NCBI Taxonomy Database (NCBITAXON) [NCBITaxon, 2022], a curated nomenclature for all organisms, The Phenotype And Trait Ontology (PATO) [Gkoutos, G. et al., 2022] to describe phenotypic properties or the Environment Ontology (ENVO)[Buttigieg et al., 2016] for the description of environments.

[27] Linked Data, `https://www.w3.org/DesignIssues/LinkedData.html`

In the Life Sciences, most vocabularies are class-based and mainly represent hierarchies. While some of them are very large in its size, for instance, NCBITAXON contains around 1,8 Mio classes with a maximum number of sub-classes of 42,960, there are also small ontologies such as the Plant Phenology Ontology (PPO) [Walls, R. et al., 2022] with only 443 classes. More sophisticated semantic rules (such as reasoning) are only supported by a few ontologies such as PATO and ENVO, most biological vocabularies focus on describing IS-A relationships. Initiatives such as the Open Biological and Biomedical Ontology (OBO) foundry[28] aims at connecting, organizing and structuring Life Sciences related entities (mainly class-based). In order to foster interoperability, all developers are committed to shared principles such as non-overlapping content, open use and collaborative environments. In practice, entities get one unique URI per context, e.g., water in chemical context (`http://purl.obolibrary.org/obo/CHEBI_15377`), but can be used in other domains. For instance, ENVO uses the same URI for water and groups it under "chemical entity". If the entity occurs in a new context, e.g. water as food, it gets a new URI (`http://purl.obolibrary.org/obo/FOODON_00002340`).

Instances in the Life Sciences represent scientific primary data such as concrete species, genes or chemical compounds. Driven by funding agencies and academic publishers, scholars are more and more urged to publish their primary data. Hence, increasingly, scientific primary data get unique identifiers such as DOIs (Digital Object Identifier)[29] that are open, persistent and machine readable. A few initiatives have already started to finally connect biological instances and classes with other web resources such as people and publications. For instance, the work by [Page, 2019] is one of the first approaches to finally build a biodiversity knowledge graph. According to the author, the biggest challenge for the development of such a knowledge graph is the alignment of the different identifiers. Numerous biological raw data are maintained in a variety of local databases that use internal identifier schemes. That hampers reuse, interoperability and citation. Hence, local identifiers need to be mapped to external identifiers (such as DOIs) that are shared by other databases.

2.3.5 Semantic Formats for Datasets

Driven by the FAIR and Linked Data principles, metadata standards are increasingly evolving towards semantic formats. In addition to Chapter 4.3, presenting metadata standards in the Life Sciences that are utilized by data repositories, this subsection briefly introduces (general) semantic metadata standards.

One of the earliest semantic metadata formats is Dublin Core in RDF/XML[30], a gen-

[28] OBO Foundry, `http://www.obofoundry.org/`

[29] DOI, `https://www.doi.org/`

[30] Dublin Core in RDF/XML, `https://www.dublincore.org/specifications/dublin-core/dcmes-xml/`

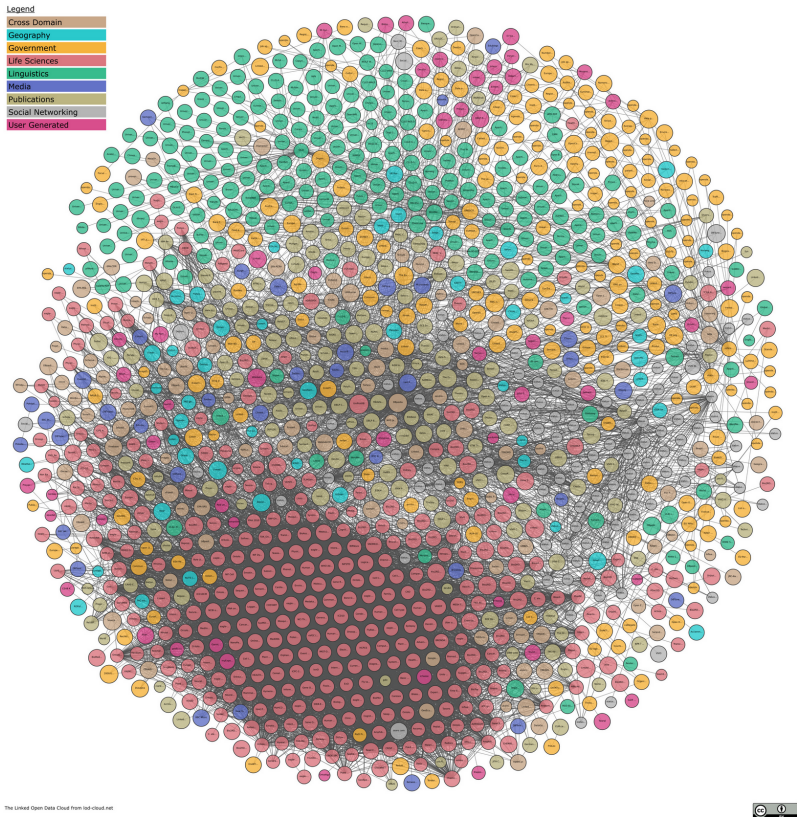


Figure 2.7: The Linked Open Data cloud in May 2020. One-third of the depicted vocabularies are from the Life Sciences. [McCrae et al., 2020]

eral metadata standard utilized in various research fields. It comprises 15 metadata fields with basic information, such as `dc:title`, `dc:description` or `dc:creator`. Further information are provided in Chapter 4.3.

Dublin Core served as the basis for further developments, such as the Data Catalog Vocabulary (DCAT)[31]. DCAT provides a data model with six main classes to describe data catalogs in the Web. A *data catalog* is a collection of *resources*, e.g., *datasets* or *data services*. DCAT was initiated by DERI and further developed by the W3C eGov Interest Group for governmental data (the group is not active anymore). Thus, DCAT is mainly utilized in governmental institutions. For instance, the European derivate DCAT-AP[32] is the leading metadata standard for datasets of the European Union, its member states and further European countries. In 2023, the European data portal[33] contains 1.5 Mio datasets from 179 catalogs and 36 countries. A quality assessment tool is also provided to check whether a dataset complies with the FAIR principles [Wentzel et al., 2023].

In contrast to scientific and governmental data repositories using the original metadata in data repositores and search applications, Google follows a different approach. It harvests the (X)HTML-based landing pages of datasets in data repositories, but only if

[31] DCAT, `https://www.w3.org/TR/vocab-dcat-2/`
[32] DCAT-AP, `https://semiceu.github.io/DCAT-AP/releases/3.0.0/`
[33] European data portal, `https://data.europa.eu`

schema.org[34] entities are contained. This semantic enrichment of (X)HTML pages allows additional filtering or supports the presentation of search results, such as information on author, location, publication year or dataset type. For applied domains, extensions such as *bioschemas.org*[35] have been developed to describe domain specific entity types, e.g., taxon or gene.

At the time of our research, in the Life Sciences, datasets in semantic formats were not used on a large scale, and vocabularies were rarely reused in metadata. However, over the last years, we observed an increasing usage of *schema.org* entities being utilized on data repositories' landing pages (see also Chapter 10).

2.4 Biodiversity Research

The term *biodiversity* is a short form of *biological diversity* and was first used at the US National Forum on BioDiversity in 1986 [Wilson, 1988]. A few years later, at the United Nations Conference on Environment and Development in Rio in 1992, the Convention on Biological Diversity (CBD) was created [Loreau, 2010] and signed by 150 countries [Convention on Biological Diversity, 1992]. The definition of the term biodiversity (Definition 2) that was formulated at this event is still valid today.

Definition 2. *"'Biological diversity' means the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems." [Convention on Biological Diversity, 1992]*

According to the CBD definition, biodiversity research can be characterized on three levels [Loreau, 2010]:

- (1) the variety of species (*taxonomic diversity*),
- (2) the *genetic diversity*
- (3) and the *ecological diversity* comprising the different and manifold ecosystems including the interactions between species and their habitats.

Nowadays, it is undisputed that human interventions cause a dramatic decline of the biological diversity. The rising population, land use or environmental pollution are only a few human actions that change the biological diversity [NKGCF, 2008] on a large scale. Species extinction, the loss of genetic information and entire ecosystems call for more research efforts to provide answers on what causes and drives this development and how to counteract [Loreau, 2010].

[34] schema.org, `https://schema.org/`
[35] bioschemas.org, `https://bioschemas.org/`

Therefore, several research projects in Germany have been established over the past decade to collect and provide data supporting decision makers in politics and society. In this work, we use data from several biodiversity research related projects that in the following are introduced in detail:

- **AquaDiva** The *CRC AquaDiva*[36] aims for a better understanding of the Earth's Critical Zone (CZ), a thin and permeable layer between the surface and the subsurface. The Critical Zone is the living environment of a variety of plants and animals including humans. Due to pollution, land use and climate change, the CZ is changing, and it is not fully clear what the consequences are of these human interventions. In particular, the subsurface, the region "below the highest density of plant roots and extending into the aquifers" [AquaDiva, 2022] has been less studied. Hence, the research question that more than 40 people in 17 research groups want to answer focuses on an improved understanding of the connections between the surface and subsurface. Special attention is given to organisms and processes influencing this environment. The *CRC AquaDiva* was established in 2014 and is primarily located in Jena, Germany.

  We collected around 90 search questions among 49 researchers on an AquaDiva workshop in 2015. We asked for full natural language questions and keywords associated with the question representing scholars' current information needs with respect to their research. No personal information was gathered. The questions were analyzed and utilized to develop an information architecture for dataset search in biodiversity research (Chapter 4).

- **idiv** The *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig* [idiv, 2019] is a large research center with about 400 employees from several universities and research institutes. It was established in 2013 and is primarily located in Halle, Jena and Leipzig. Funded by the German Research Foundation (DFG), scientists from 30 countries work on four core research questions: (1) "biodiversity patterns: What is the current state of biodiversity, and how are natural and human influences changing biodiversity in space and time?" [idiv, 2019], (2) "biodiversity processes: Which evolutionary and ecological processes create and maintain biodiversity?" [idiv, 2019], (3) "biodiversity functions: What role does biodiversity play in the functioning of ecosystems and the services that ecosystems provide for us?" [idiv, 2019] and (4) "biodiversity society: How can we protect and conserve biodiversity in managing the resources of our planet?" [idiv, 2019]. In the few years of its existence, the iDiv research center received world-wide attention with more than 2500 scientific publications in 2021. A key role plays the synthesis center sDiv. At sDiv workshops, scientists with various backgrounds come

[36] CRC AquaDiva, `http://www.aquadiva.uni-jena.de`

together and attempt to re-evaluate scattered data based on new research questions. This approach is unique in the world and attracts many international researchers to visit Leipzig. Numerous *idiv* researchers are also members of the World Biodiversity Council *IPBES (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services)*, which was founded in 2012.

We collected search questions and associated keywords at an iDiv conference in 2016. In total, four researchers provided 12 questions and keywords. The questions are introduced and analyzed in Chapter 4. In addition, we utilized ten public metadata files in a small Named Entity Recognition corpus established to evaluate the proposed text mining pipeline (Chapter 6).

- **BEF-China** The main research aim in the BEF-China[37] project was to explore Biodiversity-Ecosystem Functions (BEF) in a large and highly species-rich forest in the subtropics. In order to measure ecosystem functions such as carbon and nitrogen storage, nutrient cycling and the prevention of soil erosion, large experiments were conducted in a subtropical forest site in Jiangxi Province, China. The project was divided into 12 sub-projects which examined different aspects of ecosystem functions, e.g., primary production, plant growth and demography, woody decomposition and microbial biomass and activity. The project run from 2008 to 2016 and integrated 15 universities and research institutes from China, Germany and Switzerland.

  We utilized all 372 public metadata files to set up a test collection for dataset search (Chapter 7) with 14 questions. Ten out of 372 metadata files were also used in the QEMP corpus, a Named Entity Recognition corpus introduced in Chapter 6.

- **Biodiversity Exploratories** The Biodiversity Exploratories[38] are a large biodiversity project that have been funded by the German Research Foundation (DFG)[39] since 2006. The three sites, the *Biosphere Reserve Schorfheide-Chorin*, the *National Park Hainich* and the *Biosphere Reserve Schwäbische Alb* form the backbone to establish a framework addressing the research aims, namely, "to understand the relationship between biodiversity of different taxa and levels", "to understand the role of land use and management for biodiversity" and "to understand the role of biodiversity for ecosystem processes" [Biodiversity Exploratories, 2022]. In its sixth phase (in 2022), more than 250 members from 47 institutes in Germany and other European countries are involved in 40 projects.

  We utilized ten publicly available metadata files for the development of the QEMP corpus, a Named Entity Recognition corpus introduced in Chapter 6.

[37] BEF-China, `http://www.bef-china.de`
[38] Biodiversity Exploratories, `https://www.biodiversity-exploratories.de`
[39] DFG, `https://www.dfg.de`

- **GFBio and NFDI4Biodiversity** *The German Federation for Biological Data (GFBio)*[40] was a DFG-funded research project with the aim to establish a national infrastructure and to become a contact point for all topics around research data management of biological and environmental data [Diepenbroek et al., 2014]. Running from 2013 to 2021, the main goal was to support scholars in all steps of the data life cycle, starting from data acquisition to archiving and data publication. In 2021, GFBio was integrated into the newly established *Nationale Forschungsdateninfrastruktur - NFDI4Biodiversity*. The data portal was revised and nowadays provides access to more than 16 Mio biodiversity metadata files. The majority of this data is linked with primary research data, which can be further analyzed in an interactive visualization tool. Moreover, NFDI4Biodiversity is the national contact point for submitting biological data to ensure sustainable, long-term preservation and publication of research data. Data curators guide scholars to the submission process and forward the data to dedicated data centers. In addition, NFDI4Biodiversity organizes and supports education and training of scholars on all aspects of research data management. Associated data archives are PANGAEA[41] (environmental data), EBI/ENA[42] (genome data) and numerous German natural collections and museums such as Senckenberg Gesellschaft für Naturforschung – Leibniz Institute, Frankfurt[43], Staatliche Naturwissenschaftliche Sammlungen Bayerns – SNSB IT Center, München[44], Leibniz Institute for Research on Evolution and Biodiversity, Berlin[45], Botanic Garden and Botanical Museum Berlin, Freie Universität Berlin[46], State Museum of Natural History Stuttgart[47], Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig[48], Zoologisches Museum Alexander König[49] and the Plant Genomics and Phenomics Research Data Repository (e!DAL-PGP)[50].

We collected 69 search questions and keywords from 20 GFBio members. Some of them were visited in person at their working environment. The questions were utilized for the evaluations described in Chapter 3 as well as in the analysis and establishment of an information architecture presented in Chapter 4. In addition, eight questions were selected for the final user evaluation introduced in Chapter 8.

[40] GFBio, `https://www.gfbio.org`
[41] PANGAEA, `https://pangaea.de/`
[42] EBI/ENA, `https://www.ebi.ac.uk/`
[43] SGN, `https://www.senckenberg.de/`
[44] SNSB, `https://snsb.de`
[45] MfN, `https://www.museumfuernaturkunde.berlin/de`
[46] BGBM, `https://www.bgbm.org`
[47] SMNS, `https://www.naturkundemuseum-bw.de/`
[48] DSMZ, `https://www.dsmz.de/`
[49] ZFMK, `https://bonn.leibniz-lib.de/en`
[50] eDAL, `https://www.ipk-gatersleben.de/`

Apart from search questions, we also used GFBio/NFDI4Biodiversity metadata files in the evaluations described in Chapter 4 and Chapter 8. Ten metadata files from PANGAEA were utilized in the Named Entity Recognition corpus presented in Chapter 6.

2.5 Summary

In this chapter, we briefly introduced the research fields this work is built on. *User-centered Design* is the process to achieve a usable and user-friendly system. This goal is also called *Usability*. User-centered design is characterized by active user participation in all four steps of the developement (context of use, requirements collection, design solution and evaluation). In order to assess whether an application is usable and user-friendly, different user testing methodologies were developed. Formative evaluations aim to evaluate a system in an early stage of the development, while a summative evaluation aims to determine whether a system meets the user requirements. It is mainly conducted at the end of development cycles. There are multiple metrics to measure usability on different scales. Performance-based metrics aim to measure efficiency-based criteria, such as time on task or task success. Issue-based metrics focus on occuring usability issues that prevent users to fulfill their tasks, and self-reported metrics collect subjective and self-assessing feedback in surveys.

Information retrieval is the research field of search engines. There are several keyword-based retrieval algorithms to convert text into a mathematical representation that machines are able to understand. Evaluation is mainly focused on the accuracy of the retrieval models to determine the relevance of the returned information. A major initiative towards more user engagement in evaluations was the TREC Interactive Track, which ran from 1997 to 2002. The aim was to balance system-focused and user-focused evaluation studies.

A quite new field of research is the *Semantic Web*. Once the world wide web has emerged in the mid-nineties, there was the desire to provide and query for more meaningful information in the web. Thus, all web resources got unique identifiers that can be shared and linked, but they remain stable in the world wide web. Over the past two decades, formal models were developed that allow the modeling of the real world in a logic way and to link unique identifiers. For most domains, separate vocabularies and ontologies have been established that, in best case, are also interlinked among each other. The Life Sciences are one of the most active domains producing such terminologies. All web resources and their formally described relationships represent a network that machines can read and understand (Linked Open Data). These networks (nowadays also called Knowledge Graphs) and their applications form again new research fields being addressed at numerous conferences and scientific journals.

This work aims to support scholars in *Biodiversity Research* to find relevant datasets

for their research. Driven by the dramatic loss of biological diversity on earth, a variety of research projects have been established to understand the relationships among species and ecosystems to draw conclusions and develop approaches on how to overcome biodiversity loss. For the purpose of this work, we collected search questions and used metadata files from different biodiversity projects mainly located in Germany.

Chapter 3

Dataset Search and Improvements by Semantic Enrichment

"Having the right data is usually better than having more data."

- Christine L. Borgman in 'Big Data, Little Data and No Data', *Information scientist*

Finding relevant scientific data is a time-consuming and difficult task due to many reasons [Borgman, 2015]. Often scholars can not retrieve any data, because data do not exist, are not available or the provided search tools to do not support scholars properly (see also Chapter 1.3). In the Life Sciences, only very few studies exist that evaluate current data portals and explore the obstacles in dataset search. Therefore, this chapter addresses H_1 and aims to identify the problems in the retrieval process of current dataset search in the biodiversity domain. We also explore whether semantic techniques can help to overcome these issues.

At first, we provide an overview of related work in Section 3.1 with respect to dataset search, evaluations of current data portals and semantic search approaches in the Life Sciences. Subsequently, we evaluate the dataset search of *GFBio*, a data portal providing numerous datasets for environmental, ecology and biodiversity science (see also Chapter 2.4). In this evaluation, we focus on two aspects of the user perspective (Section 3.2): (A) We conduct a user study to determine the relevance of the search results. (B) In addition, the overall user experience is explored in a survey.

In a second study, we analyze the usefulness of semantic technologies for dataset search in biodiversity research, and we examine whether users actually benefit from a semantic search. The search results from a keyword based search are compared to the results obtained from a prototypical semantic search (Section 3.3). This second study and its results were published at *SEMANTiCS'16* [Löffler and Klan, 2016]. The findings led to the development of a first version of a semantic dataset search in the scope of the GFBio project published as demo paper at *ESWC'17* [Löffler et al., 2017].

3.1 Related Work

The more scientific articles were available for search at scientific publishers, the more the desire grew to have access to the underlying data, too. Numerous data archives in the Life Sciences have been established over the past decades. In contrast to information retrieval, which aims to find documents (unstructured textual resources), dataset search wants to retrieve datasets [Chapman et al., 2019]. We introduce studies evaluating dataset search applications in Section 3.1.1. In particular, we focus on evaluation studies that assess the relevance of datasets in a retrieval system or that explore usability aspects.

The aim in information retrieval is to find relevant documents for a specific query. Conventional retrieval algorithms can only find documents that syntactically match a user's query keywords. This is an obstacle when looking for scientific terms with different spellings and namings. For instance, information seekers might look for genes, proteins or species with common names, but data providers used scientific terms in data descriptions. Providing search applications that go beyond syntactic matches was one motivation for several semantic search approaches in the Life Sciences, which we introduce in Section 3.1.2.

3.1.1 Dataset Retrieval

User studies aiming to improve dataset search are rare. The work by [Megler and Maier, 2013] and [Castelo et al., 2021] are approaches to enhance dataset retrieval in terms of filtering for spatial and temporal areas. [Megler and Maier, 2013] developed an approach that extracts features from primary data and adds it to metadata allowing an enhanced query for particular spatial and temportal parameters. [Castelo et al., 2021] presents, besides a keyword-search, a data integration approach that permits to join columns from multiple datasets.

To the best of our knowledge, very little studies exist that evaluate dataset search applications. [Megler and Maier, 2015] and [Kalantari et al., 2020, Kalantari et al., 2021] are examples from the earth observation and oceanography domain (spatial and temporal data). The studies by [Dixit et al., 2017] and [Chen et al., 2018] are based on the Datamed portal[1], a federated approach providing biomedical datasets from 76 data repositories, and the evaluation by [Volentine et al., 2015] was conducted in the scope of DataONE[2], another federated approach collecting environmental and biological datasets from various sources and providing a single point of access to all harvested datasets.

Most studies are qualitative user studies with less than 30 participants, and a main focus is on the evaluation of the usability. For evaluation data collection, the studies

[1]Datamed, `https://datamed.org/`
[2]DataONE, `https://www.dataone.org/`

used various methods of qualitative evaluation such as thinking-aloud, questionnaires and interviews (see also Chapter 2.1).

[Megler and Maier, 2015] also assessed the relevance of the retrieved datasets. The participants were permitted to search for at least three own information needs based on their research backgrounds. They rated the Top25 results on a four-point Likert scale, the recall was estimated. The study authors report on a very high MAP value (MAP 0.96) for the Top10 results. In contrast, [Dixit et al., 2017] states that users were not able to assess the relevance, as the metadata quality of the respective datasets was too low. In addition, [Chen et al., 2018] explored the relevance of the datasets in the Datamed portal (2.3 Mio datasets in July 2017) based on the bioCADDIE benchmark [Cohen et al., 2017b], a test collection for dataset retrieval with 800,000 datasets, 15 questions and relevance judgments on the Top300 results (see also Chapter 7 for more information). Without any enhancements, the infAP (0.098), infNDCG (0.259) and P@10(0.468) (see also Chapter 2.2.3 for the metrics) were low to moderate. The Datamed portal provides an improved search based on query expansion and the UMLS[3] terminology service (see also Section 3.1.2). Utilizing this improvement, the performance increased (infAP: 0.203, infNDCG: 0.354, P@10:0.602).

Apart from [Chen et al., 2018], all other studies report on usability issues, which can be classified into two main topics.

- *Metadata quality:* The biggest usability issue in all studies were incomplete metadata [Volentine et al., 2015, Dixit et al., 2017, Kalantari et al., 2020]. For instance, users looking for biomedical datasets would like to see information on biomedical concepts, data type, data collection method, data format, data processing, sample description, date of collection or variables measured [Dixit et al., 2017]. However, this information was not present. Participants of the DataONE study [Volentine et al., 2015] suggested to have an option to open metadata in a separate tab to facilitate the comparision of multiple datasets. They also would prefer more hover functions for additional information. Users in the study of [Kalantari et al., 2020] complained about missing filters, tags being not relevant to the data and completely irrelevant results. In addition, [Megler and Maier, 2015] report on different spellings of variable names, which led to no results.

- *User interface:* A major issue with respect to the presentation of search results was the usage of acronyms, terminologies or the wording of buttons and filters that users could not understand [Volentine et al., 2015, Kalantari et al., 2020]. Another problem was inconsistent information in title and description, and users disliked the solely textual presentation [Kalantari et al., 2020]. Users from the biomedical

[3]UMLS, `https://www.nlm.nih.gov/research/umls/`

domain were missing an enhanced search input [Dixit et al., 2017] to search for multiple topics, e.g., a search for phenotypes and genes.

Based on these identified usability issues, [Kalantari et al., 2021] developed an improved user interface with more metadata information for a data portal providing spatial and temporal data. The improvements include filters based on user interests, metadata presented as a table, attributes displayed in a separate tab, and per dataset users could define use cases to describe for what purpose the data had been used in the past. In interviews, they asked eight users about their impressions. The participants particularly liked the improved titles, filters and the metadata tab. They were also very pleased about clear instructions how to open the metadata tab and a map view.

When we started our research in 2015, only two studies on user evaluations in dataset search were available. Seven years later, only very few studies have been added. The RDA survey of the Data Discovery Interest Group from 2018 [Khalsa et al., 2018] also confirms our assumption that only few data repositories evaluate their systems. In particular, evaluations in the Life Sciences (and specifically in biodiversity research) are missing. There are few studies exploring the relevance in dataset search applications, comparing different retrieval algorithms or retrieval technologies. Addressing this research gap, we conducted a relevance evaluation study in the scope of the GFBio project being introduced in the next Section 3.2.

3.1.2 Semantic Search in the Life Sciences

Semantic search is a broad field that requires a more precise classification on what data it operates and what kind of search is meant by the term 'semantic'. [Bast et al., 2016] distinguish between nine different types of semantic search approaches along two dimensions: the indexed data source and the search approach.

Data sources for indexing are either *plain text* (e.g., web documents, scientific articles, blog posts, emails, tweets or news articles) or *knowledge bases (KB)* containing structured information as triples in semantic formats such as RDF and OWL. A third category of data sources is a *combination of text and knowledge bases*. More specifically, combined data refer to text being enriched with semantic annotations. A semantic annotation is a link of a keyword to an entity (concept) in a knowledge base. The process to match keywords and entities is called *Named Entity Recognition* (NER) and *entity linking* or *entity resolution* (see also Chapter 2.2.1 for basics in Natural Language Processing). Combined data also include merges of multiple knowledge bases. The result are large knowledge graphs such as DBpedia[4] or Wikidata[5].

[4]DBpedia, `https://www.dbpedia.org`
[5]Wikidata, `https://www.wikidata.org`

With respect to the search approach, [Bast et al., 2016] differ between three categories: keyword search, structured search and natural language search. A *keyword search* is the most common search approach. Users type, in accordance to their information need, relevant terms into an input field. They obtain documents or entities (or both) that are relevant to their query. This approach is easy and fast for users, but the search is limited to a syntactic match of query keywords and terms in data sources; semantically related documents can not be obtained. In contrast to keyword search, a *structured search* aims to obtain entries from a knowledge base with query languages such as SPARQL. Formulating these structured queries is the mightiest semantic approach and allows the maximum "precise semantics" [Bast et al., 2016]. However, constructing these queries requires specific knowledge and can get very complex. Hence, it is difficult for end-users to use systems with a structured input. *Natural language* interfaces aim to overcome both limitations and permit to enter a phrase or a full natural language query. Usually, such a query starts with a W-question word (e.g., what, why), and users expect to obtain exact answers on their questions. Entering natural language is easy and convenient for users, but very difficult to understand for machines. Ambiguity is one of the biggest obstacles in these interfaces, because a term can have several meanings in different contexts.

The authors argue that this classification also contains 'grey zones' and some aspects are implied implicitly. For instance, the search result is not separately considered, but is a result of the data source (used for indexing) and the search approach. Also, 'search approach' does not only refer to the query input and explicit retrieval systems, but also considers related tasks such as informantion extraction and possible add-on search tasks.

Figure 3.1 presents an overview of all nine categories. Starting from a classical keyword search over text (upper left corner), the ultimate semantic search is a natural language search over combined data (lower right corner). Systems such as IBM Watson [Ferrucci, 2012], WolframAlpha[6] or recent developments in the Semantic Web and NLP community ([Diefenbach et al., 2020, Shen et al., 2019, Plepi et al., 2021]) are currently getting increasing attention and have brought up multiple challenges in question answering, e.g., the QALD challenge[7] or the BioASQ challenge in the Life Sciences[8]. In particular, question answering approaches based on Google's BERT language model [Devlin et al., 2019] have attracted many researchers, e.g., [Qu et al., 2019] [Wang et al., 2019, He et al., 2020]. However, returning an exact answer on a query is not the desired result for all use cases. As the aim in dataset search is to obtain relevant datasets (see also Subsection 3.1.1), the whole third column in Figure 3.1 is out of scope of this work. We also do not want to focus on systems that only search over knowledge bases. Even if there are a couple of approaches in the Life Sciences that offer

[6]WolframAlpha, `https://www.wolframalpha.com`
[7]QALD, `http://qald.aksw.org/`
[8]BioASQ, `http://bioasq.org/`

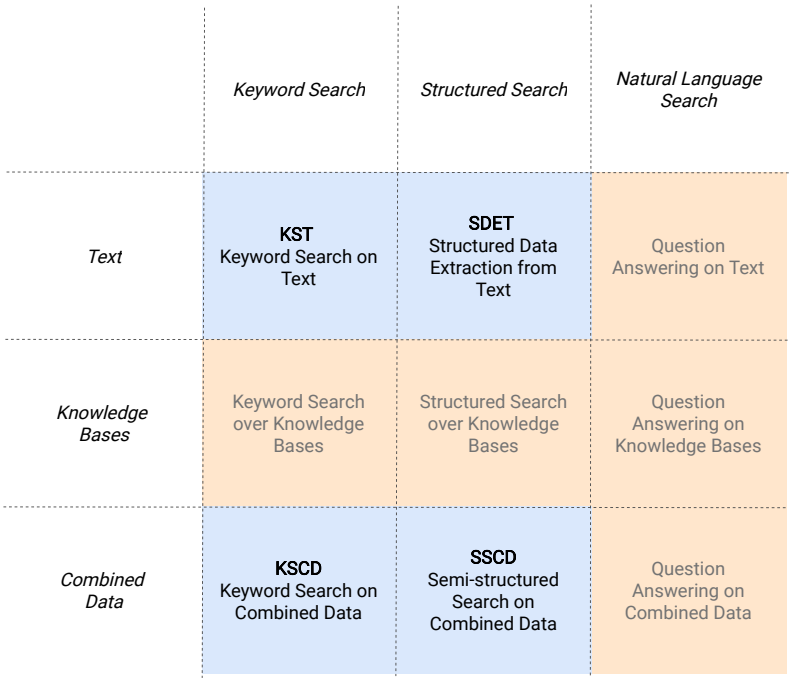


Figure 3.1: Semantic search categories based on [Bast et al., 2016] along two dimensions: data sources (row) and search approach (column). The orange colored categories are not considered in this work, only the blue colored groups are introduced.

a keyword search, a structured search or visual interfaces over biomedical knowledge bases [Belleau et al., 2008, Sy et al., 2012, Schweiger et al., 2014, Kamdar et al., 2014, Hoehndorf et al., 2015, Zaki and Tennakoon, 2017], we want to focus on systems that are based on text or combined data, because in dataset retrieval (which we aim to improve), metadata usually have at least some textual resources. Therefore, in the following list, we briefly introduce the four categories based on [Bast et al., 2016] that are relevant for the subsequent literature study on semantic search systems in the Life Sciences.

- *KST - Keyword Search on Text:* Users enter some keywords that are relevant for their information need and obtain a list of documents. The system finds and ranks documents based on frequent words occuring in the documents, or they learn the relevance from pre-trained past data. These classical search systems provide good results when users enter keywords that exist in the documents or that occur in external resources. However, more complex questions requiring to go beyond this syntactic match are not possible. Most existing search engines and data portals support this type of search. It is the most common approach and also includes search applications that automatically expand entered keywords on related terms (query expansion).

- *SDET - Structured Data Extraction from Text:* In this approach, search is not the primary goal but an add-on task. The motivation for this method is driven by the fact that users need support in getting a quick overview of a text to decide whether a

document is relevant for their information need or not. Therefore, the primary goal is to identify important terms or entities or relations in text and to highlight them in the result. This information extraction task is a pre-processing step before a search is added, or the enhanced structured query is forwarded to systems that obtain entities from KBs (Structured Search in Knowledge Bases). The extracted information is either stored in a triple store or is used to obtain additional information from one or multiple knowledge bases. The search process is performed on a classical text-based index. Most search systems in the Life Sciences fall into this category. Main entities such as genes, diseases or species are extracted from text by means of query or entity expansion. The entered keywords are expanded and the extracted entities are highlighted in the result snippets. These extractions can also include links and URIs to resources in the Linked Open Data cloud. Users benefit from this approach as it combines information extraction and search. However, determining the correct meaning of the extracted information is a challenge and often does not result in high precision and recall values. Moreover, not all information needs can be formulated in a structured format.

- *KSCD - Keyword Search on Combined Data:* The underlying data in these types of approaches are either text with entity annotations or Semantic Web data (knowledge graphs). The aim is to provide a keyword search and to return a ranked list of concepts. The result may also contain text snippets that are relevant to the query. This approach combines techniques from KST and knowledge bases. The input could be either keywords or entities. For keywords, matching URIs are looked up. For entities, a virtual document is created based on string literals or relations from the knowledge base. This virtual text is then used for a search with classical techniques from KST to retrieve relevant documents. Another approach is to perform a keyword search over text at first. Afterwards, occuring entities in the result are extracted and ranked. This is an easy approach that makes use of conventional search techniques. However, it is mostly a keyword search and therefore, it faces the same limitations as in KST approaches.

- *SSCD - Semi-structured Search on Combined Data:* As in the previous category, combined data are the data source for these approaches. The query input can be either keywords or a structured search or a combination of both. And the result are ranked lists of entities, matching documents or result snippets and information from knowledge bases. The main idea of this approach is to store text and knowledge bases either in separated indexes or in a combined-index. According to the query input, different user interfaces are required. Users benefit from this approach as it merges the advantages from keyword search over text (widley spread and well-known) and core semantic technologies (structured search - the most powerful se-

mantic query). A limitation of this approach is that a structured query input requires knowledge about the query language, which can get very complex and thus, it is difficult for end-users.

The Life Sciences, and in particular the biomedical domain, are a very active community in terms of semantic search approaches. Based on a variety of vocabularies and ontologies in the Linked Open Data cloud (see Chapter 2.3.4), numerous search approaches have been developed over knowledge bases ([Belleau et al., 2008, Sy et al., 2012] [Schweiger et al., 2014, Kamdar et al., 2014, Hoehndorf et al., 2015] [Zaki and Tennakoon, 2017]). The second large research community is inspired by NLP approaches and a series of workshops with different focus areas, e.g., BioNLP[9] (Named Entity Recognition of important biological entities, relation extraction, co-reference resolution, summarization, question answering), BioCreative[10] (gene, protein detection, relations between chemical compounds/drugs and genes/proteins, semantic indexing, provision of gold standards) and BioASQ[11] (semantic indexing and question answering). Table 3.1 provides an overview of semantic search applications in the Life Sciences with a focus on systems using text or combined data as data sources and systems providing a keyword based or structured input (the blue highlighted categories in Figure 3.1). The aim in all systems is to return textual information from publications or metadata. All systems were categorized based on their data sources, the extracted information and the supported semantic search approach.

All systems listed in Table 3.1 provide a user interface. Most developments (in particular the older ones) are not online anymore. Moreover, source code is only available for very few approaches [Pachzelt et al., 2021][13], [Chen et al., 2018][14]. Therefore, we assessed the semantic search approach based on the provided descriptions in the publications. Almost all applications use either PubMed [15] articles, a subset of PubMed articles or other biomedical publications and clinial trials as data source. Only Datamed [Chen et al., 2018] utilizes biomedical datasets from several data repositories, and BioFID [Pachzelt et al., 2021] is the only approach with textual data sources from the biodiversity domain. Systems focusing on the biomedical domain mainly extract information on genes, proteins, diseases, drugs and relations from text. In contrast, BioFID [Pachzelt et al., 2021] highlights different entities that are more relevant to biodiversity research such as taxons, habitats, geographic locations and temporal information. Most systems are approaches that extract structured information from text (SDET). Often the result snippets or the whole text provide highlightings of occurring entities, e.g.,

[9]BioNLP, `https://aclanthology.org/venues/bionlp/`
[10]BioCreative, `https://biocreative.bioinformatics.udel.edu/`
[11]BioASQ, `http://bioasq.org/`
[13]BioFID, `https://github.com/FID-Biodiversity`
[14]Datamed, `https://github.com/biocaddie`
[15]PubMed, `https://pubmed.ncbi.nlm.nih.gov`

System	Data source	Entity types	Approach	Terminologies
ESSIE [Ide et al., 2007]	ClinicalTrials.gov	-	KST, text based index, query expansion on search input	UMLS [UMLS, 2022]
Textpresso [Müller et al., 2008]	PubMed articles related to neuroscience	receptor, brain area, cellular component + relation	SDET, text based index, main categories identified and added to the index	GO [GO, 2021]
GoWeb [Dietze and Schroeder, 2009]	PubMed	genes, diseases	SDET, text-based search, retrieved snippets are enriched with NLP pipelines and ontology terms	MeSH [MeSH, 2022], GO [GO, 2021]
GeneView [Thomas et al., 2012]	PubMed articles	genes, mutations, species, chemicals, mentions of cell types, drugs, diseases, enzymes, tissues	SDET, text based index, different text mining pipelines to extract entities and relations which are stored in an RDBM, allows structured, entity-wise search	MeSH [MeSH, 2022]
Pubtator [Wei et al., 2013, Wei et al., 2019]	PubMed abstracts and articles	diseases, species, mutations, chemicals and genes	SDET, text based approach, pre-annotaion with various NLP pipelines	MeSH [MeSH, 2022], NCBITaxon [NCBITaxon, 2022], NCBIGene [NCBIGene, 2022], GNormPlus [Cai et al., 2015]
PolySearch2 [Liu et al., 2015]	Biomedical literature and databases	genes, disease, proteins, pathways, toxins, organs, MESH and GO terms	SDET, text based approach, patterns to identify relations	GO [GO, 2021], MeSH [MeSH, 2022]
DeepLife/LongLife [Ernst et al., 2016, Ernst et al., 2019, Terolli et al., 2020]	biomedial articles, clinical trials, health forum posts	genes, diseases, anatomic parts, symptoms, treatments	SDET, text based approach, entity detection and entity expansion, linar combination of TF-IDF scores (keywords, entities, entity types), rule based and probabilistic entity recognition	UMLS [UMLS, 2022]
BEST [Lee et al., 2016, Choi et al., 2012]	PubMed articles	genes, diseases, drugs, chemical compounds, transcription factors, miRNAs, toxins, pathways, mutations	SSCD, combined text and entity index, returns text and ranked entities, TF-IDF ranking, dictionary based approach for NER	GO [GO, 2021], MeSH [MeSH, 2022], FDA [FDA, 2022]
LIVIVO [Müller et al., 2017]	50 biomedical literature sources such as Pubmed and Agricola	-	SDET, text-based index, graph store (Neo4J) used in addition, TF-IDF ranking, query expansion	MeSH [MeSH, 2022], DrugBank [Wishart et al., 2018], AGROVOC [AGROVOC, 2022]
semedico [Faessler and Hahn, 2017]	PubMed abstracts + full text	Genes, Proteins, Species	SDET, text-based index, graph store used in addition (Neo4J), rule based and machine learning based NLP pipelines, TF-IDF ranking	MeSH [MeSH, 2022], GO [GO, 2021], GRO [Beisswanger et al., 2008]
LitVar [Allot et al., 2018]	PubMed	genomic variants, genes, chemicals, diseases, species	SDET, various NLP pipelines to extract entities	MeSH [MeSH, 2022]
Thalia [Soto et al., 2018]	PubMed abstracts	chemicals, drugs, metabolites, genes, diseases, proteins, species, anatomical entities	SDET, dictionary matching and conditional random fields models, query expansion	ChEBI [Hastings et al., 2016], DrugBank [Wishart et al., 2018], HMDB [Wishart et al., 2007], HGNC [HGNC, 2022], UMLS [UMLS, 2022], UniProt [Consortium, 2020], NCBITaxon [NCBITaxon, 2022], CARO [CARO, 2022]
Datamed [Chen et al., 2018]	76 biomedical data repositories	diseases, chemicals, genes, biological processes	KST, text-based index, MongoDB for extracted entity types, rule-based and machine learning-based NLP pipelines, query expansion	UMLS [UMLS, 2022], MeSH [MeSH, 2022]
BioFID [Pachzelt et al., 2021]	Digital Collection Biology (German) [12]	taxon, plant, animal, location, time	SDET, text-based index, linkage to external sources, machine learning-based NER	GBIF [GBIF, 2020], EOL [Parr et al., 2014], wikidata [Wikidata, 2022], wikipedia [Wikipedia, 2022], Geonames [Geonames, 2022], GND [GND, 2022], Wordnet [WordNet, 2022]

Table 3.1: Semantic search approaches in the Life Sciences

[Müller et al., 2017, Soto et al., 2018, Chen et al., 2018, Allot et al., 2018, Wei et al., 2019], and in some approaches, these highlights also include related information on synonyms or other expansions on hierarchical relations such as more specific terms or broader terms [Faessler and Hahn, 2017, Terolli et al., 2020]. Highlightings may also contain connections to external resources [Pachzelt et al., 2021] or the identified concepts are linked to knowledge bases [Lee et al., 2016, Wei et al., 2019]. Only one system is an approach on combined data [Lee et al., 2016]. It allows a keyword search on a mixed index containing text and semantic annotations (matching URIs from ontologies). The result are a ranked list of entities and relevant articles.

The analysis of related work reveals that systems on combined data are still very rare. In particular, there are no approaches in the Life Sciences enabling a keyword search and a structured search with query languages such as SPARQL. Outside the Life Sciences, a few systems have emerged that offer such an enhanced search on combined data [Bast and Buchhold, 2013, Bast and Buchhold, 2017, Cunningham et al., 2013] [Bontcheva et al., 2014].

However, the challenge for appropriate user interfaces remains. In order to avoid complex query languages in the user interface, more approaches are needed that combine the benefits from structured search over text (highlighting of important entities, search over categories) and knowledge bases (full semantic knowledge).

3.2 Evaluation of GFBio's dataset search

User evaluations in dataset search are rare. In order to address the need for more user studies, we evaluated the dataset search of the GFBio portal (Chapter 2.4) in a user study with six scholars in 2016. We measured the relevance of the returned datasets and conducted a user survey to determine the overall user experience and satisfaction.

3.2.1 Data Corpus

In 2015, the search index mainly consisted of datasets from an environmental data archive - PANGAEA - and selected datasets from natural collections and museums in Germany (named in Chapter 2). The search index was based on metadata, descriptive information of the primary data. As the various data archives are built upon different data structures and metadata schemes, GFBio agreed to map collection data from the collections' specific metadata schema ABCD 2.06[16] (see also Chapter 4.3) to pansimple[17], a Dublin Core-based metadata schema developed by PANGAEA. Users benefit from a unified metadata schema in the search interface, as a limited and thoroughly selected amount of metadata

[16] ABCD 2.06, `https://abcd.tdwg.org/xml/documentation/primer/2.06/`

[17] pansimple, `https://ws.pangaea.de/schemas/pansimple/pansimple.xsd`

fields allows a facted search over the different datasets [Hearst, 2006]. Listing 3.1 contains an example of a metadata file in pansimple metadata schema.

Listing 3.1: Metadata file in pansimple format (excerpt) [Frenzel et al., 2016]

```
<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>
```

The number of data records in GFBio's search index is quite dynamic and steadily increased over time. In February 2017, the search comprised 4.6 Mio datasets. In fall 2018, more than 5 Mio datasets were indexed and in March 2020, 15 Mio datasets were accessible (but 14 Mio from one data center - SNSB). In April 2022, around 17 Mio datasets were available including metadata files from the molecular data archive EBI/ENA[18]. Figure 3.2 presents a screenshot of the search in January 2016. The result items are listed in the middle of the browser window. Each item provides the title of the dataset, the data center hosting this data, a summary of provided parameters, links to the landing page of the dataset at its respective data center and a link to download the data (if the data policy allows it). The left pane displays facets for filtering, e.g., geographic region or data center. The right pane contains a map for visualizing the geographic location of the dataset if coordinates are available. Below the map, the results from GFBio's Terminology Service (TS) [Karam et al., 2016] are displayed. All search terms are sent to the TS to look for matching URI concepts in the knowledge base.

3.2.2 Relevance Evaluation

Relevance evaluations in information retrieval consist of three components: a corpus, queries and human judgments (see also Chapter 2.2.2). The corpus in this evaluation comprised all available datasets in the GFBio search in February 2016 (2 Mio datasets). In order to get genuine scholarly search interests, we asked scholars from GFBio part-

[18]EBI/ENA, `https://www.ebi.ac.uk/`

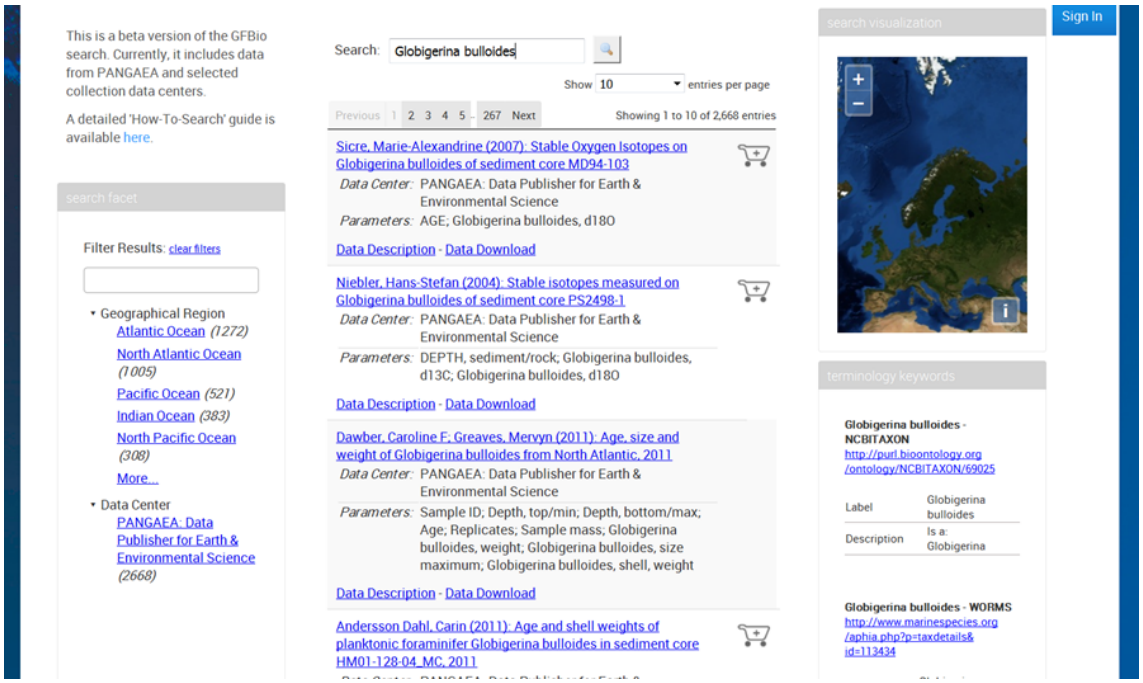


Figure 3.2: GFBio search interface in February 2016: Users can select datasets (shopping cart icon) to be displayed on the map if coordinates are available. Facets for dataset filtering are provided on the left, and results from GFBio's Terminology Service [Karam et al., 2016] are presented on the right.

ner institutions to provide search questions from their research background. In total, six researcher provided 25 search questions. In addition, we obtained search logs of the *Biodiversity Exploratories* (Chapter 2.4). We reviewed all queries and search log entries with respect to the results in the evaluation corpus. As we did not retrieve datasets for all queries, we utilized only these queries that actually led to results. The final search query pool contained 16 questions and search log entries.

Two queries of this search pool were provided to each participant. The users rated the Top 25 search results per query on a 7-point-Likert scale from '0'(irrelevant) to '6'(highly relevant). Afterwards, we asked the users to provide and rate up to two own queries. Table 3.2 presents all questions that received ratings. The asterisk denotes questions that were provided by the users during the evaluation. The question corpus was quite heterogeneous and comprised questions to obtain species-related datasets, datasets containing environmental information or measurements on specific data parameters. Concerning the complexity and granualarity, the question corpus was very diverse, too. Scholars were interested in specific data, e.g., *What can I find about 'bee'/'bees'?*, but were also looking for data to answer more complex questions such as *Which are habitat data for specific taxa?*. All user ratings were saved in csv files and are available at Zenodo [Löffler and Opasjumruskit, 2022].

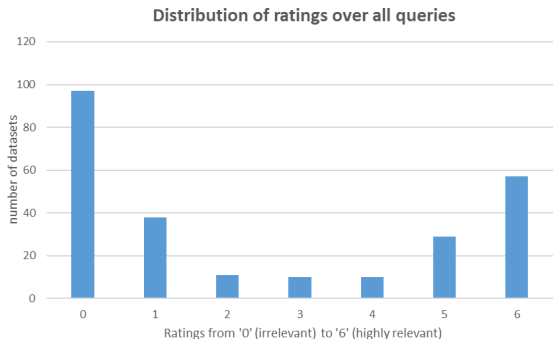
Query	User	Question	Query Terms
Q1	User 1 (Test)	What data is there for vegetation data from west africa?	vegetation data West Africa
Q2	User 1 (Test)	What data is there for root length?	root length
Q3*	User 1 (Test)	What data is there for foraminifera and benthic?	foraminifera, benthic
Q4	User 4	Which data exist for 'food webs' in the north sea?	food webs, north sea
Q5	User 4	How is the distribution of Holothuroidea (sea cucumber) in the Arctic Ocean?	Holothuroidea, Arctic Ocean
Q6*	User 4	What information is there on bird distribution in Germany?	Bird distribution Germany
Q7	User 5	What data is there for 'root length'?	root length
Q8	User 5	Which are the habitat data for specific taxa?	ecological variables, land use
Q9*	User 5	What can I find about 'bee'/'bees'?	bee
Q10*	User 5	Is there any data for 'land abandonment'?	land abandonment
Q11	User 6	How high are sulfate reduction rates at cold seeps?	cold seeps, sulfate reduction rate
Q12	User6	How has 'Goldenrod' spread over Europe?	Solidago canadensis
Q13*	User 6	Is there a global map and dataset of tree species diversity?	tree species diversity, global
Q14	User 7	How high are benthic oxygen uptake rates in den Atlantic?	Atlantic, oxygen uptake, respiration
Q15	User 7	Is there data about the leaf area index and in particular about diversity?	leaf area index, diversity
Q16*	User 7	Is the diversity of hosts influencing the diversity of plant pathogens?	plant pathogen, host, diversity
Q17*	User 7	Is the prevalence of diseases influenced by climate change?	climate change, disease
Q18	User 8	What data is there for Thysanoptera on sunflowers?	Thysanoptera, Helianthus
Q19	User 8	What data is in the repository for the 'tree of the year' 2016 (Tilia cordata)?	Tilia cordata
Q20*	User 8	What data is there on the population abundance of millipedes?	Abundance and Diplopoda
Q21*	User 8	What data is there on the abundance of Pollinators e.g. Bees Bees	Bees Bees
Q22	User 10	What data exist for microbial activities in groundwater?	microbial activities, groundwater
Q23	User 10	What data is in the repository for nutrients in soil?	nutrients, soil
Q24*	User 10	What data is in the repository on forest cover change in Africa?	forest cover change, Africa
Q25*	User 10	What data is in the repository on ocean acidification and coral bleaching	ocean acidification, bleaching
Q26	User 11	Is there data about the forest-specific diversity of vascular plants?	vascular plant, diversity, forest
Q27	User 11	Are there soil samples in Germany including measurements of pH and water content?	soil, pH, water content, Germany
Q28*	User 11	Is there data about plant traits influenced by precipitation and grazing?	plant traits, precipitation, grazing

Table 3.2: Overview of questions and search terms

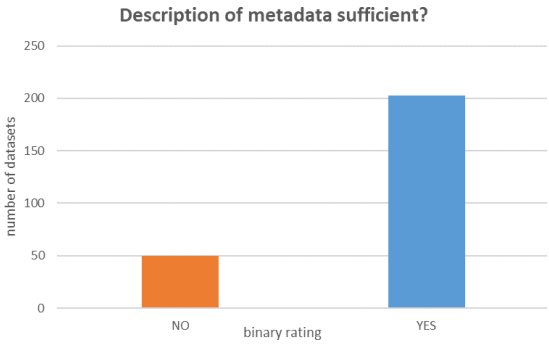
* query was not given, user provided question and keywords

MAP, relevant: all ratings > 0	MAP, relevant: all ratings > 3	nDCG
0,69	0,57	0,68

Table 3.3: Retrieval metrics over all 253 ratings



(a) Distribution of ratings on a 7-point Likert scale from 0 (irrelevant) to 6 (highly relevant).



(b) Analysis on the information presented for each dataset in the result set. Raters assessed whether the provided metadata are sufficient to judge the relevance per dataset returned for a given query.

Figure 3.3: Distribution of ratings and metadata quality results of the GFBio search study in 2016.

Results: Overall, eight domain experts from GFBio and iDiv took part in the evaluation. Seven scholars had a research background in ecology, and one scholar had expertise in zoology. In total, the scholars rated 28 queries resulting in 253 judgments. We analyzed the results with common evaluation metrics in information retrieval, namely MAP and nDCG [Manning et al., 2008] (Chapter 2.2.3). Table 3.3 presents the final metrics. The first column contains the MAP values for all ratings above 0 that are considered as relevant. The second column presents the MAP values for ratings larger than 3. The results indicate a moderate relevance of the returned datasets.

Figure 3.3 presents the distribution over all ratings and illustrates that more than 50% of all judgments got binary ratings only. Scholars either assessed a dataset as 'irrelevant' or 'highly relevant', the other Likert-scale entries were less used, in particular the medium values. We conclude, too many rating options have probably confused the participants and therefore, they 'reduced' the rating scale to an amount that was feasible for them. Moreover, the distribution also points to a large number of irrelevant results. For the analysis of irrelevant results, raters were also asked, per dataset returned, whether the information presented was sufficient to judge its relevance with respect to the given query or whether they would need the primary data in addition. The results reveal (Figure 3.3) that the provided information was sufficient to assess the relevance of a dataset.

We also determined the correlation between irrelevant results and not sufficiently described metadata (Figure 3.4). The results denote that there is no correlation. Even if datasets required no primary data to judge the relevance (e.g., Q4 and Q14), the datasets got rated as 'irrelevant'. On the other hand, datasets with no sufficient metadata did not get irrelevant ratings, e.g., Q23 and Q24. Hence, good or poor metadata quality does not lead to higher or irrelevant ratings.

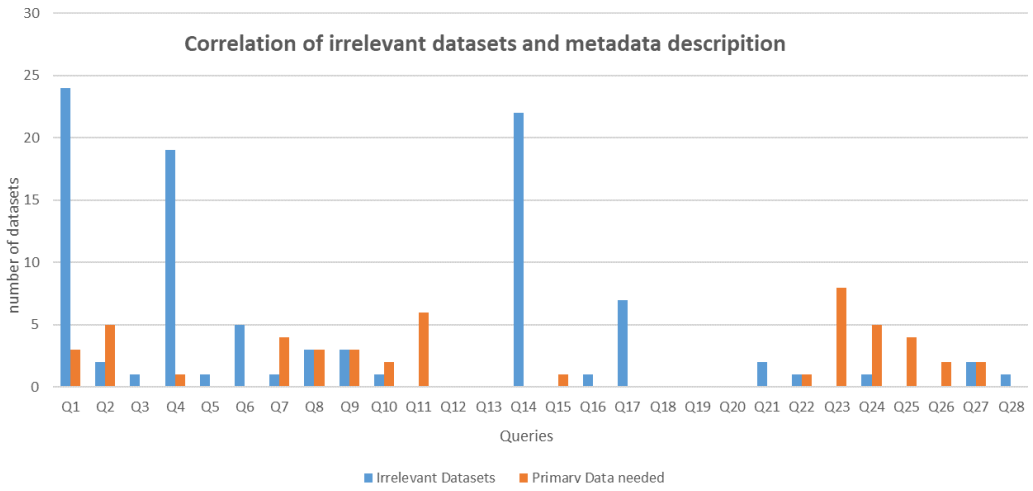


Figure 3.4: Correlation of insufficiently described datasets in the result set and irrelevant ratings.

3.2.3 User Survey

Following the relevance study, we asked the participants to take part in a survey about the overall user experience. Users rated statements grouped into four categories on a 5-point Likert scale from 'completely disagree' to 'highly agree'. For some statements, only binary ratings were provided. In total, eight scholars took part in this evaluation, assessed all 28 statements and provided qualitative feedback in seven free-text comments.

Table 3.4 presents the overall results. The focus of this survey was on the query input, controllability functions and the search result. In a fourth category, we wanted to obtain feedback on proposed improvements. We received good results with respect to the query input and the presentation of search results. Users are overall satisfied with the provided opportunities to enter a query and the information returned per dataset. One user mentioned that facets for organisms and a specific type of data are desirable. Concerning the amount of data presented in the result set, one scholar suggested to provide less information and to expand metadata only on demand. Users also evaluated suggested improvements. According to these results, users would appreciate a map with geo-reference refinement. They also favored suggestions from the system on how to improve the query, or they would like the system to provide examples and templates.

The results for control functions need further considerations. Figure 3.5 presents the results for the individual binary rated statements. Users' impression was that control functions are limited. For instance, users were not aware that quotes for exact matches are available. Some scholars had difficulties to use AND and OR to combine search terms. Surprisingly, the rating for the plural/synonym functionality got positive ratings. Plurals are indeed considered in search, but synonyms are not. Apparently, users only rated the plural function and overlooked to check whether synonyms are returned. It would have been better to split that statement into two individual statements.

With respect to the quality of the datasets returned, users had the impression that

Categories		Statements	Rating	Comments
Query input	Search box	length	3.375	1. longer search box, 2. Is it handling boolean search?
		clearly labeled	4.25	
	Search facet	availability	4.375	1. request for facets: organism, taxa, type of data
		options are useful/sufficient	3.75	
		gradual search refinement	4.375	
Search functionality	Control function	**spell check**	**2.5**	Not clear how to use it
		supports plural and synonym	3.5	
		'more like this'	**1.5**	
		store search terms and results	**2**	
	Expressiveness of query	**AND | OR**	**3**	1. Unclear if quotes are working 2. Results are not clear if these expressions work 3. bee AND/OR brazil return no result.
		quotes for exact match	**3**	
		* as a placeholder	4	
Search result	Quality	shows all necessary metadata	4	
		provides access to primary data	4	
		includes the location	3.625	
		provides access to the original landing pages of the data centers and archives	4.25	
		results are ranked by relevance	2.714	
		no duplicate results	**3**	
	Appearance	The search result page shows what was searched for and it is easy to edit and resubmit the search.	4	1. The list is a bit long and thin - a lot of scrolling down. 2. Sometimes you know from the title it is not good, perhaps a show/hide button for the metadata could be used so you can show the metadata for those that look useful - most were not
		The search results page makes it clear how many results were retrieved, and the number of results per page can be configured by the user.	4.625	
		The scope of the search is made explicit on the search results page.	3.75	
	Visualization	The handling of the visualization component is intuitive and easy.	3.875	
		The provided options for visualization are sufficient.	4	
Improvement	Proposed improvement	Showing a map with an overview of results rather than a list of results	3.75	1. Can't visualize the 'map' - do you mean geographic? That could be a useful option. 2. I prefer the list result, but map could help when you work at large scales. 3. a full map, as well as the list of results, would be an improvement. the map on it's own would not be better.
		The visualization component should provide a search refining with geo-reference data (e.g., search within a bounding box).	4.0	
	Useful hints or helps to improve the query	provides ideas for improving the query	3.375	
		provides templates, examples or hints on how to use the search effectively	3.25	
		Handles empty queries politely and provides alternative search terms	2.75	

Table 3.4: Overall results of the user survey. The users rated search statements grouped into four different categories on a 5-point Likert scale from '1' (completely disagree) to '5' (highly agree). For some statements, only binary scales (YES/NO) were provided. In order to make the values comparable, binary negative ratings were treated as 'completely disagree' and binary positive ratings were considered as 'highly agree'. The provided rating values are the average values of all ratings for this statement. Values equal or below '3' are highlighted.

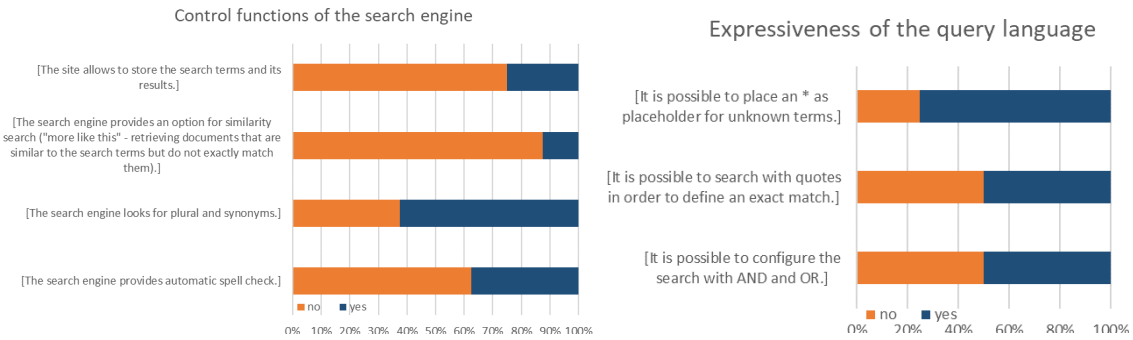


Figure 3.5: Survey results for statements concerning control functions (left) and statements concerning the expressiveness of the query language (right).

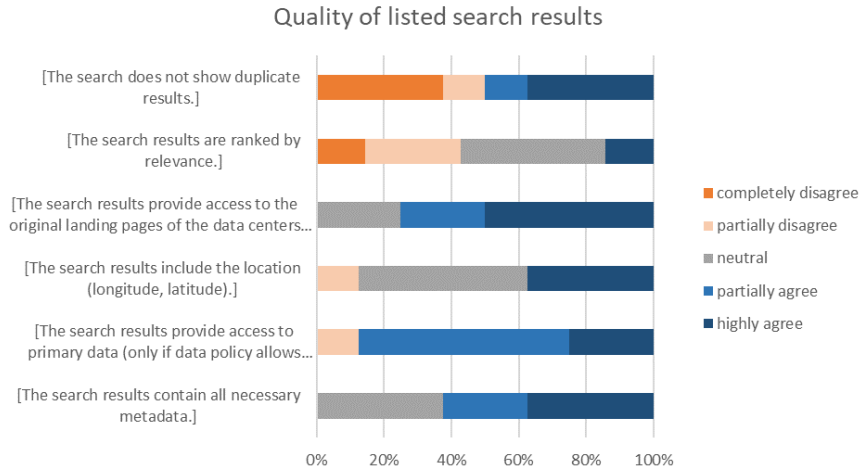


Figure 3.6: Survey result for the quality of the search result.

datasets were not sorted by relevance and contained duplicates (Figure 3.6). However, the duplicates were actually no duplicates. The large number of individual units in a collection dataset resulted in the same data descriptions but pointed to different physical objects. This picture reveals that the provided metadata information either needs to be improved or similar results have to be grouped to avoid the impression that duplicates are returned. Moreover, more relevant hits are required. This correlates with the results of the relevance evaluation. Probably, based on the moderate rating results of the prior relevance evaluation, users had the experience that the search result was not sorted by relevance.

Overall, we conclude that the users are satisfied with the usage of the query input and the presentation of search results. However, available control functions need to be better communicated. As the selected statements were a bit arbitrarly chosen, further studies are needed to analyze which control functions support users in retrieving more relevant data. Furthermore, the quality of the returned datasets need to be improved. The relevance of the search results was only moderate.

3.3 Comparison of a Keyword-Based Search and a Semantic Search

In a second user study, we aimed to explore the usefulness of semantic technologies in dataset search. We wanted to analyze whether scholars would be able to retrieve more relevant datasets in a semantic search and thus would actually benefit from semantic techniques. Therefore, we developed a prototype of a semantic search based on query expansion, and we linked key terms of a search task to concepts in domain ontologies. We expanded the search query with the obtained synonyms, alternate labels and labels obtained from further hierarchy relations. We conducted user tests with scholars, who assessed the relevance of the returned results in the semantic search and the keyword-based search (Subsection 3.3.1). In the following, we provide an extended version of our publication at *SEMANTiCS'16* [Löffler and Klan, 2016].

In order to obtain qualitative feedback, we interviewed the scholars after the relevance evaluation (Subsection 3.3.2). In total, six scholars with a background in biodiversity research took part in this time-consuming evaluation. Each visit/ phone call in December 2015 and February 2016 took around two hours. Three users were visited at their working environments and three users were contacted via telephone. Supplementary material can be found in our GitHub repository[19]. The primary data files of the user evaluation were published at Zenodo [Löffler and Klan, 2022]

3.3.1 Relevance Evaluation

As data sources, we utilized 92,856 metadata files from GFBio (randomly selected). We indexed the datasets with GATE Mímir [Cunningham et al., 2013], a search engine that allows searches over a text based and an annotation index. For these experiments, we only utilized the text based index with a boolean retrieval model and the $TF - IDF$ scoring function. Hence, we followed a basic search approach (KST), but enhanced each query in the semantic search with query expansion.

Six biodiversity scholars provided research and search questions related to their field of expertise. They also indicated relevant search terms for each query. Having both, the keywords entered for search and the underlying question, allowed us to interpret the meaning of the search terms correctly and made the query intent explicit. All queries used in this evaluation are listed in Table 3.5.

GATE Mímir provides a user interface with an own query language, which is not well suited for an evaluation with end-users. Hence, we integrated the results returned

[19]Comparative user study, `https://github.com/fusion-jena/comparision-keyword-semantic-search`

Query	Question	Search Terms
Q1	How high are sulfate reduction rates at cold seeps?	cold seeps, sulfate reduction rate
Q2	How high are benthic oxygen uptake rates in den Atlantic?	Atlantic, oxygen uptake, respiration
Q3	How is the distribution of Holothuroidea in the Atlantic Ocean?	Holothuroidea (sea cucumber), Atlantic
Q4	How high is the organic carbon content in arctic sediments?	arctic, sediments, organic carbon
Q5	Where do I find mesopelagic fish of the genus Cyclothone?	Cyclothone
Q6	How many eggs do copepods produce (e.g. eggs/female/day)?	egg production, copepoda
Q7	How variable is the oxygen concentration (e.g. in unit (mycro)mol/kg) of sea water in the mesopelagic zone (i.e. between 200-1000 m) of the global ocean?	oxygen, (mycro)mol/kg, sea water, mesopelagic zone
Q8	What data exist for Neogloboquadrina pachyderma or Globigerina bulloides?	Neogoboquadrina pachyderma, Globigerina bulloides
Q9	What data contains samples from surface water?	surface water, water sample
Q10	What are associated taxa, for example an insect and its host plant?	host (parasite), plant, insecta
Q11	What data exist for invasive grasses, e.g., Poaceae?	invasive grasses, e.g., (Poaceae)
Q12	What data is in the repository about 'climate change'?	climate change
Q13	What data is there for 'root length'?	root, length
Q14	What data exist for butterflies on oaks?	lepidoptera, quercus
Q15	What data is there for 'foraminifera' and 'benthic'?	foraminifera, benthic
Q16	What data is there for nutrients in soil?	nutrient, soil (terrestric)
Q17	What data is in the repository for the german tree of the year 2016 'Tilia cordata'?	Tilia cordata, Germany
Q18	What data exist for 'primula veris' in Germany?	Primula veris, Germany
Q19	Please show me all datasets about 'sunflowers'!	sunflower (Helianthus)

Table 3.5: Overview of questions and search terms used in the comparison of a keyword-based search and a semantic search.

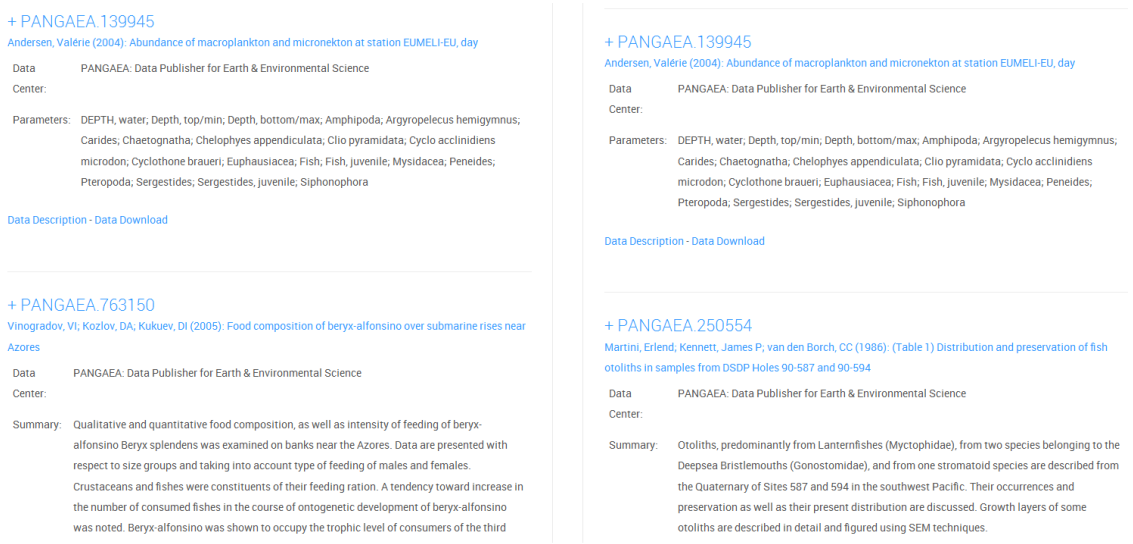


Figure 3.7: User interface displaying the two result sets based on the user-provided keywords (left) and on the expanded search terms (right)

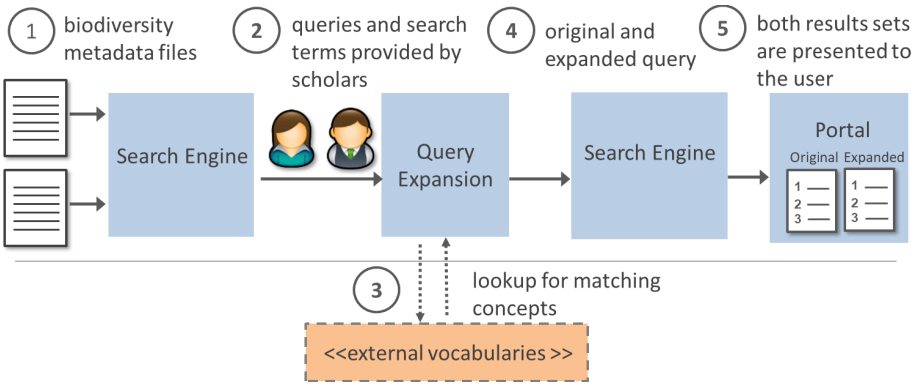


Figure 3.8: Experimental setup of the search expansion [Löffler and Klan, 2016]

by the search engine into the open-source portal Liferay[20]. In Figure 3.7, the left side of the portlet shows the results obtained from the original user-provided search terms, the right side displays the datasets that were retrieved using the semantic search. The users were not told which result set is based on which retrieval technique, however, the users quickly realized which result was based on which technology. Each search term in the semantic search was expanded with related terms obtained from GFBio's knowledge base [Karam et al., 2016] and Bioportal [Whetzel et al., 2011].

The domain experts assessed two parts of the result: (1) the relevance of the returned datasets in the keyword search result and the semantic search and (2) the expanded terms used in the semantic search to determine which expansion strategy performed best. In case (1), the users evaluated the relevance of the returned Top25 datasets on a 7-point-Likert-scale from 0 (irrelevant) to 6 (highly relevant). In case (2), ratings were provided on a binary scale.

Expansion strategies: For all provided search terms, we manually looked for matching concepts in the knowledge bases. The GFBio TS focuses on vocabularies for environmental, genetic and collection data, whereas Bioportal provides ontologies for the biological and medical domain. A preference was given to the GFBio TS since its ontologies are tailored to the datasets that were available in the GFBio portal. When a given search term could not be linked to a concept defined in one of the ontologies in the GFBio TS, we looked for entities in selected GFBio-related ontologies in Bioportal. Since geographic knowledge is not fully covered by the terminology providers, we only focused on the expansion of taxonomic information and environmental terms. Table 3.6 lists all ontologies used and the number of matching concept with search terms over all questions. For 34 out of 36 search terms we found matching concepts in the knowledge bases.

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

[20]Liferay, `http://www.liferay.com`

Source	Vocabulary	Concepts
GFBio TS	National Center for Biotechnology Information (NCBI) Organismal Classification (NCBITaxon) [NCBITaxon, 2022]	13
GFBio TS	Computer Retrieval of Information on Scientific Projects Thesaurus (CRISP) [Bair et al., 1996]	9
GFBio TS	ENVironmental Ontology (ENVO) [Buttigieg et al., 2016]	8
Bioportal	National Cancer Institute Thesaurus (NCIT) [NCIT, 2022]	3
GFBio TS	Chemical Entities of Biological Interest Ontology (CHEBI) [Hastings et al., 2016]	2
GFBio TS	Observation Ontology (OBOE) [Madin et al., 2007]	2
GFBio TS	Phenotypic Quality Ontology (PATO) [Köhler et al., 2019]	1
GFBio TS	Quantities, Units, Dimensions, and Types Ontology (QUDT) [FAIRsharing.org: QUDT, 2011]	1
Bioportal	Clinical Measurement Ontology [Shimoyama et al., 2012]	1
Bioportal	Gene Ontology (GO) [GO, 2021]	1

Table 3.6: List of ontologies used for query expansion and matched resources

sets displayed to the study participants side by side in a portal-based user interface. The overall flow is presented in Fig. 3.8.

We developed a JAVA command-line tool that expands the search terms following a certain strategy: We assumed that for a given search term all its synonyms as well as terms referring to descendant nodes, 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). For each concept with a label that matched one of the given search terms, we fetched all synonyms and direct sub-concepts (narrower concepts) and added them as expansion terms. If no descendant concepts were available, we selected the next super class (broader concept) and all sibling nodes. For species names, the genus was regarded as the most specific concept, since scientific names typically already contain the genus. Therefore, 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.

In some cases, we got a very large number of expansion terms, e.g., for insects (around $209,000$), which required to cut the number down to an amount that was manageable by the search engine. Therefore, we only utilized expansion terms for which metadata was present in the corpus. The listing below (Listing 3.2) shows an expanded set of search terms for the keyword *cyclothone* (a genus of deep sea fishes) containing an excerpt of species, synonyms and labels of sibling concepts for the term *cyclothone*.

Listing 3.2: Query expansion example

```
((cyclothone) OR (gonostomatidae) OR (diplophidae) OR (bristlemouths) OR (gonostoma
elongatum) OR (cyclothone acclinidens) OR (cyclothone pallida) OR (margrethia obtusirostra)
OR (sigmops) OR (cyclothone parapallida) OR [...] OR (sigmops bathyphilus))
```

In total, we distingued between four expansion types: synonyms, sub-classes (narrower concepts), sibling classes and super classes (broader concepts). In a first task, the

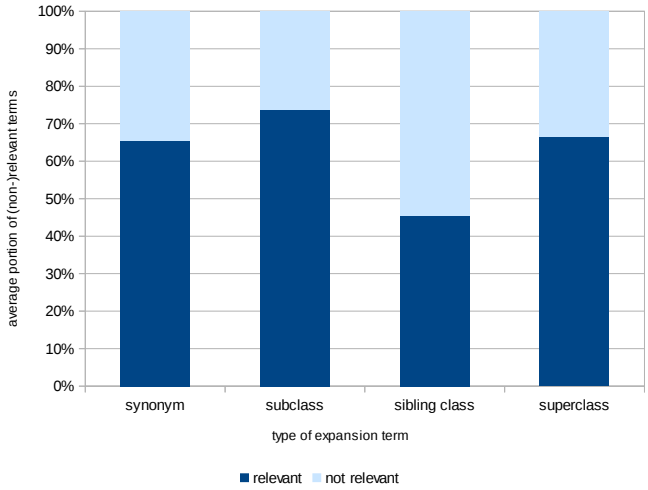


Figure 3.9: Result of the user rated expansion types [Löffler and Klan, 2016]

users had to assess the different expansion types. We showed them a list of all potential expansions per query and let them indicate which expansions were relevant with respect to their query.

Results: In total, we obtained 581 ratings from 38 queries (19 search tasks in two user interfaces). In a first analysis, we explored whether a particular expansion strategy resulted in more relevant keywords. Figure 3.9 presents the user ratings on the expansion types. The result reveals that all types of expansion are relevant from the user perspective. It is interesting to see that synonyms are not relevant per se (as assumed) and that users assessed expansions with superclasses as relevant for more than two-third of all provided terms. The aim of query expansion is to return a larger result set. It needs to be investigated whether this leads to a result with more relevant hits. Therefore, we compared the outcome of the keyword-based search with the result set of the semantic search (based on query expansion). We took a closer look on (a) the portion of relevant datasets and (b) on the user ratings.

(a) Portion of relevant datasets: We used different retrieval metrics (relative precision/ recall, MAP and nDCG, Chapter 2.2.3) to analyze the relevance from various perspectives. As the ground truth was not given, we determined the relative measures of recall and precision for the Top25, and we compared the results of the keyword search (KS) and the semantically expanded search (ES). We computed the relative precision as

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

If the result set was empty, the absolute precision was set to 0. The relative recall was

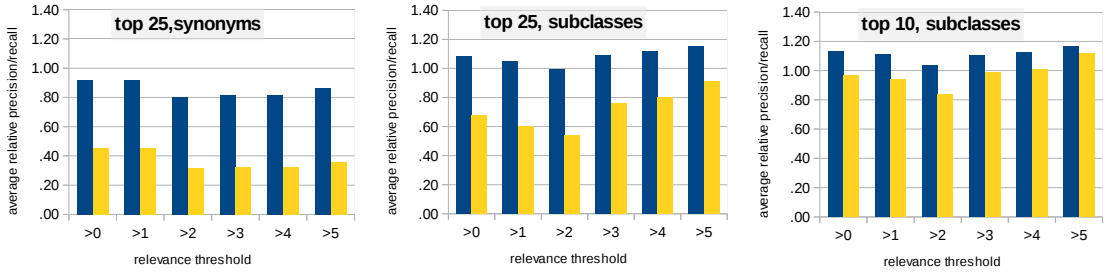


Figure 3.10: Average relative precision (blue) and recall (yellow)

MAP > 0	MAP > 3	nDCG	MAP > 0	MAP > 3	nDCG
0.74	**0.58**	0.72	**0.83**	0.55	**0.79**

Table 3.7: MAP and nDCG based on different relevance thresholds (> 0 and > 3) for the keyword (left) and semantic search (right)

determined as

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

A value of 1 denotes equal precision/ recall, a value lower than 1 points out that the precision (recall) of the semantic search is higher than the value of the keyword search. And a value larger than 1 indicates that the precision/ recall of the keyword search is higher than that of the search based on expanded keywords.

Figure 3.10 presents the relative results over all queries expanded with synonyms (left) and expansions with sub-concepts (middle and right). As we aimed to explore whether additional term expansions lead to a larger relevant result set, we omitted queries for which the users judged the keywords as irrelevant. The results were computed for different relevance thresholds. Figure 3.10 (left and middle) shows the relative precison and recall values for the Top25 results, whereas Figure 3.10 (right) refers to the Top10 datasets. The results reveal that term expansion did not increase the precision of the retrieval process (relative precision at around 1). However, for synonyms the relative precision is lower than 1, and thus, it had a slight positive effect on the relevance. For the recall, this effect is even stronger. The comparision of the Top25 and Top10 results reveal that the increase of the recall for the semantic search is larger in the Top25 results than in the Top10 results. This points to a suboptimal ranking. Looking at the MAP and nDCG values (Table 3.7), this outcome is confirmed. Even if the values for the semantic search are higher than for the keyword search, the results are not close to a value of 1 (a perfect ranking).

(b) User ratings: Besides the relevance, we also wanted to figure out whether a particular expansion strategy leads to higher user ratings. For both search applications, we

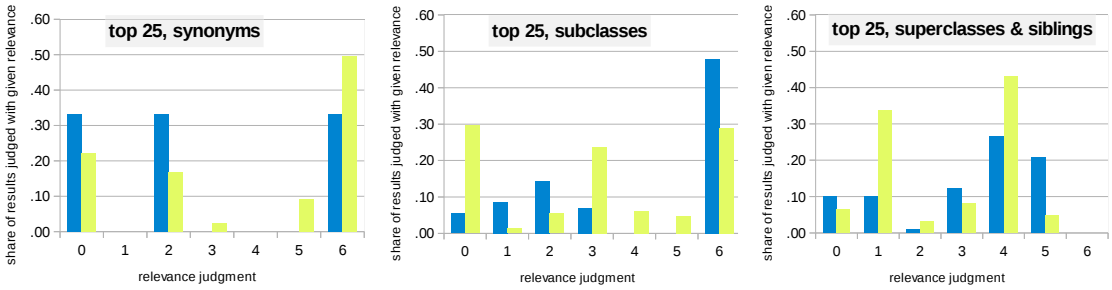


Figure 3.11: Distribution of average relevance for keyword-based (blue bars) and semantic search (green bars)

determined the fraction of the Top25 results with respect to all queries per expansion strategy (Figure 3.11). As in the relevance analysis, we only used queries with relevant keywords. The blue bars indicate the results for the pure keyword-based search, the lime green bars show the results of the semantic search using the term expansion technique. If the datasets returned are less relevant, the portion of low relevance values would increase (relevance values $0-2$). In contrast, if a particular expansion techniques result in higher ratings (relevance values $4-6$), we would observe an increase of the portions for high relevance values. For synonyms (Figure 3.11, left), we observe more highly relevant results. Hence, adding keywords with synonyms to the query increases the portion of highly relevant results. Extending the query with subclass labels (Figure 3.11, middle) results in a larger fraction of irrelevant datasets and in a larger amount of relevant datasets (relevance values $3-5$). Adding super classes and sibling labels (Figure 3.11, right) only leads to more results of low relevance (values 1 and 2).

3.3.2 Interviews

The relevance evaluation was supplemented by subsequent interviews. The questions and answers are presented in Table 3.8. The result of the third question was already analyzed in the previous section, so in the following, we summarize the remaining questions.

The answers reveal that users had the impression that the semantic search (the right tab) overall returned more relevant datasets. We can confirm this impression with the results from the relevance evaluation, however, the difference is not that large as the result from the interviews would assume. We also asked the users why they gave low ratings to datasets obtained from hierarchical terms (see Figure 3.11). 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 explanations why expanded terms appear in the result set. One user was sensitized for thoroughly described metadata and put emphasis on qualitative metadata in search. Two scholars would like to have the opportunity to query for metadata and primary

Number	Question	Answers
Q1	Overall from your perspective, which result set returned better results, the right one (Semantic Search) or the left one (keyword-based search)? Why?	left(0), right (3), undecided (tendency for the right result set) (3)
Q2	What did you dislike on the other one?	less results (2), only marine results (1), "depends on the question if sub-classes are relevant for the query or not" (1)
Q3	Are the following extended search terms appropriate for your search query?	analyzed in the previous section
Q4	Which data portals do you normally use? Why?	no further data portals used in daily research work (2), GBIF (1), Bexis (1), GeoportalBayern (1), PANGAEA (1)(extensive metadata), OBIS (1)+ ENA (1) (specific data, consistent structure, good geographic overview), EMODNet (1) (specific data, consistent structure, good geographic overview + compiled data products), SeaDataNet(1) (national oceanographic cruises, metadata available), WORMS (1) (to check metadata), MarineRegions (1) (to get controlled vocabs for geographic places), OLSVis (1) (to check search ontology terms)
Q5	Are there any kinds of queries (from your research background) where no data portal could find the right answer?	no answer (4), connection between metadata and data (e.g., return datasets with measurements from the first 20cm of sea floor) (1), "no data portal has so far been able to offer enough relevant data aboubt any question, and often the right answer is hidden in a huge amount of not relevnat answers. Specific portals are usually more satisfactory. So queries that combine/include taxonomy, geography, rate measurement" (1)
Q6	Are there any further research questions/ search questions you would like to query?	sulfate reduction rate, Were crystalline stones drilled?, What rock samples are there from the Palocene?
Q7	What do you expect from a semantic search?	Surprising results for new ideas (1), question-answering-system (1), group search terms (1), e.g., 'Artic Ocean' (Ocean), all trees (forest), cold seep (mud volcano)

Table 3.8: Interview questions, the number in brackets denotes how often this answer appeared.

data in search. Moreover, the interviews gave interesting insights what a semantic dataset search could be. The answers ranged from grouping data into meaningful categories, question answering systems with user-system interaction to "surprising results".

Discussion Our comparative study of a keyword search and a semantic search reveal that search queries expanded on synonyms and subclass labels resulted in a larger number of relevant results and also returned datasets that are highly relevant. Expansions on other relations require explanations, in particular on how the added terms are related to the original query. However, we conducted the evaluation with only a small numbers of users. Thus, more studies are needed to explore the suitability of semantic technologies in dataset search.

3.4 Summary

This chapter addressed H_1 and aimed to explore the current state of dataset search systems with respect to the retrieval approach. In addition, we investigated the suitability of semantic techniques to improve the retrieval of relevant datasets.

The results of the GFBio search evaluation reveal that the relevance of returned datasets was only moderate. That confirms the users' impression described in the problem statement 1.3) that they have difficulties in finding relevant data. Hence, the utilized,

conventional retrieval algorithms in this inspected dataset search support scholars only to some extent and do not return a large portion of relevant results.

In order get an idea whether semantic techniques may help to find more relevant datasets, we conducted a second user study in which researchers compared the results of a classical keyword-search with a semantic search based on query expansion. The outcome reveals that adding synonyms and subclass labels led to a larger number of relevant results. It is also an effective way of finding datasets that are highly relevant. Utilizing other semantic relations would need additional information on the returned datasets and their relation to the original query.

The qualitative evaluation results collected in GFBio's user survey and the post-interviews in the second study also reveal that expanding terms based on ontological criteria are not enough to support scholars effectively. Semantic techniques can support dataset search, but have to be tailored to the search purpose. As the evaluations show, scholarly search interests are quite diverse and therefore, further research is needed to better understand what scholars in biodiversity research are interested in search. In order to get answers on these questions, we analyzed search questions more deeply described in Chapter 4.

Chapter 4

Search Interests and Metadata in Biodiversity Research

"If you want a wise answer, ask a reasonable question."

- attributed to Johann Wolfgang von Goethe, *German poet*

"Metadata liberates us, liberates knowledge."

- David Weinberger in a lecture in 2008, *Philosopher*

Question corpora are common sources to determine user interests for a particular domain. Questions formulated in full natural language allow to capture the context of use and to obtain a deeper understanding of users' information needs. This detailed analysis of the target group and their research aims supports the formulation of requirements for an improved search and ensures that scholarly search interests are properly considered.

Besides questions, metadata are the other important data source in search. As Weinberger summed it up - "Metadata liberates knowledge" [Weinberger, 2008]. Most data repositories utilize metadata in their search indexes, because primary data are too heterogeneous in their formats to be integrated into the search index. Metadata occur in a variety of metadata schemas. There are general metadata standards as well as domain specific standards developed for a particular research domain. So far it is unclear what metadata standards are utilized in common data repositories for the Life Sciences. It is also not known whether scholarly information needs in biodiversity research are represented in metadata fields. Information that is not present in metadata can hardly be found. Therefore, this chapter addresses H_2 and aims to analyze whether search interests in the biodiversity domain are reflected in metadata. We asked scholars working in the field of biodiversity research to provide questions that are specific for their research. We analyzed theses questions and identified search topics that represent scholarly information needs in this domain. In an online survey, nine domain experts evaluated the proposed topics. The

results of this question corpus study are presented in Section 4.2. In a second study, we compared the identified search interests with existing metadata standards in the Life Sciences to determine whether existing metadata schemas cover scholarly information needs (Section 4.3). In a third analysis, we selected five large data repositories being relevant for biodiversity research, and we explored whether all publicly available metadata in these repositories contain relevant metadata fields identified in the previous study (Section 4.4).

The results of this chapter were published in a journal paper [Löffler et al., 2021]. In the following, we present a summarized version with a few extensions in related work (Section 4.1).

4.1 Related Work

In order to capture users' interests in search, query logs, user surveys or question corpora are important sources. In the following, we explore related work along these possible sources (Subsection 4.1.1). In addition, we provide an overview on approaches exploring the quality of metadata and data repositories (Subsection 4.1.2).

4.1.1 User Interests in Search

Query logs: Search logs provide quantitative data on user interests. nFor dataset search, oly a few studies are available. [Kacprzak et al., 2018] inspected query logs of different open data portals in the United Kingdom, Canada and Australia and determined main query topics. The result reveal that around one third of all search queries were related to Economy and Society. The authors also analyzed requests by users sent via a form on the website. Here, geospatial (77.5%) and temporal (44%) information was the most frequent mentioned topic. More than 40% of the users sent this explicit request to the data portal, because they could not find relevant data with the search application [Kacprzak et al., 2018]. Another recent study analyzed 236441 sessions from the European Data Portal (EDP)[1] over a period of one year [Ibáñez and Simperl, 2022]. The aim was to explore behavioural patterns and user profiles from internal (start on/search within the EDP portal) and external search (search via external search engines and link to result pages of the EDP portal). It turned out that the most important facets in the internal search were *countries* (the country in which the publication authority is located) and *categories* (high-level domains such as agriculture or energy). For external search sessions, the *tag* facet was the most used facet. In the Life Sciences, [Islamaj Dogan et al., 2009] inspected one month of log data with more than 58 million user queries from PubMed[2]. They randomly selected 10,000 queries for a semantic analysis. The most frequent cat-

[1] EDP, `https://data.europa.eu/en`

[2] PubMed, `https://www.ncbi.nlm.nih.gov/pubmed/`

egory over all questions was *author name* (36%) followed by *Disorder* (20%) comprising diseases, abnormalities, dysfunctions etc., and *gene/ protein* (19 %). Further main topics were abbreviations (mostly from genes/ proteins) and chemicals/ drugs.

User surveys and interviews: Another opportunity to analyze user needs are surveys and interviews. In biodiversity research, the GBIF community carried out a large user study on user needs in 2009 [Faith et al., 2013, Ariño et al., 2013]. The survey aimed to explore data gaps in the current data landscape and to obtain insights on user needs with respect to primary data. Based on the feedback from more than 700 participants, it turned out that scholare were mainly interested in species diversity, taxonomy, and phenology. For their research, scholars stated to utilize "taxon names, occurrence data and descriptive data about the species" [Ariño et al., 2013]. In the biomedical domain, Datamed [Dixit et al., 2017] interviewed 13 scholars with various backgrounds in Clinical Sciences, Biomedical Informatics and Public Health to obtain insights on various aspects of data discovery [Dixit et al., 2017]. The scientists report on their difficulties in assessing whether relevant data is available, the lack of proper metadata description and the variability of data formats which required additional effort to transform the data into formats they needed for their processes. Thus, the authors of the study conclude that "researchers would benefit from a centralized source and complete metadata documentation for finding and assessing potentially relevant datasets. Additionally, they need clear protocols to download the data, the ability to download in multiple formats, and a means to visually explore datasets" [Dixit et al., 2017].

Question corpora: Besides query logs, user surveys and interviews, question corpora are another source to determine user needs in search. In the Life Sciences, question corpora were mainly developed for text mining purposes in the medical and biomedical domain. One of the largest corpora in medicine is the Consumer Health Corpus [Kilicoglu et al., 2018], a collection of email requests (67%) received by the U.S. National Library of Medicine (NLM) customer service and search query logs (33%) of MedlinePlus, a consumer-oriented NLM website for health information. The final corpus consisted of 2614 questions and was integrated into the Medical Question Answering Task at TREC 2017 LiveQA [Abacha et al., 2017]. Six trained domain experts were involved in the annotation tasks to label information manually. The experts had to indicate named entities, e.g., problem, anatomy or measurement, and they labeled question topics such as the cause of a disease or complications (long term effects of a disease).

A common question corpus in biomedicine is the Genomics Track at TREC conferences [Hersh and Voorhees, 2009]. The topics of the retrieval tasks are formulated as natural language questions and contain pre-labeled main categories, e.g., *What [GENES] are involved in insect segmentation?*. A further large question corpus in biomedicine is the

question corpus created for the BioASQ challenge [Nentidis et al., 2017], an annual challenge for researchers working on text mining, machine learning, information retrieval, and question answering. The provided question corpus was created and annotated by a team of ten experts, selected with the goal to cover different ages and complementary expertise in the fields of medicine, biology, and bioinformatics [Polychronopoulos et al., 2013]. The experts had to provide 50 questions in English representing "real-life information needs" [Polychronopoulos et al., 2013]. However, the question providers had to follow a specific question scheme to match technical needs for question answering. For instance, the questions had to reflect a certain question type such as Yes/No or factoid questions. This restriction influenced query formulation and likely led to a bias in the question corpus.

Another question corpus in the biomedical domain is the benchmark developed for the bioCADDIE Dataset Retrieval Challenge [Cohen et al., 2017b]. This benchmark was explicitly created for the retrieval of datasets based on metadata and includes 137 questions, 794,992 datasets gathered from different data portals in XML structure, and relevance judgments for 15 questions. The question providers received instructions on how to formulate the questions, e.g., information on desired entity tyes such as data type, diseases, biological processes and species.

4.1.2 Studies on the Quality of Metadata and Data Repositories

Since the introduction of the FAIR principles [Wilkinson et al., 2016] in 2015, metadata quality and quality checks get stronger attentions at publishers, data repositories and research communities. The four principles (Findability, Accessibility Interoperability and Reuse) aim to provide guidelines for data publishers and scholars to produce and release only digital objects that enable machine processing as scholars need computational support for various tasks in daily research practice. The FAIRshake toolkit[3] [Clarke et al., 2019] was launched to allow research communitities to define criteria, to develop new or extend existing standards and to evaluate digital resources on their level of FAIRness. In particular for the biomedical domain, multiple digital ressources such as datasets, projects, repositories and workflows were assessed on their adherence of the FAIR principles. Another tool for checking the FAIRness of metadata is the system provided by [DataONE, 2022][4]. It determines whether specific metadata fields are available in metadata such as title, abstract or publication date.

Besides the more and more upcoming FAIRcheck tools, surveys among various data repositories also give insights on metadata usage and search practices. A survey by the Research Data Alliance (RDA) [Khalsa et al., 2018] reveals that almost two-third

[3]FAIRshake, `https://fairshake.cloud/`
[4]DataONE FAIRcheck, `https://www.dataone.org/fair/`

of of 98 participating data repositories utilize full metadata in their search indexes, and around 30% integrate data dictionaries or data variables. Another survey among 32 Canadian and international online data platforms aimed to analyze to what extent repositories support data storage, data transfer curation activities and other data sharing features [Austin et al., 2016]. For comparing the features, the authors propose a developed checklist. The results reveal that data portals offer different kinds of services and that a heterogenity of features exists. The authors also criticize an inconsistent compliance with standards and a lack of data repositories being certified.

Another study explored the search interfaces of six Canadian data repositories in the health domain. The authors inspected the user interfaces manually based on proposed criteria to determine the quality of the offered filter options, the metadata fields presented in the result list and the opportunity to export the metadata in various formats [Thornton and Shiri, 2021]. This study focuses solely on general metadata standards and found out that most metadata fields of the general standards were present in the analyzed data repositories. They suggest improvements such as the integration of an advanced search, the support for health specific search terms and the provision of links to related publications.

To the best of our knowledge, there is no quantitative study that analyzes scholarly information needs in conjunction with metadata schemas and metadata used at data repositories. Second, there is neither a public log analysis nor a question corpus available for biodiversity research. To enhance dataset retrieval systems, scholarly information needs are essential. Therefore, Section 4.2 introduces our corpus study with questions from biodiversity research.

4.2 Question Corpus Study

The question corpus study addresses hypothesis H_2 and aims to gather search and research questions in biodiversity research to identify important information needs (search topics, categories) in this domain. In the following Subsection 4.2.1, we describe the methodology along the paragraphs: question collection, question types, category definition and annotation process. A more detailed description is available in [Löffler et al., 2021]. The results are summarized in 4.2.2.

4.2.1 Methodology

Following the question collection and annotation procedure in information retrieval and question answering [Krallinger et al., 2015, Kilicoglu et al., 2018, Nentidis et al., 2017], we collected search and research questions in different research projects in 2015 and 2016. In several rounds, we labeled and classified the noun entities according to annotation

Question example
Does agriculture influence the groundwater?
List all datasets with organisms in water samples!
What influence do neonicotinoids have on pollinators?
How do geology and soil geochemistry affect the bacterial and archaeal nitrogen-cycling communities in soil?

Table 4.1: Example questions in the corpus.

guidelines. In order to verify the identified categories, we conducted an online survey with nine domain experts and asked them to classify important terms and phrases for all questions into provided categories.

Question collection: We asked 73 scholars from three biodiversity research related projects in Germany (CRC AquaDiva, GFBio and iDiv, see also Chapter 2.4) to provide up to five natural language questions that are specific for their research. Full natural language questions allowed us to obtain keywords in their search context. The scholars had very diverse backgrounds, such as biology (e.g., ecology, bio-geochemistry, zoology and botany) and related fields (e.g., hydro-geology). In total, we received 184 questions, a number that is comparable to related question corpora in information retrieval (e.g., bioCADDIE [Cohen et al., 2017b]). Table 4.1 provides some example questions. As not all questions were comprehensible, we omitted questions with missing verbs or misleading grammatical structures. Furthermore, we omitted occurring abbrevations, as they are fuzzy and can have several meanings. Later in the evaluation with domain experts, we let the experts decide whether to look these abbrevations up or to leave the question out. The final question corpus consists of 169 questions and is publicly available [5].

Category definition: Classifying biological terms into semantic categories is a challenging task, because a multitude of terms are fuzzy and specific to a domain. Therefore, we defined suitable categories for biodiversity research in an iterative process. In a first step, two authors of this study [Löffler et al., 2021] inspected the clean question corpus manually. Each author analyzed 50% of the corpus on her own and classified noun entities into broad categories. In a second round, both annotators discussed the identified categories in several sessions. The result are 13 categories presented in Table 4.2.

Annotation process: After this manual inspection, we evaluated the defined categories with domain experts in an online survey. Per question, the experts had to classify main artifacts (important nouns with adjectives) for each question. As we wanted to obtain an unbiased feedback in this classification task, the experts did not receive a training but only short explanations per category. They also did not label the important terms in the

[5] GitHub, `https://github.com/fusion-jena/QuestionsMetadataBiodiv`

Category	Definition	Examples
ORGANISM	all individual life forms including plants, fungi, bacteria, animals and microorganisms	foraminifera, honeybee, Poales
ENVIRONMENT	local and global environments species live in (habitats, ecosystems)	below 4000 m, ground water, city, grassland
QUALITY & PHENOTYPE	characteristics (traits, phenotypes) that can be observed, measured or computed	length, growth rate, reproduction rate
PROCESS	biological, chemical and physical processes that reoccur and transform materials or organisms due to chemical reactions or other influencing factors	nitrogen cycling, climate change, ocean acidification
EVENT	processes that appear only once at a specific time and location, such as environmental disasters	Deepwater Horizon oil spill, Tree of the Year 2016.
MATERIALS & SUBSTANCES	chemical substances and materials	oxygen, CO2, sediment, rock
ANATOMY	anatomical structure of organisms, e.g., body or plant parts, cells and genes	root, leaf, TP53
METHOD	methods and experiments leading to a certain research result	lidar measurements, observation, remote sensing, genome sequencing
DATA TYPE	results of research methods	lidar data, genome data, abundance data
LOCATION	geospatial information (without coordinates)	Atlantic Ocan, Bavaria, Lake Ontario
TIME	temporal information such as geological eras, date and time	current, Triassic
PERSON & ORGANIZATION	creators and authors of research data, research projects	
HUMAN INTERVENTION	human intervention on landscape and environment	land use, farming, fishery

Table 4.2: Identified categories after manual inspection and discussion [Löffler et al., 2021]

questions, but they had to group the provided artifacts into one of the provided categories. We added the categories OTHER to give the annotators the opportunity to provide an own category when none of the listed categories fit. We also provided the category NONE to offer the opportunity to leave an artifact out in case the annotators did not know the term or in case the term was too fuzzy.

All nine annotators (eight postdocs, one project manager) had various backgrounds in the Life Sciences and Environmental Sciences. We provided a link to an online survey to conduct the assessment on their own. Figure 4.1 presents a screenshot of the online survey. Multi-labeling was not permitted, each artifact had to be grouped into one category. If none of the provided categories were suitable, they were advised to choose 'OTHER' and to provide an alternative label. For unknown tasks, they could decide to either look the term up or to omit it. In case they did not label an artifact, the category NONE was applied automatically. We also told the annotators to spend not too much time per task and to classify the artifacts based on their expertise, as we aimed to receive an intuitive classification feedback.

4.2.2 Results

We analyzed the survey along two dimensions: the frequency of each category measuring the importance of the categories and whether there are categories missing and second, the comprehensibility and its appropriateness for biodiversity research. Besides theses responses of the experts, we also inspected the questions with respect to the expected answer and determined the existing question types.

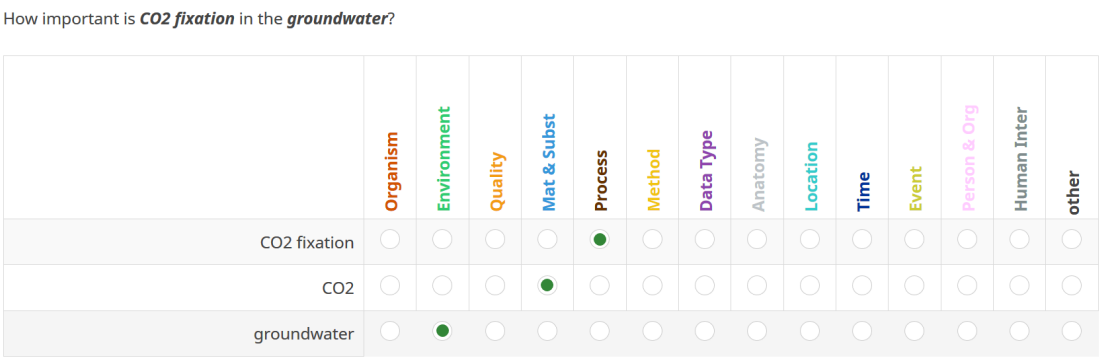


Figure 4.1: Excerpt of the classification task with experts. The annotators were only permitted to select one category per artifact [Löffler et al., 2021].

Predicative questions	**Yes/No questions**	**List questions**	**Definition questions**	**Comparative questions**
28	49	62	2	3
Explanation questions	**Association questions**	**Process questions**	**Opinion questions**	**Not Classifiable**
1	23	2	0	9

Table 4.3: Question type classification

Question types: According to [Unger et al., 2014] search questions can be classified into nine main types: Factoid questions including predicative questions, yes/no questions and list questions aim to return facts, whereas the other types point out to definitions, comparisons, associations, explanations, processes and opinions. Table 4.3 presents the analysis on the collected 169 search and research questions in the biodiversity research domain. The question type classification were performed by the same two authors that also conducted the first two classification rounds.

The dominant question type are factoid questions. This shows that scholars are mainly interested in specific information needs. The questions also reveal that researchers mainly expect search results to provide facts or information that assist in answering their information needs. The second dominant type are association questions. This result is justified, as biodiversity research is a domain exploring the relationships among species and their ecosystems. Finding associations is one major research aim. All other types occur only in less numbers or are not present (e.g., opinion questions). Concerning granularity, the question corpus shows that questions in this scholarly domain can be either of a broader scope, such as a question about ‘tree diversity’, or can be more specific such as a question about ‘benthic oxygen update rate’.

Relevance for biodiversity research: The frequency of each category is a measure for the relevance or importance of a category. Figure 4.2 presents the frequency of all categories studied. All categories were used, which is a strong indicator that all categories are relevant for biodiversity research. Seven categories (ENVIRONMENT, ORGANISM,

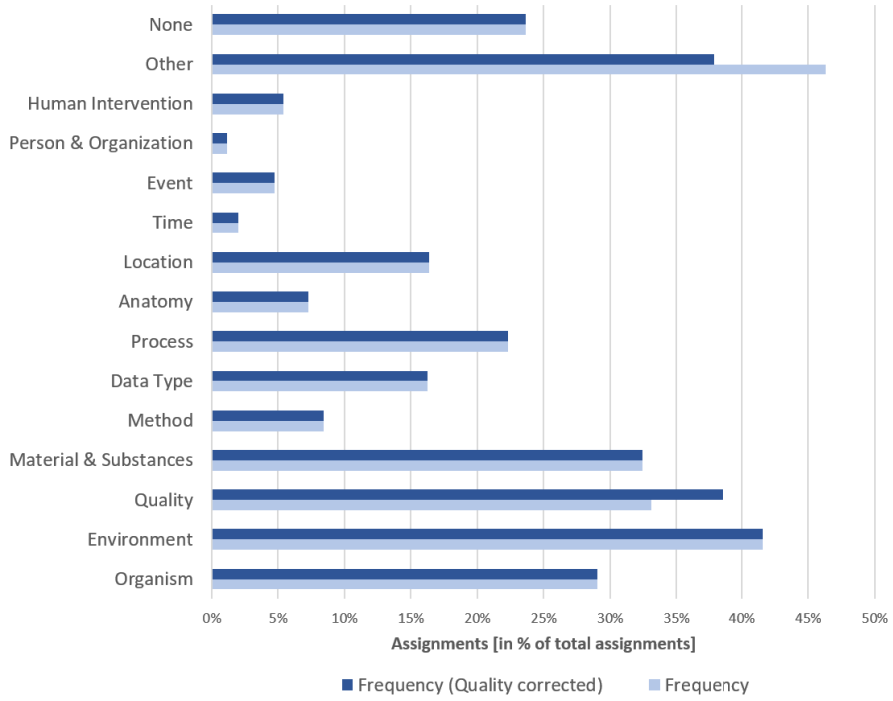


Figure 4.2: The frequency of the categories with and without QUALITY correction [Löffler et al., 2021].

MATERIAL & SUBSTANCES, QUALITY, PROCESS, LOCATION and DATA TYPE) were assigned to more than 15% of all labeled artifacts. Information that can be classified into these categories is highly relevant for biodiversity research.

However, the results also point out that several artifacts could not be classified into any of the given categories. For 46% of the artifacts, at least one expert selected the category OTHER, for 24% of the artifacts at least 24% of the scholars labeled the phrase or term with the category OTHER. This strongly indicates that categories are missing. However, at least the suggested categories represent a high domain coverage.

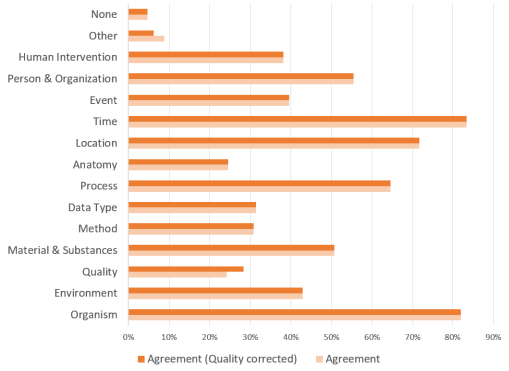
We also examined why the category OTHER was selected for so many artifacts. We inspected these cases manually and took a closer look on the comments provided. (1) Some artifacts were completly unknown to the annotators, e.g., *shed precipitation*. However, they had the impression that this phrase or term is relevant for the domain. So, they did not omit the artifact and selected OTHER. (2) A multitude of artifacts refering to data parameters, such as *soil moisture* or *oxygen uptake rate*, were labeled with the category OTHER. They were supposed to be labeled with the category QUALITY, but obviously the labeling of the category QUALITY was misunderstanding. In the comment fields, the annotators proposed alternative categories, e.g. “variable” or “parameter”. We corrected the results for these cases and added them to the QUALITY category. Thus, the results for the category OTHER decreased to 37% for one expert and 13% for two experts. (3) Another reason for the high number of ratings of the category OTHER are questions with

a broader scope. Generic terms such as *diversity*, *pattern* or *distribution* need further context to assign an appropriate category. The classification of these terms need further discussions in the research community.

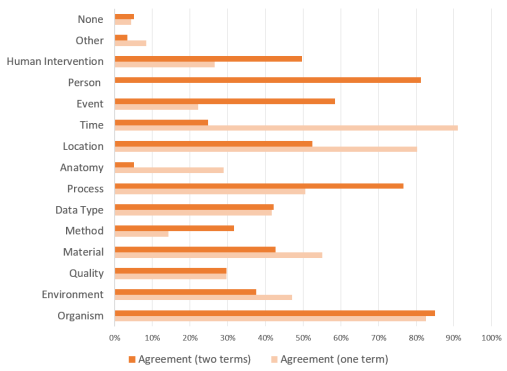
Consensus of the categories: In addition to the relevance, we explored whether the naming of the given categories is well-defined and comprehensible to domain exerts. In order to assess comprehensibility, we computed the inter-rater agreement and inter-rater reliability using Fleiss' Kappa (κ statistics) [Fleiss, 1971] and Gwet's AC [Gwet, 2008]. These measures determine how much homogeneity exists among scholars' judgments. The inter-rater reliability determines the observed agreement among raters "and then adjusts the result by determining how much agreement could be expected from random chance" [Quarfoot and Levine, 2016]. κ values occur in a range of -1 and $+1$. Values less than 0 point to poorer than chance agreement, values greater than 0 are an indicator for better than chance agreement. [Landis and Koch, 1977] recommend to consider values below 0.4 as a fair agreement beyond chance. Results between 0.4 and 0.6 point to a moderate agreement and values between 0.6 and 0.8 are an indicator for a substantial agreement. Values higher than 0.80 denote a perfect agreement. In case the distribution of the raters' scores is unbalanaced, κ statistics result in negative values, even if the observed agreement is high [Quarfoot and Levine, 2016]. To counteract this paradox, [Gwet, 2008] introduced a more robust statistic, the Gwet's AC, considering the response categories in the agreement by chance. Gwet's AC values can range from 0 to 1.

The experts' agreement over all categories point to a moderate result with a Fleiss' Kappa value of 0.48 and a Gwet's AC value of 0.51. We observe a slight increase including the QUALITY correction (0.49 for Fleiss' Kappa and 0.52 for Gwet's AC). Figure 4.3 shows the Fleiss' Kappa for the categories with QUALITY correction. The annotators were able to classify terms into the categories TIME and ORGANIZATION without any difficulties. Hence, these categories received an excellent agreement. An intermediate to good agreement is observed for the categories PERSON & ORGANIZATION, LOCATION, PROCESS, MATERIALS & SUBSTANCES and ENVIRONMENT. A fair agreement exists for the categories EVENT, HUMAN INTERVENTION, ANATOMY, DATA TYPE, METHOD and QUALITY.

We assume that the classification of some artifacts was indeed difficult, in particular for longer phrases with more than two terms. For these cases, the annotators got the instruction not to select a category. However, for 5% of the phrases two annotators did not provide a category and for 2% of the artifacts three or more annotators did not provide a classification. This correlates with our observations for the category QUALITY. For the remaining categories (EVENT, HUMAN INTERVENTION, ANATOMY, DATA TYPE, METHOD), there is no such evidence. Here, further discussions in the research community are needed.



(a) Fleiss' Kappa values per category with and without QUALITY correction.



(b) Fleiss' Kappa values per category for artifacts with one and two terms (with QUALITY correction).

Figure 4.3: Fleiss' Kappa values for the individual information categories [Löffler et al., 2021].

	Overall	**One Term**	**Two Terms**	**>= Three Terms**
$Fleiss' Kappa$	0.49	0.54	0.50	0.33
$Gwet' sAC$	0.52	0.57	0.53	0.37

Table 4.4: Annotator's agreement with QUALITY correction overall and for one term, two terms, three terms and more per artifact [Löffler et al., 2021].

Comparison of short and long artifacts Table 4.4 presents the results for artifacts with one term, two terms, three and more terms (with quality correction). Further insights can be obtained from Figure 4.3b. One-term artifacts resulted in excellent agreements (> 0.8) for the categories ORGANISM, TIME and LOCATION and a moderate agreement for ENVIRONMENT, MATERIAL, PROCESS and DATA TYPE. For the category PERSON, the results show a negative value (poor agreement) for one-term artifacts but a high agreement for two terms. This result is justified, as person names usually consist of two parts, and the corpus also does not provide person names with only one term. We observed an excellent agreement for two terms for the categories ORGANISM (species' names also consists of two terms) and PROCESS (0.76). The latter is a strong evidence that biological and chemical processes are also mainly defined by two terms, e.g., climate change, nitrogen cycling or sulfate reduction. The same applies for the category EVENT (e.g., oil spill) and HUMAN INTERVENTION (e.g., land use).

Discussion Seven out of 13 information categories received a moderate or high agreement (> 0.4). In addition, five out of these seven were frequently mentioned in the questions ($> 15\%$), namely ENVIRONMENT (e.g., habitats, climate zone, soil, weather conditions), MATERIAL (e.g., chemicals, geological information), ORGANISM (species, taxonomy), PROCESS (biological and chemical processes) and LOCATION (coordinates, altitude, geographic description) (Figure 4.4). We conclude that these classes are

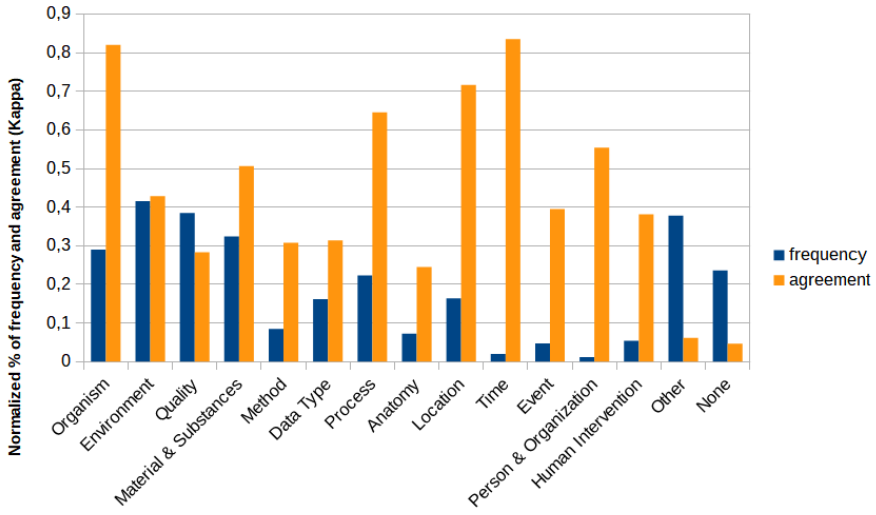


Figure 4.4: Frequency of category mentions and inter-rater agreement with QUALITY correction [Löffler et al., 2021].

important search interests for biodiversity research. For 45% out of 592 annotations, at least one scholar did not use any of the given categories but selected OTHER. This strongly indicates that either further classes are missing or the artifacts are too fuzzy to classify.

Our results show that there is room for improvement. One reason for fair or bad agreements for the categories QUALITY (data parameters measured) and DATA TYPE (nature or genre of the primary data) are incomprehensible category names. Here, further discussions in the research community are needed. In particular it should be discussed whether category names need to be changed or whether categories need to be merged.

We aimed to obtain an unbiased feedback of scholars working in the field of biodiversity research when looking for research data. Therefore, no training took place and the annotators' agreement was not excellent. Another reason for the moderate agreement values are the diverse research backgrounds of the scholars. Each research field uses its own specialized language with inconsistent and imprecise namings [Thessen et al., 2012, Ananiadou et al., 2004]. Therefore, annotating and classifying biological entities is difficult and remains a challenge.

4.3 Metadata Standards in the Life Sciences

This section introduces existing metadata standards in the Life Sciences and explores whether their elements correspond to the identified information categories of the previous section (Section 4.2).

Standard Name	Domain	Semantic Format	Examples
Dublin Core (205) [DCMI Usage Board, 2020]	general research data	Yes (RDF)	Pangaea, Dryad, GBIF, Zenodo, Figshare
DataCite (111) [DataCite Working Group., 2017]	general research data	No	Pangaea, Zenodo, Figshare, Radar
ISO19115 (47) [ISO, 2019]	geospatial data	No	Pangaea, NSF Arctic Data Center, coastMap
CSDGM (42) [FGDC, 1998]	geographic information	No	Dataverse, NSF Arctic Data Center
Darwin Core (28) [Darwin Core Interest Group, 2021]	biodiversity data	Yes (RDF)	GFBio, GBIF, VerNET, Atlas of Living Australia, WORMS
EML (26) [Jones et al., 2019]	ecological data	No	GBIF, GFBio, SNSB, Senckenberg, WORMS, NSF Arctic Data Center
RDF Data Cube (20) [W3C, 2014]	statistical data	Yes	Dryad (only RDF with Dublin Core)
ISA-Model (13) [ISA Community, 2016]	biological experiments	Yes	Data Inra, GigaDB
ABCD (11) [TDWG, 2005]	biological collection data	Yes (ABCD 3.0 in RDF)	GBIF, BioCase Network
OAI ORE (11) [Open Archives Initiative, 2014]	general research data	Yes (RDF)	Environmental Data Initiative Repository, Dryad
DCAT (9) [W3C, 2020]	data catalogs, data sets	Yes	Data.gov.au, European Data Portal
DIF (9) [Olsen and Stevens, 2010]	geospatial metadata	No	Pangaea, Australian Antarctic Data Center, Marine Environmental Data Section
CF (7) [Eaton et al., 2021]	climate and forecast	No	WORMS, NSF Arctic Data Center, coastMap

Table 4.5: Metadata standards in the (Life) Sciences obtained from re3data [re3data, 2018] and RDA Metadata Standards Catalog [Research Data Alliance, 2020]. The number in brackets denotes the number of repositories supporting the standard (provided in re3data) [Löffler et al., 2021].

4.3.1 Methodology

Metadata provide additional information about scientific primary data such as experiments, tabular data, images and acoustic files. This extra information is mostly stored in structured formats, e.g., XML or JSON, which allows a further machine processing. Metadata possess a *schema* stored in an XSD file outlining which elements and attributes exist and which of them are mandatory and/or repeatable. In order to become a metadata standard, a schema needs to be formally adopted by a standards' organization such as the International Organization for Standardization [6] or by the research community.

There are a variety of metadata standards available for the Life Sciences. Table 4.5 presents a list of 12 metadata standards obtained from re3data [re3data, 2018] and RDA Metadata Standards Catalog (v 2.0) [Research Data Alliance, 2020] in 2019. In re3data, we filtered for "Life Sciences" and received a list of 30 standards. We merged it with the RDA Metadata Standards Catalog (version 2) list and cleaned the result along the following criteria: We did not use the categories *Other* and *Repository-Developed Metadata Schema*. We omitted deprecated standards or standards with outdated information on the website, and we also did not consider standards from domains not related to the Life Sci-

[6] ISO, `https://www.iso.org`

ences. We selected all other standards that were used in at least five repositories. Further information on the selection process is available in our GitHub repository[7]. We compared the standards along the focused *domain* and its support for semantic web formats, e.g., RDF or OWL. According to the FAIR principles [Wilkinson et al., 2016], community standards, semantic formats, and ontologies ensure interoperability and data reuse. The last column contains examples of data archives supporting the standard.

The most frequently utilized metadata standard is the general metadata standard *Dublin Core*. It consists of 15 main metadata fields such as `contributor, coverage, description` and `identifier`. The second most frequently used standard is *Data Cite*, another general metadata standard with a few more metadata fields than *Dublin Core*. For instance, for the description of collection and publication date *Data Cite* offers separate elements whereas *Dublin Core* only provides one field. Examples for domain-specific standards are *ISO19115* (geospatial data) or *EML* (ecological data). The *RDF Data Cube Vocabulary* is not utilized at all. We assume, the abbreviation *RDF DC* led to some confusion and misunderstanding as DC is often used as abbrevation for Dublin Core and not 'Data Cube'.

The standards strongly differ in their granularity. General standards such as *Dublin Core* or *DataCite* provide only general fields such as `title, description` and `type`. In contrast, domain-specific standards, e.g., *EML* or *ABCD*, offer metadata fields for scientific names, methods and data parameters. However, all standards provide elements that refer to the W-questions: Who? What? Where? When? and Why? and offer elements for author or contact person, collection or publication date and geographic location. Semantic formats are also increasingly supported. *ABCD, Darwin Core, DCAT and OAI-ORE* are either fully semantic standards or are currently being transformed to semantic formats. *ISA Model* supports the integration of controlled vocabularies, and *Dublin Core* provides a additional RDF format called qualified *qualified Dublin Core*[8].

4.3.2 Results

We related the identified information categories to the available fields of the metadata schemas. The results are presented in Table 4.6. More detailed explanations and the exact matching of metadata elements and categories are availble in our GitHub repository.

There is no metadata standard or schema that reflects all information categories. As we are aware that our information model is not complete, we also did not expect to find a standard covering all interests. Apart from only one category (HUMAN INTERVENTION) all scholarly interests are covered by different metadata schemes. TIME and

[7]Metadata standard selection, `https://github.com/fusion-jena/QuestionsMetadataBiodiv/tree/master/metadataStandards`

[8]Qualified Dublin Core, `https://www.dublincore.org/specifications/dublin-core/dcq-rdf-xml/`

	Environment	Quality*	Material	Organism	Process	Location	Data Type*	Method*	Anatomy*	Human Intervention*	Event*	Time	Person
DublinCore													
DataCite													
*ISO*19115													
CSDGM													
EML													
DarwinCore													
RDFDataCube													
ISA – Model													
DIF													
CF													
ABCD													
DCAT													
OAI – ORE													

Table key

Not provided Unspecific (general element) Available (one or more elements)

Table 4.6: Overview of the identified metadata standards in the Life Sciences and corresponding information categories sorted by their occurrence frequency. The asterix marks the categories with a fair agreement (< 0.4) [Löffler et al., 2021] .

PERSON are information categories that are supported by almost all analyzed standards and frameworks. DATA TYPE and LOCATION are interests that are covered by most schemes. However, we observe a strong difference between general and domain specific standards. While general standards provide general fields, e.g., `subject`, to describe concrete search interests, domain-specific standards offer concrete metadata elements for the description of further important biological entities. For instance, *ISA Model*, *Darwin Core* and *ABCD* offer metadata elements to provide information about collected or analyzed species. In addition, *EML* supports the description of environmental information (`studyAreaDescription`) and research methods used (`methods`). The *ISA Model* is the onliest analyzed framework that provides metadata fields for the description of materials and biological and chemical processes.

The open question we address in the next section is to study whether these general or domain-specific standards are applied in data repositores.

4.4 Study on Metadata in Data Repositories

In a third analysis, we explored the usage of metadata standards in data repositories. We selected five data archives from *Nature's* list of recommended data repositories

[Nature, 2018] and inspected the metadata standards used, the utilized metadata elements in these standards, and we took a closer look on the content of descriptive metadata fields, e.g., title and description. In the following subsections, we describe the methodology and the results.

4.4.1 Methodology

We selected five data repositories, namely *Figshare*, *Dryad*, *Zenodo* (all three are general data repositories), *GBIF* (taxonomic data) and *PANGAEA* (environmental data) and parsed all available metadata from their *OAI-PMH* [9] interfaces in May 2019. *GBIF* already provides its primary data - species occurrence records - in *Darwin Core* metadata schema [Gaiji et al., 2013]. These primary data files are organized and grouped in 'datasets'. We assumed to be able to retrieve both, the metadata of these 'datasets' and the individual occurrence records. However, in the OAI-PMH interface, only the download of the metadata of the 'datasets' was possible. Hence, we could not analyze the occurrence records.

Table 4.7 presents the available standards in the inspected data repositories. Besides the already introduced metadata standards, a few more are listed in the table. "*OAI-DC* is an abbreviation for *Dublin Core*, a mandatory standard in OAI-PMH interfaces. *QCD* means *qualified DublinCore* and denotes an extended *Dublin Core* extending or refining the 15 core elements" [Löffler et al., 2021]. *Pan-MD* is a metadata schema introduced by *PANGAEA* that extends *Dublin Core* with more fine-grained geographic information such as bounding boxes or adds information on data collection, e.g., projects, parameters and methods. We ommited all standards that are mainly used for bibliographic data (*MARC21*, *MARCXML* and *METS*) and other formats and standards that are not primarily developed to describe research data (CERIF, OAI-ORE). However, we considered the repository-developed schema *Pan-MD* to analyze how data archives extend metadata standards. We took *RDF* into account, too, although *RDF* is an encoding of metadata and not a metadata standard. A closer look into the metadata revealed that repositories besides *Dublin Core* already reuse vocabularies. For instance, *Figshare* uses the VIVO Core Ontology [Corson-Rikert et al., 2012, VIVO,] and the Bibliographic Ontology [BIBO, 2016].

We parsed each downloaded metadata file with a script to determine whether a metadata field of the respective standard is present (boolean value 1) or not present (boolean value 0). The final outcome is a csv file containing dataset IDs and metadata elements used [10].

Content analysis: Apart from different metadata schemas used, datasets can also differ in its textual descriptions. Some datasets only provide a title and some keywords, other

[9] OAI-PMH, `https://www.openarchives.org/pmh/`

[10] GitHub, `https://github.com/fusion-jena/QuestionsMetadataBiodiv`

Dryad	GBIF	PANGAEA	Zenodo	Figshare
METS	EML	DATACITE3	DATACITE	CERIF
OAI-DC	OAI-DC	DIF	DATACITE3	METS
OAI-ORE		ISO19139	DATACITE4	OAI-DATACITE
RDF		ISO19139.IODP	MARCXML	OAI-DC
		OAI-DC	MARC21	QDC
		PAN-MD	OAI-DATACITE	RDF
			OAI-DATACITE3	
			OAI-DC	

Table 4.7: Available metadata schemes in the five selected data repositories in May 2019 [Löffler et al., 2021].

datasets contain an extensive abstract with further information, e.g., on the research methodolgy and parameters measured. This additional descriptive information might contain relevant data for information seekers and might only occur in longer text fields. Hence, it is not that present and visual at one glance and additional services are required to extract important biological entities. In order to determine whether additional metadata enrichments correspond to scholarly search interests in the biodiversity domain, we explored the information content in general, descriptive metadata fields with existing Natural Language Processing (NLP) tools.

In November and December 2019, we downloaded descriptive metadata fields (`dc:title`, `dc:description` and `dc:subject` in *OAI-DC* format) from all five data repositories. In addition, we gathered the keywords of the subject fields. For the subsequent NLP analysis, we used a subset of 10,000 metadata files per repository, as NLP-processing is time-consuming and requires larger hardware resources. We utilized existing NLP pipelines of the GATE framework [Cunningham et al., 2013], such as the ANNIE pipeline [Cunningham et al., 2013] and the OrganismTagger [Naderi et al., 2011] to extract geographic locations, persons, organizations and organisms from the descriptive metadata fields. Unfortunately, at the time of this study, taggers for other important entities, e.g., data parameters or habitats did not exist.

4.4.2 Results

The total numbers of downloaded metadata files per metadata schema are presented in Table 4.8. Most repositories utilize general standards, while *PANGAEA* and *GBIF* also provide metadata in domain-specific schemas. Semantic formats are also applied. *Dryad* and *Figshare* offer metadata in RDF format. Moreover, *Figshare* provides metadata also in *Qualified Dublin Core (QDC)*, an extended *Dublin Core* with additional elements to describe relations to data sources. For the further analysis, we only used metadata files with a publication date and a valid status. For instance, in *Dryad*, numerous datasets were described with the status “Item is not available”. Therefore, we omitted these datasets.

Metadata Schema	Dryad	PANGAEA	GBIF	Zenodo	Figshare
OAI-DC	186951 (142329)	383899 (383899)	44718 (42444)	255000 (255000)	3128798 (3128798)
QDC					1718059 (1718059)
RDF	186955 (142989)				3157347 (3157347)
DATACITE				1268155 (1268155)	
DATACITE3		383906 (383906)		1268232 (1268232)	
OAI-DATACITE				1266522 (1266522)	3134958 (3134958)
OAI-DATACITE3				1268679 (1268679)	
DATACITE4				1268262 (1268262)	
EML			44718 (42444)		
DIF		383899 (383899)			
ISO19139		383899 (383899)			
ISO19139.iodp		383899 (383899)			
PAN-MD		383899 (383899)			

Table 4.8: "Total number of datasets parsed per data repository and metadata schema. The numbers in brackets denote the number of datasets used for the analysis. All datasets were harvested and parsed in May 2019" [Löffler et al., 2021].

Figure 4.5 displays the percentage of metadata fields filled for each data repository and its best matching standard. More detailed information is available in our GitHub repository. For *Dryad*, the best matching standard was Dublin Core. Nine out of fifteen metadata elements were filled for most datasets (80%), e.g., `dc:title`, `dc:description` and `dc:subject`. Contact information such as `dc:contributor` or `dc:publisher` were hardly present (20%). As the EML standard does not determine any mandatory fields, we analyzed all 129 metadata elements for the available datasets in *GBIF*. The majority of elements (89 fields) were empty. General information, such as data about author, title and description, were mostly provided (80%), but the `eml:keyword` field was utilized in one fifth of all analyzed datasets. The best matching format for *PANGAEA* is its own-developed Pan-MD format. 43 out of 124 provided fields were utilized in most metadata files (80%). Examples for used metadata elements are author, project name, geographic information with coordinates, data parameters and devices. For *Zenodo*, DataCite was the metadata standard with the best fit. All mandatory fields such as identifier, creator, title, publisher and publication year were always filled. Further metadata elements, e.g., title, rights and descriptions, were also mostly present (99%). Keywords (`subject`) were only utilized in 45% of the analyzed metadata files. *Figshare's* utilized fields of Qualified Dublic Core (12 out of 17) were always filled.

Apart from the occurrence and non-occurrence of individual metadata elements, we also analyzed the semantic matching of the information categories identified in Section 4.2 and metadata fields. For each data repository and metadata format, we generated bar charts visualizing to what percentage each metadata field was filled and to what category it is related. Table 4.9 provides a summary of all inspected data repositories analyzed and their best matching standard. More information and the individual field-to-category mapping are available in our repository as supplementary material. Temporal information (TIME) and contact details, e.g., on author and/or creator (PERSON) are present in all data repositories. Most data repositories indicates the publication date. In addi-

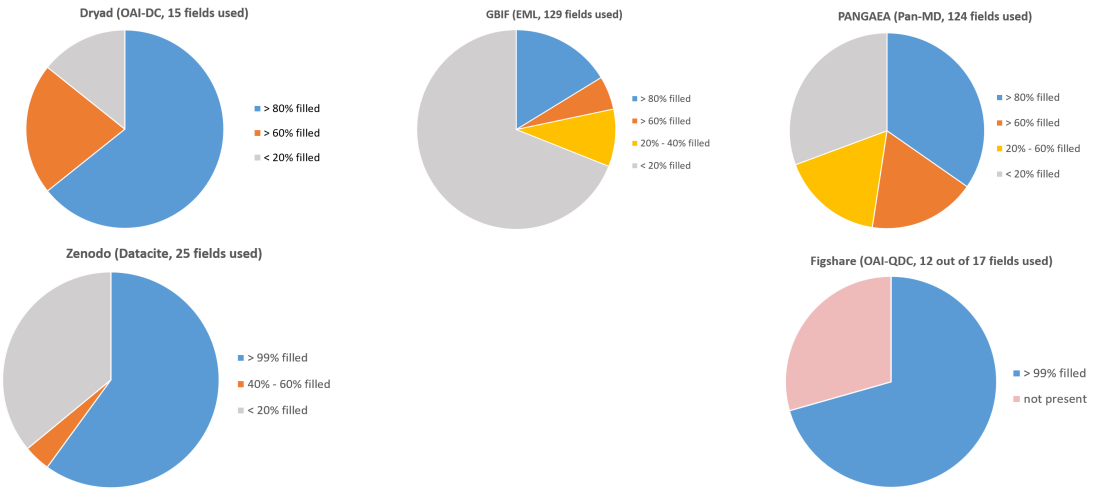


Figure 4.5: Overview of metadata elements used (best matching metadata standard) in the five data repositories analyzed [Löffler et al., 2021].

tion, *PANGAEA* also provides the collection date. Apart from *GBIF*, all data repositories provide information on data type and data formats.

Our result reveal that two data repositories partially cover the search categories with respect to the metadata standards and formats used. Multiple metadata fields of the EML metadata standard in *GBIF* correlate to several information categories, but elements that are related to information on ENVIRONMENT, ORGANISM, DATA TYPE and METHOD were rarely filled. The second repository that cover most categories is *PANGAEA* with its repository-developed standard Pan-MD. Metadata files in this format always provide data on parameters measured (QUALITY) and geographic locations. Research methods and devices utilized were also mostly present in the inspected metadata files of *PANGAEA*. For biological, chemical and physical processes as well as materials and chemical substances none of the inspected data repositories provide explicit metadata fields that are filled.

Content analysis: Besides the analysis on the usage of metadata fields, we also explored the content of descriptive metadata fields. In a first step, we determined the most frequent keywords in the metadata field `dc:subject`. Table 4.10 presents the results. All keyword lists can be found in our repository. We received an empty `dc:subject` field for 81% of *GBIF* datasets. For *Zenodo*, in only 52% of the analyzed metadata files keywords were present. In general, we observed inconsistencies in spelling across all repositories. For instance, upper and lower case as well as different spellings resulted in separate entries. The results for *PANGAEA* and *Dryad* indicate that both repositories provide marine data, such as data parameters measured and research devices used. The results for *Dryad* reveal that also terrestrial data are available, e.g., Insects (1296) (insects (180)) or pollination (471) (Pollination (170)). The keyword lists for *Zenodo* and *Figshare* point to

	Environment	Quality*	Material	Organism	Process	Location	Data Type*	Method*	Anatomy*	Human Intervention*	Event*	Time	Person
GBIF (EML)	3%			11%		35%	(8%)	18%				100%	>90%
Dryad (OAI – DC)						60%						100%	80%
PANGAEA (Pan – MD)		(>90%)				100%	100%	90%				100%	100%
Zenodo (OAI – Datacite)							100%					100%	100%
Figshare (QDC)							100%					100%	100%

Table key

Unspecific (generic element) Available (one or more elements)

Amount in brackets denotes the percentage the element is filled.

Table 4.9: Summary of the analysis on the semantic matching of information categories and the best matching standard per data repository. The categories are ranked by their frequency in the question analysis; an agreement less than 0.4 is marked with an asterisk [Löffler et al., 2021].

data repositories with a large number of collection data. We took a closer look on these results and searched in their data portals for the term 'Biodiversity'. We found out that data from the *Meise Botanic Garden* [11] occurred very often in the result set. As collection datasets consist of several individual records that are indexed separately, each occurrence record in our result set counted as hit and contained the label 'Biodiversity'. *Figshare*'s list also contains a high number for the keyword 'Biodiversity' (219022). Here, we figured out that *Figshare* harvests *Zenodo*'s data.

In addition to the most frequent keywords, we also determined Named Entities occurring in descriptive metadata fields. For this analysis, we selected 10,000 datasets per repository based on a filter strategy: For the domain-specific archives (*PANGAEA* and *GBIF*), we used 10,000 datasets randomly. Here, we assumed that all data are potentially relevant for the biodiversity research domain. For *Zenodo* and *Figshare* we selected 10,000 files with the keyword 'Biodiversity'. We know that this filter might cause a high number with only collection data due to the large number of collection datasets in these repositories. For *Dryad*, the available amount of data with the keyword 'biodiversity' was too small. Therefore, we determined a group of relevant keywords. Further details on the selection strategy can be found in [Löffler et al., 2021].

Table 4.11 presents the results of the NLP analysis. Species are mentioned in all inspected files (*Figshare* 85%, *Dryad* 36%, *PANGAEA* 32.5%, *Zenodo* 29%, *GBIF* 12%). In *GBIF*, the number of ORGANISM annotations is low, as the datasets usually provide broader descriptions of the study taxonomic terms, such as 'Family' or 'Order'. The concrete species are only mentioned in the individual occurrence records. As assumed,

[11] Meise Botanic Garden, `https://www.plantentuinmeise.be`

	PANGAEA	GBIF	Dryad	Zenodo	Figshare
	water (201102)	Occurrence (6510), occurrence (46)	Temperature (16652), temperature (15916)	Taxonomy (459877), taxonomy (105)	Medicine (1057684), medicine (240)
	DEPTH (198349), Depth(71916)	Specimen (3046), specimen (22)	Integrated Ocean Observing System (16373)	Biodiversity (458336), biodiversity (8593)	Biochemistry (1015906), biochemistry (92)
	Spectral irradiance (175373)	Observation (2425), observation (24)	IOOS (16373)	Herbarium (270110), herbarium (91)	Biological Sciences not elsewhere classified (983829)
	DATE/TIME (128917)	Checklist (589), checklist (43)	Oceanographic Sensor Data (15015)	Terrestrial (269900), terrestrial (177)	Chemical Sciences not elsewhere classified (842865)
	Temperature (118522), temperature (50)	Plantas (368), plantas (42)	continental shelf (15015)	Animalia (205242), animalia (261)	Biotechnology (792223), biotechnology (23978)
empty `dc:subject`	0	38296	15436	705730	0

Table 4.10: The most frequent keywords in the metadata field `dc:subject` [Löffler et al., 2021].

	Pangaea	GBIF	Dryad	Zenodo	Figshare
Location	9088	5718	3530	4644	**9978**
Person	3789	**6687**	3030	3201	1773
Organization	3307	1762	1486	1674	**9542**
Organism	3251	1217	3603	2891	**8542**

Table 4.11: "Number of datasets with Named Entities (out of 10,000 processed files in a reduced OAI-DC schema) per repository. Each file contains a subset of the original metadata, namely, `dc:title`, `dc:description`, `dc:subject` and `dc:date`." [Löffler et al., 2021]

Figshare metadata files contain a high number of ORGANISM annotations, as they are mainly obtained from collection data. However, other entity types do not occur in *Figshare*'s descriptive fields. *PANGAEA* metadata files provide the highest number of annotations for the categories MATERIAL (86%) and QUALITY (78%). In contrast, *Dryad* achieved the highest result for the categories ENVIRONMENT (52%) and PROCESS (67%).

Even if our results are only based on a subset of available data, and even if text-mining pipelines are mainly developed for longer textual resources and might contain false positive results (wrongly annotated terms and phrases), the outcome of this short NLP analysis shows that descriptive metadata fields contain information that is relevant for information seekers. Current NLP tools can support and leverage the extraction of biological entities from text.

4.4.3 Discussion

The outcomes of this study reveal that general metadata standards such as *Dublin Core* and *Data Cite* are widespread metadata standards used at general data repositories. In contrast, domain-specific repositories are more likely to use further domain-relevant formats such a *EML* or *Pan-MD*. For *GBIF*, we could not determine a full picture as we only analyzed the datasets and not the individual occurrence records.

The results also show that information categories representing scholarly search interests and metadata semantically poorly match. Apart from temporal information (TIME) and contact details (PERSON), important information needs such as information on environments, materials or species are not explicitly given in most inspected metadata files. To compensate for these limitations, repositories enrich metadata with keywords in general fields such as 'subject'. However, our results of the content analysis reveal that the occurring keywords in these general fields reflect scholarly information needs only to some extent. The description of research data varies in granularity and quality. Keywords can be either arbitrary chosen in different spellings, can be very broad, e.g., subject categories, or very specific such as concrete measuring methods. These poor data descriptions are one reason why searching for research data is time-consuming [Parker et al., 2016] [Ramakers et al., 2018] [Culina et al., 2018]. As most repositories use keyword-based retrieval techniques [Khalsa et al., 2018], data can only be found when the search term matches entries in the dataset. Therefore, scholars are more likely to send explicit data requests to data repositories [Kacprzak et al., 2018] or to use general search engines [Gregory et al., 2020] to find relevant research data.

4.5 Summary

This chapter aimed to analyze the coverage of scholarly information needs in biodiversity research and current metadata in data repositories. Addressing H_2, the goal was to determine whether metadata reflect search interests and to quantify the gap between information needs and metadata.

The question corpus study shows that scholarly search interests are as diverse and complex as the data. Information needs in biodiversity research go beyond the search for species and cover questions about environmental terms, geographic locations, processes and materials. Further relevant entity types are data parameters and data type. However, for these types the research community needs to discuss the namings. Categories that occur, but that are not mentioned frequently, are events (processes at a specific time and location) and information on human interventions in the environment.

Our second and third study reveal that current metadata in data repositories match scholarly search interests in biodiversity research only to some extent. Although, a variety

of metadata standards are available related to biodiversity research, only domain-specific repositories tend to use these standards. In our study, we figured out that these data-specific fields are not always filled. Here, further research is needed why this is the case. Metadata fields for describing the actual content of research data are mostly limited to descriptive fields like title, description and keywords. A second issue hampering dataset retrieval are arbitrary keywords in metadata fields. Keywords may have several meanings in different disciplines and scholars describe data from their research perspective, but information seekers search with different keywords. Only if proper aligned keywords are contained in general or descriptive fields, conventional retrieval systems are able to find relevant datasets.

In the following chapter, we introduce our concept for an improved dataset search based on these findings.

Chapter 5

Concept for a Semantic Dataset Search

"We're flooding people with information. We need to feed it through a processor. A human must turn information into intelligence or knowledge. We've tended to forget that no computer will ever ask a new question."

- Grace Hopper [Schieber, 1987], *pioneer in Computer Science*

Artifical intelligence (AI) aims to develop systems that replicate human thinking and acting. In order to achieve this "intelligence", different approaches and technologies have to be combined. One example for such a promising technology is the idea of a "semantic web", first introduced by [Berners-Lee and Hendler, 2001]. The Semantic Web belongs to the research field of symbolic AI approaches and pursues the vision of a world wide web allowing machines to 'understand' human information. Understanding in this context does not mean that machines need to interpret every sentence. Rather, human language should be presented in a formalized and structured way enabling machines to draw conclusions, or as the authors describe it "to solve well-defined problems by performing well-defined operations on well-defined data" [Berners-Lee and Hendler, 2001]. These formalized and structured data are stored in domain specific ontologies (see also Chapter 2.3). Linking words of human language to entries in ontologies enable machines to exploit additional knowledge not being present in the given resource. This semantic enrichment has the potential to enhance existing systems tremendously, for instance in search applications. As data discovery is a complex and time-consuming but necessary task in daily research practice, additional domain knowledge could support users in finding more relevant data.

In this chapter, we summarize the findings from the previous chapters (Section 5.1), list and discuss improvements (Section 5.2) and introduce a concept for a semantic dataset search (Section 5.4).

5.1 Identified Obstacles

In a first study, we explored whether scholars are able to retrieve relevant data for biodiversity research in a current data portal (Chapter 3.2). In a prototype, we also compared a keyword based retrieval approach and a semantic search approach. Our findings show that keyword based retrieval approaches result in moderately relevant results, while a semantic search approach returned more relevant datasets (H_1, Chapter 3.3). The findings also reveal that scholars need explanations to understand results going beyond keywords.

A second outcome are identified search interests in biodiversity science (Chapter 4.2). Scholars are interested in information about organisms (ORGANISM), environmental information (ENVIRONMENT), materials and chemicals (MATERIAL), involved processes (PROCESS), data parameters measured and phenotypes (QUALITY), geographic locations (LOCATION) and data types (DATA TYPE). Further relevant categories are information on events (EVENT) and human impacts on habitats (HUMAN INTERVENTION). Apart from this topical diversity, information needs can also vary in granularity. Queries can contain very specific terms, e.g., queries for species and data parameters, but can be also broader containing terms which require additional knowledge for understanding and answering, e.g., questions containing terms such as *microbial diversity, mobile organic matter, plants and insects*. Moreover, questions in biodiversity research differ also in complexity. Scholars, who ask for simple factoid questions, expect concrete facts in the result set. This type of question accounted for the largest share of the question corpus. 20% of the analyzable questions were more complex questions such as questions for definitions, explanations or association questions, e.g., *How do environmental factors impact soil moisture patterns?*. For such information needs, we imply that scholars only expect datasets containing hints rather than concrete answers.

A third outcome are results on metadata usage in data repositories. A main obstacle for dataset search in biodiversity science is the usage of general metadata standards with high-level and non-domain specific information, such as author, title, collection or publication date and citation. Explicit domain-specific information is often missing, e.g., data parameters measured, habitats or biological processes. General metadata standards are too coarse grained and reflect scholarly information needs in biodiversity research only partially (Chapter 4.3). Domain-specific repositories tend to utilize domain-specific standards in addition to the general ones. However, our study shows (Chapter 4.3) that these domain-specific fields are not always filled or that further domain relevant information is not present. Another key issue are missing controlled vocabularies. We found out that large data repositories use controlled vocabularies only partially. None of the inspected repositories link metadata to information in knowledge bases. However, our study shows that descriptive fields, such as title, description and keywords, contain relevant information, which can be enriched with text mining techniques (H_2, Chapter 4.3).

In the following, we summarize and list the identified obstacles in dataset search for biodiversity research.

A - Obstacles in Data:

- Metadata are mostly provided in fields of general metadata standards, which cover search interests only to some extent.
- Keywords in metadata are less aligned to controlled vocabularies and are not linked to available terminologies.

B - Obstacles in the Data Retrieval Process:

- Conventional retrieval approaches are focused on keyword based text retrieval, but scholarly information needs are diverse in complexity, granularity and topics and go beyond keywords.
- Most data providers use common search engines, which are based on keyword based retrieval methods.

C - Obstacles in Queries:

- Query terms in dataset search are diverse in topical coverage.
- Query terms in dataset search are diverse in granularity.
- Query terms in dataset search are diverse in complexity

5.2 Suggestions for Improvement

Based on these identified obstacles, we propose the following improvements:

5.2.1 Data Enhancements

General metadata standards often do not contain domain specific fields (*Obstacle A*). Therefore, filtering and querying for domain relevant categories are not possible. A faceted search, which allows browsing and filtering on grouped information, is only possible if information is aggregated in a pre-processing step. In search engines, these aggregations are mainly taken from existing metadata fields. If domain relevant categories are not present in metadata, they can not be searched for. Hence, domain specific information needs to be added to metadata. Apart from short information, e.g., contact details, data type or location, all metadata standards contain longer textual resources, such as

title, description and abstract. Most of them contain useful information for search, e.g., information on species observed or habitats (Chapter 4.3).

However, existing (poor) metadata can not be enhanced manually, but require automatic processes. **Implicit information** being present in longer textual fields in metadata can be **extracted with text mining techniques**. Relevant terms and categories can be identified and linked to concepts in terminologies. This additional information can be added to metadata as semantic annotations. The linkage to URI concepts also enables the extraction of further related concepts in the terminology, which can be added to metadata, too. These enrichments finally allow a category and URI-based search.

5.2.2 Enhancements in the Retrieval Process

Scholary information needs are complex, diverse and difficult to describe with a few keywords. Classical retrieval models based on keywords are therefore less suited for dataset search (*Obstacle B*).

Concept based retrieval models are a promising approach to counteract the drawbacks of exact matches with keywords (Chapter 3.1.2). Concept or entity based retrieval utilizes URIs in the search process instead of keywords (see also Chapter 7). This broadens the search on all semantically related terms being represented by this URI, e.g., synonyms, common names or different languages. Moreover, the usage of URIs enables the population of further domain knowledge into the retrieval process. The search could be expanded on related concepts addressing the need for a search beyond keywords, e.g., by including concepts higher in hierarchy. In order to allow a search within domain-specific categories, entity types need to be extracted and indexed, too.

5.2.3 Query Enhancements and User Interface

Query Enhancements: Search interests are usually expressed in a few keywords, which reflect a range of topics on different levels in complexity and granularity (*Obstacle C*).

To achieve a concept based retrieval, besides the data, query keywords also need to be linked to entries in ontologies. An URI based query enables an URI based search and also allows the expansion on further relevant semantic relations. **Expanding user queries on related terms** can help to overcome the obstacles of a keyword based search. However, expansions can also lead to the opposite result, in particular when flooding the result set with data that is semantically related but not relevant for the information seeker. Therefore, only tailored expansions will lead to more relevant results.

User Interface: The literature studies about existing semantic search systems in the Life Sciences (Chapter 3.1.2) and user studies about dataset search (Chapter 3.1.1) as

well as our own findings on a question corpus (Chapter 4.2) reveal that scholars need different opportunities to enter information needs. Scholarly search interests in biodiversity research are diverse in granularity and can vary between specific interests in facts (e.g., butterflies on calcareous grasslands), but they can also be broad und fuzzy (insects [ORGANISM] on grassland [ENVIRONMENT]). However, in our question study, we showed that information needs can be classified into a limited amount of entity types.

Based on these findings, we propose a **structured search input** that provides the opportunity to search within the identified search categories. In addition, our findings in the user evaluations comparing a keyword based and a semantic search (Chapter 3.3) reveal, that user interfaces with search results going beyond keywords need **explanations**. Scientists are no experts in semantic relations and scientific terms of a determined domain. Therefore, the returned datasets in a search summary should provide explanations on why they are displayed and what relation their entries have to the originally entered key terms. Following the suggestions for user interfaces by [Shneiderman et al., 2016], user interfaces need to provide an overview first and details on demand. In order to get a fast overview of what is inside a dataset, **text hightlights and labels** drawing the user's attention on important terms or facts are helpful improvements in search result presentations. They can also be utilized to explain search results by highlighting the query terms or text snippets being relevant to the query terms.

For dataset search, there are some additional requirements in contrast to a search over articles or web pages with respect to the user interface. Dataset evaluation studies 3.1.1 show that major obstacles in current data portals, besides incomplete metadata (addressed in Section 5.2.1), are missing information in the result presentation. The studies also show that improved user interfaces increase user satisfaction and system acceptance. In particular, users appreciated full metadata descriptions, details on demand and aligned terminologies. The **Research Data Alliance's (RDA)** data discovery group on dataset search confirms these **suggestions for improving data discoverability**. They published ten recommendations for data providers to enhance dataset search and the user's search experience [Wu et al., 2019]. The recommendations comprise suggestions with respect to the query input ("Provide a range of query interfaces to accommodate various data search behaviours" [Wu et al., 2019]), the presentation of search results ("Make it easier for researchers to judge relevance, accessibility and reusability of a data collection from a search summary" [Wu et al., 2019]) including highlightings of search terms and machine readable access for interoperability ("Follow API search standards and community adopted vocabularies for interoperability" [Wu et al., 2019]). Novel user interfaces in dataset search should consider these RDA recommendations in the development process.

5.3 Requirements

In the following, we sort the above mentioned improvements into functional requirements for the proposed semantic dataset search. Functional requirements describe what concrete tasks or functions a system should fulfill. In contrast, non-functional requirements describe how these functions should be achieved. They comprise system relevant characteristics that are not function specific.

Functional Requirements:

- *R1 search for datasets beyond keywords:* As keyword based retrieval is a major obstacle in finding research data in biodiversity research (addressing Subsection 5.2.1 and 5.2.2), the proposed system should provide datasets in the result set that go beyond a syntactic keyword match, but that are semantically related to the search terms.

- *R2 category based search:* Although scholarly user interests are very diverse, they can be grouped into a limited set of categories. Therefore, the proposed system should provide the opportunity to search within information categories (addressing Subsection 5.2.3).

- *R3 comprehensible search result:* If the search is extended on semantically related information, it is necessary to provide explanations on the relationship of information in the result set and the search input. Users need to understand why a certain dataset appears in the search result (addressing Subsection 5.2.3). Highlights in the search result support users in getting a quick overview of what is inside a dataset and what relation the dataset has to the query terms.

Non-Functional Requirements

- *R4 easy to use:* Complicated user interfaces with many different settings confuse users and reduce satisfaction in usage [Bakalov et al., 2011]. Therefore, the proposed system should follow the mantra of [Shneiderman et al., 2016] and needs to provide an overview first and details on demand.

- *R5 adherence of user-centered design principles:* In order to properly address R3 and R4, users need to be involved into the entire development process.

5.4 Proposed System

Our literature study about semantic search approaches in the Life Sciences (Chapter 3.1.2) shows that search systems on combined data (text and URIs) are still rare. However, they are the most promising approach towards a semantic search [Bast and Buchhold, 2013, Bast and Buchhold, 2017, Cunningham et al., 2013, Bontcheva et al., 2014]. In particular, these systems are based on a concept based retrieval model rather than keywords, or they combine several different indexes, e.g., tokens and URIs. To address *Requirement R1*, we decided to continue work in this research direction and to build the retrieval on a **concept based model**.

The decision for a concept based model requires URIs, in the underlying metadata and the queries. Therefore, in a pre-processing step, important **keywords** in descriptive metadata fields have to be **linked to entries in ontologies**. These metadata analysis should also determine the entity type information of the extracted terms. The result of that pre-processing step are semantic annotations being added to metadata with information on (1) entity type, (2) URI and (3) start and end node. The same text extraction and entity linking process has to take place on query side.

Semantic search approaches on combined data still lack of an appropriate user interface (Chapter 3.1.2) and often require to know the correct URIs. As this hampers users with no background knowledge in Semantic Web in using semantic search systems, a keyword based input should be offered, too. In case of keywords, appropriate URIs are determined and inserted into SPARQL templates. In order to address *Requirements R2 and R3*, the **category information** of the semantic annotations in metadata needs to be **added to the search index** to allow a category based search in the user frontend.

These considerations lead to three major subsequent components. Here, we only briefly introduce the main idea of each component. They are individually introduced in the following chapters.

1. *Pre-processing Component:* At first, metadata are loaded from a datastore and are analyzed by various NLP pipelines. Important terms and phrases in descriptive metadata fields such as title, description and abstract are extracted and classified into the identified information categories (Chapter 4.2). Furthermore, the terms are linked to selected terminologies in a knowledge base that are relevant for biodiversity research. The outcome are semantic annotations (with information on the entity type, the URI and start and end node), which are added as additional information to metadata (Chapter 6).

2. *Retrieval Component:* The retrieval component is responsible for the actual search and ranking. A semantic index on both, URIs and tokens ensures a URI-based retrieval by default, but allows a keyword search in case no URI can be linked to a

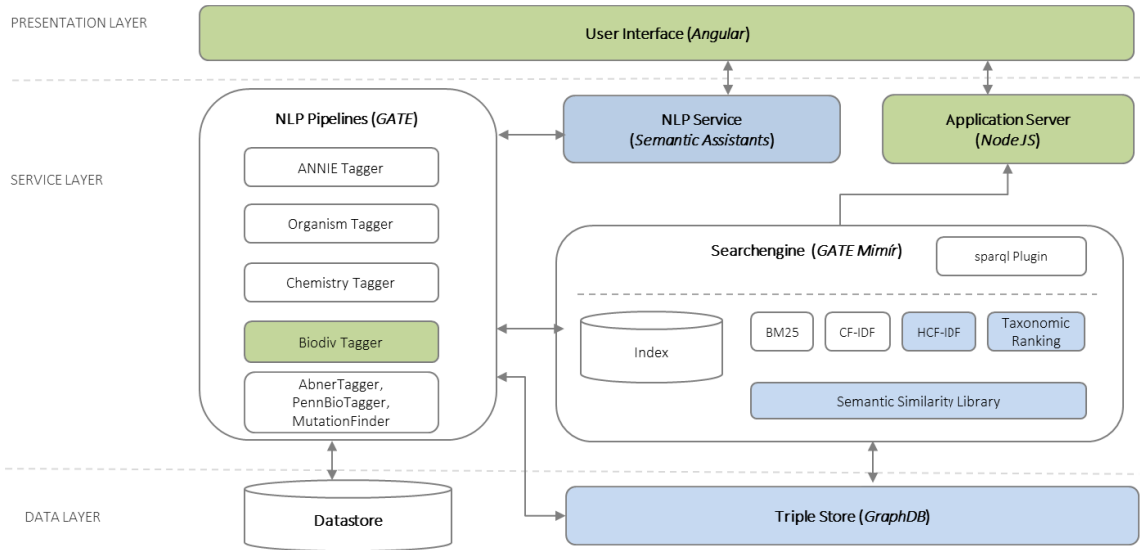


Figure 5.1: Architecture of the proposed system. The blue color denotes components that we integrated into the system, but have been proposed in related work. All green colored components are the new components introduced and developed in the scope of this thesis.

term. We analyze whether existing entity based retrieval approaches are sufficient and whether additional semantic relations enhance dataset search (Chapter 7).

3. *Presentation Component:* The user interface denotes the third major component and consists of a UI frontend and a middleware handling the requests to the search engine (Chapter 8). The UI component takes the user query, links the entered keywords to entities in the knowledge base and generates an expanded entity query (based on templates) being sent to the search engine.

The overall idea of the proposed system is to utilize semantic technologies in the entire system. Instead of keywords, URIs (concepts) should be used for matching datasets and search queries. Along the semantic search classification of Bast [Bast et al., 2016], the proposed system falls into the category *SSCD - Semi-structured Search on Combined Data*. The idea is to operate on combined data sources, text and knowledge bases. Concerning the search approach it should be possible to search with plain text (keywords) and entities. The expected outcome are relevant metadata files and not entries in knowledge bases. However, semantic relations obtained from knowledge bases should be considered in the retrieval process.

Figure 5.1 presents the overall architecture with three layers: data, service and presentation. In the data layer, the original metadata files as well as the annotated metadata files are hosted in a datastore. A strong focus lies on the integration of controlled vocabularies, in particular on ontologies in semantic formats such as RDF and OWL. Therefore, a knowledge base (triple store) hosting various Life Science ontologies serves as knowledge source for the other components. We introduce the used ontologies in Chapter 6, Section 6.3.1. The service layer consists of the pre-processing component with the NLP

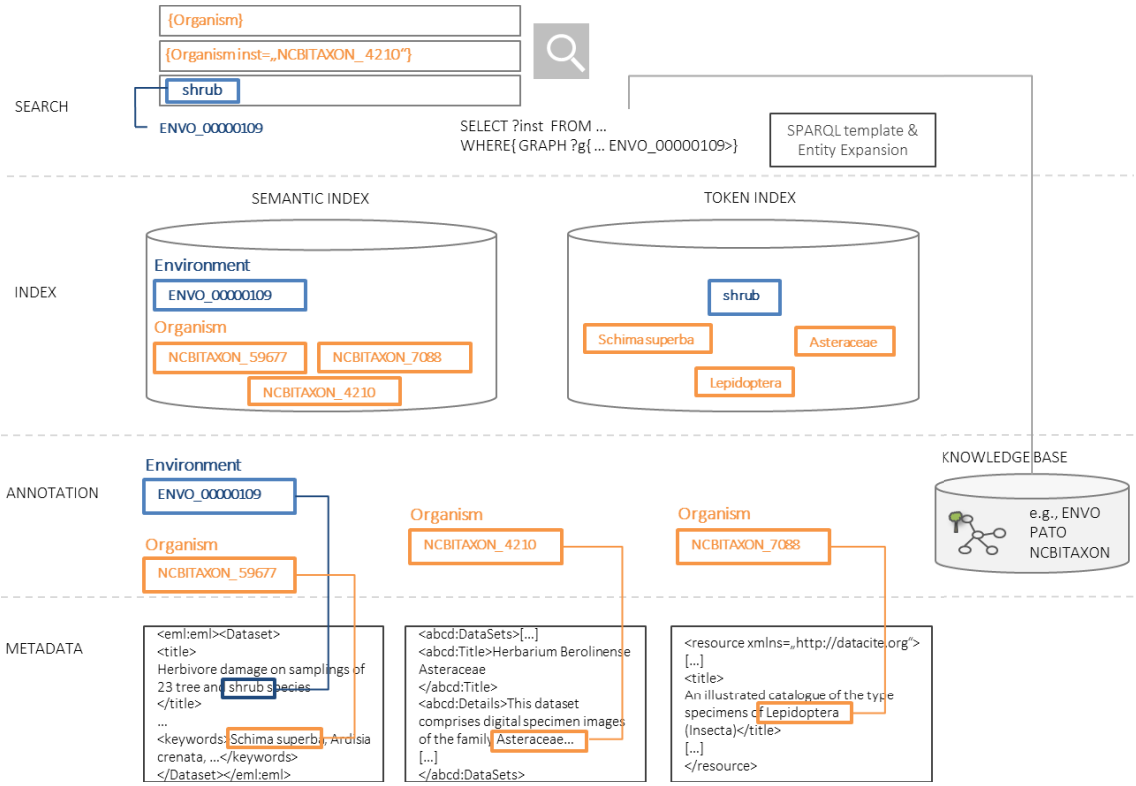


Figure 5.2: Data perspective of the proposed approach. Independent of the metadata schema of the original datasets important entity types and entities are extracted and indexed. This URI-based index approach finally allows a search beyond keywords. Excerpts of datasets (from left to right):[Assmann and Schuldt, 2008, Curators Herbarium B, 2020, Zahiri et al., 2021]

pipelines, the search engine and two additional services providing access to NLP pipelines via a REST interface, and the application server encapsulating the access to the search index. The presentation layer is the user frontend of the search. This modular architecture allows the replacement of individual components or services.

From the data perspective (Figure 5.2), the main advantage of an URI-based index is its independence of the metadata schema of the datasets. While former federated approaches [Löffler et al., 2017] map different metadata schemas to one schema, the recognition of entity types, entities and an URI-based index allows a search over datasets with various metadata schemas. Relevant semantic relations can be either obtained during the annotation process and can be added to the index, or they can be integrated into a SPARQL query in the actual search phase. Another advantage is the search over keywords, categories and URIs. Users are able to search with keywords that are mapped to URIs and SPARQL queries in the backend, or they can retrieve datasets within a domain category. For experts, it is also possible to search for a specific URI in a category. However, as not all scholars are familiar with ontologies and URIs, supportive filtering and search approaches should be provided in the user interface. More detailed information on the suggested user interface are presented in Chapter 8.

5.5 Use Cases

To illustrate the enhancements of our proposed system, we describe two use cases, which are very difficult or time consuming in data portals with conventional keyword based retrieval techniques and which we would like to improve. We could not utilize the use cases of Chapter 1.2, because not enough data are available in the GFBio data portal, the reference metadata corpus we utilized in the final usability evaluation (Chapter 8). Therefore, we selected similar use cases of the collected question corpus (Chapter 4.2).

- **Use Case 1:** How variable is the oxygen concentration of sea water of the global ocean?

  To find suitable datasets for this question, scholars would need to search for various data parameters with the search term 'oxygen' or the chemical formular 'O2' and 'sea water', 'sea water' or 'ocean'. Moreover, the key term 'sea water' can be grouped into two entity types. It can occur in the context of materials or as an environment. An enhancement would be if scholars could specify the context with a category search. The results should consider various spellings and synonyms.

- **Use Case 2:** What microbial communities such as bacteria occur in groundwater?

  For this use case, scholars need to think about microbial communities, e.g., specific bacteria species. The key term 'groundwater' can also occur in different spellings. Therefore, an improvement in search would be if the search expands the result on more specific terms. For the query term 'bacteria', it would be helpful to automatically return bacterial species. For the term 'groundwater', various spellings and synonyms should be considered.

5.6 Summary

In this chapter, we summarized the findings from the previous chapters and introduced our concept for a semantic dataset search. The main focus is on a proper integration of controlled vocabularies into the whole search process. We propose a system that consists of three main components: (1) a data pre-processing component that analyzes the datasets syntactically and semantically (Chapter 6), (2) the search engine that indexes and ranks the datasets (Chapter 7) and (3) a user interface that takes the query input and presents the search result (Chapter 8).

Chapter 6

Semantic Annotation of Biodiversity Metadata

"All knowledge is connected to all other knowledge. The fun is in making the connections."

- attributed to Arthur Aufderheide, *Physician and pioneer in paleopathology*

Text mining pipelines allow the extraction of semantic information, e.g., persons, locations and temporal information. In recent years, more and more text mining approaches have emerged that also support the linkage to concepts in terminologies. We assume, enriching metadata with this additional knowledge enhances dataset descriptions and thus dataset retrieval.

Therefore, this chapter introduces the pre-processing component of the proposed semantic dataset search (Chapter 5.4), a text mining pipeline extracting and linking important entities in biodiversity metadata to concepts in a knowledge base. In order to evaluate this pipeline, we created a manually annotated metadata corpus (gold standard) and compared the automatically created annotations from the pipeline with the manual annotations. This gold standard is the very first manually labeled metadata corpus for the biodiversity domain. At first, we present related work in Section 6.1. Afterwards, we introduce our manually created metadata corpus in Section 6.2. In Section 6.3, we summarize the existing textmining pipelines and ontologies that represent the identified information categories from Chapter 4. Moreover, we introduce our text mining pipeline *BiodivTagger* and present the evaluation results in Section 6.4.

This work has been published at the Language Resource and Evaluation Conference (LREC2020) [Löffler et al., 2020]. The following chapter contains the results from this publication, but provides extensions in related work and methodology.

6.1 Related Work

The term *semantic annotation* can have different meanings in scientific literature. Some publications only refer to the identification of entity types [Campillos et al., 2018, Kilicoglu et al., 2018], others denote the linkage to entity types *and* ontology classes as semantic annotation [Balog, 2018, Jovanović and Bagheri, 2017, Maynard et al., 2017]. Therefore, in the following subsections, we distinguish between *simple semantic annotations* when talking about annotations that solely extract entity types (categories) and *extended semantic annotations* that, in addition to the entity type information, also include the URI to an ontology concept.

6.1.1 Semantic Annotation in the Life Sciences

In the Life Sciences, a large variety of taggers were developed to extract information in particular from medical text such as diseases, treatments, chemical compounds, genes, enzymes and proteins [Hawizy et al., 2011, Pyysalo and Ananiadou, 2013], [Cunningham et al., 2013, Campillos et al., 2018, Jovanović and Bagheri, 2017]. A majority of these approaches focus on simple semantic annotations [Hawizy et al., 2011, Pyysalo and Ananiadou, 2013, Cunningham et al., 2013, Campillos et al., 2018]. However, more and more services are launched that also provide linkage to URI concepts. For instance, *Bioportal*'s annotator [Tchechmedjiev et al., 2018] offers a graphical interface and programmatic access to its hosted terminologies. Users and developers obtain a list of matching ontology concepts for entered keywords. In case of several matches for a given keyword, all possible matching concepts are returned, disambiguation is not provided. A large step towards precise and distinct extended semantic annotations is the approach by [Kim et al., 2019]. Their service is based on a recently introduced new language model for the biomedical domain (BioBERT) [Lee et al., 2019]. The returned extended semantic annotations contain a classification into one of the given categories, e.g., genes, diseases, drugs/chemicals and species, and also provide links to knowledge bases (if available), e.g., MESH [MeSH, 2022] and NCBI Taxon [NCBITaxon, 2022]. However, the recognition of biological entities remains challenging as the scientific language in the Life Sciences is fuzzy and can provide inconsistent namings [Thessen et al., 2012] [Ananiadou et al., 2004]. The outcomes of the study by [Gurulingappa et al., 2010] reveal that the usage of ontologies is suitable for Named Entity Recognition (NER) tasks. Utilizing multiple ontologies achieves a good coverage of ontology concepts in biomedical literature.

Taggers and terminology services for biodiversity research: While a multitude of Named Entity Recognition tools are available for the biomedical domain, only very few

approaches study the requirements and needs for biodiversity research. Some taggers were developed for the extraction of taxonomic information and species names, e.g., [Naderi et al., 2011, Pafilis et al., 2013] and *TaxonFinder* [1]. The recognition of morphological characters or phylogenetic attributes succeeded with approaches introduced by [Balhoff et al., 2010, Eliason et al., 2019]. [Naderi et al., 2011] and [Balhoff et al., 2010] provided extended semantic annotations with linkage to NCBITaxon [NCBITaxon, 2022] and PATO [Gkoutos, G. et al., 2022]. Bioportal [Whetzel et al., 2011] is a large terminology service in the Life Sciences offering a multiude of ontologies for the biomedical domain. In contrast, GFBio's terminology service (GFBio TS) [Karam et al., 2016] was introduced to offer programmtic access to terminologies dedicated to biodiversity research, e.g., ENVO [Buttigieg et al., 2016] or PATO [Gkoutos, G. et al., 2022]. The GFBio TS has a stronger focus on collection data and marine data. Therefore, further vocabularies such as Catalogue of Life (COL) [Bánki et al., 2022] and WORMS [Horton et al., 2022] were added to the repository. Disambiguation and entity type detection are not provided. A newer service developed by [Dimitrova et al., 2020] is a tagger looking for matching entities from ENVO [Buttigieg et al., 2016] and PATO [Gkoutos, G. et al., 2022].

6.1.2 Available Gold Standards in the Life Sciences

The development of text mining pipelines is heavily influenced by available manually labeled text corpora. These so-called gold standards are necessary to assess the accuracy of automatically generated annotations of text mining pipelines in comparision to manually labeled annotations (see also Chapter 2.2.1). Two main workshop series in the last decade encouraged the emergence of different labeled corpora in the Life Sciences, namely BioNLP[2] and BioCreative VI[3].

BioNLP has a stronger focus on the extraction of genes, proteins and diseases including extended semantic annotations. For example, one outcome of the BioNLP workshops is the CRAFT corpus [Cohen et al., 2017a], a resource that consists of 97 full-text articles with 100.000 concept annotations from the biomedical domain. In the first version, the corpus only provided simple annotations for genes, the latest version (version 5.0) already includes concept annotations from various *OBO Foundry*[4] ontologies, an initiative that aims to interlink ontological concepts, e.g., Chemical Entities of Biological Interests (ChEBI) [Hastings et al., 2016], NCBITaxon [NCBITaxon, 2022] and the Protein Ontology (PRO) [Natale et al., 2017].

BioCreative is more oriented towards the chemical domain and relation extraction. One result of these workshops is the CHEMDER corpus [Krallinger et al., 2015]. It con-

[1] TaxonFinder, `http://taxonfinder.org/`

[2] BioNLP, `http://bionlp-corpora.sourceforge.net/index.shtml`, `https://aclanthology.org/venues/bionlp/`

[3] BioCreative, `https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/`

[4] OBO Foundry, `http://www.obofoundry.org/`

tains 85,000 manually labeled chemical annotations in 10,000 PubMed abstracts. Curators with expertise in Chemistry were trainined in the labeling of various chemical types such as formulas (e.g., O2), chemical families (e.g., ketolides) and abbrevations (e.g., ATP). Approaches for relation extraction in the chemical domain are for instance gene-disease relations [Bravo et al., 2015, van Mulligen et al., 2012] and chemical-protein relations [Peng et al., 2018]. All these approaches offer simple annotations, however, terminologies were partially used to determine the correct entity type.

A further known corpus is the GENIA corpus [Kim et al., 2003]. In the original version, the GENIA corpus provides 96 582 manual annotations in 2000 PubMed abstracts. The annotations are classified into 47 bio-medical categories obtained from an own-developed GENIA ontology. A corpus with textual resources in multiple languages is the Mantra GSC corpus [Kors et al., 2015]. It comprises scientific abstract titles, drug labels and biomedical patent claims being decomposed in units (titles or sentences). Per source and language (English, French, German, Dutch, Spanish) 100 units were selected for the final corpus. In total, 12 annotators were involved to label ten semantic types such as anatomy, chemicals and drugs, disorders, devices, phenomena and procedures. Pre-annotations were given from UMLS [UMLS, 2022]. The final extended semantic annotations contain the entity type information and the respective UMLS concepts. The above mentioned corpora are only a few out of a multitude of available labeled resources for the biomedical domain. Recent research approaches in this domain are more and more focused on the construction of Knowledge Graphs [Xu et al., 2020], a subsequent research aim once entity recognition and relation extraction have been successfully processed.

Gold standards for biodiversity research: The domain coverage of the above mentioned corpora mainly comprises drugs, genes, diseases and chemical compounds. These categories address information needs in biodiversity research only partially. For biodiversity research, we need resources with labeled species, environmental terms, data parameters, materials and processes. Only very few corpora exist that are suited for this purpose.

Species are the most prominent semantic category being covered in a couple of corpora. For instance, Linnaeus [Gerner et al., 2010] provides species mentions in 100 full-text articles from PubMed. The annotations also provide NCBITaxon [NCBITaxon, 2022] identifiers. Species-800 [Pafilis et al., 2013] is another corpus focused on the annotation of organisms. It provides taxon entity annotations from 800 PubMed abstracts and the species mentions were linked to the corresponding NCBI identifiers [NCBITaxon, 2022]. The Bacteria Biotope (BB) [Deléger et al., 2016] corpus was created in the scope of the BioNLP workshop series. PubMed titles and abstracts were manually labeled by seven annotators with simple bacteria mentions (no linkage to URI concepts) and their locations such as habitats or geographic information. The COPIOUS corpus [Nguyen et al., 2019]

is another gold standard corpus with simple annotations of the following entity types: geographical locations, habitats, temporal expressions and person name entities. The corpus uses pages from the Biodiversity Heritage Library (BHL) [Gwinn and Rinaldo, 2009], an open access library with legacy literature on biodiversity research. For the German language, [Ahmed et al., 2019] introduced the BIOFid corpus, a manually annotated corpus of 969 scientific articles for biodiversity research with general, simple entity type annotations, such as person, organization, time, geographic information and taxon names.

The literature study reveals that, to the best of our knowledge, no gold standard for research metadata exists. Moreover, for biodiversity research, there are only corpora based on legacy documents. Due to this lack of labeled corpora, in the past decade, only very few tools were introduced to extract important entities and phrases for biodiversity research. Therefore, in the next sections, we introduce our labeled metadata corpus and a novel text mining pipeline that allows the automatic extraction of biological entities, e.g., processes, environmental information or data parameters.

6.2 A Metadata Gold standard for Biodiversity Research

In the Life Sciences, a variety of metadata standards were developed for different purposes and sub-domains (Chapter 4). The proposed metadata corpus aims to reflect this heterogeneity and diversity of biological metadata standards. Therefore, we selected five data repositories and project databases with various metadata formats, namely *Dryad* [5] (a generic data repository) in *Dublin Core* [6] format, *PANGAEA* [7] (a data archive for environmental data) in an extended Dublin Core format called *PanMD* [8] and three project-related portals and databases such as *BEFChina* [9] (a joint Chinese-German-Swiss biodiversity research project) in *EML*[10] format, *iDiv* [11] (The German Centre for Integrative Biodiversity Research Halle-Jena-Leipzig) and the *Biodiversity Exploratories* [12] (a large, long-running functional biodiversity research project in Germany), both providing an own XML based metadata schema (see also Chapter 2.4). For each data repository, ten representative files were utilized for the annotation task. An excerpt of a metadata file from BEFChina in EML format is shown in Listing 6.1.

[5] Dryad, `https://https://datadryad.org`
[6] Dublin Core, `https://dublincore.org/`
[7] PANGAEA, `https://www.pangaea.de/`
[8] PanMD, `https://wiki.pangaea.de/wiki/Metadata`
[9] BEFChina, `http://china.befdata.biow.uni-leipzig.de/`
[10] EML, `https://knb.ecoinformatics.org/tools/eml`
[11] iDiv, `https://idata.idiv.de/`
[12] Biodiversity Exploratories, `https://www.bexis.uni-jena.de`

Listing 6.1: Metadata file in EML (excerpt) [Germany and Erfmeier, 2019]

```
<eml:eml [...]><dataset id='577'>
<alternateIdentifier>http://china.befdata.biow.uni−leipzig.de/datasets/577</alternateIdentifier>
<title>Main Experiment: Seedling addition experiment − growth and biomass data</title>
<creator id='markus_ger'>
<individualName>
<givenName>Markus</givenName><surName>Germany</surName>
</individualName>
[...]
<abstract>
<para>While coexistence in plant communities is frequently explained
by effects of resource niche partitioning, the Janzen−Connell (J−C)
hypothesis is an alternative approach that has been assumed as a
major ecological mechanism explaining high species richness levels,
in particular, in tropical forest ecosystems. [...]</para>
</abstract>
<keywordSet>
<keyword>janzen connell</keyword>
<keyword>Main Experiment</keyword>
<keyword>seedling addition</keyword>
<keyword>seedling performance</keyword>
[...]
<keyword>Leaves_Dam</keyword>
<keyword>Leaves_Dead</keyword>
<keyword>Damage_pro</keyword>
<keyword>Biomass_Above</keyword>
<keyword>Biomass_Below<
[...]
</dataset></eml>
```

Annotation guidelines: The annotation guidelines were built on the guidelines presented in our previous work [Löffler et al., 2021]. We only labeled noun entities and adjectives, as the focus for this corpus is on biological entities, relations are omitted. Based on the identified main search interests in biodiversity research (Chapter 4), we concentrated on the labeling of phenotypic qualities and characteristics that can be measured or observed (QUALITY), environmental terms (ENVIRONMENT), chemical compounds and materials (MATERIAL) and biological, chemical and physical processes (PROCESS). We did not consider categories that are not only specific for biodiversity research, are not frequently mentioned and for which corpora already exist, e.g., ORGANISM, LOCATION, PERSON, TIME. The first letters of the used categories (Q - QUALITY, E - ENVIRONMENT, M - MATERIAL, P - PROCESS) formed the name of the novel corpus - *QEMP*.

For each compound nouns (nouns with two words, e.g., rain forest), we determined its domain relevance. In case a term or phrase is relevant to biodiversity research such as “climate change” or “ocean acidification”, we labeled the two words as a whole. The first word in these compound nouns describes nouns more closely. We considered these nested entities, when they represent a biological term. For instance, “ocean” additionally

got the annotation [ENVIRONMENT]. To facilitate the decision whether a term is domain specific, the annotators were permitted to look up the phrases in additional sources, e.g., *BioPortal* [13] or *BiodiversityA-Z* [14]. Furthermore, the annotators had the opportunity to use the labeled questions of our previous research [Löffler et al., 2021]. Abbreviations and units were omitted, but multi-labeling was permitted, as for some terms several meanings applied, e.g., "biomass" [MATERIAL, QUALITY].

Annotation process: Four PhD students in computer science annotated the metadata corpus manually. A postdoc in biology gave advice concerning the category definitions and complex biological terms. We utilized the text mining framework GATE [Cunningham et al., 2013] as annotation tool. GATE supports manual annotation tasks and offers functions to merge different annotation sets into a final gold standard and to compute inter-annotator agreement (IAA) measures. A main challenge was to achieve a proper training of the group, because not all team members had expertise with biodiversity data. Hence, we decided to have three training phases over a period of two months. In a *trial round*, the annotators got familiar with the categories and labeled five files individually. The outcomes were discussed in the group and the annotation guidelines were refined accordingly. In the *pilot phase*, we formed two annotator pairs. Each pair double-annotated four metadata files. The results were discussed, and the annotation guidelines were finalized. For the main phase, each annotator team labeled 25 metadata files.

Evaluation of the annotation process: The average inter-rater agreement measures per data repository are presented in Table 6.1. The detailed results are available in our GitHub repository[15]. Usually, IAA is computed with Cohen's Kappa (two annotators) or Fleiss Kappa (> two annotators) [Fleiss, 1971]. However, the usage of κ statistics requires the counting of missing values. For instance, annotators might label different mentions. If the second annotator labeled the term, it would count as non-entity for the first annotator. It is not clear and defined how many terms this non-entity comprises, but Kappa metrics can only be computed if the amount of these negative cases is quantified. Therefore, in Named Entity Recognition tasks, other metrics are preferred such as precision, recall and f-Measure (see also Chapter 2.2.3) from information retrieval [Hripcsak and Rothschild, 2005]. Here, one annotator set is used as "key" set, whereas the other annotator set is used as reference set. At first, some statistics need to be computed: the exact matches (both annotators labeled that term or phrase with the exact start/end offset), partial matches (annotations are overlapping but not identical in span), the missing phrases in annotator set A and the spurious annotations from annotator set B. Precision, recall and f-Measure are computed as stated in Section 6.4.

[13] BioPortal, `https://bioportal.bioontology.org/annotator`
[14] BiodiversityA-Z, `https://www.biodiversitya-z.org`
[15] BiodivTagger, `https://github.com/fusion-jena/BiodivTagger/tree/master/QEMP`

	Precision	Recall	F-Score
BExIS	0.53076	0.53182	0.4423
BEFChina	0.77537	0.67081	0.65824
PANGAEA	0.60154	0.63092	0.61244
Dryad	0.41434	0.68081	0.47059
idiv	0.57993	0.7409	0.6221

Table 6.1: Average inter-annotator agreement measures per data repository.

The F-score values in Table 6.1 are low to moderate. Even with a thorough training, biological entities are difficult to classify and fuzzy. Terms for which the meaning or classification was unknown or difficult were collected in a list. The postdoc in biology looked through these terms and phrases and took the final annotation decision. The expert's decisions were added to the final metadata annotations.

Corpus statistics: The statistics of the QEMP corpus are introduced in Table 6.2. Most annotations were classified into the category QUALITY (37.1%), representing data parameters measured. This large portion is not surprising as metadata of PANGAEA, iDiv, BExIS and GBIF usually contains information on measured variables. Out of 5357 total annotations, 26.22 % of the labeled terms and phrases were grouped into the category ENVIRONMENT and 31.9% into MATERIAL. PROCESS annotations occurred in 4.6% of the annotated tokens. Processes are less represented in metadata fields. Therefore, data providers and data curators can only utilize general fields to describe biological, chemical or physical processes (see Chapter 4.4.2). We assume, if concrete metadata fields are missing, data providers and data curators are not aware of adding this additional information to metadata.

Figure 6.1 illustrates the distribution of entity types over all tokens (left) and the number of annotated tokens per data repository (right). Most tokens were annotated for BEFChina files. The EML metadata schema used by BEFChina provides fine-grained metadata fields for various aspects in data collection, data structure and methodology. Most of these elements are filled, however, we also noticed duplicate text in abstracts and methodology descriptions. Thus, BEFChina files are larger and provide more text to annotate, which resulted in this large amount of annotations. Concerning the distribution of the categories per entity types, Figure 6.1 (left) emphasizes the diverse content of the data in the repositories. Information on materials and chemical compounds are mainly present in metadata files from BEFChina and PANGAEA, environmental terms are the most frequent category in iDIV, BExIS and Dryad. Data parameters measured can be found in all data repositories. However, per data repository less than 10% of the available tokens were annotated. Most tokens in the sentences are either too general and not domain-specific, are to fuzzy and difficult to classify or do not fall in any of the provided categories.

	Dryad	PANGAEA	BEFChina	BExIS	iDiv	Total
All Token	3943	14069	47388	22817	4640	92857
ENVIRONMENT	71	135	643	390	166	1405
MATERIAL	17	283	1132	254	27	1713
PROCESS	50	74	45	43	39	251
QUALITY	72	202	1359	232	123	1988
Total	**210**	**694**	**3179**	**919**	**355**	**5357**

Table 6.2: QEMP corpus statistics: the total number of annotated tokens per entity type and data source.

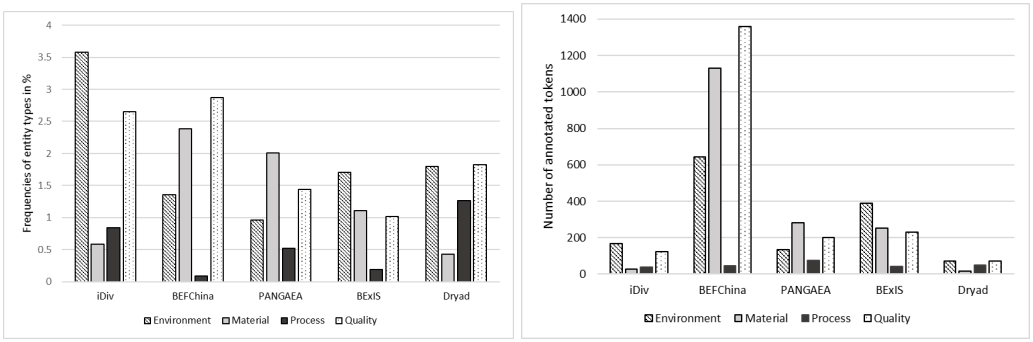


Figure 6.1: Distribution of entity types over all tokens per data repository (left) and the number of annotated tokens per data repository (right).

The QEMP corpus is publicly available in our GitHub repository[16]

6.3 A Text Mining Pipeline for Biodiversity Metadata

The overview of related work in Section 6.1.1 shows that only very few text mining approaches exist for the recognition of relevant entity types in biodiversity research. Table 6.3 summarizes the available taggers for each category. Non Life Sciences specific entity types such as Location, Person and Time are already covered by a variety of text mining approaches. The detection of species is also possible with at least two taggers. Both approaches also link the extracted species to the identifier in the NCBI Taxonomy. All other relevant categories for biodiversity research are only partially covered or not covered at all. Therefore, our methodology focuses on these partially or not yet covered entity types, namely ENVIRONMENT, QUALITY, PROCESS, MATERIAL and DATA TYPE.

Our approach for entity type extraction is based on ontological gazetteers [Saggion et al., 2007] using knowledge bases to form large gazetteer lists. One or several ontologies are assigned to categories. In case an ontology covers more than one category, we assign selected concepts of the ontology to an entity type. Lookup services match terms and entities of these selected terminologies. Thus, a resulting annotation re-

[16]BiodivTagger and QEMP corpus, `https://github.com/fusion-jena/BiodivTagger`

Category	Tagger	Coverage
ORGANISM*	OrganismTagger [Naderi et al., 2011] SPECIES [Pafilis et al., 2013]	species only, no hierarchy considered, linkage to NCBITaxon ID
ENVIRONMENT*	Pensoft [Dimitrova et al., 2020]	no category recognition, ENVO [Buttigieg et al., 2016] entity lookup
LOCATION*	ANNIE [Cunningham et al., 2013] CoreNLP [Manning et al., 2014] NTLK [Bird and Loper, 2004] SpaCy [17] Stanza [Qi et al., 2020] GeoNames [18] DBpediaTagger [Sateli and Witte, 2015]	extraction of geographic names or geographic gazetteer lists (GeoNames) linkage to LOD
Process*	-	
QUALITY*	Pensoft [Dimitrova et al., 2020]	no category recognition, PATO [Gkoutos, G. et al., 2022] entity lookup
MATERIAL*	ChemicalTagger [Cunningham et al., 2013]	chemical elements and compounds
DATA TYPE*	-	
ANATOMY	AnatomyTagger [Pyysalo and Ananiadou, 2013]	anatomical entities, e.g., organs, tissue, cell (no URIs)
Methodology	-	
Event	-	
Human Intervention	-	
PERSON	ANNIE [Cunningham et al., 2013] CoreNLP [Manning et al., 2014] NTLK [Bird and Loper, 2004] SpaCy [19] Stanza [Qi et al., 2020]	names for humans and organizations
TIME	ANNIE [Cunningham et al., 2013] CoreNLP [Manning et al., 2014] NTLK [Bird and Loper, 2004] SpaCy [20] Stanza [Qi et al., 2020]	temporal information

Table 6.3: Important entity types in biodiversity research and available taggers. The asterisk denotes the categories that were mentioned frequently in the question corpus study (Chapter 4.2).

ceives an URI and a category (the category assigned to the ontology or the ontology part). The following subsections introduce the selected ontologies (Subsection 6.3.1), present the architecture (Subsection 6.3.2) and describe the evaluation (Subsection 6.4). Finally, we discuss the results in Subsection 6.4.1.

6.3.1 Ontology Selection

A comprehensive knowledge graph for biodiversity is still missing (Chapter 2.3.4). Numerous ontologies have been developed in biology and biomedicine in the past decade. Most of them are available in open data repositories such as *Bioportal* [Whetzel et al., 2011] or the *GFBio Terminology Service*[21]. Initiatives such as the *OBO Foundry*[22] aim to provide terminologies that are interlinked and adhere to several principles including open use and strictly-tailored content. This well-structured domain knowledge can be used as ontological gazetteers in information extraction tasks [Saggion et al., 2007]. We set up an own local triple store instead of using existing services such as *Bioportal* or *GFBioTS* to be able to integrate new ontologies and to avoid dependencies during the development process. Table 6.4 presents the final categories considered in the pipeline and their

[21]GFBio TS, `https://terminologies.gfbio.org`
[22]OBO Foundry, `http://www.obofoundry.org/`

Category	Terminology (:starting node)	Description
ORGANISM	NCBI Taxonomy [NCBITaxon, 2022]	organisms in the public sequence databases
ENVIRONMENT	ENVO [Buttigieg et al., 2016] :environmental feature (ENVO_00002297), :environmental condition (ENVO_01000203), :environmental system (ENVO_01000254) :immaterial entity (BFO_0000141)	description of environmental entities
QUALITY & PHENOTYPE	PATO [Gkoutos, G. et al., 2022] :cellular quality (PATO_0001396), :molecular quality (PATO_0002182), :physical quality (PATO_0001018), :process quality (PATO_0001236), :morphology (PATO_0000051)	phenotypic qualities
	TO [Cooper et al., 2017]	phenotypic traits in plants
	FLOPO [Hoehndorf et al., 2016]	ontology of phenotypes reported in Floras
	ENVO [Buttigieg et al., 2016] :quality (PATO_0000001)	description of environmental entities
	PPO [Walls, R. et al., 2022] :plant phenological trait (PPO_0002000), :specifically dependent continuant (BFO_0000020)	phenology of individual plants and populations of plants
PROCESS	ENVO [Buttigieg et al., 2016] :process (BFO_0000015)	description of environmental entities
	UBERON [Mungall et al., 2012] :processual_entity (UBERON_0000000)	integrated cross-species anatomy ontology representing a variety of entities
	OBI [Bandrowski et al., 2016a] :process (BFO_0000015)	ontology of investigations, the protocols and instrumentation
	PO [Cooper et al., 2013] :plant structure development stage(PO_0009012)	links plant anatomy, morphology and growth and development to plant genomics data
	REX [OBOFoundry, 2022] :physical and chemical process (REX_0000001)	ontology of physico-chemical processes
	GO [Ashburner et al., 2000, Cooper et al., 2017] :biological_process (GO_0008150)	vocabularies for the annotation of gene products with respect to their molecular function, cellular component, and biological role
MATERIAL & SUBSTANCE	ChEBI [Hastings et al., 2016]	chemical compounds of biological relevance
	ENVO [Buttigieg et al., 2016] :environmental material (ENVO_00010483)	description of environmental entities

Table 6.4: Terminologies that match the defined search categories. If no start node is provided, the entire ontology reflect the category.

corresponding ontologies. We omitted categories which are not specific to the Life Sciences (LOCATION, PERSON, TIME). We only analyzed ontologies from OBO-Foundry to ensure the linkage to other vocabularies. We also only considered ontologies that are still maintained and which are provided in full semantic formats such as RDF and OWL. For the category DATA TYPE only very few ontologies exist, e.g., the Biological Collections Ontology (BCO)[23]. The entry `information content entity, IAO_0000030` provides subnodes that only partially reflect the category DATA TYPE such as journal articles, graphical layouts, and graphs. Therefore, we decided not to consider the category DATA TYPE in our text mining pipeline.

We concentrated on the detection of environmental terms such as habitats or environmental features (ENVIRONMENT), biological, chemical and physical processes (PROCESS), chemical compounds and materials (MATERIAL), phenotypic qualities and characteristics that can be measured or observed (QUALITY). The ENVO ontology [Buttigieg et al., 2016] was used to extract environmental terms. Due to its wide scope, ENVO also contains processes, materials and quality concepts. Therefore, we only considered entities starting from nodes addressing our definition of *Environment* (Chapter 4.2.1).

[23]BCO, `https://github.com/BiodiversityOntologies/bco`

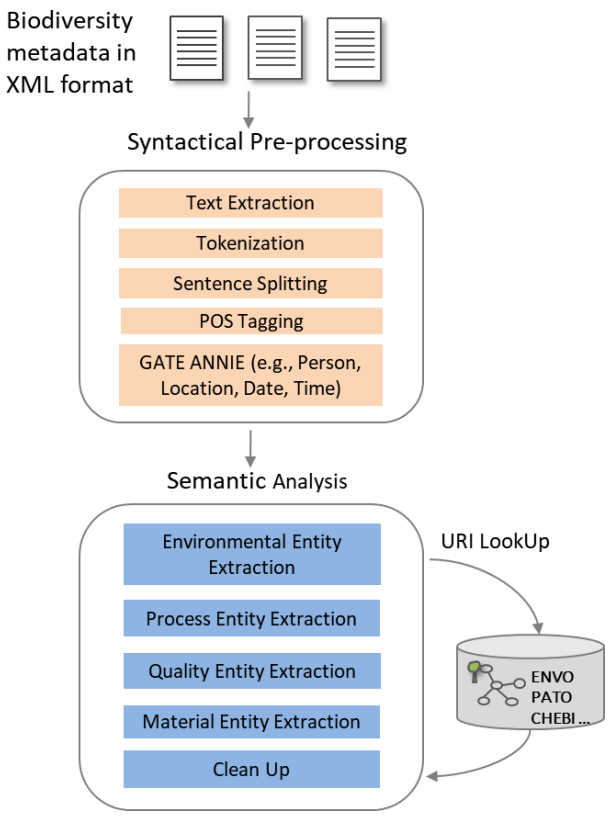


Figure 6.2: Overall flow with all processing resources executed sequentially [Löffler et al., 2020].

For PATO[24] and the detection of phenotypic qualities we also only used concepts from specific nodes as PATO also contains concepts that are too broad for our purpose. We utilized multiple ontologies to cover process entities, e.g., ENVO, Gene Ontology (GO)[25] and UBERON[26]. In order to extract materials, we used the Chemical Entities of Biological Interest (ChEBI)[27] and nodes from ENVO describing environmental materials.

6.3.2 Architecture

We build our approach on the text mining framework GATE [Cunningham et al., 2013] offering basic text mining functions and various plugins for the Life Sciences. The overall workflow is illustrated in Figure 6.2.

In the first syntactical phase, we use GATE's in-build processing steps: At first, the XML structure of the metadata files is analyzed and all textual resources are extracted. Subsequently, tokenization, sentence splitting and Part-Of-Speech (POS) tagging take place. The result are tokens with annotations containing information on occuring noun entities, verbs and adjectives. We also lemmatize the document's text to utilize all inflected forms of nouns (singular vs. plural).

[24]PATO, `http://purl.obolibrary.org/obo/pato.owl`
[25]GO, `http://geneontology.org/`
[26]UBERON, `http://uberon.org`
[27]ChEBI, `https://www.ebi.ac.uk/chebi/`

Following the syntactic pre-processing steps, the semantic analysis is executed. It also represents our own contribution. Each category (entity type) consists of large ontological gazetteer lists. GATE's Large Knowledge Base (LKB) Gazetteer plugin allows access to remote knowledge bases via a SPARQL[28] interface. Instead of using external providers for access to terminologies, we downloaded and host all ontologies in an own *GraphDB*[29] triple store. Therefore, we are not depended on possible maintenance downtimes or ontology upgrades occuring regularly at external terminology providers. All SPARQL queries are stored in text files. The LKB plugin uses these queries and sends them to the SPARQL interface of the *GraphDB* triple store. The result are lists with ontology concepts including URIs and labels.

Listing 6.2: SPARQL query excerpt to retrieve concepts and subconcepts for entity type *Environment* [Löffler et al., 2020]

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix obo: <http://purl.obolibrary.org/obo/>
prefix oboInOwl: <http://www.geneontology.org/formats/oboInOwl#>

SELECT DISTINCT ?la ?entity
FROM NAMED <http://gfbio-git.inf-bb.uni-jena.de/BIODIV/ENVO>
WHERE {
    #environmental system
    { GRAPH <http://gfbio-git.inf-bb.uni-jena.de/BIODIV/ENVO>{
        ?entity rdfs:subClassOf* <http://purl.obolibrary.org/obo/ENVO_01000254>.
        { ?entity rdfs:label ?la.}
        UNION
        { ?entity oboInOwl:hasRelatedSynonym ?la.}
        UNION
        {?entity oboInOwl:hasExactSynonym ?la.}
      }
    }
   #environmental feature
   UNION{
      GRAPH <http://gfbio-git.inf-bb.uni-jena.de/BIODIV/ENVO>{
       ?entity rdfs:subClassOf* <http://purl.obolibrary.org/obo/ENVO_00002297>
       { ?entity rdfs:label ?la.}
       UNION
       {?entity oboInOwl:hasRelatedSynonym ?la.}
       UNION
       {?entity oboInOwl:hasExactSynonym ?la.}
      }
   }
   #environmental condition
   UNION {[...]}
   #immaterial entity
   UNION {[...]}
}
```

The SPARQL queries also integrate semantic relations with the same meaning such as `hasRelatedSynonym` and `hasExactSynonym`. Transducers match the received ontolo-

[28] SPARQL, `https://www.w3.org/TR/rdf-sparql-query/`

[29] GraphDB, `http://graphdb.ontotext.com/`

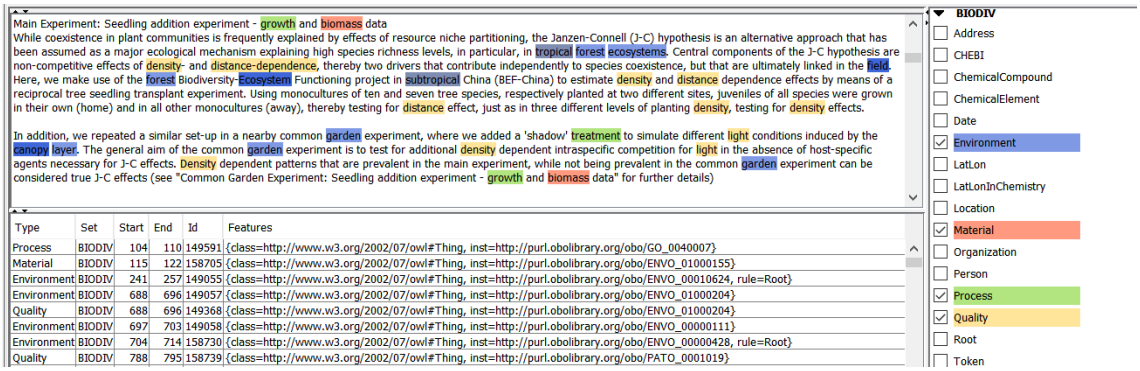


Figure 6.3: Screenshot of GATE's graphical editor presenting a dataset from BEFChina [Germany and Erfmeier, 2019] with semantic annotations and their matching URI concepts [Löffler et al., 2020].

gical lists against the document's tokens. A positive match leads to a semantic annotation with a link to the corresponding resource URI. This look up also considers different constraints such as case-insensitive, word order-sensitive and all possible string matches. We utilize GATE's Chemical Tagger [Cunningham et al., 2013] for the extraction of chemical compounds and elements. Again, by means of the LKB, the annotated tokens are linked to concepts in the CHEBI ontology. ENVO also provides a large amount of environmental materials. Therefore, the Material LKB gathers these concepts via a SPARQL query and the transducer adds them to the annotation MATERIAL. Multiple URI concepts are permitted per annotation. One example is "grassland" being linked to eight URI concepts. This also includes sub-classes, because the sub-class "prairie" contains "grassland" as synonym. In a cleaning step at the end, we eliminate Named Entities such as Person, Location, Organization, Date, Time and Address (extracted by GATE's default ANNIE pipeline [Cunningham et al., 2013]) as they would lead to mismatches. Further erroneously annotated elements such as latitude and longitude (e.g., 'N' is labeled with 'nitrogen') are also removed.

6.4 Evaluation & Results

Figure 6.3 provides a screenshot of GATE's graphical interface and illustrates the resulting semantic annotations. The right panel shows the created annotations for the various categories, whereas the bottom panel presents more detailed information on the individual annotation such as the URI concepts. We ran the BiodivTagger pipeline on the QEMP corpus and compared the pipeline outcome with the QEMP gold standard.

Measurements: Frequently used metrics in information retrieval tasks are precision, recall and f-score [Manning et al., 2008]. For the evaluation of ontology-based information extraction tasks, [Maynard et al., 2006] proposed adapted metrics based on precision

and recall. At first, statistics of 'correct' (exact match), 'missing' (no URI found), 'spurious' (entry in the ontology but not labeled in the gold standard) and 'partial' (ontological concepts cover the tokens in text only partially) matches are computed. Hence, precision, recall and f-Measure in ontology-based information extraction tasks are defined as follows:

$$\text{Precision} = \frac{Correct + \frac{1}{2}Partial}{Correct + Spurious + Partial} \tag{6.1}$$

$$\text{Recall} = \frac{Correct + \frac{1}{2}Partial}{Correct + Missing + Partial} \tag{6.2}$$

$$\text{F-Measure} = \frac{(\beta^2 + 1)P * R}{(\beta^2 P) + R} \tag{6.3}$$

"β denotes the weighting of precision versus recall. If precision and recall should be weighted equally, β is 1. In order to put more emphasis on the precision, β is set to 0.5. If it is set to 2, the recall is twice as weighted as the precision." [Löffler et al., 2020]

The overall performance can be determined with macro and micro measurements. The macro measurement is a single value representing the averaged desired metric with an equal weight to all entity types. In contrast, micro measurement considers the corpus as one large document and is prefered in unbalanced datasets [McCowan et al., 2004].

All metrics introduced above were implemented in a Python script. We processed all corpus files in a batch mode and counted a correct match in case the annotation possessed the correct entity type and at least one URI concept. A partial match was found when the category was correct but the span of the ontology concept and the labeled term partially matched. In case a labeled term in the gold standard did not receive a URI concept, we counted it as 'missing'. The pipeline annotated multiple terms additionally to the gold standard. We considered that as 'spurious' in the evaluation. All scripts and results are available in our Github repository [30].

Outcomes: The precision is a measure to determine how many system generated annotations are correct. In contrast, the recall denotes how many of the expected annotations were actually returned. In information retrieval tasks, the precision is usually considered to be more important. However, in information extraction tasks both metrics have the same importance as the recall implicitly describes the ontological coverage.

We provide the overall results per entity type in Table 6.5. We achieved precison values between 0.423 and 0.589 and recall values between 0.38 and 0.74. Both metrics are moderate as the pipeline annotated lots of additional terms. For instance, for processes the pipeline found twice the number of annotations (spurious) as correct matches.

[30] BiodivTagger, `https://github.com/fusion-jena/BiodivTagger`

	Correct	Missing	Spurious	Partial	Precision	Recall	F-Score
ENVIRONMENT	647	720	559	22	0.536	0.474	0.503
PROCESS	89	148	123	9	0.423	0.38	0.4
MATERIAL	1111	591	1016	2	0.522	0.653	0.58
QUALITY	1414	472	974	75	0.589	0.74	0.656
Macro					0.518	0.562	0.535
Micro					0.549	0.625	0.585

Table 6.5: Results for all documents per category (equal weight of precision and recall) [Löffler et al., 2020].

Moreover, multiple process terms could not be linked to an URI, which resulted in low recall values. For the entity type PROCESS various ontologies are already integrated into the pipeline. However, our results reveal that the ontological coverage is low. The recall values for environmental terms are low, too, as matching terms occur in the ontology only with a different category. One example is the term 'soil'. It got the label ENVIRONMENT in the gold standard, but the ENVO ontology provides an entry under the node 'environmental material' (MATERIAL). The QEMP corpus consists of multiple datasets with soil measurements. Therefore, the number of correct annotations is low and the number of missing annotations is high. In our GitHub repository, we provide a lists of terms that are (from our perspective) misclassified and for which no URI could be found. The BiodivTagger achieved good evaluation results for the recall values of the categories MATERIAL and QUALITY. This reveals that a majority of terms in the gold standard could be linked to concepts in ontologies.

We also determined precision and recall values per data repository (Table 6.6). This also allows the analysis of various dataset length as datasets from *Bexis* and *BEFChina* are longer in terms of size and descriptions than datasets from *PANGAEA*, *Dryad* and *iDiv*. The results show no significant difference between shorter and longer datasets. For *idiv* datasets, eight out of 10 *idiv* files come from the same research group, and the overall ontological coverage in these files were good. This resulted in slightly higher metric values compared to the other repositories. Taking a closer look on the differences of the recall values between biodiversity projects (*idiv*, *BEFChina* and *BExIS*) and data repositories, the evaluation results show that the project repositories achieved slightly higher recall values than the data archives. As the projects are focused on current research questions in biodiversity research, our selection of ontologies is right and confirms their suitability for semantic annotation of biodiversity research data.

6.4.1 Discussion

This study aimed to explore the suitability of biological ontologies for information extraction and NER tasks. Due to the large amount of available biological terminologies, our assumption was that this domain knowledge should be covered by multiple domain onto-

	Correct	Missing	Spurious	Partial	Precision	Recall	F-Score
Dryad	128	76	86	3	0.574	0.589	0.563
PANGAEA	404	268	322	14	0.524	0.603	0.551
BExIS.	571	308	696	26	0.476	0.619	0.514
BEFChina	1921	1169	1418	52	0.566	0.657	0.596
idiv	237	110	150	13	0.614	0.741	0.652

Table 6.6: Evaluation results per data repository [Löffler et al., 2020].

logies. Maintenance and interoperability are essential in ontology management. Hence, we only utilized terminologies being organized by large research communities. This ensures interlinkage among each other and the strictly tailored scope of each terminology.

Basically our assumption is confirmed. Most of the studied entity types being relevant for biodiversity research achieved good recall values. This reveals an overall good coverage of biodiversity terms in available existing ontologies. However, our approach heavily depends on a specific ontology version. All SPARQL queries would require updates in case ontology nodes are moved, renamed or eliminated. Thus, all used ontologies are provided locally in the pipeline as cached versions to countact this problem.

Nevertheless, all metrics leave room for improvement. For instance, there is a large number of missing terms. In particular, for biological, chemical and physical processes the ontological coverage is low. This calls for more discussions in the biodiversity research community to extend existing terminologies. Moreover, misclassifications in the ontologies also resulted in decreased recall values. The precision values are only moderate, as we received a multitude of spurious annotations from the ontological gazetteers. For instance, too broad terms such as “position” or “content” got annotated by the pipeline. We already excluded specific nodes of the ontologies in the SPARQL statements to avoid these spurious broad annotations. However, this needs further analysis and revision. One further solution could be the integration of additional machine learning approaches to increase the precison values, but this would require a larger labeled training corpus.

6.5 Summary

The overall aim of this chapter was to develop and evaluate an approach for the extraction of relevant biological entities from metadata (H_3, second part). Due to the large amount of available terminologies in the Life Sciences, one requirement was to make use of this domain knowledge in the extraction process.

The outcomes of this chapter are two-fold. At first, a metadata gold standard was developed. Fifty metadata files from five data repositories and project databases relevant for biodiversity research were manually labeled with four entity types, namely ENVIRONMENT, PROCESS, MATERIAL and QUALITY. Second, a novel text mining pipeline (BiodivTagger) was developed identifying these categories based on ontolo-

gical gazetteers. For that purpose, multiple ontologies from the OBO Foundry initiative were analyzed and integrated into Large Knowledge Base Gazetteers (LKB) via complex SPARQL queries. Per category various different ontologies were either fully or partially (starting from a specific node) added and cached as local knowledge bases. The pipeline performs a look-up per category in these cached knowledge bases and links the ontology entries with the tokens occuring in metadata in case of a successful match. Thus, semantic annotations are created consisting of concept URIs, spans with start and end nodes and the category information.

The evaluation of the BiodivTagger pipeline shows moderate results. We achieved good results for the recall values for the categories MATERIAL and QUALITY (data parameters), but moderate precision and recall values for the other entity types. However, it is the very first text mining pipeline for the extraction of important entity types for metadata in biodiversity research. Moreover, the selection of the currently used ontologies is right and suitable for semantic annotation tasks. Now, further work is needed to improve the current approach, e.g., with additional machine learning techniques or ontology extensions.

Chapter 7

Semantic Dataset Retrieval

"Intelligence is not the ability to store information, but to know where to find it."

- attributed to Albert Einstein, *Physicist and Nobel prize winner*

Search engines are the major component in retrieval applications. Improving search engines towards results going beyond keywords is the main goal in the research field of semantic search. One approach for such as enhanced search are entity-based retrieval models. Data sources and queries are represented by entities enabling a search over text and knowledge bases.

After the presentation and evaluation of the pre-processing component (Chapter 6) analyzing and annotating metadata with important biological entity types and URIs of the LOD cloud, we now present retrieval component, the second component of our proposed semantic dataset search. We explore various entity expansion techniques, various entity-based retrieval models, and we also introduce a novel test collection for the evaluation of dataset retrieval. At first, we present related work about entity-based retrieval approaches and available test collections in the Life Sciences (Section 7.1) followed by some preliminary considerations (Section 7.2). Afterwards, we describe a small user study to gather requirements for a semantic retrieval system in biodiversity research (Section 7.3) and present the a novel test collection for dataset retrieval in Section 7.4. The expansion strategies are introduced in Section 7.5, and the explored retrieval models are described in Section 7.6. We present the evaluation setup in Section 7.7 and the results in Section 7.8.

The test collection for dataset search in biodiversity research was published as short paper in the RIO Journal [Löffler et al., 2021].

7.1 Related Work

Entity-based retrieval is part of research in the field of semantic search [Balog, 2018, Bast et al., 2016]. Entities are usually represented by Internationalized Resource Identifiers (URIs). According to Balog [Balog, 2018], the word 'entity' is mainly used in terms of Named Entities such as geographic locations (e.g., "Atlantic Ocean") or persons (e.g., "Jane Goodall"). Abstract objects such as data parameters (e.g., "length", "growth rate") or vegetation layers (e.g., "shrub layer") are mainly referred to as 'concepts'. In this work, we use the words 'entity' and 'concept' synonymously. In contrast to the classical bag-of-words representation of text such as in the Vector Space Model [Salton et al., 1975] or BM25 [Sparck Jones et al., 2000] (Chapter 2.2.2), entity-based retrieval models utilize entities extracted from keywords as representations of text. Words are linked to entities by entity linking approaches [Guo et al., 2009, Meij et al., 2011, Hasibi et al., 2015, Blanco et al., 2015, Carmel et al., 2014]. These linkages are utilized to incorporate additional related information from knowledge bases to documents and queries. That allows for a richer textual representation of query terms by including synonyms, or in case of disambiguation explicit knowledge can be used to select the correct sense. In the following, we present related work in the field of entity-based retrieval models for document retrieval. At first, we focus on non-domain specific entity-based retrieval approaches (Subsection 7.1.1), followed by entity-based retrieval approaches in the Life Sciences (Subsection 7.1.2). Afterwards, we introduce available test collections in the Life Sciences in Subsection 7.1.3.

7.1.1 Entity-Based Retrieval and Ranking

Driven by the emergence of knowledge graphs such as DBpedia[1] and Wikidata[2], research on entity-based retrieval has increased. One of the earliest approaches was proposed by Goosen et al. [Goossen et al., 2011] employing the Vector Space Model on ontological concepts (CF-IDF) in a news personalization system. The term 'bag-of-entity' representation was introduced by Xiong et al. [Xiong et al., 2016] aiming at a simple entity-based model. They suggested two ranking scores: Coordinate Match (COOR) (documents were ranked by the number of query entities contained in a document) and Entity Frequency (EF) computing a document score based on the frequency of query entities occuring in the document.

These models showed promising results but did not consider any additional knowledge about the represented entity in the query. A first entity-based retrieval approach exploiting further semantic relations was the model presented by [Waitelonis et al., 2015]. Inspired

[1] DBpedia, `https://www.dbpedia.org/`
[2] Wikidata, `https://www.wikidata.org`

by the General Vector Space Model (GVSM) multiplying document and query vectors with measures of semantic relateness [Tsatsaronis and Panagiotopoulou, 2009], Waitelonis extended the GVSM and replaced terms with entities. Document and query vectors were expanded with semantic relations. For each entity, the set of its classes from the YAGO[3] knowledge base was incorporated into the vectors (e.g., entity: `dbp:Neil_Armstrong`, classes: `yago:Astronaut`, `yago:Person`). The model distinguished between exact matches and related matches and gave more credit to exact matches. To determine if two entities are an exact or a related match, the semantic similarity between entities was computed with Resnik's relatedness measure [Resnik, 1999]. This result served as weight to express the relevance of a class to its entity. [Nishioka and Scherp, 2016] also suggested a similar approach assuming that knowledge bases have hierarchical structures. Therefore, they proposed HCF-IDF, a hierarchical ranking utilizing the information on which hierarchy level a concept and its neighbouring concepts were located in an ontology or taxonomy. This idea avoided providing too high weights to generic concepts and was developed for short text messages in social media.

A novel approach using entity embeddings to determine features for ranking was proposed by [Xiong et al., 2017c]. In order to improve the ranking of scientific publications on the popular platform *Semantic Scholar*[4], they constructed a knowledge graph from publications indexed in *Semantic Scholar* and Freebase[5] containing entities, descriptions, context and relations to authors and venues. Entity embeddings were trained from the graph structure and were used to determine the semantic relatedness between query and document entities. By means of two pooling steps the entity matches were used as ranking features. Finally, a learning to rank model was applied to compute the final document score. A variety of subsequent approaches exploited this idea to improve the model to combine entity linking and entity-based ranking [Xiong et al., 2017b], to incorporate words and entities [Xiong et al., 2017a] and to study the importance of entities [Li et al., 2019].

7.1.2 Entity-Based Retrieval and Ranking in the Life Sciences

The Life Sciences can build on a very active research community providing and curating domain knowledge in ontologies and controlled vocabularies. In May 2020, one third of the 1,301 datasets in the LOD cloud came from the Life Sciences [McCrae et al., 2020], in particular the biomedical domain. Hence, a couple of retrieval approaches have been developed for scientific articles, clinical trials and health forum posts exploiting this additional knowledge.

[3] YAGO, `https://yago-knowledge.org/`

[4] Semantic Scholar, `https://www.semanticscholar.org/`

[5] Freebase was a large knowledge graph, but its development was stopped in 2016. It is partially integrated into Wikidata.

Most approaches were based on query expansion [Dietze and Schroeder, 2009], [Stoyanovich et al., 2011, Sy et al., 2012, Berlanga et al., 2015, Löffler et al., 2017], [Faessler and Hahn, 2017, Ernst et al., 2019] linking query keywords to ontological concepts and considering synonyms [Dietze and Schroeder, 2009, Löffler et al., 2017], [Faessler and Hahn, 2017] or further hierarchically related concepts such as hyponyms (sub-classes, narrower concepts, descendants) and hypernyms (super classes, broader concepts or ancestors) [Stoyanovich et al., 2011, Sy et al., 2012, Berlanga et al., 2015]. Other approaches also considered the entity type in the document vector [Ernst et al., 2016, Ernst et al., 2019, Terolli et al., 2020] to improve the retrieval of scientific publications, clinical trials and health forum posts. In the query and documents vector they used TF-IDF weights of the query keywords, entities and related categories (entity types). [Terolli et al., 2020] gave evidence that entity expansion outperforms classical keyword search. Focused entity expansion based on a greedy method further improved the retrieval of clinical trials and health forum posts.

As query expansion increases recall [Manning et al., 2008] but potentially lowers precision, the consideration of hierarchy relations requires additional approaches to improve the precision. One option is to compute the semantic similarity between query and document concepts [Stoyanovich et al., 2011, Sy et al., 2012] and to consider not only exact matches but also all descendants. [Berlanga et al., 2015] proposed a language model based on the knowledge resource that aims at the identification of concepts that are related to the target corpus, and [Ernst et al., 2016] used boost factors to give more credit to the original keywords. Entities and categories obtained smaller boost factors.

All entity-based approaches in the Life Sciences are focused on the retrieval of scientific articles, clinical trials or health forum posts. In contrast, dataset search aims to retrieve scientific data based on metadata. Metadata descriptions vary greatly in their scope. In common data archives such as *Dryad*[6], *EBI/ENA*[7] or *Zenodo*[8] metadata descriptions can be very rich or sparse. Entity-based expansion and retrieval might alleviate the current problems in dataset search such as different spellings and the usage of different terms in data descriptions. However, studies are missing that explore entity expansion approaches for dataset retrieval in the Life Sciences and that compare different semantic ranking approaches.

7.1.3 Test Collections in the Life Sciences

As very few approaches are available providing semantically annotated questions or data with entities in the Life Sciences, we introduce all publicly available test collections in the Life Sciences based on scientific articles and datasets.

[6]Dryad, `https://datadryad.org`
[7]EBI/ENA, `https://www.ebi.ac.uk/`
[8]Zenodo, `https://zenodo.org/`

The Genomics Track Challenge [Hersh and Voorhees, 2009] was based on PubMed articles and natural language questions. The questions were marked with entity types being relevant for the biomedical domain such as [PROTEINS], [GENES] or [DISEASES] (e.g., "What [GENES] are involved in the melanogenesis of human lung cancers?"). The provided binary human assessments denoted the relevance of a document to a question. The annual BioASQ Challenge [Tsatsaronis et al., 2015] has a focus on Question Answering [Unger et al., 2014] and consists of three competition tasks. The first one comprises the extraction of entities. In the second task, the participants have to convert natural language questions into semantic formats such as RDF triples [9]. Finally, in the last task, the aim is to find the exact answer to a natural language query. The document corpus consists of PubMed articles, and the topical focus in the questions is on biomedical entity types such as diseases, genes, proteins, species and drugs.

To the best of our knowledge, the only available test collection for dataset search is the bioCADDIE Test Collection [Cohen et al., 2017b]. The data corpus comprises 794,000 biomedical metadata files from various data repositories. The provided 137 questions with information needs from biomedicine were created by domain experts. The experts got question templates with the advice to consider specific entity types, such as data type, disease type, biological processes and organisms. The data corpus was indexed in multiple search engines, and 15 questions were selected for the final test collection. For each question, two runs in each search engine were performed, and the results were merged. Annotators with biomedical expertise judged the relevance of the top ranked results and determined whether a dataset was relevant, partially relevant or not relevant.

To the best of our knowledge, there is no test collection available for dataset search in biodiversity research. Therefore, we developed a novel test collection for this domain being presented in Section 7.4.

7.2 Preliminary Considerations

Search queries with tailored entity expansion resulted in more relevant scientific publications, clinical trials and health forum posts than search queries without entity expansion [Terolli et al., 2020]. This also correlates with our study on the comparison between a keyword search without query expansion and a semantic search in which keywords were linked to entities. The queries were expanded with synonyms and further related terms obtained from the hierarchy (Chapter 3.3). However, expanding only the query terms does not represent a semantic full text search. In order to allow a search over entities and full text, entities and their respective types need to be added to the index. Concept-based retrieval techniques introduced in Subsection 7.1.1 enable such a search on entities and full text. Entities with URIs also allow the incorporation of further semantic relations,

[9] RDF Primer, `https://www.w3.org/TR/rdf11-primer/`

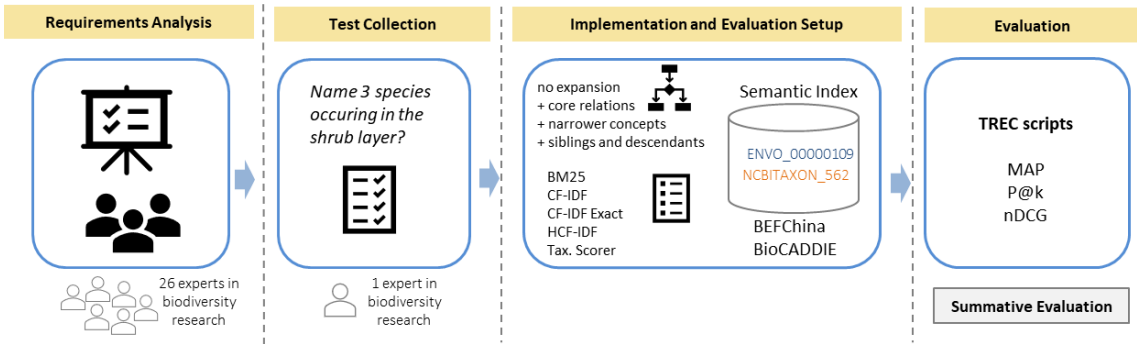


Figure 7.1: Overall flow of the retrieval study.

such as `subClassOf` (hierarchy relations) or `partOf` (core relations). The BFO ontology [Smith et al., 2005] and the Relation Ontology (RO)[10] are upper ontologies providing a basic set of semantic core concepts and relations. These core relations are utilized in multiple Life Science related ontologies, e.g., in the ontologies of the OBO Foundry initiative[11], a framework of terminologies being interlinked among each other and containing only strictly-tailored content. Hence, our main idea for the retrieval component is to combine these different considerations in one component. We aim to:

1. Explore different entity expansion strategies with various hierarchy and core semantic relations on datasets for biodiversity research.

2. Explore different entity-based retrieval models for dataset search in biodiversity research.

We introduce them in detail in Section 7.5 and Section 7.6. Both research questions address H_4 and Requirements *R1*, a dataset search beyond keywords, and *R2*, a search in domain specific categories (Chapter 5.3). Following Requirement *R5* (user-centered design), we conducted a user study and collected user preferences for entity expansion and the ranking of datasets in the search result. Due to the lack of existing test collections for dataset search, we also created a novel metadata test collection for biodiversity research in cooperation with a domain expert in biodiversity research.

Figure 7.1 presents the overall flow of the following study. At first, we collected preferences on expansions and rankings of datasets among scholars (Section 7.3). Next, we generated a novel metadata test collection for biodiversity research (Section 7.4). Based on these insights, we studied various expansion strategies (Section 7.5) and entity-based retrieval models (Section 7.6) being suitable for dataset search in the Life Sciences. Our implementation is based on the GATE framework [Cunningham et al., 2013], and the evaluation was carried out on our novel dataset test collection and on the bioCADDIE

[10] RO, `https://oborel.github.io/`
[11] OBO Foundry, `http://www.obofoundry.org/`

test collection (Section 7.7). The results of this summative evaluation and the discussion are described in Section 7.8.

7.3 User Study for Requirement Analysis

In order to better understand user preferences for entity-based retrieval in dataset search, we conducted a user survey among biodiversity experts from November 2018 to March 2019.

7.3.1 Methodology

Presenting scenarios to a target group is a common technique to capture functional user requirements [Sommerville, 2021]. Usually, it is difficult for users to specify requirements. Confronting them with concrete examples from their daily working life is therefore easier to identify obstacles and improvements. As biodiversity research is very heterogenous research field and scholars' backgrounds and expertise vary greatly (Chapter 4.2), we decided to provide search scenarios on a higher, general level and to tell participants to think about concrete examples from their own research perspective. However, to enhance imagination and to illustrate theses general search scenarios, we listed some concrete search cases as examples.

In an online survey, 26 biodiversity experts from collections, museums and further biodiversity research related scientific projects (in the scope of the GFBio [12] project) assessed four general search scenarios with five questions each. Three questions focused on possible expansion strategies, ranking preferences and further topical recommendations in search and provided pre-defined answers users had to select. Two questions were open questions and allowed the participants to give qualitative feedback. All questions were optional and participants could skip questions they did not want to answer. For some questions, it was also permitted to provide multiple answers. The five questions were provided as follows: The four scenarios comprised searches for organisms, environmental terms, data parameters and materials. These categories are the most frequently mentioned topics in the analyzed search questions (Chapter 4.2).

1. What primary data and metadata do you need in the result set when searching for organisms (environmental terms | data parameters | materials)? *Possible answers: [exact matches, synonyms, more specific terms, broader terms]*

2. What further information should be integrated in the search result? *[open question]*

[12] GFBio, `https://www.gfbio.org`

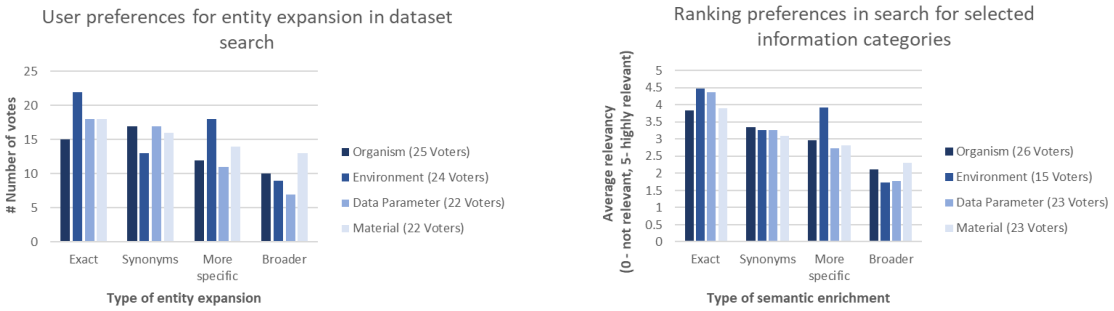


Figure 7.2: User preferences with respect to entity expansion (left) and ranking (right) in search results.

3. What primary data and metadata do you want to be ranked on top when searching for organisms (environmental terms | data parameters | materials)? *Possible answers: [exact matches, synonyms, more specific terms, broader terms]*

4. Would you be interested in recommendations with related information when searching for organisms (environmental terms | data parameters | materials)? *Possible answers: [environmental information | species | data parameters | processes | materials]*

5. What further information should be recommended when searching for organism (environmental terms | data parameters | materials)? *[open question]*

7.3.2 Results

Figure 7.2 (left) presents users' assessments on entity expansion preferences. The participants have the assumption that all types of hierarchical **entity expansions** are relevant for dataset search in biodiversity research. However, there are slight different preferences for the four given categories. When searching for datasets with organisms, users would prefer to see results containing synonyms (17 votes) and exact matches (15 votes) over datasets with more specific concepts (12 votes) or broader concepts (10 votes). In contrast, in a search for environmental terms, biodiversity experts would like to see exact matches (22 votes) and narrower concepts (18 votes) on top than datasets containing synonyms (13 votes) or broader concepts (9 votes). For dataset search scenarios with a focus on data parameters, users would prefer to obtain datasets with exact matches (18 votes) and synonyms (17 votes) than datasets with narrower concept (11 votes) or broader concepts (7). For a search for materials, the values differ only slightly between 18 votes for exact matches and 13 votes for datasets containing broader concepts. It stands out that for materials an expansion with broader concepts obviously matters more than for other scenarios. In the scenario on materials, we provided the following concrete examples: *What*

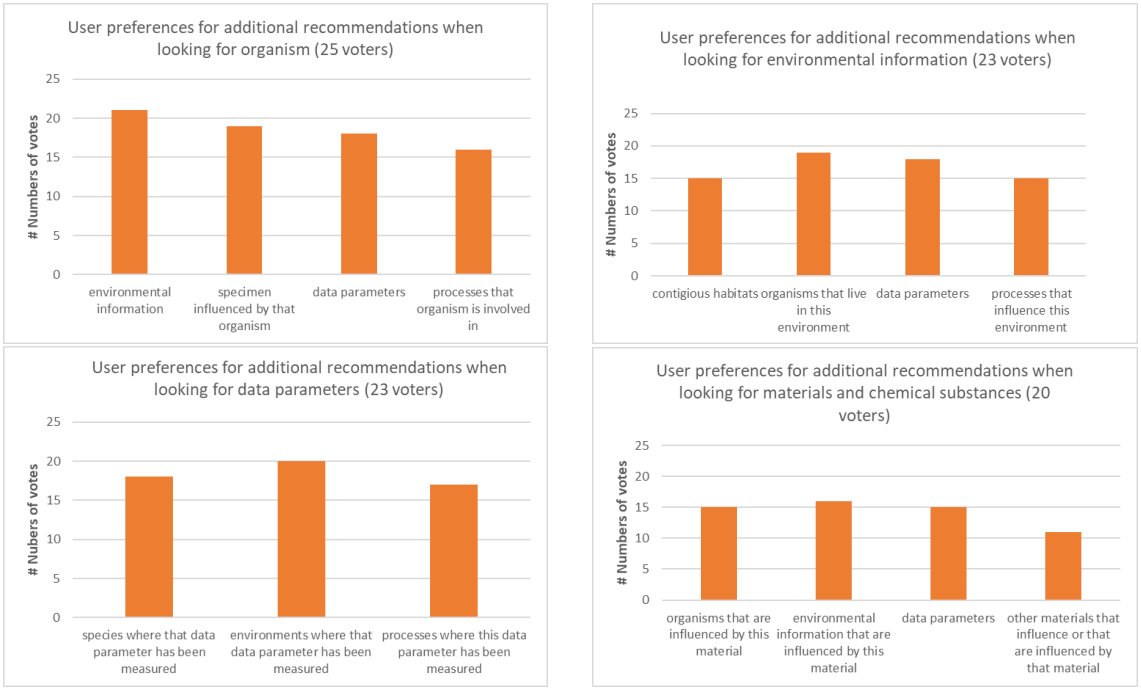


Figure 7.3: Results of user preferences with respect to recommendations on further related contextual information.

primary data and metadata do you need in the result set when searching for material substances (chemical compounds)? (a) exact search terms, e.g., 'neonicotinoid insecticide', (b) synonyms such as chemical formulas, different descriptions, (c) more specific results, e.g., 'acetamiprid', 'clothianidin', (d) broader terms such as 'neurotoxin'. Scholars answered that they would like to see results containing concepts from a higher hierarchy level such as 'neurotoxin' and not only datasets containing synonyms or narrower concepts.

Figure 7.2 (right) shows a decreasing linear progression. **Ranking preferences** for all categories got the highest values for exact matches, followed by synonyms and descendant concepts. Expansions with broader concepts should be listed at the end of search results. The only exception are environmental terms. Only 15 experts took part in the voting for this category. We assume that participants from the GFBio project dropped here as they are more familiar with collection data. However, these 15 participants all agreed that expansions with descendants should be higher ranked than synonyms.

Per search scenario, we also asked the scholars whether they would like to receive **recommendations** on further related contextual information with respect to the search terms (Figure 7.3). For instance, when looking for datasets with species, we wanted to know whether it would be helpful for scientists to see supplement information on the environment in which species live or relations to other species. It turned out that all provided recommendation options are relevant for biodiversity research. However, the frequency values for all given recommendation options are very close to each other.

Search Scenario	Additional search preferences	Recommendations
Organism	"Common terms and family name", "additional information on the species is welcome.","First descriptor is important information to validate species names.", "distribution","geographical locations, ecosystems","Essential: selector on data quality (spatial accuracy; presence vs. presence/absence inventory; time interval; ...)", "What sort of data (e.g. abundance, presence, collected specimen, etc.)? Where/Who/When collected.", "Conservation patterns if any and ecological specificities"	"Source (Observation, literature, ...)", "When applicable, short notice on ecosystem function of e.g. larvae abundance", "distributional ranges (e.g., with map), closest relatives, information of possible hybridisation/ different ploidy levels within the species, refugium during ice ages, elevational range, Ellenberg Indicator values for plants.", "distribution of species","Resource preferences, habitat preferences, etc.", "population dynamics and trends, invasive alien species...", "When proposing environmental parameters (e.g. bioclimatic conditions) some background should be available on whether the information comes from actual field observations, or instead from laboratory measurements, or other sources"
Environment	"phytosociological information needed. elevational range and exact location (e.g., North America or South Africa) would refine the term grassland.","key species", "If part of a classification (e.g. if it is a class of a given land-cover layer), background on the classification scheme (minimum: list of classes, and bibliographic reference) should be available", "location (country or continent)" , "specific values (incl. ranges) for specific parameters. I.e. for terrestrial habitats: habitat types, vegetation, soil parameters, climate etc."	"phytosociological information needed. elevational range and exact location (e.g., North America or South Africa) would refine the term grassland.", "elevation, climate parameters", "Maybe how these habitats are linked to/classified in global data layers of habitat cover/land use", "land cover/use, climate date", "role in N (nitrogen), P (phosphorus), K (potassium) or carbon cycle", "Mean annual precipitation, mean annual temperature","(range) values for specific parameters"
Data parameters	"units of the measurments and the range of measurements recorded previously. definitions of the terms (sometimes not clear what is reported exactly)", "Parameters from individual or community sampling? For example, root biomass might be from 1 plant, or from a soil core taken in a habitat."	"as mentioned the range of the measurements. Time of the day/ phenological stage when the measurement has been recorded","used method - how the parameter was measured", "soil parameters", "Are there repeated measures over a given time, or one-time measure only?"
Materials	"for medicial use is needed or effects of the compounts as well as derivates", "Maybe concentrations or measurement of the amount?", "scientific (peer reviewed) results, or authoritative opinions clearly flagged"	"just the effects on other organisms, possible uses, other products related to that compound.", "suspicions for which no scientific evidence is yet available.", "See previous answer on environmental conditions (do they come from field observations? Laboratory analysis? Other? This piece of information should be available, with supporting literature)", "Reason product is used in a given environment."

Table 7.1: Selected user suggestions for topical extensions in dataset search

We also collected qualitative feedback in comment fields. User had the opportunity to provide feedback on topical extensions in search. Table 7.1 presents selected user comments sorted into additional search preferences (semantic extensions users would like to see in a search result but that are not listed in the previous answer options) and recommendations (information that would be helpful for data exploration and that could be expanded on demand). For organisms, user prefer to get more information on species, such as "common terms", "family name" or the "first descriptor". Depending on the scientific background and research questions, criteria also play an important role, e.g., "selector on data quality", geographic location and distribution. As further relevant and related content users listed "distribution", threats and nutrition of organisms. For the second search scenario - environments, the suggestions for the search preferences vary greatly and range from "location" and "elevation" to "habitat type", "vegetation" and "soil parameter". Recommendations could be information on "land cover/use", climate parameters and "the role of N (nitrogen), P (phosphorus), K (potassium) and carbon cycle". In the scenario about data parameters, the participants would like to see information on the "units of measurements and the range of measurements recorded" and whether the measured data come from individuals or communities. As additional topic they would like to get information about the frequency of measurement (one-time or repeated measures). For the search scenario about materials, users emphasized the need for synonyms as "many chemicals have common names or manufactured brand names". The survey results are publicly available [Löffler et al., 2022a].

Overall, all proposed entity expansion and ranking preferences seem to be relevant for scholars in biodiversity research. The relevance of the different entity expansion strategies vary slightly between the individual categories. For the ranking preferences, the result is a bit clearer. The stronger preference is given to exact matches followed by synonoyms. However, the frequency values for narrower and broader concepts are also not too far away and therefore can not be neglected. Hence, we conclude, that a more detailed and summative evaluation is necessary to confirm or to correct users' assessments.

7.4 BEF-China Test Collection

We utilized all publicly available metadata files of the BEF-China (Section 2.4) project as data corpus of the test collection. The BEF-China project aims to explore Biodiversity-Ecosystem Functions (BEF) in a species-rich forest in the subtropics [Bruelheide et al., 2011, Bruelheide et al., 2014]. The collected primary data are described by 372 metadata files in EML metadata schema[13]. An excerpt of a BEF-China metadata file can be found in the publication [Löffler et al., 2021].

[13]EML, `https://eml.ecoinformatics.org/`

Question number	Question	# datasets being relevant to this question
Q1	Name 3 species that occur in the shrub layer.	16
Q2	Find 3 plant species where root lengths (depth) have been considered	1
Q3	Find 3 datasets from oaks where nitrogen content have been measured.	18
Q4	Find 3 datasets where dry weights from conifers have been measured.	5
Q5	Which nutrients occur in soil?	20
Q6	Identify all parameters that are correlated to soil depth.	24
Q7	Which taxa associated with tree species have been found, for example, insects on host trees?	46
Q8	Which soil samples in BEF-China data show a low pH value?	6
Q9	Does tree diversity reduce competition?	40
Q10	Do the soil carbon concentrations increase with soil depth?	6
Q11	Are there data about the leaf area index (LAI) and, in particular, in combination with diversity?	8
Q12	How has tree height been measured in BEF-China experiments?	25
Q13	How does the nitrogen cycle interact with water?	20
Q14	How significant is the role of throughfall as water input to the forest floor?	4

Table 7.2: BEF-China question corpus [Löffler et al., 2021]

In the novel test collection, we aimed to ensure two requirements: First, we wanted to use search questions reflecting real world information needs from biodiversity research, and second, the data corpus need to contain datasets being relevant to the selected questions. Based on these considerations, we selected six questions from the question corpus introduced in Chapter 4. As six questions are too few questions, we created question templates of the question corpus (e.g., <PROCESS> influences <ENVIRONMENT>) and generated eight more questions focusing on ecosystem functioning. Table 7.2 lists the final questions and datasets being relevant to the question.

Besides the questions and a data corpus, human assessments are necessary in a test collection to determine what dataset is relevant or not relevant to a question. We asked the data manager of the BEF-China project to provide these assessments. He went through all datasets and 14 questions, which resulted in 5208 binary relevance judgments (14 questions x 372 datasets). 239 of these 5208 relevance judgments were marked as relevant or partially relevant. We provide the relevance judgments in the TREC benchmark data format *<Question number> <Iteration> <Dataset#> <Relevance judgment>*[14].

[14]BEF-China test collection, `https://github.com/fusion-jena/befchina-test-collection`

Ontology	Object property	Number
ENVO [Buttigieg et al., 2016]	partiallySurroundedBy (ENVO:01001307)	35
	surroundedBy (RO:0002219)	23
PATO [Gkoutos, G. et al., 2022]	pato#increased_in_magnitude_relative_to	108
	pato#decreased_in_magnitude_relative_to	107
CHEBI [Hastings et al., 2016]	hasRole (RO:0000087)	34119
	chebi#has_functional_parent	13419
UBERON [Mungall et al., 2012]	partOf (BFO:0000050)	15313
	developsFrom (RO:0002202)	2180
GO [Ashburner et al., 2000, GO, 2021]	partOf (BFO:0000050)	10893
	regulates (RO:0002211)	7048

Table 7.3: The two most frequent occuring object properties of selected knowledge resources in the Life Sciences.

An example entry looks as follows:

1 0 161 1

Datasets being not mentioned for a question are considered to be not relevant and therefore are not listed in the relevance file.

7.5 Entity Expansion Strategies

Knowledge resources are terminologies describing the knowledge of a specific domain. Most of them are organized in tree-based (hierarchical) structures and consist of is-a or subClassOf relations. In addition, some of these terminologies provide further semantic relations between entities (object properties). We analyzed nine ontologies with respect to the available object properties and their frequencies. An excerpt of the two most frequent object properties in selected ontologies in the Life Sciences is provided in Table 7.3. The full object property metrics of these ontologies are in our GitHub repository[15]. So far it has not been studied whether the usage of these relations or the usage of the hierachical structure of the ontology leads to more relevant results in entity expansion.

Basic semantic relations between entities are defined in the Basic Formal Ontology (BFO) [Smith et al., 2005], an upper-level ontology that describes core entities and their relations. It is a non domain-specific ontology and only provides basic constructs to develop domain-specific ontologies. The Relation Ontology (RO)[16] is based on BFO 2.0 [Smith et al., 2015] and provides in its core set[17] foundational relations that can be used to define domain and range axioms in domain-specific ontologies. Examples for such basic relations are for instance *hepatic cell* (BTO:0000575) that isPartOf (BFO:0000050) *liver* (BTO:0000759) or *calcium atom* (CHEBI:22984) hasRole (RO:0000087) *macronu-*

[15]Ontology property metrics, https://github.com/fusion-jena/entity-retrieval-evaluation/tree/main/Expansion/ontology_relation_metrics

[16]RO, https://oborel.github.io/

[17]RO core, https://github.com/oborel/obo-relations/wiki/ROCore

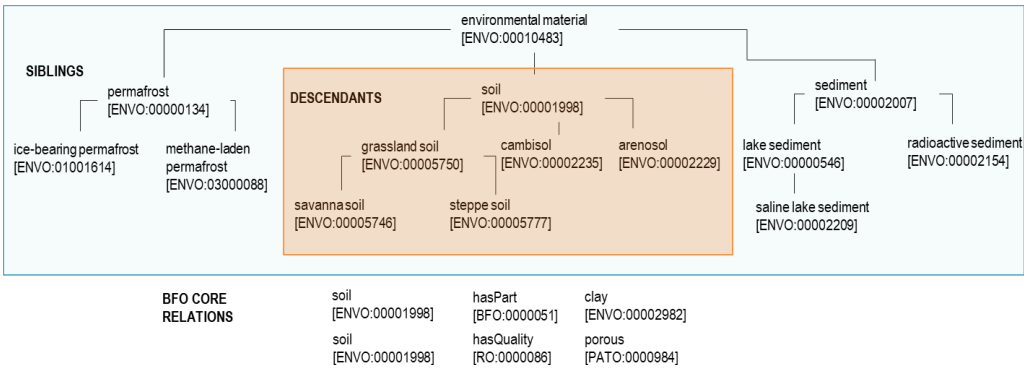


Figure 7.4: Excerpt of the Environment Ontology (ENVO) for the concept 'soil' with its hierarchical and core relations.

trient (CHEBI:33937). As search is usually an ad-hoc task consisting of less keywords, entity expansion on this additional implicit knowledge, which users might possess but do not enter explicitally, could lead to more relevant results. Hence, as first expansion strategy we utilize all object properties defined in the core set of the Relation Ontology including object properties such as `isPartOf` (BFO:0000050), `hasRole` (RO:0000087) and `locatedIn` (RO:0001025). In addition to the **BFO core relations** we also want to explore hierarchical expansions, because ontologies in the Life Sciences mainly consist of hierarchy relations. They are determined with the object property `rdfs:subClassOf`. As informations needs in biodiversity research can contain terms of a broader scope such as 'vegetation' or 'bacteria', expansions on **descendant concepts** and **siblings and their descendants** might return more relevant datasets in the result set. An example for all expansion strategies utilized in the following evaluation is provided in Figure 7.4. The baseline is the original query keyword linked to a corresponding URI in an ontology. An entity expansion contains the original concept URI + the entities resulting from the respective semantic relation.

7.6 Evaluated Entity-based Retrieval Models

We aimed to make use of the large amount of domain knowledge in the Life Sciences. Therefore, we selected entity-based retrieval approaches from related work (Section 7.1) that enable a flexible usage of different semantic relations of a concept in the retrieval process and that differ between exact matches (exact URI in query and metadata) and related matches (URIs in query and metadata are related, e.g., a child concept or a parent concept). We only considered entities. We omitted the original keywords, and we did not use entity type information, because [Terolli et al., 2020] have shown that entity type information did not result in an improved retrieval.

CF-IDF The adapted concept-based TF-IDF weighting scheme, *CF-IDF*, was introduced by [Goossen et al., 2011] for the recommendation of news items. In CF-IDF, entities or concepts of knowledge bases are linked to keywords. The frequency of each concept within a document and across the entire corpus determines a weight of interest in this document. This approach permits to include synonyms instead of using the plain keyword in the weighting scheme. We used this idea for metadata files as presented in Equation 7.1. The weight of a concept c in a metadata file m is composed of the *concept frequency* $cf_{c,m}$ (Equation 7.2) and the *inverse metadata frequency* idf_c (Equation 7.3). The *inverse metadata frequency* denotes how often a concept c appears in all metadata files $|M|$ in the corpus.

$$w_{cf-idf}(c,m) = cf_{c,m} * idf_c \tag{7.1}$$

$$cf_{c,m} = \frac{n_{c,m}}{l_m} \tag{7.2}$$

$$idf_c = \log \frac{|M|}{|m : c \in m|}, \tag{7.3}$$

There are a number of variants for the computation of the *concept frequency* [Manning et al., 2008]. We computed the *concept frequency* as presented in Equation 7.2 where $n_{c,m}$ denotes the number of occurrences of a concept c in metadata file m and l_m denotes the character length of metadata file m. This adaption mitigates the problem of higher concept frequency values in longer documents [Manning et al., 2008].

HCF-IDF The hierarchy-based retrieval model HCF-IDF [Nishioka and Scherp, 2016] was developed for recommending social media items such as short text messages based on user interests. Metadata are usually also short text passages, and Life Science vocabularies are mainly hierarchy based. Hence, HCF-IDF is worth considering for dataset search. Each weight $w_{hcf-idf}(c,m)$ for a concept c in a metadata file m is computed by a BellLog function $BL(c,m)$ [Kapanipathi et al., 2014] and the classical IDF part that denotes how often the concept appears over all metadata (Equation 7.4). The BellLog function (Equation 7.5) is a recursive computation that weights the frequency $cf_{c,m}$ of a concept c based on its hierarchy level l in the knowledge base where $FL(c) = \frac{1}{log_{10}(nodes(h(c)+1))}$. *Nodes* returns the number of concepts that occur on the same level as the given concept, $h(c)$ denotes the hierarchy level of a given concept c in a knowledge base and $C_l(c)$ is the set of child concepts of concept c at hierarchy level l.

$$w_{hcf-idf}(c,m) = BL(c,m) * \log \frac{|M|}{|m : c \in m|} \tag{7.4}$$

$$BL(c,m) = cf_{c,m} + FL(c) * \sum_{c_j \in C_l(c)} BL(c_j,m) \tag{7.5}$$

Linked data enabled General Vector Space Model [Waitelonis et al., 2015] introduced a model that takes taxonomic relations into account. It used DBpedia as knowledge graph in which entities are instances of classes (`rdf:type` property) and sub-classes and super classes are identified with `rdfs:subClassOf` property. As sub-class relations dominate Life Science vocabularies, we also selected this retrieval model for our analysis of metadata files. The weight for a given term vector $\overrightarrow{t}$ or a weight $w_{tax}(e,m)$ in a metadata file m is computed by the entity vector $\overrightarrow{e}$ of the linked entity and the set of its related classes (concepts) $c(e)$ (Equation 7.6). The weight $w_{tax}(c,e)$ determines the influence of a related concept c on entity e. It represents a semantic relatedness measure and is computed by Resnik's information content *IC* [Resnik, 1999]. α_e and α_c are factors (α_e,$\alpha_c \in [0,1]$:) that determine how strongly exact matches of the query are preferred over matches of related entities (Equation 7.7).

$$w_{tax}(e,m) = \alpha_e e + \alpha_c \frac{v}{|v|}, v = \sum w_{tax}(c,e) \cdot c \tag{7.6}$$

$$\alpha_e = \frac{\alpha}{\sqrt{\alpha^2 + (1-\alpha)^2}}, \alpha_c = \frac{1-\alpha}{\sqrt{\alpha^2 + (1-\alpha)^2}} \tag{7.7}$$

The final scoring of retrieved datasets for a query q followed the *overlap score measure* [Manning et al., 2008] and was determined by the sum over all entity weights $w_{tax}(e,m)$ in a metadata file m (Equation 7.8).

$$score_{q,m} = \sum_{e \in q} w_{tax}(e,m)) \tag{7.8}$$

7.7 Evaluation Setup

We aim to explore entity-based dataset retrieval along two dimensions: (1) entity expansion on core and hierarchical relations and (2) different entity-based retrieval models. The overall evaluation architecture is depicted in Figure 7.5. For our experiments we utilized two metadata corpora from the Life Sciences. In a pre-processing phase, important biological terms in the metadata files were linked with concepts from Life Science ontologies with NLP pipelines. Afterwards, all annotated metadata files were added to a semantic index. In a second phase, prepared and manually annotated search tasks were sent to

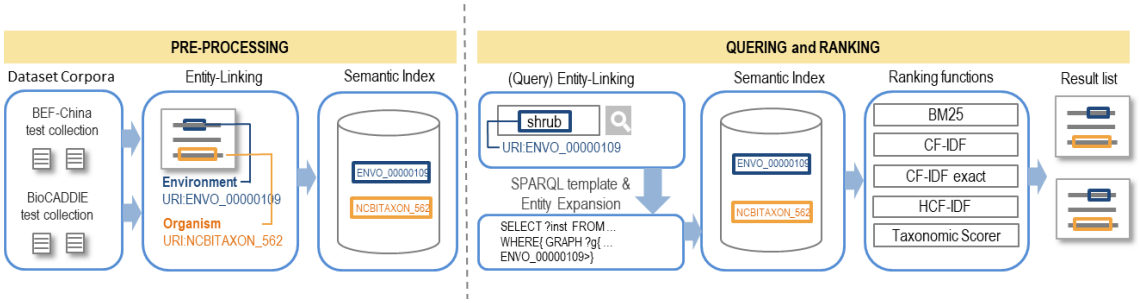


Figure 7.5: The overall architecture of the retrieval component.

Test Collection	Question
BEF-China	Name 3 species that occur in woodland area.
	Are soil samples in BEF China data acidic?
bioCADDIE	Find protein sequencing data related to bacterial chemotaxis across all databases.
	Find all data types related to inflammation during oxidative stress in human hepatic cells.

Table 7.4: Example search tasks from the test collections.

the semantic index. The retrieval process included SPARQL queries in the backend to expand the query on further related entities. We explored different entity-based models. Finally, the retrieved result list was evaluated with the gold standards provided in the test collections. In the following, we describe all steps in detail.

Metadata test collections: Test collections in information retrieval providing metadata as data source are rare. We utilized the two metadata test collections in the Life Sciences that are publicly available. As first benchmark, we used the BEF-China test collection with 372 metadata files and 14 questions we developed for dataset search in biodiversity research [Löffler et al., 2021]. The BEF-china test collection only provides binary relevance judgments. The metadata files being relevant for a query also include partial results. As second test collection, we selected the bioCADDIE dataset [Cohen et al., 2017b], a test collection with 15 questions and 794992 metadata files from the biomedical domain. The relevance judgments in the bioCADDIE corpus are given as a three-point Likert-scale: 0-non relevant, 1-partially relevant, 2-relevant. In the annotation guidelines, a dataset is only considered as relevant if it provides all key terms and an answer on the question (or if the mentioned search terms are related to each other). The annotators were requested to judge a dataset as partially relevant if only partila key terms are contained in the dataset or if the key terms are present but without any relationship. Example questions for both test collections are presented in Table 7.4.

Entity type	Tagger	Ontology
Organism	OrganismTagger [Naderi et al., 2011]	NCBITaxon [NCBITaxon, 2022]
Environment	BiodivTagger [Löffler et al., 2020]	ENVO [Buttigieg et al., 2016]
Material	BiodivTagger [Löffler et al., 2020]	ENVO [Buttigieg et al., 2016] CHEBI [Hastings et al., 2016]
Quality	BiodivTagger [Löffler et al., 2020]	ENVO [Buttigieg et al., 2016] PATO [Gkoutos, G. et al., 2022] FLOPO [Hoehndorf et al., 2016] PPO [Walls, R. et al., 2022] TO [Cooper et al., 2017]
Process	BiodivTagger [Löffler et al., 2020]	ENVO [Buttigieg et al., 2016] UBERON [Mungall et al., 2012] REX [OBOFoundry, 2022] PO [Cooper et al., 2013] GO [Ashburner et al., 2000, GO, 2021] OBI [Bandrowski et al., 2016b] INO [Özgür et al., 2016]

Table 7.5: Entity linking for the BEF-China corpus: In the taggers (middle column), we utilized the ontologies listed in the right column to ensure a good domain coverage for each entity type.

Entity linking in metadata: All noun entities in the metadata files of both corpora were linked with entities. We used the GATE framework [Cunningham et al., 2013] and various existing NLP pipelines and ontologies from the OBO Foundry[18] collection to extract entities and their URIs.

For the BEF-China corpus we extracted five entity types with the OrganismTagger [Naderi et al., 2011] and the BiodivTagger [Löffler et al., 2020]: Organism, Environment, Materials (including chemicals), Process (biological, chemical and physical processes) and Quality (data parameters and phenotypes). Table 7.5 provides an overview of the selected taggers and ontologies.

The bioCADDIE metadata files were annotated with taggers available in the GATE framework and the OrganismTagger [Naderi et al., 2011] to extract entity types such as organisms, genes and proteins (Table 7.6). For entity linking, we added additional steps in the NLP pipeline to connect the identified terms and phrases with URIs in ontologies. In case a phrase could not be linked to an URI, at least the annotation was created with a default URI as we did not want to loose the entity type information. For the extraction of materials (e.g., chemical compounds), anatomical entities and diseases, we utilized an extended version of the BiodivTagger[Löffler et al., 2020] with additional ontologies. Entity linking in these taggers is mainly based on look-up services with a SPARQL query considering not only exact matches in the `rdfs:label` attribute but also synonyms.

In addition to the exact URIs, we also added URIs of super classes linked via a subclass relationship with the entity class in the knowledge base. Therefore, a keyword finally possessed two types of annotations: (a) *inst* annotations that denote the URI of concepts exactly matching the semantics of a keyword (including synonyms) and (b) *broader*

[18]OBO Foundry, `http://www.obofoundry.org/`

Entity type	Tagger/ Methodology	Ontology
Organism	OrganismTagger [Naderi et al., 2011]	NCBITaxon [NCBITaxon, 2022]
Genes & Proteins	AbnerTagger [Cunningham et al., 2013] PennBioTagger [Cunningham et al., 2013]	NCIT [NCIT, 2022] GO [Ashburner et al., 2000, GO, 2021] OGG [He et al., 2014]
Material	BiodivTagger [Löffler et al., 2020]	ENVO [Buttigieg et al., 2016] CHEBI [Hastings et al., 2016] NCIT [NCIT, 2022]
Process	BiodivTagger [Löffler et al., 2020]	ENVO [Buttigieg et al., 2016] UBERON [Mungall et al., 2012] REX [OBOFoundry, 2022] PO [Cooper et al., 2013] GO [Ashburner et al., 2000, GO, 2021] OBI [Bandrowski et al., 2016b] INO [Özgür et al., 2016] MOP [Batchelor, C. et al., 2022] NCIT [NCIT, 2022]
Anatomy	Dictionary LookUp	UBERON [Mungall et al., 2012] BTO [Gremse et al., 2011] CL [Diehl et al., 2016] NCIT [NCIT, 2022]
Disease	Dictionary LookUp	DOID [Kibbe et al., 2015] SYMP [Schriml et al., 2021] HP [Köhler et al., 2019] NCIT [NCIT, 2022]

Table 7.6: Entity linking for the bioCADDIE corpus

annotations providing linkage to related concepts located higher in the hierarchy of the knowledge base (Figure 7.6). Later in search, these two types of annotations allowed a flexible and fast semantic querying of either exact entity URIs or querying for descendant concepts containing *broader* annotations. The source code for both pipelines is available in our GitHub repository[19].

Semantic indexing: All annotated metadata files were indexed with GATE Mímir[20] [Cunningham et al., 2013], a search engine that enables the indexing of semantic annotations and allows the integration of SPARQL statements into the search query. In case of semantic annotations, GATE Mímir creates an annotation index per entity type and indexes the defined features (in this case the URIs) of each annotation. Plain keywords are indexed in token indexes. Based on the query and the existence of semantic annotations in the query, GATE Mímir performs a search on either the token or the semantic index or on both.

[19]Entity Retrieval, `https://github.com/fusion-jena/entity-retrieval-evaluation`

[20]The bioCADDIE corpus contains some large files (> 1MB) for which the annotation process resulted in a timeout. It was also not possible to split these files into smaller portions. Therefore, we decided to omit these nine files. The indexed bioCADDIE corpus consists of 794983 metadata files.

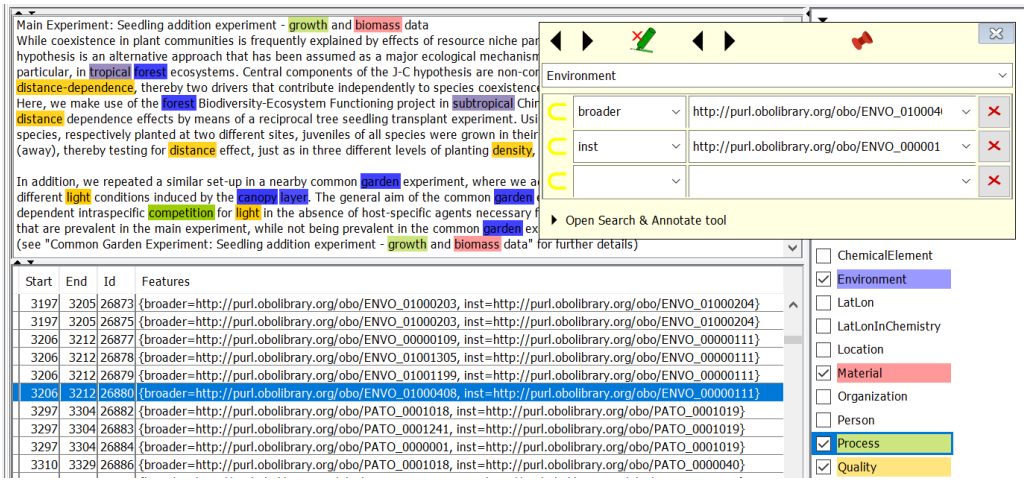


Figure 7.6: GATE screenshot with *inst* and *broader* annotations for a dataset of the BEF-China corpus [Germany and Erfmeier, 2019]. *inst* annotations denote exact URI linking of a keyword including synonyms and *broader* annotations link the keyword to concepts that are located on a higher hierarchy level in the knowledge base. Each annotation is classified into domain specific categories, e.g., the keyword *forest* is an ENVIRONMENT annotation.

Entity linking of search tasks and entity expansion: The original queries from the test corpora were manually annotated and linked to concepts that best reflect their semantics in ontologies from OBO Foundry. These linked queries were our baseline (simple URI search = no expansion). As a first expansion strategy we utilized all object properties defined in the core set of the Relation Ontology with respect to a query term's corresponding concept. We also explored hierarchical concepts and expanded the query on descendant nodes as well as on sibling concepts and their descendants. In order to simulate an adaptive system, we also determined the best expansion strategy per query and evaluated the scorers across this adapted expansion strategy. Listing 7.1 provides an example query with an expansion on descendant nodes being sent to the search engine. Keywords are not contained in the queries but only entities from respective ontologies matching the keyword. We combined the original concepts from the query with a logical AND, but added the concepts obtained from expansion strategies with OR. Based on the given annotation in the query, GATE Mímir either searches for matches in the *inst* (exact match) or *broader* (related match) annotations. The same query could also be performed with a SPARQL statement (Listing 7.2)[21]. As SPARQL statements at run time can take some seconds to come back (depending on the complexity of the statement), we decided to move this entity expansion to the pre-processing step and to add hierarchy relations as semantic annotations to the metadata.

[21] In order to obtain only ancestor concepts from the selected ontology (ENVO in this case) and not from the upper level ontology BFO a statement for the ancestor label has to be added. BFO concepts do not possess labels.

```
{Organism} AND {Environment inst = "http://purl.obolibrary.org/obo/ENVO_00000109"}
OR {Environment broader="http://purl.obolibrary.org/obo/ENVO_00000109"}}
```

Listing 7.1: Example of a search query expanded with descendants.

```
{Organism} AND {Environment sparql=
SELECT DISTINCT ?parent
FROM NAMED <http://example.org/BIODIV/ENVO>
WHERE{GRAPH ?g{
    VALUES ?child{<http://purl.obolibrary.org/obo/ENVO_00000109>}
    ?child rdfs:subClassOf* ?parent.
    ?parent rdfs:label ?parentLabel}}}
```

Listing 7.2: Example of a search query using a SPARQL statement.

Ranking functions: In GATE Mímir, datasets are retrieved with a boolean retrieval model (provided by the underlying MG4J framework [Boldi and Vigna, 2005]) and ranked in a second step. GATE Mímir permits to add own scoring plugins to enable the implementation of user-defined ranking strategies. Therefore, we implemented all ranking functions presented in Section 7.6 via appropriate scorers. GATE Mímir includes a TF-IDF and BM25 scorer. In case the query only contains entities (annotations with URIs), both ranking algorithms already work concept based. Hence, BM25 uses entities with URIs instead of keywords. We added a variation of CF-IDF (called *CF-IDF exact*) to consider only hits with URIs occurring in the query. This gives more credit to matches with URIs, in case the entity linking did not work properly and could only extract a semantic type but not a concept in an ontology. We implemented the scorers CF-IDF exact, HCF-IDF and Taxonomic Scorer as strict URI-based scoring functions to evaluate a full URI-based entity retrieval system. In case no URI could be linked, the annotation is not considered in scoring. For the computation of the semantic similarity between two concepts (needed in the taxonomic ranking of [Waitelonis et al., 2015]), we used the SML library from Harispe [Harispe et al., 2015]. We evaluated this scoring function with the similarity measure IC: Zhou/Resnik, which performed best in the evaluation of [Waitelonis et al., 2015]. Supplementary information are in the GitHub repository[22].

7.8 Results

In the following, we present our results with respect to entity expansion (Section 7.8.1) and the evaluation of the ranking functions (Section 7.8.2). The results are publicly available [Löffler et al., 2022b]

[22]Entity Retrieval, `https://github.com/fusion-jena/entity-retrieval-evaluation`

7.8.1 Entity Expansion

The final results are presented in Table 7.7 and Table 7.8. We computed Mean Average Precision (MAP), Precision@K, and normalized Discounted Cumulated Gain (nDCG) [Manning et al., 2008] for all returned metadata files with the latest TREC script[23]. TREC only considers binary judgments even if a more fine grained scale (e.g., 'partially relevant' and 'relevant') is provided [24]. All metrics were computed with the '-c' parameter which penalizes for queries with no results. While the P@k metric measures the precision at a specific cut-off rank per query (equal weight per rank), the MAP is a measure of the precision over all queries and P@k values until a cut-off rank. MAP and nDCG compute the precision over all retrieved documents, but nDCG gives more credit to top ranked documents. In addition to these classical metrics we used the evaluation script from the Precision Medicine Track[25] to determine the inferred average precision (infAP) and inferred normalized Discounted Cumulated Gain (infNDCG) [Yilmaz et al., 2008] for the bioCADDIE corpus. The inferred measures are more robust measures for benchmarks with incomplete relevance judgments. Since bioCADDIE's relevance judgments are based on depth-k pooling, the relevance judgments might be incomplete as only top k documents retrieved by the systems were judged and the rest of the documents were deemed non-relevant[26].

BEF-China: The manually annotated 14 questions resulted in 41 entities from seven ontologies. Concerning the expansion strategies, adding descendant entities to the query resulted in the highest precision values in the BEF-China corpus (Table 7.7, Figure 7.7). This holds for all tested algorithms. In comparison to the simple URI queries without any expansion (nDCG: 0.2248 - 0.2697), the expansion with descendants resulted in nDCG values between 0.3646 to 0.4156. The second best strategy was the expansion with sibling nodes and their descendants. This strongly indicates that hierarchy expansion leads to more relevant results. However, against our expectations, the expansion with core concepts from the BFO ontology did not result in more relevant hits. We obtained 44 additional entities from core relation expansion (based on 11 entities in the original query providing core relations). We think that these expansions against our assumption go too far away from the original query term and therefore do not result in more relevant hits. A simple ANOVA test per column (expansion strategy) and per nDCG, MAP and P@k values results in $p < 0.01$ for all three metrics. This give clear evidence that the differences

[23]TREC script, `https://trec.nist.gov/trec_eval/`

[24]TREC Data, `https://trec.nist.gov/data/reljudge_eng.html`

[25]Precision Medicine Track, `https://trec.nist.gov/data/clinical/sample_eval.pl`

[26]The computation for the inferred measures requires a 5-column 'qrels' file (instead of 4 columns in the usual TREC scripts) with the relevance judgments. As this was not provided by the challenge organizers, we added the 4th column (identification which category/stratum a document comes from) ourselves. The adapted 'qrels' file is in our repository.

Table 7.7: Results for the BEF-China test collection, fixed number of queries (q=14)

	plain URI query			**+ core relations**			**+ all descendants**			**+ siblings and descendants**			**best case per question**		
	MAP	P@10	nDCG	MAP	P@10	nDCG	MAP	P@10	nDCG	MAP	P@10	nDCG	MAP	P@10	nDCG
BM25	0.1982	0.2143	0.2598	0.1842	0.2000	0.2513	0.2849	0.3643	0.3912	0.2443	0.3000	0.3861	0.2898	0.3786	0.4092
CF-IDF	0.2144	0.1929	**0.2697**	0.2088	0.1929	**0.2675**	0.3133	0.3357	0.4150	0.2778	0.3071	0.4095	0.3115	0.3500	0.4234
CF-IDFExact	0.2155	0.1929	0.2679	0.2115	0.1857	0.2671	0.3139	0.3357	0.4123	0.2811	0.3071	0.4085	0.3128	0.3571	0.4239
HCF-IDF	0.1416	0.1143	0.2248	0.1342	0.1000	0.2159	0.2450	0.3000	0.3646	0.2226	0.2500	0.3633	0.2277	0.2643	0.3704
Tax. Scorer	0.2154	0.1929	0.2679	0.2104	0.1857	0.2659	0.3178	0.3500	**0.4156**	0.3031	0.3286	**0.4189**	0.3192	0.3714	**0.4294**

Table 7.8: Results for the bioCADDIE test collection [a], fixed number of queries (q=15)

	simple URI query			**+ core relations**			**+ all descendants**			**+siblings and descendants**			**best case per question**		
	MAP	P@10	nDCG	MAP	P@10	nDCG	MAP	P@10	nDCG	MAP	P@10	nDCG	MAP	P@10	nDCG
BM25	0.0537	0.4733	0.1165	0.0538	0.4600	**0.1182**	0.0896	0.5533	**0.1742**	0.0959	0.4800	**0.1935**	0.1241	0.6067	**0.2352**
CF-IDF	0.0560	0.4867	**0.1187**	0.0526	0.4400	0.1158	0.0778	0.4800	0.1671	0.0635	0.3800	0.1754	0.0914	0.4600	0.2148
CF-IDFExact	0.0520	0.4800	0.1151	0.0490	0.4333	0.1132	0.0697	0.4400	0.1614	0.0385	0.2600	0.1480	0.0697	0.3867	0.1922
HCF-IDF	0.0420	0.3667	0.1065	0.0410	0.3333	0.1076	0.0600	0.4200	0.1525	0.0282	0.2200	0.1168	0.0431	0.3067	0.1439
Tax. Scorer	0.0520	0.4800	0.1151	0.0486	0.4333	0.1121	0.0713	0.4533	0.1633	0.0416	0.2800	0.1500	0.0714	0.4067	0.1942

[a] underlined numbers denote that the expansion strategies could not be executed for one query and were assessed with 0.0

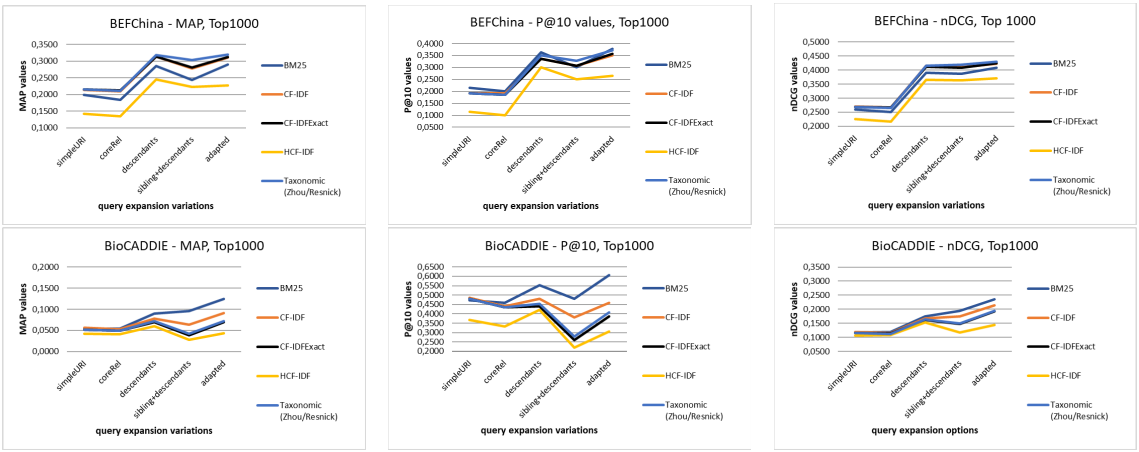


Figure 7.7: TREC metrics for BEF-China (top) and bioCADDIE (bottom)

for the various entity expansion strategies are statistically significant.

bioCADDIE: In total, 92 entities were manually extracted from 13 ontologies for 15 questions. In terms of the expansion strategies, adding descendant nodes to the query resulted in more relevant results across all scoring functions (Table 7.8). This strongly indicates that entity expansion with descendants in general has a positive effect in dataset search in the Life Sciences. What stands out is that the expansion with sibling nodes and all descendants also resulted in more relevant hits (higher MAP and nDCG values than for descendants) for the BM25 scorer. However, they are not top scored (P@10 is lower for sibling + descendants than for descendants). Here, further investigations are needed whether this holds only for this specific domain. Despite the fact that the biomedical domain has a large amount of available ontologies including BFO relations, the expansion with core relations did not result in more relevant hits. We obtained 20 additional entities from the core relation expansion (based on given 18 entities providing core relations). Again, we think that these expansions with core relation entities go too far away from the original search intent. An adaptive expansion (best case per question) strategy further improves the retrieval of datasets (Table 7.8). Using an adaptive expansion strategy leads to a further improvement of 18% for the nDCG values and 28% in terms of MAP values. The statistical analysis with a simple ANOVA per column (expansion strategy) results in $p \leq 0.05$ (MAP). This shows that the results for the entity expansion strategies in this corpus are statistically significant, too.

7.8.2 Ranking Functions

BEF-China: If no expansion strategy was applied, the CF-IDF resulted in the highest nDCG values (0.2697), but the more hierarchy expanded concepts were added to the query, the better the Taxonomic Scorer performed (nDCG 0.4156 for the expansion with descendants and nDCG 0.4189 for the expansion with siblings and their descendants).

The difference in the MAP and nDCG values between the BM25 and Taxonomic Scorer is larger for the expansion with siblings and their descendants (column 'siblings and descendants', Table 7.7) than in the expansion with narrower concepts (column 'descendants', Table 7.7). However, the statistical analysis with a simple ANOVA per row (ranking functions) results in p values > 0.05 (MAP:0.13, P@10:0.51, nDCG:0.78)

bioCADDIE: The results for the bioCADDIE test collection reveal a different picture (Table 7.8, Figure 7.7). Here, the classical scorers provide higher MAP, P@10 and nDCG values in contrast to the scorers considering semantic relations. In particular, the differences between the scorers are larger in the queries expanded with siblings and their descendant than in the other strategies. The statistical analysis with a simple ANOVA per row for one metric results in p values > 0.05 (MAP:0.02, P@10:0.008, nDCG:0.49).

This result shows that we can not conclude that a particular scorer leads to an improved ranking for all expansion strategies. Here more studies are required to investigate whether more queries and search tasks are needed or whether there are other dependencies.

7.8.3 Discussion

For the bioCADDIE corpus, we compared our results with the bioCADDIE retrieval challenge [Roberts et al., 2017]. To do so, we used the results from the adapted expansion strategy with a BM25 scorer. Our infNDCG (0.4196) is very close to the median and mean infNDCG of the bioCADDIE challenge (best run per participant) with 0.423 and 0.425. Taking into account that we combined core concepts with a logical AND to ensure the correct interpretation of the original query, which resulted in 0 results for three queries (Figure 7.8), our result is notable. For the top scored datasets (Figure 7.8, P@10) for 9 out of 15 queries our P@10 values are either equal or higher than the mean values over all participants of the bioCADDIE challenge. For the infNDCG values measuring the long tail of the search result set for 7 out of 15 queries our system (BM25 scorer, adapted expansion strategy) returns more relevant hits than the mean of the bioCADDIE participants.

The results reveal that entity expansion improves dataset retrieval in the Life Sciences. In addition to the findings by [Terolli et al., 2020], we showed that in particular the expansion on descendant nodes leads to more relevant results for both corpora. The expansion with descendants in the BEF-China corpus resulted in a larger difference to the baseline than in the bioCADDIE corpus. More studies on further corpora and more questions are needed to analyze whether these differences are domain-specific, query-dependent or due to a different ontology coverage of the corpus.

In comparison to the results of the user survey and expansion and ranking preferences (Section 7.3), the results of both test collections show that the expansion with concepts of

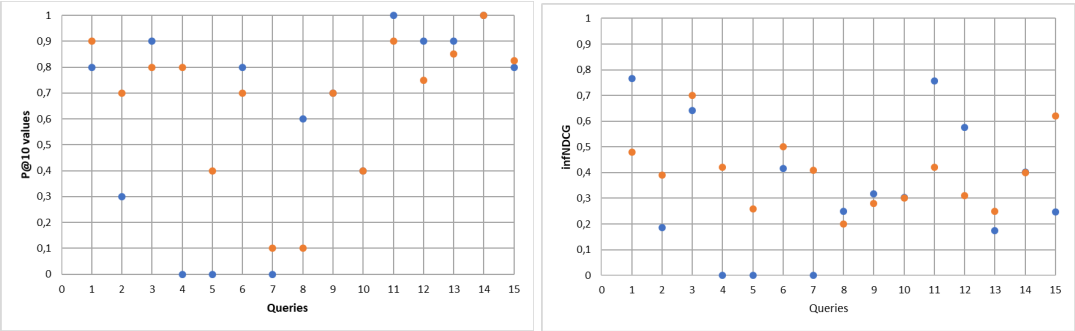


Figure 7.8: Comparision of bioCADDIE participants' results (orange) with our best run (BM25, adapted = best case per question) (blue) per question: P@10 values (left) and infNDCG values (right)

the hierarchy led to more relevant results. This confirms users' assessment that not only exact matches are relevant in dataset search but also related hierarchical entities.

Our work has a limitation concerning the analysis on the different entity-based ranking functions. The number of questions provided in the test corpora are very small (14 and 15). According to the annual TREC competitions[27], at least 50 queries or more are needed to evaluate different ranking scorers.

Overall, our values in all measures are very low. Here, only an in detail error analysis would help to determine the reasons. An error analyis by [Bast et al., 2018] showed that in semantic search documents can be found that are relevant but that are not part of the given relevance judgments (ground truth). In this case, these additional documents were not counted as positive match and thus lowered the final evaluation metrics. Whether this also holds for our results needs to be analyzed with domain experts. In particular for the biomedical corpus, additional expertise in biomedicine is needed, which we will address in future work.

7.9 Summary

This chapter introduced the retrieval component of our proposed semantic dataset search. It aimed to explore various entity expansion strategies and entity-based retrieval models addressing H_4. Our solution is based on the GATE framework with a semantic index consisting of an annotation index with relevant entity types for biodiversity research and URIs extracted from metadata. Concerning the proposed entity expansion, it turned out that the expansion with descendant concepts of the terminology returned more relevant datasets. Moreover, our results show that the expansion on broader releations such as siblings and their descendants resulted in more relevant datasets with respect to the queries without expansion, too. Hence, we can confirm H_4. Tailored entity expansion with hierarchy relations lead to more relevant results with respect to a plain entity-based retrieval without

[27] TREC, `https://trec.nist.gov/`

any expansion. Concerning the various retrieval models and the consideration of the different semantic relations in scoring, we can not draw any conclusions, as the variability in our result set was too large. Therefore, more studies and larger test collections with more questions are needed to explore the differences in entity-based retrieval models.

Chapter 8

User Interfaces for Dataset Search and Usability Evaluation

> *"Usability is about people and how they understand and use things, not about technology."*
>
> - Steve Krug in 'Don't Make Me Think' [Krug, 2013], *user experience professional*

A variety of user interfaces (UI) and frameworks are available for search applications, but very few user studies exploring their usability are publicly available. User studies are time-consuming and personnel-intensive, which are very likely the main causes for the current lack of available studies in research. However, user studies give valuable insights on attitudes, expectations and usage. Therefore, the awareness that the success of a novel system depends on functionality, design AND *usability* must be increased.

The aim of this chapter is to address the needs for more insights on the usability of semantic search systems. User studies about a dataset search (Chapter 3.2) and the exploration of search interests in biodiversity research (Chapter 4.2) have already been introduced. Now, as last component of our proposed dataset search (Chapter 5.4), we present the user interface and its development and evaluation with users. As an easy to use and transparent interface is essential for the acceptance and effectiveness of a novel system, it is crucial to involve users not only in testing and evaluation but also in the design and development process.

Following the user experience design principles (Chapter 2.1), we present the overall approach for the development and evaluation of the user interface of our semantic dataset search in Section 8.2 after a section about related work (Section 8.1) introducing existing user interfaces and discussing evaluation frameworks and user studies on semantic search systems. We also present the results of a first formative evaluation with a focus group and a paper prototype. Afterwards, we introduce the architecture and implementation of our proposed user interface in Section 8.3 followed by the user evaluation setup in Section

8.4. The results of the evaluation on the usability of our proposed user interface are finally presented in Section 8.5.

A first version of the implemented user interface has been published at the S4biodiv workshop [Shafiei et al., 2021] as demo paper.

8.1 Related Work

Scholarly information needs are as diverse and heterogenous as data. Therefore, retrieval systems must meet these challenges and must adapt to these enhanced search requirements. The RDA Data Discovery Paradigms Interest Group[1] addressed this need for improved dataset retrieval systems and published ten recommendations for enhanced dataset search and user experience [Wu et al., 2019]. For these guidelines, the authors collected 79 data discovery use cases, conducted heuristic evaluations and interviewed scholars. The recommendations include among others (1) multiple search inputs to support different information needs, (3) comprehensible search summaries to judge the relevance, accessibility and reusability of datasets, (5) bibliographic information, (7) consistency in the presentation and (10) usage of API search standards and domain terminologies.

The available user studies for dataset search have already been introduced in Chapter 3.1.1. In the following section, we present existing work on user interfaces for semantic search applications in the Life Sciences (Section 8.1.1). We do not only focus on dataset search, but we expand the analysis on systems offering scientific literature. Afterwards, we introduce evaluation frameworks and evaluation studies exploring the usability in semantic search systems (Section 8.1.2). As very few user studies are available for semantic search systems in the Life Sciences, we expand the view on evaluation studies about all kinds of semantic search systems.

8.1.1 User Interfaces in Semantic Search Systems in the Life Sciences

The Semantic Web community is a very active community in terms of novel interfaces over Linked Data. Therefore, most existing approaches of user interfaces for a semantic search are visualizations and user interfaces to browse and filter data in triple stores containing ontologies or knowledge graphs. [Koho et al., 2016, Vega-Gorgojo et al., 2016a, Khalili et al., 2016, Khalili et al., 2018, Ikkala et al., 2022]. Current metadata files are primarly available in half-structured formats, such as JSON or XML with longer textual resources. Hence, approaches over Linked Data are not a complete solution, but semantic full text approaches with combined search indexes on text (documents or metadata) and URIs are needed. In this research domain, only very few user interfaces for end users exist. One system was Broccoli, a semantic full text search proposed by

[1] RDA DDP IG, `https://rd-alliance.org/node/52248/outputs`

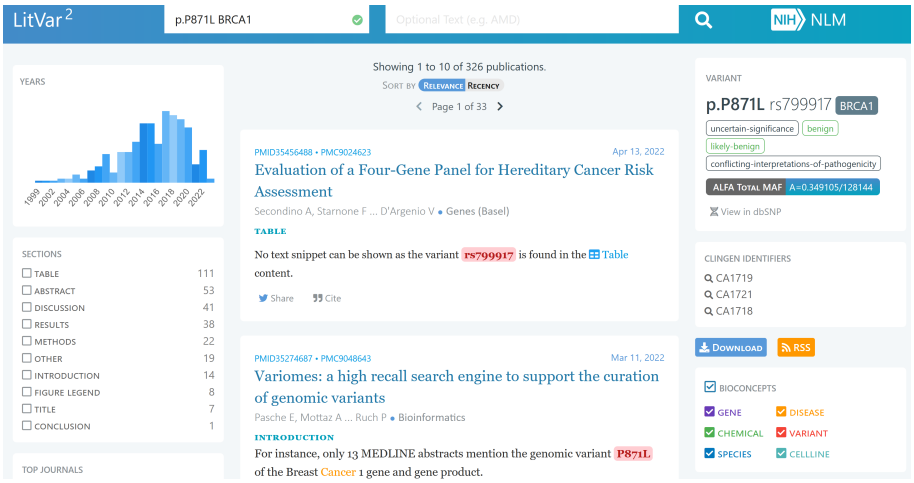


Figure 8.1: Screenshot of the semantic search LitVar [Allot et al., 2018]. Biological concepts are highlighted in different colors. On demand users can get further information on these highlighted terms, e.g., in which organisms this variant occurs and links to the respective concept in a terminology.

[Bast and Buchhold, 2013]. The system allowed a search over Wikipedia articles and enabled filtering over facets from the knowledge base YAGO [Suchanek et al., 2008]. The full query was visualized as a tree and provided linkage to all matching entities and classes. It permitted a search with keywords and ontological concepts. Queries were interpreted as a tree with relations. The system supported two types of relations: relations obtained from the ontology and 'occurs-with' relations denoting the existence of a term, instance or class in a sentence. The successor QLever[2] [Bast and Buchhold, 2017] provides an improved efficient semantic index on combined data with a SPARQL interface, which is not suitable for end users.

Due to the lack of search interfaces over text and Linked Data, we took a closer look again on semantic search approaches in the Life Sciences, which are mainly based on text indexes but provide query expansion (KST) or information extraction techniques with additional search services (SDET) (Chapter 3.1.2). For the following overview (Table 8.1), we only selected systems that are publicly available as demonstration. We inspected them concerning two main aspects in search, (1) the query input (e.g., natural language interface/ one input field/ form-based search) and filtering options such as facets and (2) the presentation of search results with a search summary containing information on relevance, data access and reusability. In our comparision study between a keyword search and a semantic search, we figured out that scholars need explanations on the extended search result in the semantic search [Löffler and Klan, 2016]. Therefore, as a third aspect (3), we analyzed whether the search applications offer supplementary information on the search result.

Keyword search and faceted search are supported by all systems. Narrowing the

[2]QLever, `https://qlever.cs.uni-freiburg.de`

System	Sources	Query Input/ Facets	Search Summary	Explanations
PolySearch2 [Liu et al., 2015]	Biomedical literature and databases	keyword input and type search, e.g., drugs, disease, gene and proteins	search term and matching entities with synonyms shown, number of hits per document	highlightings of semantic categories on landing page, legend for highlighting
LIVIVO [Müller et al., 2017]	50 biomedical literature sources such as Pubmed	keyword input, facets for publication years, subjects, document types, databases and languages	search terms highlighted in title, data source present, subjects shown as icons	-
LitVar [Allot et al., 2018]	PubMed articles	keyword input and named entities on variants of genes, facets for publications years, article sections and journals	search term + further biomedical terms highlighted, different colors for different categories, data source and citation present	descriptions on matched entities with linkage to terminology concepts
Thalia [Soto et al., 2018]	PubMed articles	keyword input and search for named entities in domain categories, facets for year, journal, author, type	search terms highlighted in search result	matching entities listed and sorted into entity types (chemicals, diseases, drugs, genes) in separate boxes
Datamed [Chen et al., 2018]	76 biomedical data repositories	keyword input, facets for data types, repositories and accessibility	search terms and synonyms highlighted, data source present	full search query and used synonyms displayed in a separate box
BioFID [Pachzelt et al., 2021]	fulltext from the Digital Collection Biology (German) [3]	keyword input, factes for database, publication year and authors	search terms and semantic categories (Taxon, Plant, Animal, Location, Time) in different colors highlighted, data format and citation available	per entity additional information from Wikipedia displayed in a separate box

Table 8.1: Comparision of functionalities in the user interface of selected semantic search approaches in the Life Sciences.

search on specific domain categories (either in separate input fields or with a query language) is only provided in [Liu et al., 2015, Soto et al., 2018, Allot et al., 2018]. Information on where the datasets or articles come from are mostly present in search summaries. Moreover, the retrieved entries in the result set contain textual highlightings to give users a quick overview on matched query terms. In particular, the text highlightings in PolySearch2 [Liu et al., 2015], LitVar [Allot et al., 2018] (Figure 8.1) and BioFID [Pachzelt et al., 2021] (Figure 8.2) distinguish between search terms and further relevant keywords belonging to different domain categories. Further important information on the data format is only available in [Pachzelt et al., 2021]. Bibliographic data are provided in [Pachzelt et al., 2021] and [Allot et al., 2018]. The search summary in the user interface of [Liu et al., 2015] displays the numbers of hits per document, and in [Müller et al., 2017] subject areas are visualized with icons. Information on semantic categories are available in the search summaries of [Pachzelt et al., 2021, Allot et al., 2018]. With respect to explanations, some systems provide explanations on the matching entities, either in separate boxes or dialogs with information about the used ontology, class and concept ID. The most comprehensive explanations offers LitVar [4]. Each search term or biological keyword occuring in the result documents is highlighted and clickable. Upon request, users get more information on the highlighted term, e.g., a link is provided to browse to the respective ontology or terminology. In Datamed [Chen et al., 2018], a sep-

[4]LitVar, `https://www.ncbi.nlm.nih.gov/research/litvar2/`

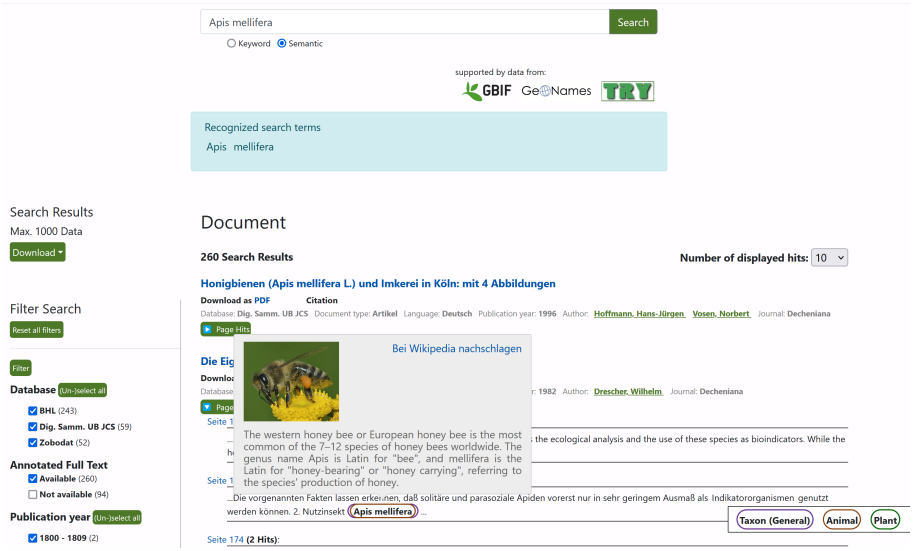


Figure 8.2: Screenshot of BioFID's semantic search [Pachzelt et al., 2021]. Biological entities are highlighted, and users can obtain more information from external resources on demand.

arate box shows the full search query being sent to the search engine. In addition, another box gives information on the synonyms utilized in the query. Further ontological information such as data about used ontologies is not provided.

As discussed above, semantic dataset search applications are hardly present. Most of the introduced systems provide a semantic search over literature. Only very few systems are available offering a search in domain categories. Moreover, explanations on matched entities are rare and not used at all in the introduced dataset search applications. Therefore, our study focus is primarily on different search inputs in a semantic dataset search, including a search in domain categories, and explanations to better understand the search result.

8.1.2 Evaluation Frameworks and User Studies in Semantic Search

The evaluation of semantic search systems is mainly influenced by classical Information Retrieval (IR) evaluation approaches [Elbedweihy et al., 2015]. The Cranfield model with a test collection, a document corpus and relevance judgments (Chapter 2.2.3) is still present in evaluation campaigns in semantic search. We first present approaches about general semantic search evaluation frameworks followed by a paragraph introducing concrete studies with user engagement.

Semantic search evaluation frameworks: One of the first evaluation initiatives was the SemSearch workshop series in 2010[5] and 2011[6] with a focus on ad-hoc object retrieval over Linked Data. The evaluation used the Billion Triples Challenge Dataset

[5]SemSearch 2010, `http://km.aifb.kit.edu/ws/semsearch10/`
[6]SemSearch 2011, `http://km.aifb.kit.edu/ws/semsearch11/`

[Herrera et al., 2019], and the test collection was constructed by human judges using Amazon's Mechanical Turk crowdsourcing marketplace[7]. The expected input of the semantic search was a set of keywords and the output was a ranked list of concepts retrieved from the Billion triple dataset. Classical IR metrics (MAP, P@k and nDCG, see also Chapter 2.2.3) were utilized to assess the relevancy of the returned list of URIs. Participating systems mainly used classical retrieval approaches and made only very little use of additional data from the knowledge base to improve the retrieval or to expand the queries [Elbedweihy et al., 2015]. Other initiatives are the annual Question Answering over Linked Data (QALD) challenge[8] and the Entity track at TREC[9]. The QALD uses multiple RDF datasets and additional knowledge sources as data sources, and queries are formulated as natural language questions or keywords. The search result are correct answers or SPARQL queries retrieving these answers. In contrast, the Entity track aimed to return results from text and Linked Data. It operated on the ClueWeb 2009 web corpus, a corpus of web pages with entities (not linked to existing knowledge bases but own IDs). The information need was provided in a semi-structured format with a given query entity, a type (mapped to the DBpedia[10] ontology) and an information need in free text. The Entity track consists of two main tasks: (1) the extraction of ClueWeb entities related to the target type in the information need including documents supporting the query and (2) a list of entities related to the ClueWeb query entity in the associated knowledge base (Sindice [Campinas et al., 2011]).

All of the introduced initiatives used constructed queries over knowledge bases or text with general data (non domain specific data). They also did not involve users. An example for an evaluation approach towards real informations needs was the Question Answering track at TREC[11]. It aimed to retrieve exact answers on user questions from the medical domain. The information were obtained from web pages of the National Institutes of Health (NIH). However, even if real user's information needs were utilized in this campaign, the study only measured the relevancy of search results and did not include any active user studies. To the best of our knowledge, there was only one semantic search evaluation initiative that considered user aspects. The SEALS evaluation framework [Wrigley et al., 2010] favored a two-phase evaluation with an automatic phase and a user-in-the-loop (UTIL) phase. In the automatic phase, performance based measures (Chapter 2.1.2) were determined, such as execution success on search queries, results (e.g., relevancy of search results) and time to receive the result. In the second phase, users got search tasks and had to enter queries in the respective query language. This additional task could get challenging depending on the complexity of the query language. There-

[7] MTurk, `https://www.mturk.com/`
[8] QALD, `https://qald.aksw.org/`
[9] Entity Track, `https://trec.nist.gov/data/entity.html`
[10] DBpedia, `https://www.dbpedia.org/`
[11] LiveQA, `https://github.com/abachaa/LiveQA_MedicalTask_TREC2017`

fore, besides the execution success und returned results, in this phase, further measures were added, such as the underlying query (in the tool's internal format, e.g., SPARQL) and user-specific statistics, e.g., the time required to obtain the answer, demographics and the System Usability Scale (SUS) as well as an in-depth satisfaction questionnaire. Unfortunately, no documents and code are available anymore for the SEALS evaluation framework. An existing evaluation campaign focusing on the usage of semantic web techniques in NLP approaches is the International Workshop on Semantic Evaluation[12]. This initiative addresses several NLP tasks such as Named Entity Recognition, disambiguation or the identification of social attitudes, but there are no Information Retrieval tasks.

Due to this lack of evaluation frameworks for semantic search systems, user studies for semantic search evaluations follow their own evaluation strategy. Most of the introduced studies below are inspired by the TREC Interactive Search track (Chapter 2.2.3) including user tasks and questionnaires.

User studies on semantic search systems: One of the earliest user study on four different user interfaces was the work by Kaufmann et al [Kaufmann and Bernstein, 2007]. The authors implemented four different query interfaces and setup a search over the Mooney Natural Language Learning Data, a knowledge base with geographical information about the US and their logical representations [Tang and Mooney, 2001]. The aim was to explore the usefulness of Natural Language Interfaces and different query languages. 48 participants carried out four search tasks per system and assessed the usability with the SUS questionnaire (see also Chapter 2.1.2). The system, which required full English questions as query input, got the highest SUS score (75.73) and also resulted in the highest success rate. Users "appreciated the freedom of the query language" [Kaufmann and Bernstein, 2007].

[Elbedweihy et al., 2012] extended these studies and explored different user types (experts and casual) and their preferences for various query input formats (natural language in free form, natural language with a controlled query language, form based and graph based) on the same data source. Ten casual users and ten experts (unfortunately, the study does not report on what 'expert' exactly means, e.g., domain expert, expert in search, semantic web expert) had to perform search tasks in five different user interfaces (one natural language based system in free form, one natural language system with a controlled query language, one form-based and two grap-based tools). The expected result were exact answers and not documents. While casual users preferred the form-based query interface, expert users favored the graph-based approach. The user interface with the controlled query language provided most support for casual users, but it limited experts in expressing their information needs [Elbedweihy et al., 2012].

Another comprehensive user study on a semantic search was conducted by

[12] SemEval, `https://semeval.github.io/`

[Bontcheva et al., 2012]. The authors compared a keyword search and a semantic search. The semantic search was based on a combined index (scientific documents from the environmental domain, DBpedia[13] and GeoNames[14]). 17 subjects performed four search tasks per system (A/B testing) and filled out a SUS questionnaire afterwards. The results reveal that in the semantic search, users were more often able to complete the tasks. In addition, the semantic search reached a higher SUS score (72.3).

An evaluation with users and different visual query interfaces over a semantic dataset was the study by [Vega-Gorgojo et al., 2016b]. Following the TREC IIR methodology with an A/B testing, 15 users got four search tasks in random order in two semantic search systems over an RDF dataset with administrative and financial information of Norwegian companies. Per system, the participants performed four different tasks. Before and after each search tasks they had to answer a few questions in a questionnaire. After each system, the filled out a post-session questionnaire (not SUS) and finally, at the very end, an exit questionnaire. Their results confirm the outcome of [Elbedweihy et al., 2012]. The form-based system PepeSearch [Vega-Gorgojo et al., 2016a] was preferred by casual users, whereas OptiqueVS [Soylu et al., 2013], a graph-based user interface, received better assessments from expert users.

Apart from user tests and questionnaires, other usability methods such as eye-tracking also give valuable insights on a user's behaviour. The study by [Liu et al., 2017] examined domain experts' gaze interactions in a search over biomedical documents enhanced with MeSH [MeSH, 2022] terms. In total, four user interfaces with different approaches for the integration and display of MeSH terms were provided. For example, in one user interface, MeSH terms were collected for all documents on top of the search result page, while in another interface MeSH terms were displayed per document. The 32 participants had to execute four search tasks in only two of the four user interfaces. For each task, users had only a limited time (7 min) to perform the search. The authors measured how often people looked at specific areas of interests (AOI), such as title, author, abstract and MeSH terms. They also collected information on users' domain background and their usage of search applications. The results reveal that the level of domain expertise affects eye gaze behaviour. For instance, "users with a high level of domain knowledge are less likely to attend to the element of title" [Liu et al., 2017]. Another correlation of background knowledge and gaze behaviour had also been found. Participants who indicated to use search engines frequently payed more attention to author information and MeSH terms in the result set as subjects who use search applications not frequently [Liu et al., 2017].

The presented studies mainly focus on the analysis of the query input and the user interface overall. There is no study examing individual components in the search result presentation. In addition, we still do not know whether the current findings also hold for

[13]DBpedia, `https://www.dbpedia.org/`
[14]GeoNames, `https://www.geonames.org/`

dataset search (all presented studies were conducted over scientific documents or Linked Data). In particular, it is unclear to what extent additional semantic information obtained from knowledge bases to enhance the search result should be displayed in the user interface explicitly. Do users want to be informed about this additional knowledge and should matching entities be presented in a prominent way (e.g., with links to the respective ontology) or do users prefer a less explanatory user interface? In order to answer these questions, we developed two user interfaces and conducted a usability study with scholars being presented in the following sections.

8.2 Study Overview and Formative Evaluation

In this section, we introduce preliminary considerations and the overall goals of the study. We also present the overall study procedure and the results of the formative evaluation (focus group and paper prototype).

8.2.1 Preliminary Considerations

In Chapter 3.1.1, the introduced evaluation studies on dataset search applications report on shortcomings concerning the metadata quality, e.g., poor information on domain specific concepts, data variables measured, data type, data format, data collection process and unaligned terminologies. In addition, the study by [Dixit et al., 2017] about a dataset search and the study by [Elbedweihy et al., 2012] on a semantic search also underline the need for a category based search.

Therefore, we utilized our text mining pipeline *BiodivTagger* (Chapter 6) and the *OrganismTagger* [Naderi et al., 2011] to enhance the quality of metadata files and to extract relevant biological entities from selected terminologies. This additional information was added as semantic annotation to metadata and enabled an entity-based search. It also allowed a search beyond keywords including synonyms, common names or alternate labels provided in the matching URI (aligned vocabulary). Moreover, the question corpus study (Chapter 4) reveals that user interests can be very broad. In Chapter 7, we utilized the matching entities of the metadata files for the expansion on descendant nodes and obtained an higher amount of relevant results. Therefore, a search beyond keywords in the proposed novel semantic dataset search includes descendant nodes and enables a hierarchy search (*addressing R1 of the functional requirements* introduced in Chapter 5.3).

In addition to entities, the taggers also return information on domain specific entity types such as Organism, Quality, Environment, Material, Process. These categories reflect domain specific information needs in biodiversity research (Chapter 4). This type of information was added to metadata, too, allowing a category based search (*addressing R2 of the functional requirements* introduced in Chapter 5.3).

Concerning an improved user interface for dataset search, we mainly followed the recommendations of the RDA Data Discovery Interest Group [Wu et al., 2019]. In particular, we concentrated on the above mentioned five recommendations focusing on improvements in the user interface. The information in brackets denotes how we addressed the recommendations in the implementation:

- *(1) - Provide multiple access points to find data:* Form-based input based on domain specific categories and a classical one input field

- *(3) - Make it easier for researchers to judge relevance, accessibility and reusability of a data collection from a search summary:* Labels are provided per dataset in the result summary to highlight general dataset characteristics such as publication date, data center, accessibility. Title, author and publication year are highlighted and linked to the landing page of the data repository hosting the dataset. Data parameters measured and abstract can be expanded if available. Query terms or phrases matching the query terms are highlighted in the search result.

- *(5) - Enable sharing and downloading of bibliographic references:* As it is not a strong focus in the evaluation, only a simple citation information is shown.

- *(7) - Strive for consistency with other repositories:* Result summary presentation according to GFBio[15], PANGAEA[16], Zenodo[17]

- *(10) - Follow API search standards and community adopted vocabularies for interoperability:* usage of OBO-Foundry terminologies (Chapter 6), REST API for search results with a Swagger[18] documentation

Four out of the other five recommendations can be primarily implemented by the data providers. Due to conceptional reasons, we did not consider the second recommendation (R2) for providing different entry points in a user interface, e.g., facets. Different query inputs and filtering options would interfere with each other in an evaluation, as both address the input of a user's information need. Therefore, we decided to leave facets our for this study.

In addition to these recommendations, the novel user interface also includes explanations on the search result (addressing R3, a comprehensible search result). Our findings in a first semantic search study reveals (Chapter 3.3) that users need more explanations on the search results when the search goes beyond keywords. In particular, highlightings are a simple but effective approach to emphasize the query terms or related terms and phrases

[15]GFBio, `https://www.gfbio.org/`
[16]PANGAEA, `https://pangaea.de/`
[17]Zenodo, `https://zenodo.org/`
[18]Swagger, `https://swagger.io/`

in the result datasets. Furthermore, matching entities to the query terms or other biological entities not explicitly given in the query terms are highlighted. On demand users can obtain more information. Full query explanations give insights on the query interpretation and the exact query being sent to the search engine.

In order to address *R4 - non-functional requirements* introduced in Chapter 5.3, we conducted a summative user study to explore the usability of the novel semantic dataset search. We focused on the comparision of a form-based user interface and user interface with one input field as existing studies report that casual users prefer a form-based input [Elbedweihy et al., 2012] and that scholars like to search in topics [Dixit et al., 2017]. We also explored what kinds of explanations on system (e.g., full query) and background information (matching entities, terminologies) are desired in a semantic dataset search. Therefore, users were not only involved in the summative evaluation (*R5 - non-functional requirements*), but participated actively in the development process being described in the next subsection.

8.2.2 Study Goal and Evaluation Flow

The overall goal of this study was to develop an improved user interface for a semantic search over metadata files and domain vocabularies. Following a user-centered design, domain experts were involved in the design and development process. We aimed to study two query inputs, a form-based approach as it performed best in previous studies for casual users and a one input field as users are familiar with this kind of search. Therefore, we developed several user interfaces and evaluated the systems with scholars working in the field of Life Sciences and Environmental Sciences. The main focus in the evaluation is on the usability of two search components: the query input and the presentation of explanations in the search summary. We limited the evaluation on these two parts, as usability evaluations usually examine user interfaces as a whole. In particular, it has not been studied so far what kind of explanations are benefical for dataset search. In a semantic search, the additional challenge is to figure out whether information on utilized terminologies and matching URIs confuse users who are not familiar with semantic technologies. From the research projects we are involved in, we know that most scholars in the Life Sciences and Environmental Sciences are casual search users with little experience in taxonomies and ontologies. However, as FAIR data management is getting increasing attention in academia, it would be helpful to better understand what additional information should be present in a dataset search going beyond keywords. For the user evaluation, we mainly follow the TREC Interactive Search track (Chapter 2.2.3) guidelines with search tasks and questionnaires. However, we adapted that concepts for our needs. Further information can be found in Section 8.4.

Study goal: Based on the above mentioned prelimary considerations, we formulate the following research goals: (A) We want to explore whether users prefer a form-based search input with given domain-specific categories or whether they favor the classical single input field search. Second, (B) we want to analyze whether users prefer a specific explanation strategy (less information versus explicit information on matching entities, used terminologies and full queries). These high level goals need to be further specified with concrete usability criteria. Usability is characterized by five components: Learnability, Efficiency, Memorability, Errors and Satisfaction (Chapter 2.1). Here, we only address four out of five components as Memorability is a characteristic that can only be measured when the system is shown to users at least a second time. Therefore, we focus on the remaining four components and add further characteristics for the second goal. The information in brackets gives first ideas on how to measure this study goal being further explained in Section 8.4.

(A) Determine whether users prefer a form-based input or a classical one input field for a semantic dataset search

- *Efficiency and Effectiveness of Use* (conduct user tasks):
  - *Completeness* (measurement of task success/ performance based metrics): Can users achieve their goals?
  - *Efficiency* (measurement of task time/ performance based metric): In what time?
  - *Navigation* (numbers of clicks/ performance based metric): How many clicks are necessary to reach the goal?
  - *Errors* (number of occuring errors or usability issues/ issue based metric): Do errors occur that prevent users from achieving their tasks?
- *Easy to Learn* (measured by questionnaires/ self-reported metric and observation/ issue based metric): Does the system architecture match users' mental model? Can users start doing their search task?
- *Satisfaction* (measured by questionnaires/ self-reported metric): What components do users like? What do they dislike?

(B) Explore whether the presence of entity information from knowledge bases and full query information supports the comprehensibility of search results in a semantic dataset search or whether it is disturbing and distracting

- *Comprehensibility* (questionnaire/ self-reported metric and observation/ issue based metric): Do users understand the provided information on additional knowledge resources?

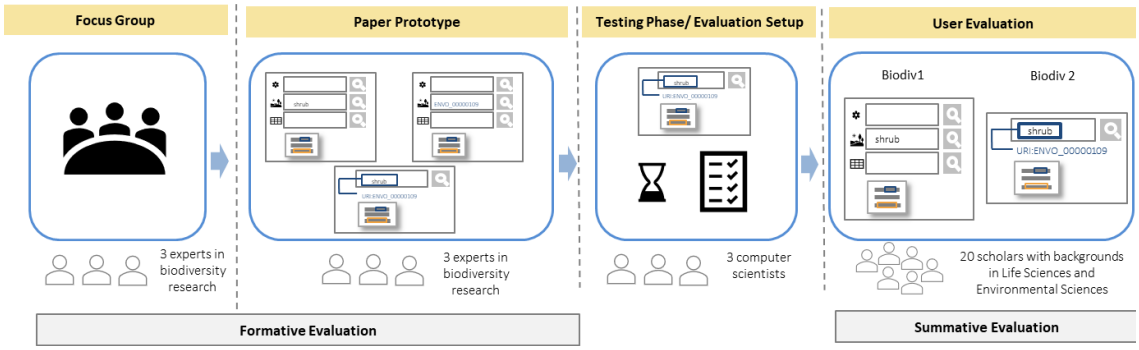


Figure 8.3: Overall development and evaluation flow: In a focus group, we discussed attitudes and expectations for a semantic dataset search. Based on these discussions, we provided three clickable paper prototypes to specify and revise the functions needed in the user interface. The focus group discussed and decided on the final user interfaces. After the implementation, system tests and evaluation setup took place in parallel. In a fourth phase, we conducted a user study with 20 domain experts.

- *Completeness* (questionnaire/ self-reported metric): Do users require more information on the provided data or is the given content sufficient?
- *Satisfaction* (questionnaire/ self-reported metric): What additional information as explanation do users like? What do they dislike?

In order to achieve these research goals, we discussed and implemented several user interfaces in close conjunction with domain experts. The overall development and evaluation flow is presented in Figure 8.3.

Evaluation and development flow: In a focus group, first insights on previous search experiences, attitudes and expectations were discussed with three experts from various projects in biodiversity research. Afterwards, based on these discussions and the preliminary considerations, three user interfaces were developed as clickable paper prototypes. They were again discussed with the focus group. Once the final decisions for two user interfaces were made, the interfaces were developed and tested. In parallel, we already setup the evaluation with surveys and user tasks following the guidelines of the TREC Interactive Track. Finally, the usability evaluation with 20 domain experts took place over a period of six weeks in June and July 2022.

8.2.3 Focus Group Meetings

A focus group is a qualitative research method to collect attitudes, expectations and prior experience in a group meeting with around 6-9 target users [Nielsen and Norman, 2022]. The discussion is usually only around a certain (focused) topic and provides the opportunity to get fast feedback on a product. Therefore, we utilized this method to gather insights before the development of the user interfaces has started.

Expert	Issue/ Use Cases
1	"I wanna search for biome" (= return all datasets for this biome)
1	"I wanna search for a species name" (but what they actually mean is a higher level in the taxonomy, e.g., a genus or a family and they think the system is already smart enough to return all narrower species)
1	"I wanna search for a location." (return all datasets for this location – data collected/observed in that area)
1	"How many of these datasets have replicats?" (Has it been already reproduced or reused?)
2	datasets with data of family Apidae on given plots
2	abundance data of wild bees (listed in the red list) in Baden-Württemberg, extract the list of protected species from those datasets, get coordinates or location information about the site where the bees were found and the name of the persons identifying the bee species
2	search for datasets describing silvicultural management strategies and management intensity of grasslands, search for soil data and climate data of grasslands

Table 8.2: Use cases collected in the focus group meetings.

The focus group consisted of three postdocs with a background in biology and expertise in biodiversity research and a moderator (the thesis author). We did not ask more scholars, as we wanted to have various insights from different projects, and we wanted to keep the pool of scholars being suitable for the subsequent evaluation as large a possible. The three focus group members did not take part in the final user evaluation. We conducted three online sessions of two hours each in winter 2021/22. The focus in the first meeting was on general attitudes and expectations, in the second and third meeting we discussed the clickable paper prototypes.

Experience and expectations: We asked the expert group if they have own experiences with dataset search applications or semantic search approaches. If so, we also asked for concrete examples, problems or any other occuring issues. None of the three experts had own experiences in dataset search or semantic search approaches. However, all of them are in close contact to scholars who need to search for datasets frequently. Two of three experts are data curators, and one Post-Doc is a scientific coordinator of a research data management unit. Two experts report on some issues or use cases in dataset search presented in Table 8.2. The listed issues are similar to the search interest introduced in Chapter 4 and also reflect the diversity with respect to granularity (broad questions such as data about a specific biome versus detailed questions, e.g., data on specific species) and topic (species (ORGANISM), biome (ENVIRONMENT), abundance data, climate data (DATA TYPE), given plots (LOCATION)).

To the question about expectations on a semantic search, the experts answered that users "don't care about the technology" and they would believe that the system returns all concrete species when looking for a higher taxa level, e.g., bacteria or insect. Another

	A	B	C
Input type	form-based input (categories)	form-based input (categories)	one query field
Input terms	keywords	keywords - entities	keywords
Auto-completion	search index	terminology index	search index
Entity Linking	automatic	by user	automatic
Entity expansion	automatic entity expansion on descendants	no automatic entity expansion but only on demand	automatic entity expansion on descendants
Query Explanations	no URIs, no ontology information	URIs and ontologies shown	query interpretation is shown, explanations without URIs and ontologies but visible on demand
Full query	not shown, only a simple query	fully query displayed including SPARQL	excerpt of query presented, full query on demand

Table 8.3: Proposed user interfaces - concepts

Focus group member said that it depends on the background and experience of users: "If users have the experience that the system only returns an exact match they don't expect more detailed results. However, if you want to get an overview, e.g., to see what data are in the repository, search filters help to start the search."

Concerning improvements, the experts agreed that a search over domain categories would improve the search experience. In general, they found that existing data portals are already a big step. Ten years ago no data portal was available and scholars searched for literature first. Only in a second step they looked for associated data. Nowadays, scholars go to e.g., Web of Science[19] or GenBank (for genoms, now part of the International Sequence Database Collection (INSDC)[20]), and a few scholars have also started using the Google Dataset Search[21].

Paper prototype: Based on the insights of the first meeting, we proposed three user interfaces as clickable paper prototype and discussed these sketches with the Focus Group in two further sessions. Table 8.3 presents the concepts for the proposed user interfaces (UI). All three UI concepts have in common that the presentation of search results adheres the RDA recommendation for dataset search. The main difference between the three UIs is the display of semantic information. In **UI A** (Figure 8.4), the idea is to hide the semantic enhancements and to link keywords to entities automatically to provide no information on utilized terminologies and URIs and to present only a simple query explanation. The query terms are entered into several pre-defined categories (form-based approach), which are considered in the query, and the auto-complete function presents the

[19]Web of Science, `https://clarivate.com/products/web-of-science/`
[20]INSDC, `https://www.insdc.org/`
[21]Google Dataset Search, `https://datasetsearch.research.google.com/`

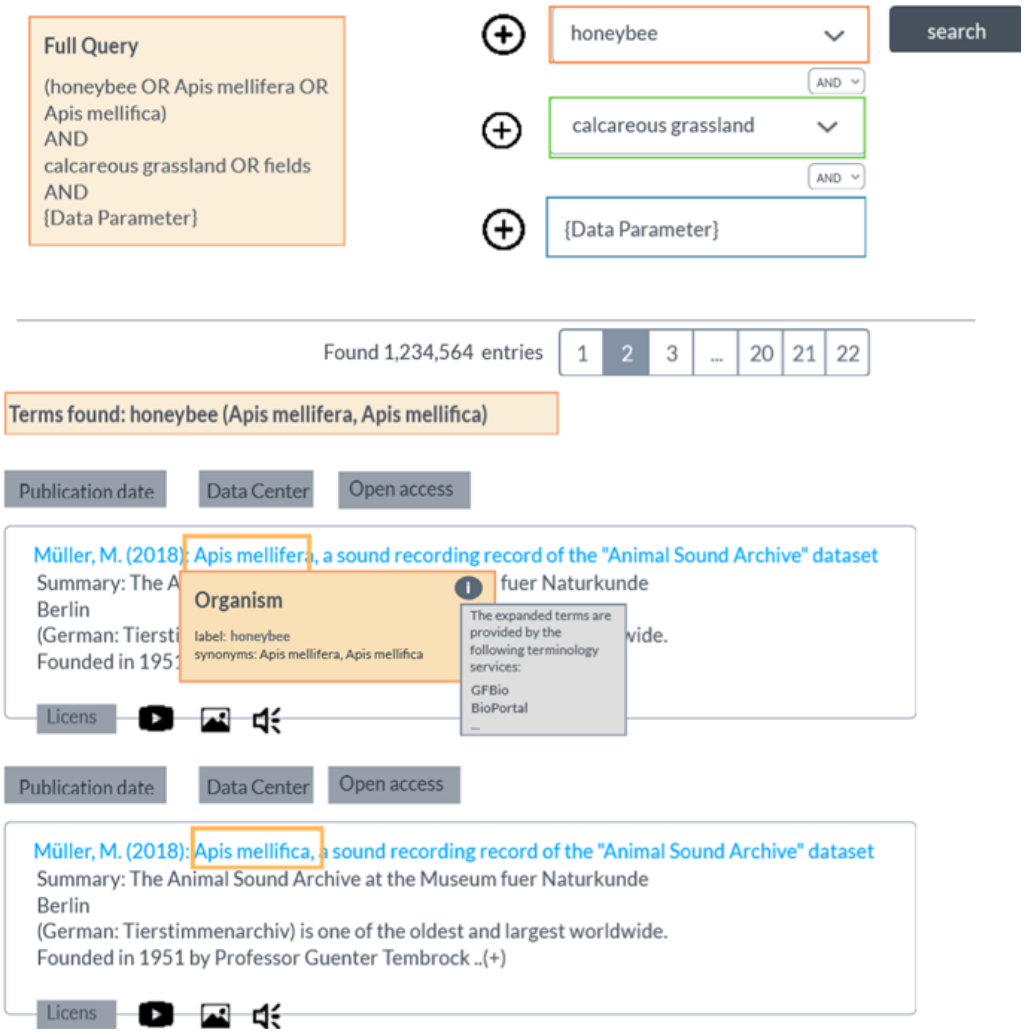


Figure 8.4: Paper prototype of user interface A

terms from the search index and is provided per category input field. This user interface takes all the burden from users to deal with semantic information and is supposed to be the most convenient approach. **UI B** also follows a category-based input, but in contrast, users have to link the typed keywords per category explicitly to suggested entities from the terminology index, or users have to provide URIs explicitly. There is also no automatic expansion on descendant nodes, but users would have the opportunity to select an expansion strategy. All semantic information are presented in detail including matching entities and ontologies as well as a full query with SPARQL statements. This UI gives users the most freedom to steer the semantic search process but requires some semantic web knowledge. The third suggested UI, **UI C** (Figure 8.5), provides only one input field and is similiar to all existing dataset search systems. It follows a hybrid approach with an automatic entity linking and entity expansion. The matching entities are shown in a query interpretation section and only an excerpt of the full query is shown to the user. On demand users can expand the full query.

The focus group's feedback on the UI A was quite positive. In contrast, they assessed

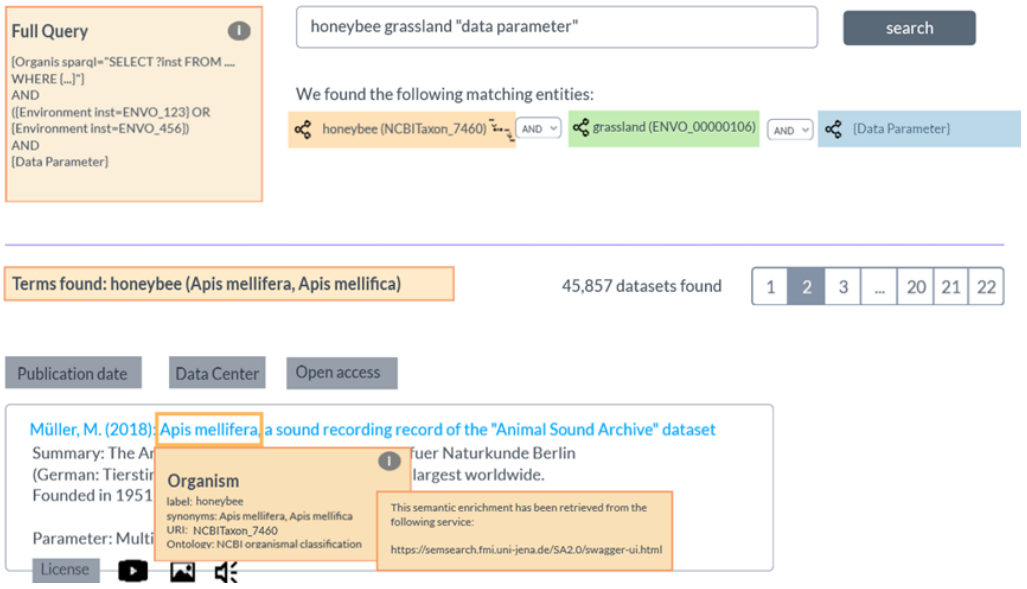


Figure 8.5: Paper prototype of user interface C

the second interface (UI B) as too complex with too much knowledge needed to fulfill a search task. The third UI (UI C) got a mixed feedback from the focus group. The experts were unsure whether matching URIs confuse or delight users. Originally, we wanted to conduct the evaluation with two user groups, scholars who are not familiar with the Semantic Web and scholars who have already some experience in semantic web technologies (at least in its usage, e.g., working with ontologies). All three experts of the focus group agreed that it would be very difficult in recruiting a sufficent number of scholars with expertise in the Life Sciences and Semantic Web. Therefore, we finally decided to go for a 'within subject' evaluation with only one user group, namely scholars with background in the Life Sciences and Environmental Sciences. Due to these decisions, we agreed to omit the second UI and to implement only UI A (Figure 8.4) and C (Figure 8.5) with the following adaptions in UI C: Full query information is shown at once. Matching URIs and ontologies are presented explicitally and not on demand including links to the respective ontology entries. These adaptions make the distinction between the two UIs clearer.

8.3 System Architecture and Implementation

The overall architecture of the user interface component is presented in Figure 8.6. We named the system *[Dai:Si]* in accordance to the phonetic relationship to the terms 'dataset search'. We present a new version targeted at functions needed for the user evaluation[22]. An earlierer version of [Dai:Si] has been presented at the S4biodiv workshop [Shafiei et al., 2021]. The backend consists of the local terminology service introduced

[22][Dai:Si] Semantic Search, `https://github.com/fusion-jena/daisi-semantic-search`

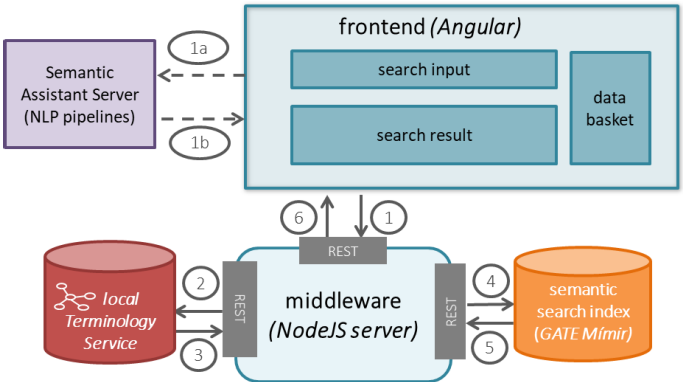


Figure 8.6: *[Dai:Si]*'s architecture and overall flow.

in Chapter 6.3.2 and a semantic index over metadata files generated with GATE Mímir (Chapter 7). In order to link query terms to concepts in ontologies, we use an updated version of the Semantic Assistants framework [Witte and Gitzinger, 2008]. In this new version based on a spring boot application, GATE pipelines are exposed via REST interfaces[23]. Currently, three pipelines are available (GATE ANNIE [Cunningham et al., 2013], BiodivTagger [Löffler et al., 2020] and OrganismTagger [Naderi et al., 2011]). The middleware was implemented with *NodeJS*[24], and the frontend is an *Angular*[25] application. The *NodeJS* server provides REST APIs to communicate with the backend applications.

As modularity is a main characteristic of [Dai:Si], the domain and business unit are separated from functional components. This allows an easy replacement of a search index or permits to add further search indexes. Each search index needs an own module being added to the middleware. In order to utilize the frontend components, the search index needs to support specific metadata fields according to the data model. More details are described on our GitHub page. Replacing the terminology service is also possible. The middleware provides a configuration file with different settings such as the URL to the terminology service, the search index and the term suggestion service. To the best of our knowledge, there are no protocols or standards for terminology services. Hence, replacing the service may cause adjustments to the middleware, in particular concerning the API requests and responses.

Both implemented UIs consist of two main components: a search input and a search result. A data basket is provided to collect datasets during search and to download the whole basket. To start a search, users enter keywords into the query input field (1), either per category or in the one input field (see paragraph below for the characteristics of the two different UI). For the one input field query, the query terms are sent to the Semantic Assistants first to obtain the entity types and if available URIs (1A - 1B). Then, the query terms are forwarded to the terminology service (2). Per query term matching URIs are

[23] SA2.0, `https://github.com/fusion-jena/semantic-assistants-2.0`
[24] NodeJS, `https://nodejs.org/en/`
[25] Angular, `https://angular.io/`

determined and returned to the middleware (3). Afterwards, these matching URIs are utilized to form full queries being sent to the semantic search index (4). The result (5) is returned to the frontend, and the datasets are displayed (6).

Dataset corpus preparation and indexing: For the user evaluation study we used selected metadata files from GFBio as data corpus. To ensure that appropriate data are available for the search tasks, we only downloaded metadata from the GFBio search API with relevant query keywords and expanded terms. As the aim of our semantic dataset search was to receive datasets beyond keywords, but GFBio only provides a keyword search, we followed an exhaustive query expansion process to obtain not only datasets that exactly match occuring terms in the search tasks but also synonyms and more specific terms. Therefore, we annotated 13 search tasks manually with categories (Organism, Environment, Data Parameter, Process, Material) and URIs from relevant OBO Foundry ontologies. These terminologies were the same as for the development of the BiodivTagger. We used the URIs to form a SPARQL query and to expand the search on narrower concepts. These additional labels and synonyms were added to the original keywords, and the expanded queries were sent to the GFBio search API. As these queries got very large, we cut them into blocks and ran the queries subsequently. We downloaded only the first 100 datasets per subquery to receive a corpus size that is managable and fast to index. We finally downloaded 52.000 metadata files. The code for the metadata download[26] and query expansion[27] are publicly available.

We annotated the obtained GFBio metadata files with the OrganismTagger and the BiodivTagger. This resulted in annotations for the following entity types: Organism, Environment, Data Parameter, Process, Material. We extended the taggers on GATE's Semantic Enrichment Processing Resource to add all parent nodes of each entity as semantic annotation (`subClassOf*` relations), too. This resulted in larger files in size, however, it allows a faster search at run time. SPARQL queries at runtime (e.g., in the search interface) can get very complex, and it might take several seconds or even minutes to return a result. Hence, to ensure a fast search result within a few seconds, we added hierarchy relations such as `subClassOf` relations in the pre-processing phase and used templates linked to these 'broader' annotations in the search phase. In our previous study in Chapter 7, we could not confirm a benefit for the consideration of semantic relations in search. Therefore, we utilized the CF-IDF ranking as retrieval model, which treats all semantic relations equally.

User interface A - Biodiv 1: Figure 8.7 presents a screenshot of [Dai:Si]'s UI A named 'Biodiv 1'. Keywords are entered per given category. Each keyword is looked up in the

[26] GFBio Metadata, `https://github.com/fusion-jena/GFBioMetadata`

[27] Query Expansion, `https://github.com/fusion-jena/semantic-search-usability-analysis/tree/main/data_corpus_preparation/QuestionExpansion`

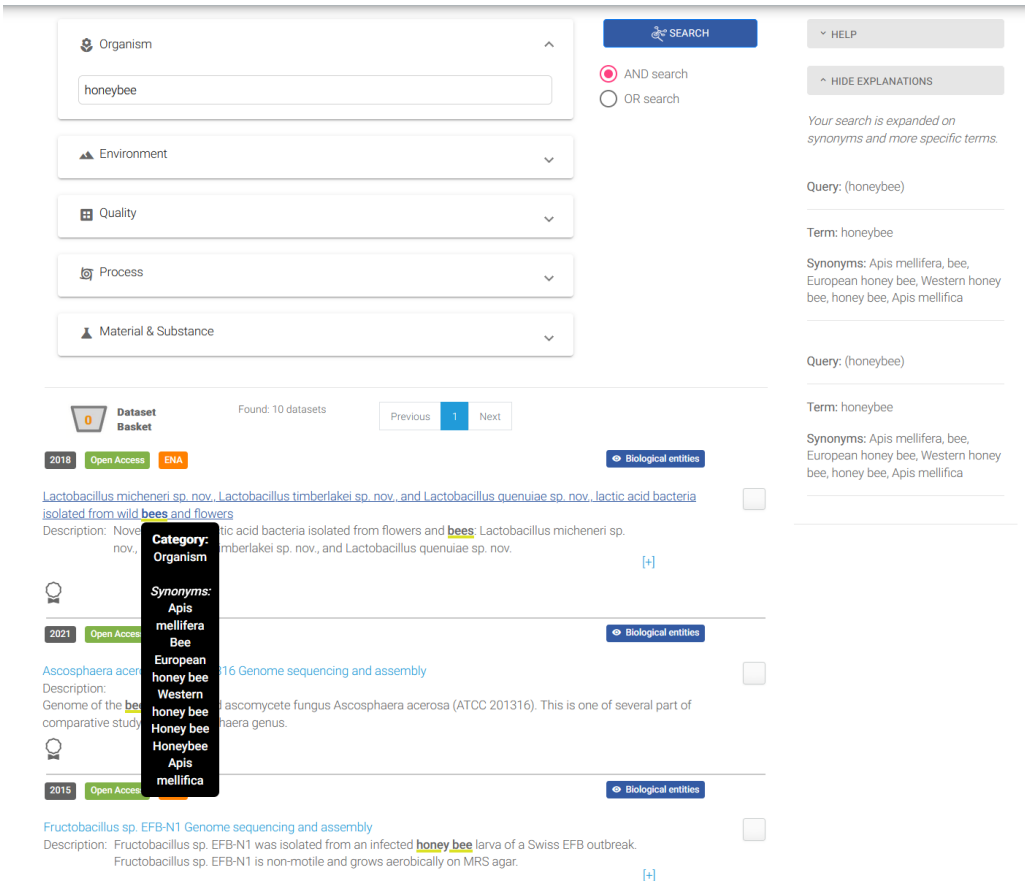


Figure 8.7: User interface Biodiv1 representing a category based input with no explicit semantic information.

terminology service. The matching entities and the selected categories are utilized in the final query to the index. We also consider the keywords as tokens in the query, ensuring a keyword search in case no matching entity can be found. However, also for a keyword search the selected categories are considered. The full query for Figure 8.7 is presented in Listing 8.1.

```
(((((‘‘honeybee’’ IN {Organism}) OR (‘‘honeybee’’ OVER {Organism}) OR (‘‘honeybee’’ AND {Organism}))
    OR ({Organism broader=‘‘http://purl.obolibrary.org/obo/NCBITaxon_7460’’} OR {Organism inst=‘‘
    http://purl.obolibrary.org/obo/NCBITaxon_7460’’})))
```

Listing 8.1: Full query for the search ‘honeybee’.

The result set contains datasets with the exact match of query terms as well as synonyms or if available further narrower concepts. The query terms are highlighted in bold fonts in the result set. In case URIs can be found for a keyword, query terms are also underlined in green and users can get more information by hovering over them. An explanation dialog displays the category and synonyms or alternate labels, URIs are missing in this user interfaces. In addition to the mouseover explanations, users also have the opportunity to get further information by expanding the ‘explanation’ tab. It shows a shortened query and again lists all query terms and their synonyms. The button ‘biological

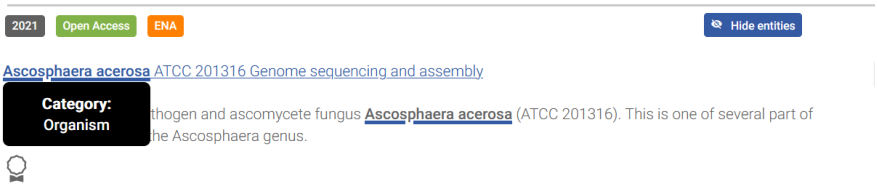


Figure 8.8: Screenshot with highlightings from the biological entity function in Biodiv 1.

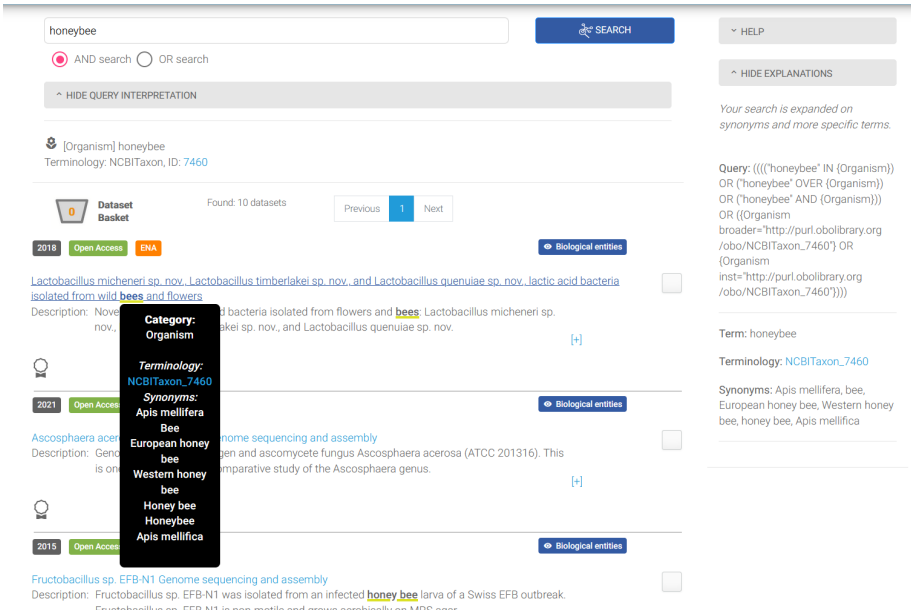


Figure 8.9: User interface Biodiv 2 with displayed URIs and terminology links.

entities' enables an additional highlighting of biological terms. The whole textual dataset information is sent to the Semantic Assistants. The system calls the BiodivTagger and OrganismTagger and returns a highlighted text to the user with the identified biological entities, e.g., data parameters, environmental terms, process, materials and species. These highlightings do not necessarily correspond to the search terms. To visualize this difference to the user, the biological entity highlights are underlined in blue color. Figure 8.8 presents a screenshot of a dataset result and active biological entity highlighting. Again, further semantic information such as URIs are missing in this highlighting.

User interface C - Biodiv 2: The second user interface, Biodiv 2 (Figure 8.9), provides a classical one query input field. The entered keywords are sent to the Semantic Assistants first to obtain categories and URIs. In addition, the keywords are also sent to the terminology service to increase the likelihood to find URIs. The obtained categories and URIs are presented to the user in a 'query interpretation' tab and are utilized to form the final query to the search index. The example in Figure 8.9 shows that the result is the same as for Biodiv 1 and hence is the same full query. The highlightings in Biodiv2 also provide URIs, in addition to alternate labels and synonyms. The same applies for

the explanation tab and biological entities function. All URIs are also linked to external terminology services (Ontobee[28] and Bioportal[29]) to provide the opportunity to look for further information. Instead of a simple query, this user interface provides full query information.

8.4 Usability Study

We followed the TREC-9 guidelines[30] for an A/B testing (Chapter 2.2.3) with user tasks and questionnaires. We asked 22 scholars to take part in the study and resulted in 20 final user sessions. Each participant had to do eight search tasks in different ordering. Ten users started with Biodiv 1, and the other ten users started with Biodiv 2. This setup allowed a within subject study to analyze whether there are differences for these two systems within the group. All 20 subjects had a background in the Life Sciences or Environmental Sciences and spoke and understood English. We did not ask for age and sex, but we requested some statistical information at the end of the study (graduation, research background, usage of dataset search applications).

Experimental design: The participants' task was to find datasets which are relevant for a given search task within a maximum of five minutes. Before and after searching the scholars were asked if they expect specific content in the datasets and how certain they are about it. It was a within-subject design and every subject had to search on all the questions. The questions were provided in pseudo-random fashion, with 16 variations to ensure that each task is performed at a different position (1st through 8th) for each system. Hence, a complete round of the experiment required 16 subjects. However, to have a backup we conducted 20 user sessions. For the last four users, we kept the same ordering of questions but with opposite systems. In the result section, we provide the results for 16 and 20 users, respectively. The final experimental matrix is provided in Table 8.4.

We visited most subjects at their working places, namely iDiv, Leipzig[31] | Senckenberg, Frankfurt[32] | Georg-August-Universität Göttingen[33] | TU Ilmenau[34] | BGBM Berlin[35] | UFZ Halle[36] | DSMZ Braunschweig[37], and conducted the study with four parti-

[28]Ontobee, `https://ontobee.org`
[29]Bioportal, `https://bioportal.bioontology.org`
[30]TREC Interactive Track, `https://trec.nist.gov/data/t9i/t9i.html`
[31]iDiv, `https://www.idiv.de/`
[32]Senckenberg, `https://www.senckenberg.de`
[33]Uni Göttingen, `https://www.uni-goettingen.de`
[34]TU Ilmenau, `https://www.tu-ilmenau.de/`
[35]BGBM, `https://www.bgbm.org`
[36]UFZ Halle, `https://www.ufz.de/`
[37]DSMZ, `https://www.dsmz.de/`

Subject	Block 1	Block 2
1	Biodiv 2: 4-7-5-8	Biodiv 1: 1-3-2-6
2	Biodiv 1: 3-5-7-1	Biodiv 2: 8-4-6-2
3	Biodiv 1: 1-3-4-6	Biodiv 2: 2-8-7-5
4	Biodiv 1: 5-2-6-3	Biodiv 2: 4-7-1-8
5	Biodiv 2: 7-6-2-4	Biodiv 1: 3-5-8-1
6	Biodiv 2: 8-4-3-2	Biodiv 1: 6-1-5-7
7	Biodiv 1: 6-1-8-7	Biodiv 2: 5-2-4-3
8	Biodiv 2: 2-8-1-5	Biodiv 1: 7-6-3-4
9	Biodiv 1: 4-7-5-8	Biodiv 2: 1-3-2-6
10	Biodiv 2: 3-5-7-1	Biodiv 1: 8-4-6-2
11	Biodiv 2: 1-3-4-6	Biodiv 1: 2-8-7-5
12	Biodiv 2: 5-2-6-3	Biodiv 1: 4-7-1-8
13	Biodiv 1: 7-6-2-4	Biodiv 2: 3-5-8-1
14	Biodiv 1: 8-4-3-2	Biodiv 2: 6-1-5-7
15	Biodiv 2: 6-1-8-7	Biodiv 1: 5-2-4-3
16	Biodiv 1: 2-8-1-5	Biodiv 2: 7-6-3-4
17	Biodiv 2: 4-7-5-8	Biodiv 1: 1-3-2-6
18	Biodiv 1: 4-7-5-8	Biodiv 2: 1-3-2-6
19	Biodiv 2: 8-4-3-2	Biodiv 1: 6-1-5-7
20	Biodiv 1: 8-4-3-2	Biodiv 2: 6-1-5-7

Table 8.4: Experimental matrix of the user evaluation

cipants in our lab at the FSU Jena. In a moderated live session, we collected quantitative and qualitative feedback. The moderator guided the participants through the whole study. The questionnaires were provided in an online form (LimeSurvey[38]) to facilitate the analysis. We provided a laptop (Windows NT 10.0, Win64), a mobile keyboard and a mouse. All search tasks and surveys were carried out in a Mozilla Firefox browser (Version 102.0). Thus, all participants took part in the study under the same technical conditions. We explicitly decided not to use the participants normal working environment, as different browsers and screens could have had an influence on the impression and satisfaction of the displayed search results.

The total evaluation time per user was around 120 minutes. The search part of the experiment took around 80 minutes. Each task was around ten minutes: One minute before the search to answer the questionnaire, five minutes to find the datasets, and two minutes after the search to answer questions about the result. Another two minutes were scheduled to go to the next question and to report usability issues. The participants had to add appropriate datasets to the data basket by selecting them. The moderator stopped the time per task manually. After five minutes at the latest, the task was stopped and the users had to download the basket.

The non-search part took around 40 minutes. We scheduled 25 minutes for the post system questionnaires (SUS score, two systems, ten minutes each) and the exit questionnaire with questions about a comparision between the two interfaces and some statistic questions. Another 15 minutes in the beginning were utilized for some introductory ex-

[38]LimeSurvey, `https://www.limesurvey.org`

Search tasks in the user evaluation
1.) What data are in the repository for Foraminifera (forams, single-cell organisms) in the benthic zone (water layer in the ocean floor)?
2.) How variable is the oxygen concentration of sea water of the global ocean?
3.) What data exist for Poales (invasive grasses), e.g., Poaceae (grass family)?
4.) How high are sulfate reduction rates at cold seeps (cold vents, areas in the ocean floor where hydrocarbon-rich fluids are leaking)?
5.) What data are in the repository on ocean acidification or coral bleaching?
6.) What data exist in the repository for bacteria in the groundwater?
7.) What data exist for Lepidoptera (butterflies, moths) on oaks (Quercus)?
8.) What data in the repository contain samples from surface water?

Table 8.5: Search tasks in the user evaluation

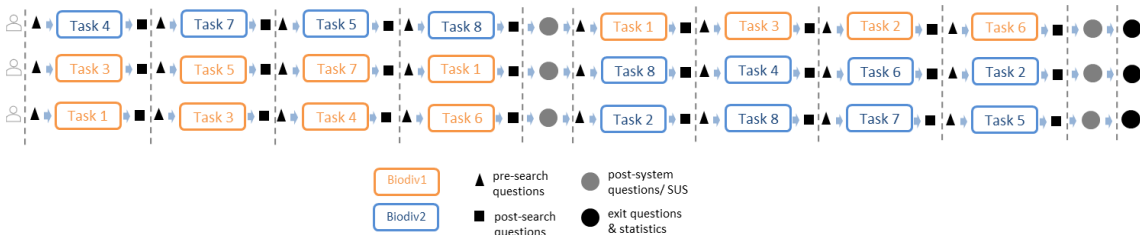
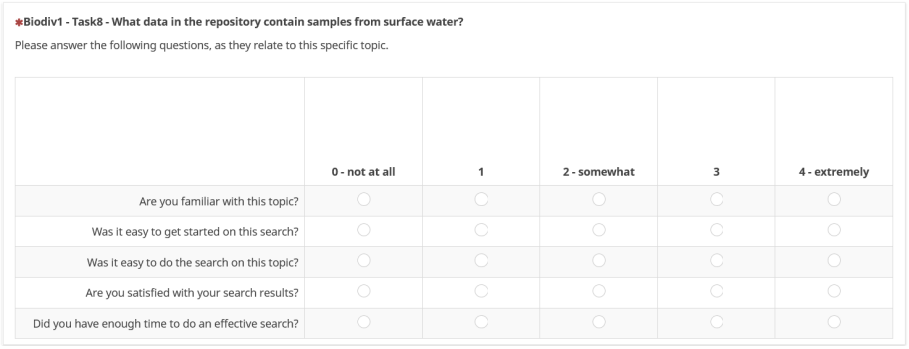


Figure 8.10: Overview of the overall survey flow for the first three users.

planations (welcome, purpose of the study) and some training information. We wanted to minimize the 'training effect', so that participants can do and start the search immediately. We told the participants that the first interface needs keywords and categories as input, while the second user interface recognizes the categories on its own. We also explained that both interfaces provide a search beyond keywords and that the system automatically adds synonyms and more specific terms of a query term. We briefly explained the highlighting: bold font denotes search terms and its relative terms, a green underline allows a further inspection of synonyms and the biological entity function highlights additional biological terms (not necessarily related to the search term) in blue. We did neither explain the individual categories nor the differences in the explanations.

Search tasks: The largest group of question types in the question corpus study were factoid questions (Chapter 4). Thus, we solely selected search tasks of this question type from that corpus as it denotes the highest demand on information needs. Furthermore, we also looked for questions where data is available in GFBio. In addition, the questions also had to match the studied entity types relevant for biodiversity research, namely Organism, Environment, Material, Process, Quality. The final eight search tasks are presented in Table 8.5. We provided brief explanations on the main keywords, but we permitted the subjects to look for further descriptions in external services such as Wikipedia[39] if needed. We also encouraged them to search with other keywords than the provided ones.

[39]Wikipedia, `https://www.wikipedia.de/`



*Biodiv1 - Task8 - What data in the repository contain samples from surface water?

Please answer the following questions, as they relate to this specific topic.

	0 - not at all	1	2 - somewhat	3	4 - extremely
Are you familiar with this topic?	○	○	○	○	○
Was it easy to get started on this search?	○	○	○	○	○
Was it easy to do the search on this topic?	○	○	○	○	○
Are you satisfied with your search results?	○	○	○	○	○
Did you have enough time to do an effective search?	○	○	○	○	○

Figure 8.11: The post task questions provided after each search task.

Questionnaires: The users were guided through all questionnaires and search tasks in a survey. For each user, we prepared a separate survey according to the experimental matrix (Table 8.4). The overall survey flow is shown in Figure 8.10. Before each search task, the users were asked if they expected a certain content in the datasets (in case they were familiar with the research topic) and how sure they were about their answers. After each search task, they were asked why the selected datasets were relevant, and they had to answer questions on satisfaction, ease of use, ease of learnability and time (Figure 8.11). The questions were mainly taken from the NIST questionnaires, but were adapted to dataset search. The full questionnaires are provided at Zenodo [Löffler et al., 2022c].

After using each system, users had to fill out a SUS questionnaire. We adapted the ten statements according to the topics we wanted to explore. Therefore, the questionnaire consisted of two questions on the query input, three questions on the default highlightings, three questions on the biological entities highlighting and two questions on the provided query explanations. However, we followed the SUS schema with five positive and five negative statements.

Data collection and metrics: We utilized a mixture of performance-based metrics, self-reported metrics and issue-based metrics (Chapter 2.1.2) for the user study. Performance-based metrics were primarily used for the measurement of the efficiency of use. Occurring errors and usability issues (issue-based metrics) were collected orally in the live sessions as well as in comment functions in the surveys. The different questionnaires before and after each search task as well as the SUS questionnaires after each system provided self-reported metrics to measure satisfaction, learnability, comprehensibility and completeness. Table 8.6 presents the evaluation model with the evaluation criteria and the measurements.

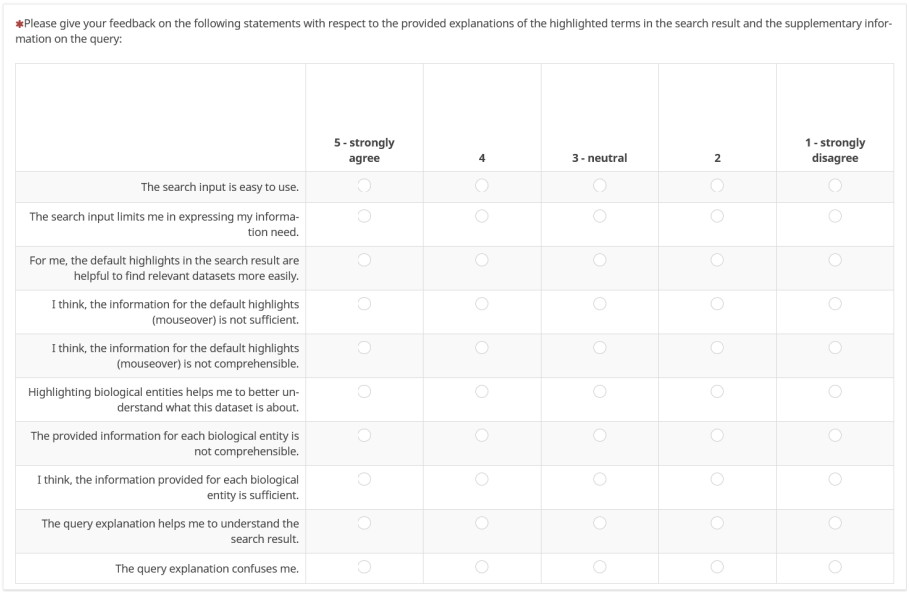
(A) Determine whether users prefer a form-based input or a classical one input field for a semantic dataset search

Criteria		*Measurement*
Efficiency of Use		User tasks: 4 search tasks per user interface, 5 min to solve the task
	Task success	Level of Success • complete success: at least 3 datasets are downloaded • partial success: 1 or 2 datasets are downloaded • failure: no datasets found/ downloaded or time over
	Task time	Per task, the time is manually stopped by the moderator
	Navigation	Log file analysis: the number of clicks counted per user interface until the download button of the data basket is pressed
	Errors	Errors and issues are collected orally in the live sessions and in comment functions in the surveys, they are classified into four categories: functional (technical issues), content (questions unclear or data missing), comprehensibility (confusing descriptions or functions), presentation (highlightings, icons, colors, fonts)
Easy to Learn		Questionnaaires after each search task, exit questionnaire, observation
Satisfaction		Questionnaaires after each search task, SUS questionnaire, exit questionnaire, orally described impressions during live sessions

(B) Explore whether the presence of entity information from knowledge bases and full query information supports the comprehensibility and acceptance of a semantic dataset search or whether it is disturbing and distracting

Criteria	*Measurement*
Comprehensibility of highlightings and explanations	SUS questionnaire, observation during live sessions
Completeness of highlightings	SUS questionnaire
Satisfaction of highlightings and explanations	SUS questionnaire
Usage of biological entity highlighting	Log file analysis: numbers of clicks on 'biological entities' button

Table 8.6: Evaluation Model presenting the explored criteria and their measurements


*Please give your feedback on the following statements with respect to the provided explanations of the highlighted terms in the search result and the supplementary information on the query:

	5 - strongly agree	4	3 - neutral	2	1 - strongly disagree
The search input is easy to use.	○	○	○	○	○
The search input limits me in expressing my information need.	○	○	○	○	○
For me, the default highlights in the search result are helpful to find relevant datasets more easily.	○	○	○	○	○
I think, the information for the default highlights (mouseover) is not sufficient.	○	○	○	○	○
I think, the information for the default highlights (mouseover) is not comprehensible.	○	○	○	○	○
Highlighting biological entities helps me to better understand what this dataset is about.	○	○	○	○	○
The provided information for each biological entity is not comprehensible.	○	○	○	○	○
I think, the information provided for each biological entity is sufficient.	○	○	○	○	○
The query explanation helps me to understand the search result.	○	○	○	○	○
The query explanation confuses me.	○	○	○	○	○

Figure 8.12: The SUS questionnaire provided after each system.

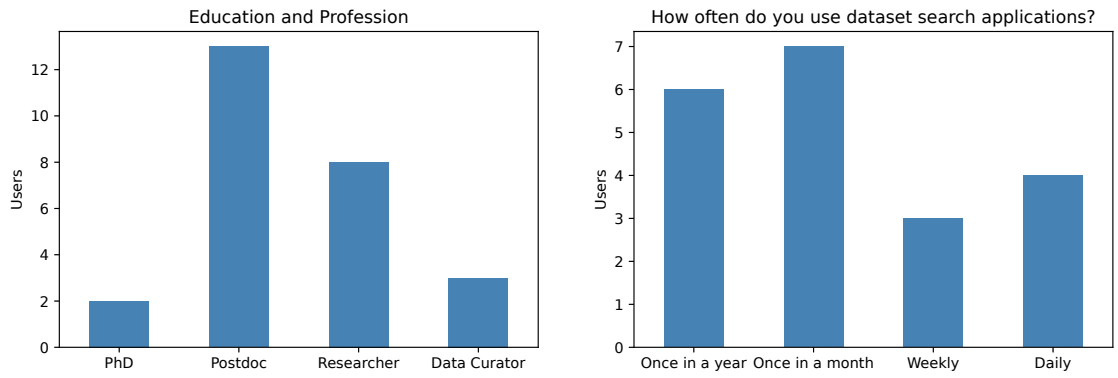


Figure 8.13: Statistical information on the participants.

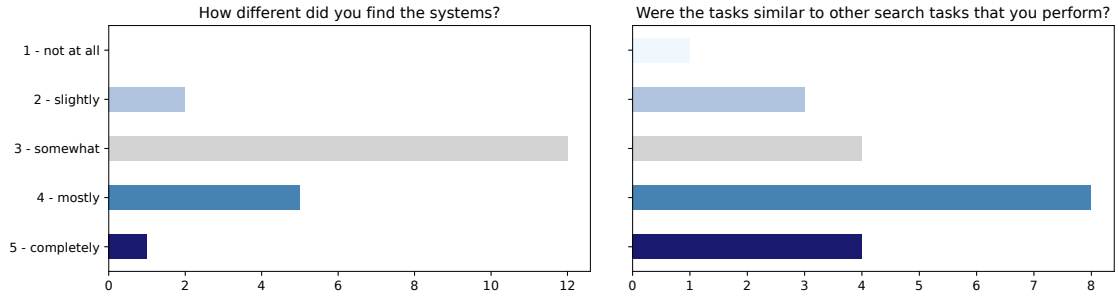


Figure 8.14: Results on the search experience (questions taken from TREC Interactive Track).

8.5 Results

The results of the surveys and observations as well as the task success were compiled into one large csv file. By means of a Jupyter notebook[40], we analyzed the individual results. The code is publicly available at GitHub[41].

Figure 8.13 presents some statistical information. All 20 participants are experienced scholars in the Life Sciences or Environmental Sciences with background in Botany, Ecology and Biodiversity. Two-third of them use dataset search application only from time to time, e.g., monthly or once in a year. Seven scholars indicated to look for datasets frequently (daily or weekly). The provided search tasks were similar to search tasks from the scholars' daily research practise (Figure 8.14). Thus, the selected search tasks represent real information needs.

The overall SUS scores are above average for both interfaces (68 for Biodiv 1 and 71 for Biodiv 2). This denotes an overall good usability for both interfaces as a value higher than 68 is desired. Figure 8.15 gives more insights on the dispersion and shows that in particular for Biodiv 2 the dispersion is large. The values for Biodiv 2 are in a range from 42 to 97. Looking at the results for each user, the result indicates that 11 users gave higher ratings for Biodiv 2 than for Biodiv 1. However, the SUS scores alone do not give a full

[40] Jupyter notebook, `https://jupyter.org/`
[41] Analysis, `https://github.com/fusion-jena/semantic-search-usability-analysis`

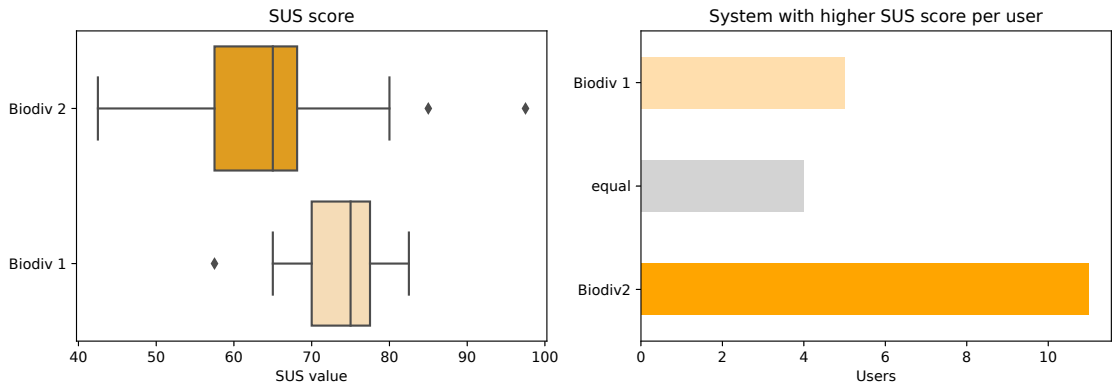


Figure 8.15: SUS score overall (left) and over all users (right)

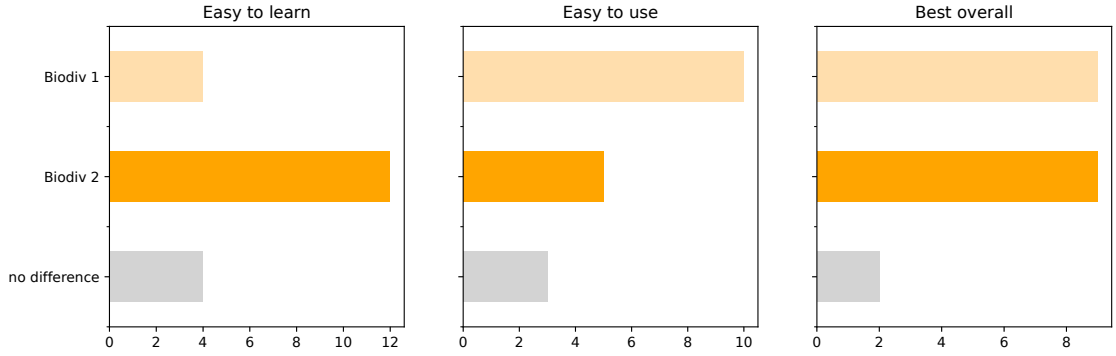


Figure 8.16: Results on the ease of use, learnability and best overall (Exit questionnaire)

picture. The results of the exit questions on the ease of use and learnability (Figure 8.16) show that Biodiv 1 was not so easy to learn but easier to use. Probably, that is the reason why there is no winner as 'best overall'. This result also correlates with the answers on the open exit questions on what users liked and disliked. Users liked the pre-defined categories in Biodiv 1, because they "helped to narrow down the search criteria and to get pertinent results." The term suggestion per category was also emphasized positively. However, a couple of participants also a liked Biodiv 2, because it is "easier, as it goes straight forward without thinking about categories" and it "is more general and helped in searching for topics that I was unfamiliar with".

8.5.1 Search Input

The overall results for task success and task time are depicted in Figure 8.17. In Biodiv 2, more participants were able to find 3 datasets being relevant for the search task than in Biodiv 1. In addition, the result for Biodiv 2 contains less failure cases. Concerning the time, the differences are small. The average times are 225 ms for Biodiv 1 and 204 ms for Biodiv 2, respectively.

A more detailed picture provides the task-based analysis. For almost all tasks users were able to find 3 datasets in both user interfaces (Figure 8.18). For only two tasks, the

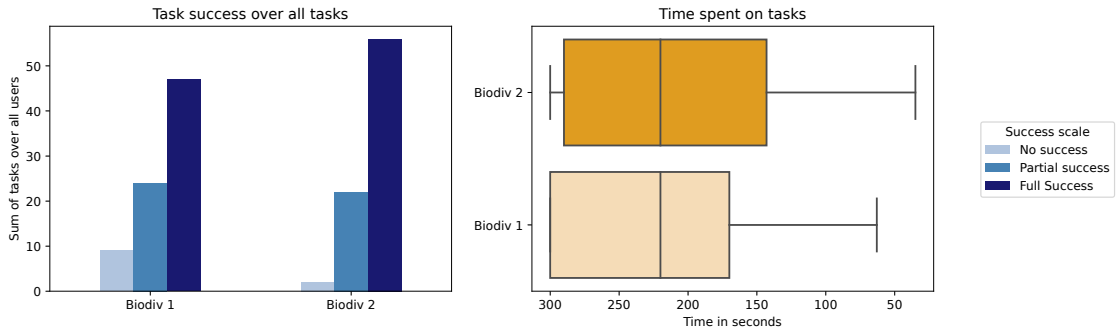


Figure 8.17: Task success and time in seconds.

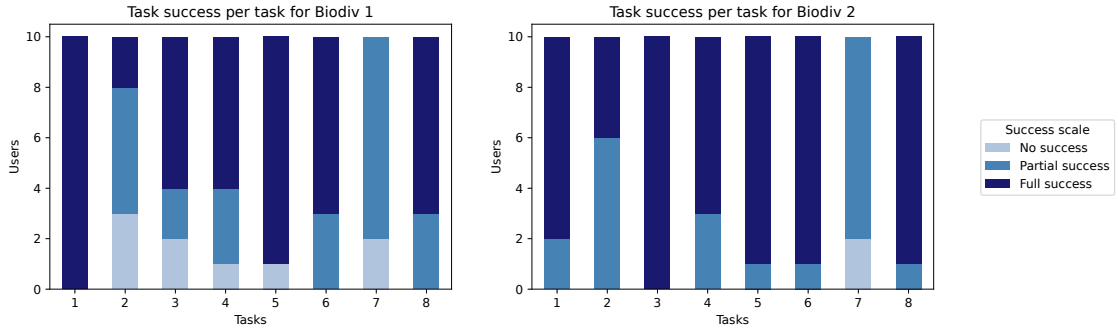


Figure 8.18: Task success per task.

task success is low. For task 7 (Lepidoptera on oaks), no participants found 3 relevant datasets. Looking at the orally reported and written comments gives further insights on these cases. Several users complained that the returned datasets for task 7 "are not relevant" or that only partial data is contained in the datasets, e.g., "Quercus contained in datasets but no butterflies?" The same applies for task 2 (oxygen concentration of sea water). Not many users were able to find three relevant datasets in both user interfaces. Here, the scholars argued again that "a few relevant datasets" were returned, e.g., the system "picks up carbon but I entered oxygen". Obviously, for these two tasks, the corpus did not provide sufficient data. Another circumstance to consider in the analyis of the task success is that most users were no experts in the proposed search tasks (Figure 8.19). As biodiversity research is a broad and heterogenous topic, this was also not to be expected. Therefore, with respect to the task success, the search interface with a single input field and automatic category recognition is more suitable for unknown search topics.

The total numbers of clicks for both user interfaces are 448 for Biodiv 1 and 456 for Biodiv 2, respectively. As these values are very close together, we can not draw any conclusion for the efficiency concerning the total clicks.

Table 8.7 lists the most occurring issues mentioned by at least three users and classified into four categories. Only one of the presented issues prevented users from fulfilling the search task. The search with three or more terms did not work at all time. As we had a few other network issues during the live session, e.g., WiFi connection lost, we

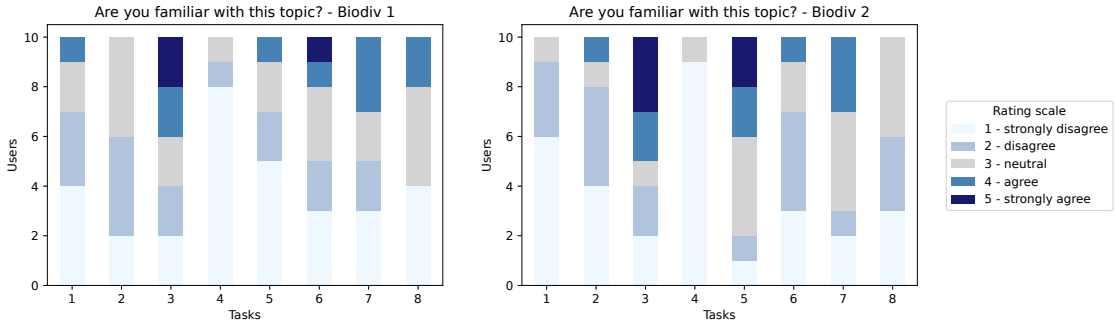


Figure 8.19: Topical expertise per task.

Functional	**Content**	**Comprehensibility**	**Presentation**
technical issues	*data or question related issues*	*confusing issues*	*issues concerning highlightings, colors, fonts, icons, buttons*
[Biodiv1] Delete button (Clear function) to delete all entries in the field would be good	Search returned unrelated results (various tasks)	Confirmation after "basket delete" is missing	highlighting of query terms missing
Sorting the search result by database, type and relevance, time would be an improvement	Not all Poales are invasive grasses (task 3)	"invasive" – unclear which category to choose	miss-highlighting or too many highlightings
Biological highlighting didn't work for all datasets	Too many similar datasets (task 4)	not clear how the results are sorted	
Search did not come back when entering more than 3 search terms	partially relevant data in the datasets (task 2)		
Facets should come up with the search result	less relevant data available, e.g., 'Quercus contained in datasets but butterflies?' (task 7)		

Table 8.7: Orally reported and written comments on errors and usability issues sorted into four categories. For each group, up to five issues are listed, reported by at least three subjects. If not stated in brackets otherwise, all listed issues occurred in both interfaces.

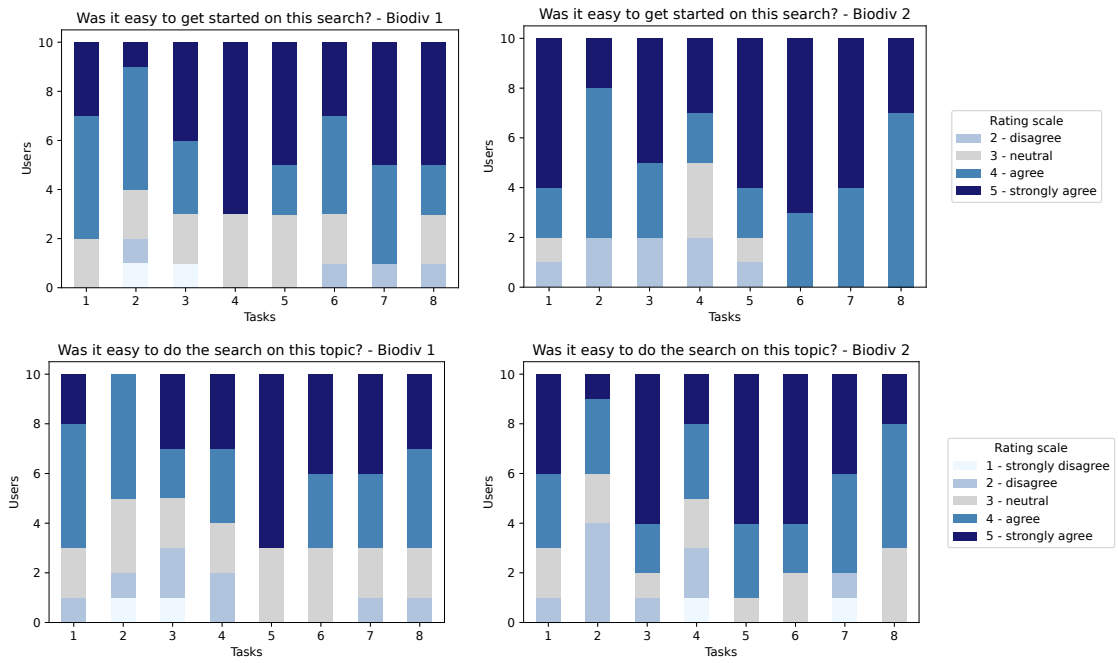


Figure 8.20: Learnability (top row) and easy of use (bottom row) per task.

assume that this mentioned issue is also related to the sometimes instable network connection. In general, both interfaces are capable to perform a search with more than three search terms. However, two classical usability issues have been detected. Several subjects missed a 'delete button' in Biodiv 1 to clean all input fields at once for the next search. And second, a confirmation statement is missing after the data basket has been deleted. Apart from problems or confusing issues, the participants also indicated further necessary improvements for a productive usage. Several times users requested a sorting by data repository or data type. This also addresses the reported issues on the unclear sorting in the comprehensibility column. Moreover, once the first search results are shown, users would like to have facets for narrowing the result. The participants also reported multiple times that the highlightings of query terms are missing, which is a hint that they had used this explanation and missed it in case it was not present.

Concerning the learnability, the results in Figure 8.20 (top row) reveal that the ratings for Biodiv 2 are slightly higher than for Biodiv 1. Hence, it is easier for users to get started with Biodiv 2. However, the results for the ease of use (Figure 8.20 (bottom row)) are balanced between the two user interfaces. Users were able to do the tasks with Biodiv 1, but it took some time to get familiar with the UI. These answers are also reflected in the results of the statements about the search input in the SUS score. On the statement "The search is easy to use", Biodiv 2 got the highest rating almost twice as high as the ratings for Biodiv 1. However, the sums of all high ratings (4 - agree and 5 - strongly agree) over all users are close together (Biodiv 1: 18 users, Biodiv 2: 19 users). The results for the overall satisfaction per task are displayed in Figure 8.21. The high ratings (4 - agree and

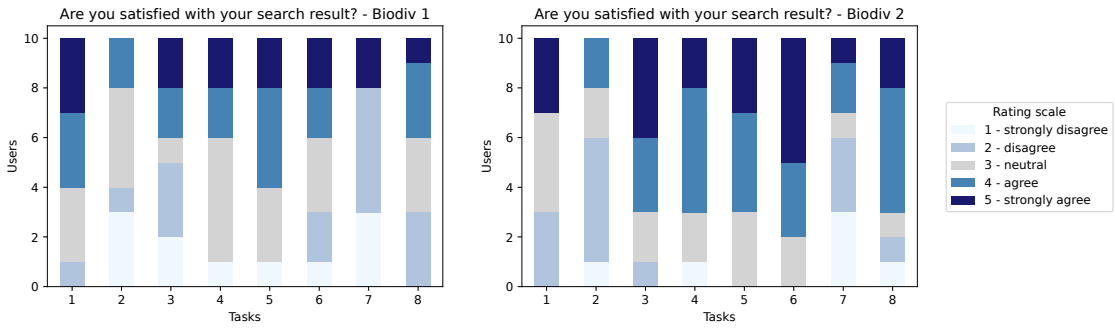


Figure 8.21: Satisfaction per task.

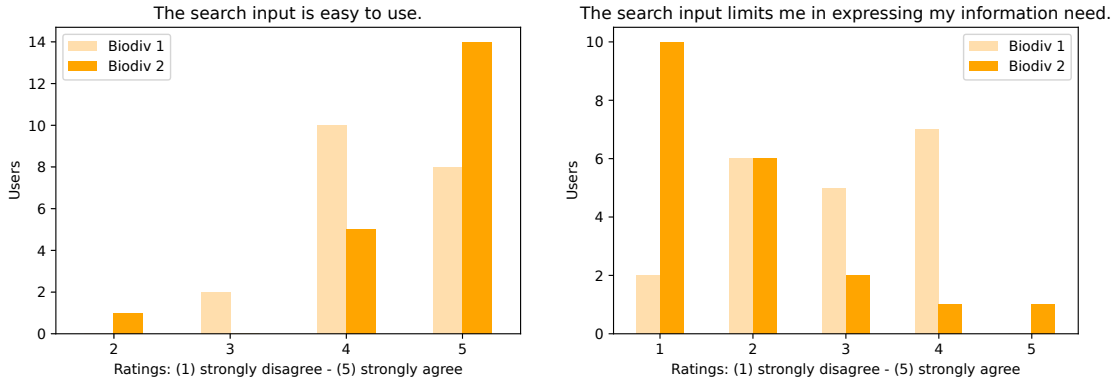


Figure 8.22: SUS questions to the search input.

5 - strongly agree) occur more often for Biodiv 2 than for Biodiv 1. Again, the answers on the statement “The search input limits me in expressing my information need” in the SUS questionnaire show that users tend to agree more with that statement for Biodiv 1 than Biodiv 2.

8.5.2 Highlightings and Explanations

For the majority of users (Biodiv 1: 15, Biodiv 2: 13), highlightingss of query terms and related terms are helpful (ratings of 4 and 5) in both interfaces (Figure 8.23 (left)). In most cases, the provided information is sufficient. However, in some datasets the highlightings were missing (see also Table 8.7) and users were a bit confused when they were not present. The participants also noticed some mis-highlightings and complained about too many terms being displayed in bold font. Hence, they gave a medium rating. The presented information for the highlighting of query terms is comprehensible for the majority of scholars in both interfaces (Figure 8.23 (right)). However, due to the mentioned issues several users gave only medium ratings, and some users also wish to get more explanations on the highlighted query terms.

Based on our observations, only 50% of the participants used the biological entities function to get further explanations of the biological terms mentioned in the datasets. In the logs, we counted 101 clicks on the button ‘biological entities’ for Biodiv 1 and 75

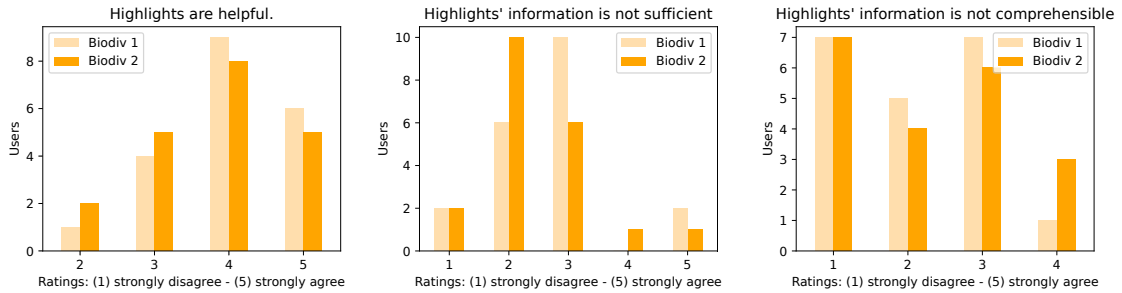


Figure 8.23: SUS statements on the default highlighting of query terms in bold font.

clicks for Biodiv 2. When rating the statements in the SUS questionnaire, some participants looked back into the user interface to figure out what this function is doing in order to be able to rate it. One user also gave an oral feedback that he "always give neutral ratings (3) when he is not familiar with something or when he has not used a function yet". In general, the participants liked the biological entities function in both interfaces (Figure 8.24 (left)). However, for datasets with longer text, e.g, abstract or description, the call to the Semantic Assistants takes a few seconds to come back. In worst case, the function did not return a result. Thus, we counted that as a functional issue (Table 8.7). The presented information for each biological entity is comprehensible (Figure 8.24 (middle)), however, it leaves room for improvement. For Biodiv 2, more than twice users as for Biodiv 1 indicated that the provided information is not sufficient (Figure 8.24 (right)). During the live sessions, for both user interfaces, we noticed that users looked up terms in Google[42] or Wikipedia[43] (which was permitted) to get further information. In particular, this happened for terms that were not highlighted or for which the biological entities function did not return a result. Hence, there is a demand for additional explanations, such as detailed descriptions for a term. Concerning the negative statements on the comprehensibility, for both kinds of highlightings, default (query terms and their relatives, Figure 8.23 (middle)) and biological entities (Figure 8.24 (middle)), the participants gave low ratings, which means they disagreed with the statements and the highlightings are comprehensible. The differences between both UIs in terms of the lowest ratings ((1) strongly disagree and (2) disagree) for the comprehensibility of the highlightings are too small to conclude which kind of explanation (the simple explantion with no semantic information or the extended explanation with terminology information and URI) is more useful. However, together with the results on the completeness and our observations, we infer that there is a tendence that additional terminology information and URIs do not confuse users and more additional information is desired.

Similar to the biological entities function, we also observed low usage of the query explanation tab. Thus, the ratings for the query explanations (Figure 8.24) are not that

[42]Google, `https://www.google.de/`
[43]Wikipedia, `https://de.wikipedia.org`

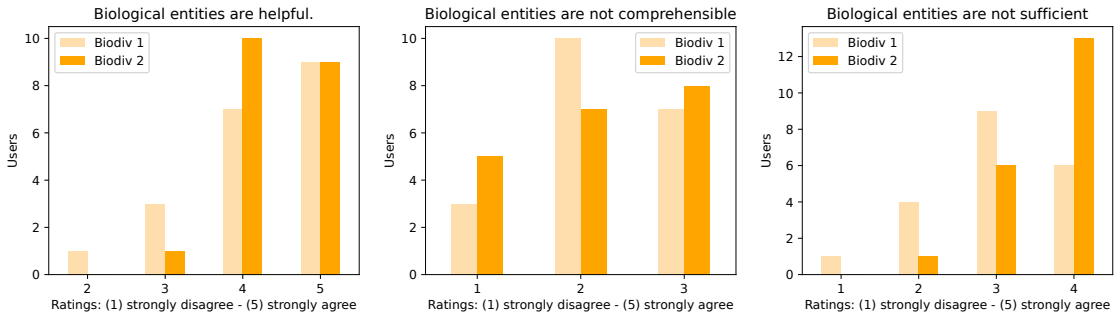


Figure 8.24: SUS statements on the biological entities function.

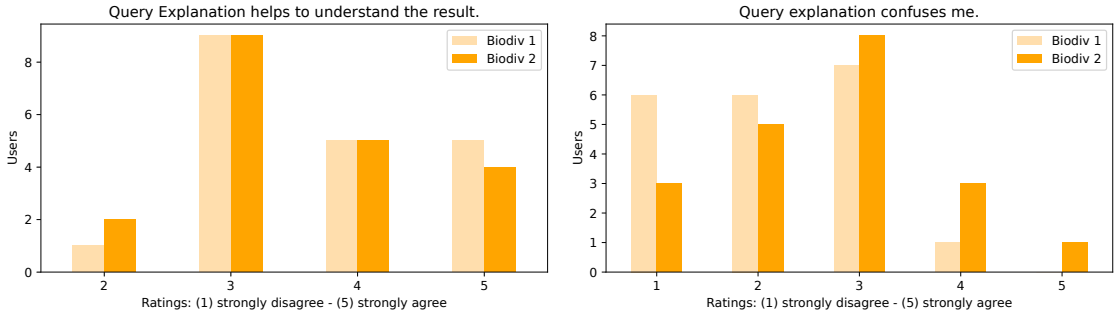


Figure 8.25: SUS statements on the query explanation

high and most users again gave a medium rating for the statement on the helpfulness. The ratings for the second statement on the comprehensibility show a slight preference for the simpler query explanations in Biodiv 1 without the full SPARQL information.

8.5.3 Discussion

Our presented results are based on the responses of 20 users. For four search tasks in two user interfaces, the official TREC evaluation matrix needs only 16 subjects. Therefore, we also computed all results for 16 users and provide the figures in the GitHub repository [44]. Figure 8.26 presents the SUS score and the responses on the exit questions for 16 users. The result reveals that for 16 users the dispersion in the SUS score is as large as for 20 users and more users resulted in a higher SUS score in Biodiv 2 than in Biodiv 1. However, the results for 16 users on the exit questions show that in terms of the ease of use Biodiv 1 was preferred. Therefore, the overall results for 16 subjects correspond to the outcomes with 20 subjects.

Hence, concerning the research goal to explore the different search inputs in a semantic dataset search (form-based with pre-defined categories versus a single input field with automatic category detection) we conclude that overall both interfaces are suitable for dataset search. For unknown search tasks, as provided in this study, a user interface with a single input field and automatic category detection is more suitable. For search

[44] Analysis, `https://github.com/fusion-jena/semantic-search-usability-analysis`

topics users are familiar with, a category-based search could be a good solution as well. However, for this assumption further long-term studies are needed. The highlighting of query terms and related terms are in general helpful. The provided semantic information on terminologies and URIs in Biodiv 2 did not confuse the subjects but were appreciated. In particular, the highlightings of the biological entities function were useful. However, further additional resources, such as links to Wikipedia, might be helpful to allow users to obtain further descriptions immediately. In terms of the query explanations our results show that full query information with SPARQL statements can be omitted. Simple query information on extended terms is sufficient.

The reported issues during the live sessions and in the comments show that further functions are needed to use our system in production. The most desired functions in both user interfaces are additional facets to narrow down the results after a search and filters to sort the datasets according to different metadata fields such as data repository, data type or collection date.

With respect to other studies in semantic search [Elbedweihy et al., 2012] [Vega-Gorgojo et al., 2016b], our results reveal that in a semantic dataset search both user inputs are proper search accesses. It depends on the expertise of the scholars what interface to take. For search tasks they are not familiar with (as provided in the study), the task success was slightly higher in the interface with one input field. However, users overall liked both proposed interfaces and gave positive feedback on the form-based category search.

Our study also provides some limitations. For at least two tasks, the corpus consisted of too little appropriate datasets influencing the task success. As we took care that all tasks are performed in both interfaces in the same number, it does not influence the overall result. In addition, some functions are still prototypical and not fully functional such as the highlighting of search terms and the biological entity function. However, the users' feedback on the highlightings and in particular on the biological entity function are overall positive. Future studies are needed to explore semantic dataset search in long-term usage allowing insights on daily working context with own information needs.

8.6 Summary

The last component of our proposed semantic dataset search focused on the comparision between a structured, form-based search input based on categories and a classical one input field. Second, we explored different explanation approaches. In particular, we wanted to know whether semantic information about utilized terminologies and URIs confuse users or whethers they appreciate this additional information. Both research questions address H_5.

We found out that the choice of the search input depends on the expertise of the schol-

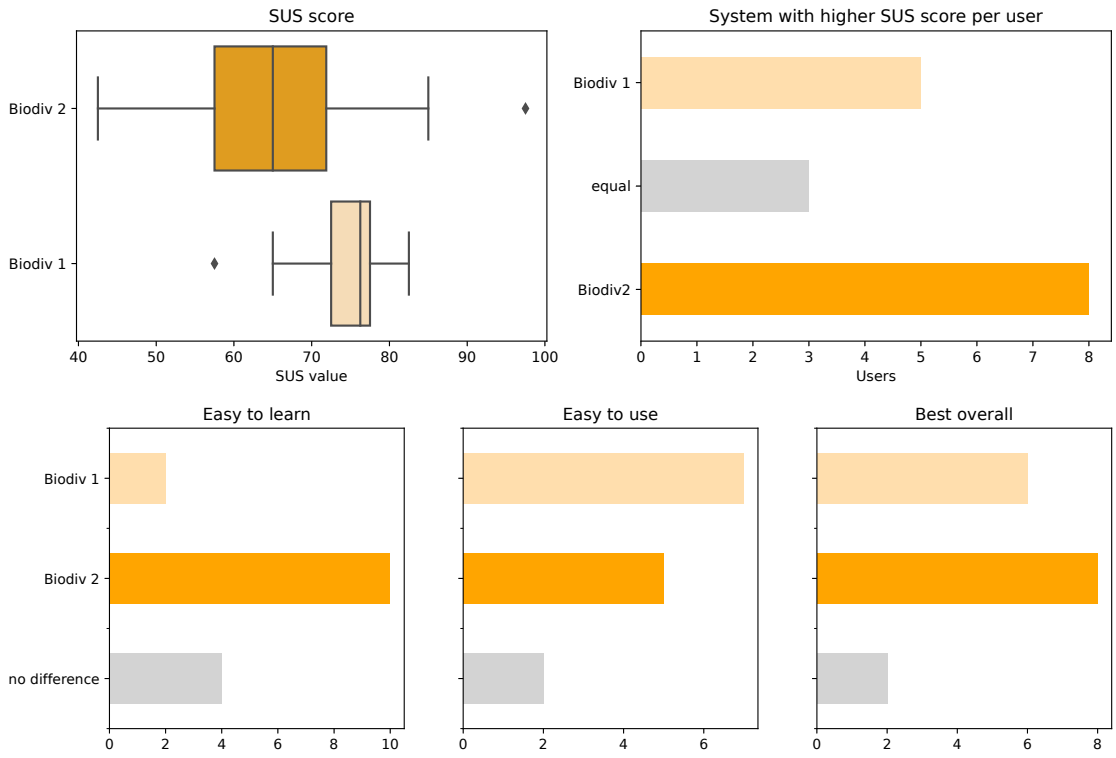


Figure 8.26: SUS score for 16 users (top row) and results of the exit questionnaire for 16 users (bottom row))

ars and how familiar they are with the search topics. In case users are no experts in the topic for which they need datasets, a classical one input field with an automatic detection of semantic categories is preferred. However, users gave positive feedback on a form-based search input with the opportunity to explicitely provide the semantic category. Here, further research with long-term studies is needed to explore the suitability in daily research practise. Therefore, our hypothesis has not been confirmed. However, we are also not completly wrong with our assumption. Concerning the proposed explanation approaches, our results reveal that full query explanations are not necessary but simple explanations on extended terms and utilized categories are sufficient. In terms of text highlightings, we conclude that it is highly recommended to highlight search terms and related terms in search results. On demand, users also like to obtain additional information such as the semantic category and linked terminologies and URIs. Moreover, further integrated resources such as Wikipedia could be helpful to provide additional descriptions on unknown terms.

Chapter 9

Summary and Discussion

> *"Some people call this artificial intelligence, but the reality is this technology will enhance us. So instead of artificial intelligence, I think we'll augment our intelligence."*
>
> - attributed to Virginia M. Rometty, *CEO of IBM, 2012 - 2020*

The emergence of the Semantic Web allows to link research artifacts, experiments, code and publications. Every single part of research can now be described in a way enabling machines to access data, to find data and hence to reuse data. The Semantic Web allows insights into data we have not seen before, and we can draw conclusions that were not possible before.

This work aimed to analyze the current obstacles in dataset retrieval. It was also targeted to develop a novel semantic dataset search supporting scholars in biodiversity research to find more relevant research data and to enable a semantic search beyond keywords. This chapter summarizes the findings (Section 9.1) and discusses the results (Section 9.2).

9.1 Contributions

In this work, we addressed two objectives and five hypotheses. In the following, we repeat the objectives, hypotheses and describe briefly how we addressed these research aims and what we have achieved. Limitations are presented in Subsection 9.1.1.

Objective 1: The first objective (O1) aimed to identify obstacles in current dataset search systems based on the three components in a retrieval system (questions, data, retrieval approach).

Hypothesis H_1 focused on the current issues in dataset search applications concerning the retrieval approach and the accuracy of the returned results. We wanted to figure

out whether scholars in biodiversity research are able to find relevant datasets in a current dataset search system. In a summative evaluation (Chapter 3.2), eight domain experts assessed 28 search tasks from biodiversity research. The scholars answered a questionnaire afterwards. This allowed us to get further qualitative feedback. In a second user study (Chapter 3.3), we explored the differences in a keyword search with no query expansion and a semantic search with query expansion including an adapted hierarchical expansion strategy. This gave us insights on the overall usefulness of semantic technologies for dataset search.

The evaluation of GFBio's dataset search resulted in moderate MAP and nDCG values (Chapter 3.2). Hence, users are able to retrieve relevant datasets in a current keyword based data portal with conventional retrieval models such as TF-IDF, but the results leave room for improvements. Therefore, our study is a contribution towards quantifying the current problems in dataset search. In general, users' impressions of a difficult and time-consuming search introduced in the problem statement are confirmed. However, the GFBio project, with PANGAEA hosting the search index, already aimed to improve metadata descriptions, and therefore, the project utilized a domain-specific metadata standard (panSimple) with domain-specific metadata fields being mostly filled (see also Chapter 4.4). We think, that these improvements already enhanced dataset retrieval and thus, the evaluation did not result in poor MAP and nDCG values as one could assume based on the descriptions in the problem statement. Concerning the user interface, users were overall satisfied with the query input and the presentation of search results, but the given control functions need to be better communicated. The comparitive study between a keyword and a semantic search (based on query expansion but including hierarchical terms) resulted in a larger number of relevant datasets with search queries expanded on synonyms and subclass labels, and the returned datasets are also highly relevant. Other semantic relations, in particular expansions on siblings and their descendants, require further explanations. Therefore our study reveals that the usage of terminologies and ontologies enhance the retrieval process in dataset search significantly.

The results of the comparative study have been published as a poster paper at *SEMANTiCS'16* [Löffler and Klan, 2016]. The findings of this study were incorporated into a first version of a semantic dataset search in the scope of the GFBio project, published as demo paper at *ESWC'17* [Löffler et al., 2017].

In **Hypothesis H_2**, we aimed to analyze information needs and metadata in biodiversity research. In order to get insights on what scholars in biodiversity research are interested in for dataset search, we gathered search questions in three biodiversity research related projects (Chapter 4.2). We inspected the questions manually, developed an information model and classified the noun entities into domain categories. These proposed categories were evaluated with nine domain experts. The scholars had to sort the prelabeled noun entities and adjectives into the given categories. This formative evaluation

(card sorting) aimed to verify the completeness and comprehensibility of the suggested domain classes and thus to verify the information model. In order to study whether these information needs are reflected in metadata, we analyzed metadata standards (Chapter 4.3) and utilized metadata fields in five large data repositories (Chapter 4.4). In a content study, we also explored whether existing text mining pipelines are able to identify important entity types for biodiversity research and to enhance and improve metadata descriptions.

One major outcome is a question corpus with 169 search questions collected in three biodiversity research related projects (Chapter 4.2). Based on these questions, we grouped noun entities into domain categories and evaluated the proposed classes with domain experts. Seven out of 13 proposed domain categories turned out to be highly relevant (Environment, Quality, Material, Organism, Process, Location and Data Type) in scholarly search interests. The study also showed that information needs in biodiversity research are very diverse. They have a wide range in terms of complexity and can change over time. However, a large portion of artifacts and phrases in the search questions could not be grouped into any of the given categories. Hence, our results also show that the determined categories do not form a complete picture and further discussion in the community are needed to extend, to merge or to rename categories. Based on the main existing metadata standards for the Life Sciences, we analyzed metadata in five large data repositories on what metadata standards they utilize and whether these metadata fields contain scholarly information needs. Out results reveal that general data repositories tend to use only general metadata standards. Domain-specific data repositories are more likely to utilize domain specific standards. This result might not be surprising, however, what stands out is that even if a domain specific metadata standard is utilized, the respective fields are not always filled. Moreover, in general standards, the field `dc:subject` can be utilized to provide domain specific information. However, even those fields are not always filled. In our content study, we showed that descriptive fields such as `dc:subject, dc:title and dc:description` indeed contain hidden information being relevant for information seekers. Therefore, we conclude, that current metadata reflect scholarly search interests only to some extent. A major problem are insufficiently described metadata. The results of Hypothesis H_2 (Chapter 4) were published in a journal paper [Löffler et al., 2021].

With the results of H_1 and H_2 **Objective 1** is achieved. Obstacles in data are sparse and insufficient descriptions. Information needs in biodiversity research are diverse, have a wide complexity range and can change over time. Concerning the retrieval process, it turned out that conventional keyword based retrieval approaches result in moderate accuracy values and semantic expansions help to retrieve more relevant datasets. In the user interface, improvements are needed in terms of explanations of search results and control options.

Objective 2: The second objective (O2) focused on the design and development of an improved dataset search taking scholarly information needs into account, describing scientific data appropriately and returning relevant results being comprehensible and reusable. Three hypotheses were related to this goal.

Hypothesis H_3 argued that descriptive metadata fields, such as title or abstract, contain useful information implicitly, reflecting information needs of the respective domain. This hidden data could be extracted and semantically enriched with text mining techniques. We analyzed this hypothesis by means of a text mining pipeline *BiodivTagger* representing the pre-processing component of our proposed semantic dataset search (Chapter 6.4). For the evaluation, we created a metadata gold standard with manually labeled categories (Chapter 6.2). Extracting domain specific entity types and URIs was also the first step towards a search beyond keywords (*Requirement R1*) and a search within categories (*Requirement R2*).

The inter-annotator agreement measure of the labeled metadata corpus QEMP showed that biological entities are fuzzy and difficult to classify. Even though a thorough training of the annotators took place, the f-score values were low to moderate (Chapter 6.2). The novel text mining pipeline *BiodivTagger* was developed based on rules and look-up services in selected domain specific terminologies. The evaluation results with the QEMP corpus show moderate results (Chapter 6.4). For the categories MATERIAL and QUALITY we achieved good results in terms of the recall values, but for the other entity types the precision and recall values point to a moderate result. However, our *BiodivTagger* pipeline is the very first text mining pipeline for the extraction of important entity types and URIs for metadata in biodiversity research. We showed that domain ontologies (or selected nodes) are a good starting point for semantic annotations of metadata and that hidden information in metadata fields can indeed be extracted with text mining pipelines. Future work is needed to improve the current approach, e.g., with additional rules, machine learning techniques or ontology extensions (see also Chapter 10). The analysis and results of Hypothesis H_3 (Chapter 6) were published as conference paper at LREC2020 [Löffler et al., 2020].

In **Hypothesis H_4**, we aimed to evaluate various concept based retrieval models. We analyzed whether an entity based retrieval model considering semantic relations in ranking returns more relevant results than a concept based retrieval model using no additional semantic relations in ranking. Therefore, we implemented several entity based retrieval models and proposed various entity expansions based on core and hierarchy relations of the respective domain terminologies (Chapter 7.7). Based on the results of the pre-processing component, we established a semantic index over semantic annotations (URIs and entity types), which also addressed *Requirements R1 and R2*. For the evaluation of biodiversity metadata, we created a test collection with 14 questions and binary human judgments for 372 metadata fields from the BEFChina project (Chapter 7.4). In order to

evaluate the models and expansions in another Life Sciences related domain, we utilized the bioCADDIE test collection with 15 questions and about 792.000 datasets.

Besides the test collection, the implementation of three existing entity based retrieval models being publicly available as GATE Mímir plugins is another contribution (Chapter 7.7). The evaluation on different entity expansion strategies with two dataset corpora (BEFChina and bioCADDIE) showed significant increased Precision values for the entity expansion with descendant nodes of the linked ontologies (Chapter 7.8). The second best entity expansion strategy was the expansion with sibling nodes and their descendants with respect to the baseline without any expansion. In contrast to our assumption, the expansion with core relations such as `partOf` did not result in more relevant datasets. Concerning the inspected entity retrieval models, we could not draw any conclusion that a particular scorer leads to an improved ranking for a specific entity expansion strategy. The values for all metrics were very close, and the number of questions provided in the test corpora were less than usually provided at TREC competitions. We conclude that more studies are required to investigate whether more queries and search tasks are needed or whether there are other dependencies. The novel dataset test collection (Chapter 7.4) was published as short paper in the RIO journal [Löffler et al., 2021].

H_3 and H_4 both addressed *Requirements R1 and R2* (search beyond keywords and a search within domain specific categories). Both requirements were achieved by using a semantic index over URIs and categories. Due to an entity based retrieval model using URIs, a search beyond keywords includes synonyms and alternative labels if this information is provided in the ontology. In addition, entity expansion allows the integration of further semantic relations. The proposed architecture based on the GATE framework even enables SPARQL queries at run time. However, as SPARQL queries are too complicated for end users and can take several seconds or minutes to come back, we utilized templates in the frontend and SPARQL statements (to obtain related hierarchy concepts) in the pre-processing phase.

Hypothesis H_5 addressed the usability in a semantic dataset search (Chapter 8.4). The research aim was to figure out whether users are more efficient in a search over semantic categories in comparison to a semantic dataset search offering a one input field as search input (*Requirement R3*). In addition, we also wanted to explore various explanation strategies (*Requirement R4*). In particular, we were interested whether users appreciate semantic information such as URIs and terminology names or whether they prefer simpler explanations. To verify this hypothesis, we developed two user interfaces on top a semantic index and conducted a summative evaluation in a user study with 20 participants, eight search tasks and surveys before and after each search task, after each system (SUS questionnaire) and at the end of the survey (Chapter 8.4).

One outcome is a modular frontend framework called [Dai:Si] to easily replace or extend the search index, terminology service or the frontend technology (Chapter 8.3). It is

publicly available, and a first version is now integrated into the services of NFDI4Biodiversity[1]. Another contribution is a comprehensive user study with 20 scholars (Chapter 8.5). The usability evaluation revealed that for unknown search tasks users prefer the classical one input field search. However, the results for the SUS score and task success of the second proposed user interface with a category based search were very close to the results for the one input field search. Several users also emphasized the advantages of a category based search in oral or written comments. Hence, further long-term user studies are needed to explore dataset search applications in daily research practice. In terms of the search input, our hypothesis has not been confirmed, but we are also not completely wrong. The second focus in the user study was on the exploration of different explanation strategies. Our results reveal that simple explanations on extended terms and utilized categories are sufficient. Highlights of search terms and related terms in the search results are very helpful for users to easily recognize the relationship of the originally entered keywords and the datasets in the result set. Users were not confused by information about the linked terminology and URI. Moreover, further integrated resources such as Wikipedia could be helpful to provide additional descriptions on unknown terms. The first version of [Dai:Si] was published as demo paper at the s4biodiv workshop in 2021 [Shafiei et al., 2021].

Hypothesis H_5 addressed *Requirements R3 and R4*, a comprehensible search providing explanations and an easy-to-use search interface. As both user interfaces reached high SUS scores above the state-of-the-art threshold of 68 and explanations are given by highlightings and additional functions, both requirements are fulfilled.

Therefore, as H_3 - H_5 are addressed, **Objective 2** is also achieved: We proposed and developed an improved dataset search. We ensured a modular structure replace individual components easily. However, the pre-processing component is a domain specific solution. Setting the same infrastructure up for a new domain would require to establish a domain specific information model. Suitable terminologies and taggers need to be found or developed. The retrieval component is capable to replace or to add further domain categories, and a modular structure in the user frontend is also given. The search index only needs to provide specific fields of the data model (described in the documentation on GitHub).

Use Cases: In Chapter 5, we introduced two use cases that we aimed to improve with our semantic dataset search. Figure 9.1 provides screenshots of the first use case in Biodiv 1, where users have the opportunity to sort the key term ‘sea water’ into different categories. Depending on the selected classification (Material or Environment), the system considers this user input in the query and only returns datasets with ‘sea water’ in the specified context. A demonstration of the second use case is depicted in Figure 9.2. A

[1] GFBio@NFDI4Biodiversity, `https://search.gfbio.org/`

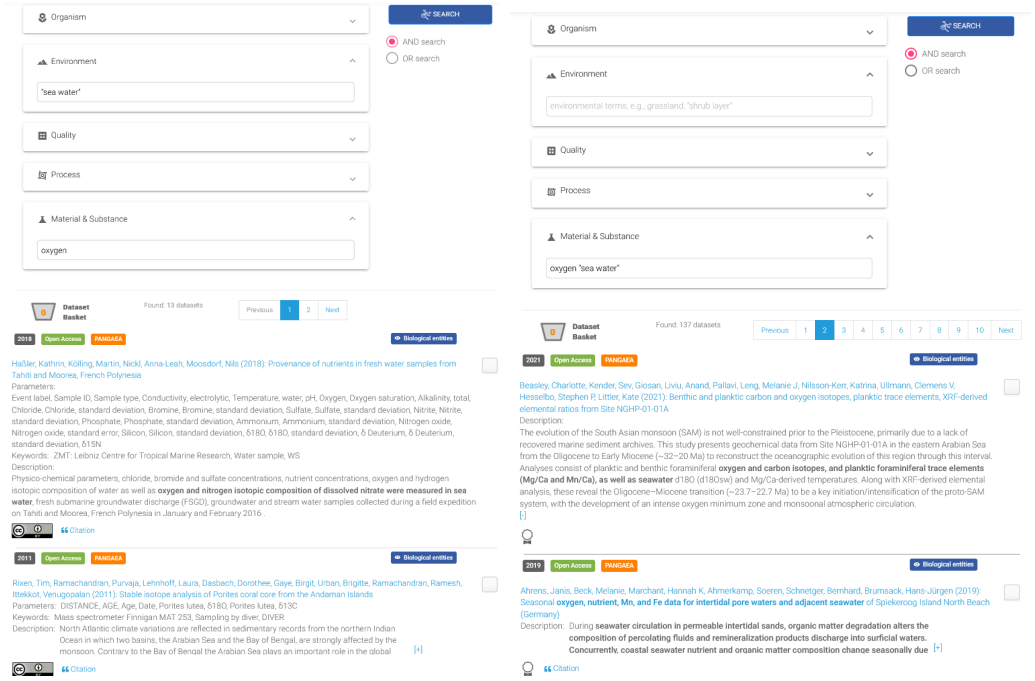


Figure 9.1: Screenshots for the first use case. A search with the key term 'sea water' returns different results for the category Material and Environment and also considers the spelling 'seawater'.

query for 'bacteria' now returns concrete bacterial species and takes the burden from the user to recall or to look up the concrete species names.

Taking a look at the **overall Hypothesis**,

It is possible to analyze, describe and enrich scientific datasets and scholarly information needs so that scholars are able to retrieve, understand, and reuse scientific datasets with an improved retrieval system.

Our achievements reveal that domain-specific semantic annotations based on the LOD cloud help to enhance metadata. These improvements support scholars in finding, understanding and reusing scientific data in biodiversity research.

9.1.1 Limitations

Our proposed semantic dataset search approach has some limitations, which we discuss in more detail below.

Pre-processing Component: The current text mining pipeline BiodivTagger does not contain a disambiguity component. Thus, in case several entries in ontologies match, we take all suggested entities. Furthermore, the gold standard we introduced was only built on a small size of 50 metadata files. In a new study, we developed a second improved metadata gold standard with a larger amount of metadata files (150 metadata files) and

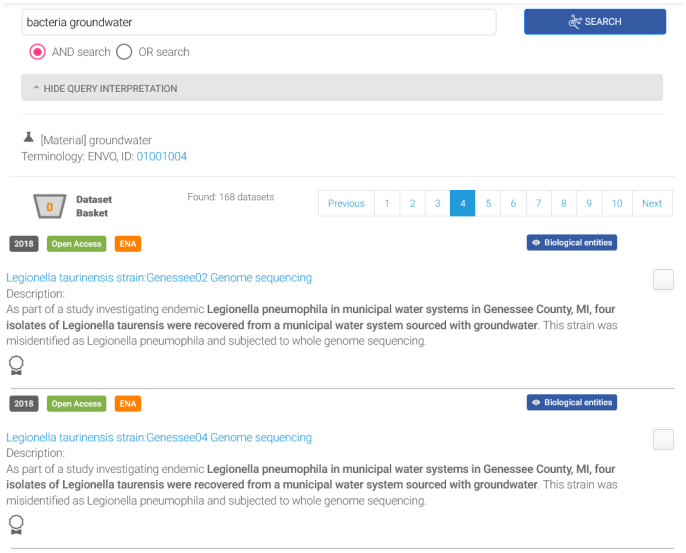


Figure 9.2: Screenshot for the second use case. A search with the key term 'bacteria' returns concrete bacterial species.

six domain categories being labeled [Abdelmageed et al., 2022]. In order to support machine learning tasks, we also provide this novel gold standard in the BIO-scheme format. However, what is still completely missing are metadata gold standards with labeled entity types and URIs (see also Chapter 10).

Retrieval Component: Updates in ontologies are currently not considered in our approach. In case a concept is moved, deleted or points to a different entity type, it is also necessary to update the semantic annotations. A second issue is the graph computation in memory. During the boot process, all required ontologies are loaded into the memory to allow a fast access in the retrieval process (for HCF-IDF and the TaxScorer). However, even then the computation can take several seconds or minutes. Another limitation is the increased metadata size. Adding semantic annotations including hierarchy information to metadata increases the file size, which also increases the index time. In worst case, metadata files can grow from a few KB to several MB.

Presentation Component: Concerning the implementation, the main issue reported by the users were missing highlightings or wrongly highlighted phrases. Second, the highlighting of biological entities is still an unstable function. A limitation in the evaluation is the fact that we did not study data reuse. It would be interesting to explore for what purpose the discovered datasets are reused and to quantify the benefit of a semantic dataset search.

9.2 Discussion

Based on Bast's categorization of semantic search systems [Bast et al., 2016], the utilized search engine GATE Mímir is a "Semi-Structured Search on Combined Data". Due to a semantic index of URIs and entity types, it allows the usage of keywords, entity types and SPARQL statements in the query. The results are text snippets from datasets and matching information from knowledge bases. As SPARQL queries are too complex for end users, we developed a frontend framework providing a keyword based search over a GATE Mímir semantic index. The keywords are linked to URIs, and the URIs are added to templates enabling a search including descendant nodes.

The proposed classification of semantic search systems by [Bast et al., 2016] has some limitations. For instance, it does not distinguish between query input and retrieval approach (Chapter 3.1.2). The user interface and the complexity of information needs are also not considered. Therefore, we propose the following extended dimensions for the classification of semantic search system being inspired by [Bast et al., 2016] and [Unger et al., 2014]. We take the two dimensions (search approach and data sources) introduced by [Bast et al., 2016], but we split the search approach into two groups, query input and retrieval approach. The work by [Unger et al., 2014] focuses on question answering, a specific type of semantic search systems, and suggests three characteristics: question/answer type (nature and semantic of question and answer), data sources (unstructured, semi-structured, structured) and semantic complexity (e.g., disambiguity). In addition to data sources, overlapping with [Bast et al., 2016], we add the dimensions question/answer type and complexity to our proposed classification. As a completely novel dimension, we incorporate explanations. Explanations are essential for comprehensibility in dataset search, in particular to understand whether a result item belongs to the originally entered query terms or not. We categorize each dimension along three levels. Figure 9.3 presents a radar chart with all six dimensions including a classification of our proposed system to what extend it supports the individual variables.

According to [Bast et al., 2016], *Query input* can be grouped into keyword search, structured search (type based, specific query languages or formal query languages such as SPARQL) and natural language search. Our system supports a keyword search, however, the underlying system GATE Mímir also allows a structured search using entity types and SPARQL sub queries. Hence, our system can be considered as an approach supporting keyword and structured search.

The dimension *Data* describes the structure of data sources and comprises completely unstructured data (e.g., text) or partially structured data with unstructured text fields (e.g., metadata, documents in XML format), structured data (knowledge resources in RDF and OWL format) and a combination of both unstructured/partially structured and structured data (e.g., documents and linked entities) [Bast et al., 2016]. Our system is mainly de-

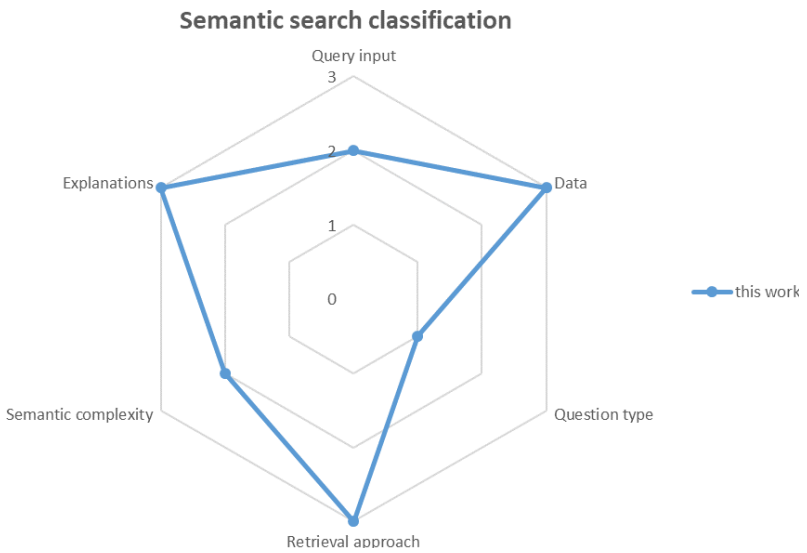


Figure 9.3: Variables influencing a search process.

signed for semi-structured metadata with short text passages as it occurs in title or abstract metadata fields. Metadata and queries are enriched with domain categories and entities. The output are datasets and information on utilized terminologies and entities. Therefore, our system supports combined data.

Question and *Answer type* characterize the question on the level of difficulty. According to [Unger et al., 2014], we distinguish between simple factoid questions, definition questions and further more sophisticated questions such as association questions or opinion questions. Based on the findings in Chapter 4.2, we recognized that most questions in biodiversity research are factoid questions expecting datasets containing the desired information. Our system is not designed as a full question-answering-system and therefore can not provide a perfect answer. Therefore, we classify our approach not on a high level for this variable but on a level mainly supporting factoid questions returning datasets that either contain the required answer or that provide hints to answer the question.

In the *Retrieval approach*, according to [Bast et al., 2016], we differ between keyword based retrieval, keyword based retrieval but with ontology support (query-expansion or entity highlighting in result set) and fully concept based ranking. Our approach is fully concept based and integrates further relations through additional SPARQL queries. The result set contains metadata files additionally enriched with concepts from knowledge bases.

Inspired by [Unger et al., 2014], we add the *Complexity of a search task* as another dimension. It comprises the comprehensibility and interpretability of an information need. As a search does not aim to provide an exact answer (as in question answering), we do not consider all suggested subclasses by [Unger et al., 2014] but only take the subclass 'semantic complexity', which describes the "complexity of the domain, the amount of ambiguity and vagueness in the question, as well as data hetero- geneity" [Unger et al., 2014].

As this classification is difficult to rank, we propose the following three levels based on the suggested subclasses: low complexity (no or only one subclass partially considered in the proposed system), medium complexity (two or more dimensions partially or fully considered), high complexity (all dimensions fully addressed). Our system does not support disambiguation, but offers a search over heterogenous data in biodiversity research, a highly complex domain, we assign our system to a medium level.

The outcomes of the first user evaluation of our semantic search prototype (Chapter 3.3) indicate the importance of *Explanations* in the user interface. Incorporating additional knowledge into the search process requires improved guidance and explanations why a result is returned and how it is related to the original query. This dimension reflects the degree of explanations in the user interface. We differ between explanations on the search query, highlights of search terms and related terms in the result set and additional highlightings such as further relevant entities. Our system fulfills all three kinds of explanations.

With this work, we proved the feasability of integrating additional semantic knowledge into the retrieval process. We also showed that domain knowledge improves dataset search significantly. In order to achieve these goals, we combined results and findings of several research fields in computer science, namely text mining, information retrieval and semantic web. In the future, similar to biodiversity research, new innovations and technological breakthroughs in computer science will only happen when interdisciplinary thinking and research is conducted. A lot of challenges remain, which we address in the next chapter.

Chapter 10

Future Work

"The Web as I envisaged it, we have not seen it yet. The future is still so much bigger than the past."

- Sir Tim Berners-Lee [Silva, 2009], *Inventor of the World Wide Web and winner of the Turing Award*

Ontologies and knowledge graphs are increasinly becoming the backbone of FAIR data management in academia and industry. This accomplishment is based on semantic web technologies and belong to the research field of symbolic AI, the classical, rule-based artifical intelligence. In some areas, the outcomes of symbolic AI have been overtaken by subsymbolic approaches such as deep learning and neural networks, e.g., language models such as BERT achieve impressive accuracy values in various NLP tasks. The same also applies to learning-to-rank approaches in information retrieval. However, only the classical, symbolic AI approaches enable reasoning, the inference of logical implications, and are closer to human thinking. In the future, we assume that both, subsymbolic and symbolic AI, will be combined to form declarative, unbiased intelligent applications such as semantic dataset search. In the following chapter, we point to future research directions for different topics addressed in this work. In particular, we focus on novel ideas to improve research data, approaches in text mining and information retrieval to link and combine symbolic and subsymbolic AI approaches, further ideas for improving and evaluating user interfaces in dataset search, and we give an outlook on extended functions in search such as personalization.

Improvements of research data: Our research on scholarly search interests and metadata (Chapter 4) reveal that general metadata standards reflect scholarly information needs only partially. Hence, we recommend data providers to utilize domain specific metadata standards when available. Since a couple of domain standards provide a large number of metadata fields, it is necessary to determine mandatory and optional

metadata fields being aligned to relevant domain categories. In addition to domain specific metadata standards, it is also essential to support scholars in the data submission process. In particular, data curation services are needed to describe data with proper terminologies and identifiers. Linked Open Data [Heath and Bizer, 2011] permits to link datasets, code and publications fostering an improved data findability and data reuse. In the long run, we assume that metadata standards will develop towards full semantic formats such as RDF. One example is ABCD 3.0[1], a standard for collections and museums in biodiversity research. Approaches to map CSV, TSC and XML to RDF already exists[2] [Dimou et al., 2014].

Initiatiaves such as *schema.org*[3] and domain specific variants, e.g., *bioschemas.org*[4] could be another improvement for landing pages at data archives. The idea of schema.org is to provide meaningful entities with unique identifiers inside the HTML tags (instead of keywords) to label authors, organizations, datasets, geographic and temporal information in web pages. These enrichments enhance data crawling for external search engines, e.g., Google's dataset search[5] as well as data findability. We assume that this extension will become popular in the near future due to its easy integration. In 2023, already three of our five studied data repositories, namely GBIF, DRYAD and PANGAEA, provide *schema.org* and/or *bioschemas.org* entities on their landing pages.

Improvements in the pre-processing component: The proposed text mining pipeline BiodivTagger (Chapter 6.3) is based on simple look-up services in specific domain ontologies or ontology nodes. A further disambiguation is currently not provided. Hence, in case several entries in an ontology match, we import all matching URIs. A disambiguation component could compare the descriptions of matching entries in ontologies with the given text. The best matching entries could be ranked on top. The score of this ranking process could determine whether an entry is considered as semantic annotation or not. Other approaches for Named Entity Recognition utilize language models such as BERT [Devlin et al., 2019] to extract the correct entity type, which requires large manually annotated corpora. Recently, we already proposed a larger metadata corpus over 150 metadata files for biodiversity research [Abdelmageed et al., 2022]. This novel gold standard also includes species annotations, the annotations are provided in a format suitable for machine learning tasks, and we also extended the manual annotations on relations to leverage relation extraction tasks and finally knowledge graph creation. However, gold standards with URIs are still missing. As gold standards with URIs are a very time-consuming task, silver standards with automatic semantic annotations (by means of

[1] ABCD 3.0, `https://abcd.tdwg.org/3.0/`
[2] RML, `https://rml.io/specs/rml/`
[3] schema.org, `https://schema.org`
[4] bioschemas.org, `https://bioschemas.org/`
[5] Google Dataset Search, `https://datasetsearch.research.google.com/`

look-up services) and manual corrections might be a solution.

Improvements in the retrieval component: Besides classical NLP tasks, the BERT language model was successfully applied to learning-to-rank approaches [Yates et al., 2021]. Similar to learning approaches for Named Entity Recognition, large ground truth information is needed to develop or adapt such models for other domains. However, if the ground truth is either large enough or provides the necessary data covering the respective domain, learning-to-rank approaches could be an opportunity to overcome the current technical limitations (see Chapter 9.1.1). Once more relations are added to existing ontologies, it would be also interesting to study further semantic relations, e.g., for exploratory search or recommendation.

Ideas for further user studies: In our study (Chapter 8), the search tasks were unknown to the subjects. Therefore, a long-term study would be helpful to analyze the usage in daily working context with own search terms. In addition, an eye-tracking study could supplement the current results on behavioural information. Furthermore, data reuse was not a main focus in our evaluation. This could be also explored in a long-term study. As dataset search is a broad field, and we mainly focused on biodiversity research, further research opportunities would be to study other application domains. Moreover, exploring explanations in other semantic search approaches (not specificly tied to dataset search) could also be a topic for future research.

Towards question answering: In the current system, users enter keywords and obtain datasets with either partial or full information on the desired information need. As this is the expected behaviour of search engines, a full question answering system was out of scope of this work. Nevertheless, future research might develop towards intelligent agents and assistants giving complete and exact answers on a natural language question. This requires improved concepts in the query input, such as a question understanding component and a translation of the question into semantic triples with machting entities. Afterwards, SPARQL queries can be created to find the answer or to retrieve relevant entities from the knowledge base. As supplementary information or provenance data, datasets supporting the answers could be displayed.

Towards personalization: Searching for data is an active tasks where users have to take some time to find relevant information. In contrast, recommending data does not require human participation. Based on a user profile, relevant data could be proactively presented or sent to consumers. From the technical point of view, content based recommendations work similar as search engines [Pazzani and Billsus, 2007]. Instead of search queries, user profiles are utilized to find relevant data. Extending our approach to content

recommendations requires semantic user models with URIs as interests. In collaboration with the Semantic Software Lab in Montréal, Canada, we developed an approach to construct semantic user models out of research publications [Sateli et al., 2017]. We took publications as input and extracted competences with various NLP techniques including a syntactic and a semantic analysis. We also linked the competences to entries in the LOD cloud by means of the LODTagger[6]. Finally, we exported the competences to a triple store[7]. In order to provide more transparency for users on what is extracted and stored in their user profile, we also developed a visualization component [Löffler et al., 2020]. The application displayed all extracted competences in different visualizations, e.g., table, force-directed graph and cord chart, in an ordered sequence. It also provided provenance information and listed all publications from which the competences were extracted. However, extracting and storing personal data is not liked by everybody. Due to security concerns, people hestitate in providing personal data. Therefore, future work could also concentrate on decentralized approaches, such as SOLID [8]. SOLID inverts the idea of storing personal data. Their vision is to store and manage personal data in secure pods and only on approval, personal information are forwarded to applications and third parties.

As we have shown, there are many open research questions that still need to be addressed in semantic dataset search. In particular, interdisciplinary approaches are demanded as well as further application domain specific research testing semantic technologies in practice. Only if we can identify and close these research gaps, a search in the web of data[9] will be successful.

[6]LODTagger, `https://www.semanticsoftware.info/lodtagger`
[7]LODExporter, `https://www.semanticsoftware.info/lodexporter`
[8]SOLID, `https://solidproject.org/`
[9]W3C, `https://www.w3.org/standards/semanticweb/data`

List of Figures

Listings

List of Tables

Bibliography

[ISO, 2018] (2018). ISO IEC 25062:2006, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Common Industry Format (CIF) for usability test reports.

[Abacha et al., 2017] Abacha, A. B., Agichtein, E., Pinter, Y., and Demner-Fushman, D. (2017). Overview of the Medical Question Answering Task at TREC 2017 LiveQA. Technical report, TREC LiveQA 2017.

[Abdelmageed et al., 2022] Abdelmageed, N., Löffler, F., Feddoul, L., Algergawy, A., Samuel, S., Gaikwad, J., Kazem, A., and König-Ries, B. (2022). BiodivNERE: Gold standard corpora for named entity recognition and relation extraction in the biodiversity domain. *Biodiversity Data Journal*, 10:e89481.

[AGROVOC, 2022] AGROVOC (2022). Food and Agriculture Organization of the United Nations (FAO): AGROVOC Multilingual Thesaurus. `https://agrovoc.fao.org/`. accessed on April 22, 2022.

[Ahmed et al., 2019] Ahmed, S., Stoeckel, M., Driller, C., Pachzelt, A., and Mehler, A. (2019). BIOfid dataset: Publishing a German gold standard for named entity recognition in historical biodiversity literature. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 871–880, Hong Kong, China. Association for Computational Linguistics.

[Allot et al., 2018] Allot, A., Peng, Y., Wei, C.-H., Lee, K., Phan, L., and Lu, Z. (2018). LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Research*, 46(W1):W530–W536.

[Ananiadou et al., 2004] Ananiadou, S., Friedman, C., and Tsujii, J. (2004). Introduction: named entity recognition in biomedicine. *Journal of Biomedical Informatics*, 37(6):393 – 395. Named Entity Recognition in Biomedicine.

[Antoniou et al., 2012] Antoniou, G., Groth, P., van Harmelen, F., and Hoekstra, R. (2012). *A Semantic Web Primer, third edition*. The MIT Press.

[AquaDiva, 2022] AquaDiva (2022). The collaborative research centre aquadiva. http://www.aquadiva.uni-jena.de/About.html. accessed on Feb 19, 2022.

[Ariño et al., 2013] Ariño, A. H., Chavan, V., and Faith, D. P. (2013). Assessment of user needs of primary biodiversity data: Analysis, concerns, and challenges. *Biodiversity Informatics*, 8(2).

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25:25–9.

[Assmann and Schuldt, 2008] Assmann, T. and Schuldt, A. (2008). Dataset: CSPs: Herbivore damage on saplings of 23 tree and shrub species in the CSPs, https://data.botanik.uni-halle.de/bef-china/datasets/147.

[Austin et al., 2016] Austin, C. C., Brown, S., Fong, N., Humphrey, C., Leahey, A., and Webster, P. (2016). Research Data Repositories: Review of Current Features, Gap Analysis, and Recommendations for Minimum Requirements. *IASSIST Quarterly*, 39(4):24.

[Baeza-Yates and Ribeiro-Neto, 2008] Baeza-Yates, R. and Ribeiro-Neto, B. (2008). *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison-Wesley Publishing Company, USA, 2nd edition.

[Bair et al., 1996] Bair, A. H., Brown, L. P., Pugh, L. C., Borucki, L. C., and Spatz, D. L. (1996). Taking a bite out of crisp. strategies on using and conducting searches in the computer retrieval of information on scientific projects database. *Computers in nursing*, 14:218–24; quiz 225–6.

[Bakalov et al., 2011] Bakalov, F., König-Ries, B., Hennig, T., and Schade, G. (2011). Usability study of a semantic user model visualization for social networks. In *Workshop on Visual Interfaces to the Social and Semantic Web (VISSW2011), Co-located with ACM IUI 2011, Feb 13, 2011, Palo Alt*.

[Balhoff et al., 2010] Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., Lundberg, J. G., Mabee, P., Midford, P. E., Westerfield, M., and Vision, T. J. (2010). Phenex: Ontological Annotation Of Phenotypic Diversity. *PLOS ONE*, 5(5):1–10.

[Balog, 2018] Balog, K. (2018). *Entity-Oriented Search*, volume 39 of *The Information Retrieval Series*.

[Bandrowski et al., 2016a] Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., Fan, L., Fostel, J., Fragoso, G., Gibson, F., Gonzalez-Beltran, A., Haendel, M. A., He, Y., Heiskanen, M., Hernandez-Boussard, T., Jensen, M., Lin, Y., Lister, A. L., Lord, P., Malone, J., Manduchi, E., McGee, M., Morrison, N., Overton, J. A., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Schober, D., Smith, B., Soldatova, L. N., Stoeckert, C. J. J., Taylor, C. F., Torniai, C., Turner, J. A., Vita, R., Whetzel, P. L., and Zheng, J. (2016a). The Ontology for Biomedical Investigations. *PloS one*, 11:e0154556.

[Bandrowski et al., 2016b] Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., and Bug, B. e. a. (2016b). The Ontology for Biomedical Investigations. *PloS one*, 11:e0154556.

[Bast and Buchhold, 2013] Bast, H. and Buchhold, B. (2013). An index for efficient semantic full-text search. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 369–378, New York, NY, USA. Association for Computing Machinery.

[Bast and Buchhold, 2017] Bast, H. and Buchhold, B. (2017). Qlever: A query engine for efficient sparql+text search. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 647–656, New York, NY, USA. Association for Computing Machinery.

[Bast et al., 2016] Bast, H., Buchhold, B., and Haussmann, E. (2016). Semantic search on text and knowledge bases. *Foundations and Trends® in Information Retrieval*, 10(2-3):119–271.

[Bast et al., 2018] Bast, H., Buchhold, B., and Haussmann, E. (2018). A quality evaluation of combined search on a knowledge base and text. *KI - Künstliche Intelligenz*, 32(1):19–26.

[Batchelor, C. et al., 2022] Batchelor, C. et al. (2022). Molecular Process Ontology: Processes at the molecular level, `https://obofoundry.org/ontology/mop.html`. `https://obofoundry.org/ontology/mop.html`. accessed on 16.06.2022.

[Beisswanger et al., 2008] Beisswanger, E., Lee, V., Kim, J.-J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., and Hahn, U. (2008). Gene regulation ontology (gro): design principles and use cases. *Studies in health technology and informatics*, 136:9–14.

[Belleau et al., 2008] Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., and Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41:706–16.

[Benjelloun et al., 2020] Benjelloun, O., Chen, S., and Noy, N. (2020). Google Dataset Search by the Numbers. In Pan, J. Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., and Kagal, L., editors, *The Semantic Web – ISWC 2020*, pages 667–682, Cham. Springer International Publishing.

[Berlanga et al., 2015] Berlanga, R., Nebot, V., and Pérez, M. (2015). Tailored semantic annotation for semantic search. *Journal of Web Semantics*, 30:69–81. Semantic Search.

[Berners-Lee et al., 2005] Berners-Lee, T., Fielding, R., and Masinter, L. (2005). Uniform resource identifier (uri): Generic syntax. `https://www.ietf.org/rfc/rfc3986.txt`, accessed on 3.1.2020.

[Berners-Lee and Hendler, 2001] Berners-Lee, T. and Hendler, J. (2001). Publishing on the semantic web. *Nature*, 410(6832):1023–1024.

[BIBO, 2016] BIBO (2016). Bibliographic Ontology (BIBO) in RDF. `https://www.dublincore.org/specifications/bibo/bibo/`. accessed on 01.09.2023.

[Biodiversity Exploratories, 2022] Biodiversity Exploratories (2022). Research for biodiversity. (`https://www.biodiversity-exploratories.de`). accessed on 27.02.2022.

[Bird and Loper, 2004] Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

[Blanco et al., 2015] Blanco, R., Ottaviano, G., and Meij, E. (2015). Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 179–188, New York, NY, USA. Association for Computing Machinery.

[Boldi and Vigna, 2005] Boldi, P. and Vigna, S. (2005). MG4J at TREC 2005. In Voorhees, E. M. and Buckland, L. P., editors, *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*, number SP 500-266 in Special Papers. NIST. `http://mg4j.di.unimi.it/`.

[Bontcheva et al., 2012] Bontcheva, K., Kieniewicz, J., Aswani, N., Wallis, M., and Andrews, S. (2012). User feedback report on the envilod semantic search interface. Technical Report, EnviLODproject deliverable (2012). Technical report.

[Bontcheva et al., 2014] Bontcheva, K., Tablan, V., and Cunningham, H. (2014). *Semantic Search over Documents and Ontologies*, pages 31–53. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Borgman, 2015] Borgman, C. L. (2015). *Big Data, Little Data, No Data - Scholarship in the Networked World*. MIT Press.

[Bravo et al., 2015] Bravo, A., Piñero, J., Queralt-Rosinach, N., Rautschka, M., and Furlong, L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1):55.

[Brooke, 1996] Brooke, J. (1996). *Usability Evaluation In Industry*, chapter SUS - A quick and dirty usability scale. CRC Press.

[Bruelheide et al., 2011] Bruelheide, H., Böhnke, M., Both, S., Fang, T., Assmann, T., Baruffol, M., Bauhus, J., Buscot, F., Chen, X.-Y., Ding, B.-Y., Durka, W., Erfmeier, A., Fischer, M., Geißler, C., Guo, D., Guo, L.-D., Härdtle, W., He, J.-S., Hector, A., Kröber, W., Kühn, P., Lang, A. C., Nadrowski, K., Pei, K., Scherer-Lorenzen, M., Shi, X., Scholten, T., Schuldt, A., Trogisch, S., von Oheimb, G., Welk, E., Wirth, C., Wu, Y.-T., Yang, X., Zeng, X., Zhang, S., Zhou, H., Ma, K., and Schmid, B. (2011). Community assembly during secondary forest succession in a chinese subtropical forest. *Ecological Monographs*, 81(1):25–41.

[Bruelheide et al., 2014] Bruelheide, H., Nadrowski, K., Assmann, T., Bauhus, J., Both, S., Buscot, F., Chen, X.-Y., Ding, B., Durka, W., Erfmeier, A., Gutknecht, J. L. M., Guo, D., Guo, L.-D., Härdtle, W., He, J.-S., Klein, A.-M., Kühn, P., Liang, Y., Liu, X., Michalski, S., Niklaus, P. A., Pei, K., Scherer-Lorenzen, M., Scholten, T., Schuldt, A., Seidler, G., Trogisch, S., von Oheimb, G., Welk, E., Wirth, C., Wubet, T., Yang, X., Yu, M., Zhang, S., Zhou, H., Fischer, M., Ma, K., and Schmid, B. (2014). Designing forest biodiversity experiments: general considerations illustrated by a new large experiment in subtropical china. *Methods in Ecology and Evolution*, 5(1):74–89.

[Buckley and Voorhees, 2004] Buckley, C. and Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 25–32, New York, NY, USA. Association for Computing Machinery.

[Buttigieg et al., 2016] Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., and Mungall, C. J. (2016). The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, 7(1):57.

[Bánki et al., 2022] Bánki, O., Roskov, Y., Döring, M., Ower, G., Vandepitte, L., Hobern, D., Remsen, D., Schalk, P., DeWalt, R. E., Keping, M., Miller, J., Orrell, T., Aalbu, R., Adlard, R., Adriaenssens, E. M., Aedo, C., Aescht, E., Akkari, N., and Alfenas-Zerbini, P. e. a. (2022). Dataset: Catalogue of Life Checklist (Version 2022-07-12). Catalogue of Life. https://doi.org/10.48580/dfpz. https://doi.org/10.48580/dfpz.

[Cai et al., 2015] Cai, Y., Wei, C.-H., Kao, H.-Y., and Lu, Z. (2015). GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed Research International*, 2015:918710.

[Campillos et al., 2018] Campillos, L., Deléger, L., Grouin, C., Hamon, T., Ligozat, A.-L., and Névéol, A. (2018). A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Language Resources and Evaluation*, 52(2):571–601.

[Campinas et al., 2011] Campinas, S., Ceccarelli, D., Perry, T., Delbru, R., Balog, K., and Tummarello, G. (2011). The Sindice-2011 Dataset for Entity-Oriented Search in the Web of Data.

[Darwin Core Interest Group, 2021] Darwin Core Interest Group (2021). Darwin core. https://dwc.tdwg.org/, accessed on 11th July 20222.

[DataCite Working Group., 2017] DataCite Working Group. (2017). Datacite. https://schema.datacite.org/meta/kernel-4.1/, accessed on 17th July 2022.

[DCMI Usage Board, 2020] DCMI Usage Board (2020). Dcmi metadata terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ accessed on 17th July 2022.

[FGDC, 1998] FGDC (1998). Content Standard For Digital Geospatial Metadata (CSDGM), Vers. 2. https://www.fgdc.gov/metadata/csdgm-standard, accessed on 17th July 2022.

[ISA Community, 2016] ISA Community (2016). ISA Model and Serialization Specifications. https://isa-specs.readthedocs.io/en/latest/index.html.

[ISO, 2019] ISO (2019). Iso 19115-1:2014, geographic information — metadata. https://www.iso.org/standard/53798.html, accessed on 17th July 2022.

[Open Archives Initiative, 2014] Open Archives Initiative (2014). Ore specifications and user guides. http://www.openarchives.org/ore/1.0/toc.

[TDWG, 2005] TDWG (2005). Biodiversity Information Standards: Access To Biological Collection Data (ABCD), Version 2.06. Biodiversity Information Standards (TDWG). `http://www.tdwg.org/standards/115`, accessed on 13th July 2022.

[Carmel et al., 2014] Carmel, D., Chang, M.-W., Gabrilovich, E., Hsu, B.-J. P., and Wang, K. (2014). Erd'14: Entity recognition and disambiguation challenge. *SIGIR Forum*, 48(2):63–77.

[CARO, 2022] CARO (2022). CARO: The Common Anatomy Reference Ontology. `https://obofoundry.org/ontology/caro.html`. accessed on Jan 4, 2022.

[Castelo et al., 2021] Castelo, S., Rampin, R., Santos, A., Bessa, A., Chirigati, F., and Freire, J. (2021). Auctus: A Dataset Search Engine for Data Discovery and Augmentation. *Proc. VLDB Endow.*, 14(12):2791–2794.

[Chapman et al., 2019] Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.-D., Kacprzak, E., and Groth, P. (2019). Dataset search: a survey. *The VLDB Journal*.

[Chen et al., 2018] Chen, X., Gururaj, A. E., Ozyurt, B., Liu, R., Soysal, E., Cohen, T., Tiryaki, F., Li, Y., Zong, N., Jiang, M., Rogith, D., Salimi, M., Kim, H.-e., Rocca-Serra, P., Gonzalez-Beltran, A., Farcas, C., Johnson, T., Margolis, R., Alter, G., Sansone, S.-A., Fore, I. M., Ohno-Machado, L., Grethe, J. S., and Xu, H. (2018). DataMed – an open source discovery index for finding biomedical datasets. *Journal of the American Medical Informatics Association*, 25(3):300–308.

[Chin et al., 1988] Chin, J. P., Diehl, V. A., and Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, page 213–218, New York, NY, USA. Association for Computing Machinery.

[Choi et al., 2012] Choi, J., Kim, D., Kim, S., Lee, S., Lee, K., and Kang, J. (2012). BOSS: context-enhanced search for biomedical objects. *BMC Medical Informatics and Decision Making*, 12(1):S7.

[Clarke et al., 2019] Clarke, D. J. B., Wang, L., Jones, A., Wojciechowicz, M. L., Torre, D., Jagodnik, K. M., Jenkins, S. L., McQuilton, P., Flamholz, Z., Silverstein, M. C., Schilder, B. M., Robasky, K., Castillo, C., Idaszak, R., Ahalt, S. C., Williams, J., Schurer, S., Cooper, D. J., de Miranda Azevedo, R., Klenk, J. A., Haendel, M. A., Nedzel, J., Avillach, P., Shimoyama, M. E., Harris, R. M., Gamble, M., Poten, R., Charbonneau, A. L., Larkin, J., Brown, C. T., Bonazzi, V. R., Dumontier, M. J., Sansone, S.-A., and Ma'ayan, A. (2019). Fairshake: toolkit to evaluate the findability, accessibility, interoperability, and reusability of research digital resources. *bioRxiv*.

[Cohen et al., 2017a] Cohen, K. B., Verspoor, K., Fort, K., Funk, C., Bada, M., Palmer, M., and Hunter, L. E. (2017a). *The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain*, pages 1379–1394. Springer Netherlands, Dordrecht.

[Cohen et al., 2017b] Cohen, T., Roberts, K., Gururaj, A. E., Chen, X., Pournejati, S., Alter, G., Hersh, W. R., Demner-Fushman, D., Ohno-Machado, L., and Xu, H. (2017b). A publicly available benchmark for biomedical dataset retrieval: the reference standard for the 2016 bioCADDIE dataset retrieval challenge. *Database*.

[Consortium, 2020] Consortium, T. U. (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489.

[Convention on Biological Diversity, 1992] Convention on Biological Diversity (1992). `https://www.cbd.int/convention/`. accessed on 03.11.2022.

[Cooper et al., 2017] Cooper, L., Meier, A., Laporte, M.-A., Elser, J. L., Mungall, C., Sinn, B. T., Cavaliere, D., Carbon, S., Dunn, N. A., Smith, B., Qu, B., Preece, J., Zhang, E., Todorovic, S., Gkoutos, G., Doonan, J. H., Stevenson, D. W., Arnaud, E., and Jaiswal, P. (2017). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research*, 46(D1):D1168–D1180.

[Cooper et al., 2013] Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T. Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., and Jaiswal, P. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant & cell physiology*, 54:e1.

[Corson-Rikert et al., 2012] Corson-Rikert, J., Mitchell, S., Lowe, B., Rejack, N., Ding, Y., and Guo, C. (2012). *The VIVO Ontology*, pages 15–33. Springer International Publishing, Cham.

[Croft et al., 2009] Croft, B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition.

[Culina et al., 2018] Culina, A., Baglioni, M., Crowther, T. W., Visser, M. E., Woutersen-Windhouwer, S., and Manghi, P. (2018). Navigating the unfolding open data landscape in ecology and evolution. *Nature Ecology & Evolution*, 2(3):420–426.

[Cunningham et al., 2013] Cunningham, H., Tablan, V., Roberts, A., and Bontcheva, K. (2013). Getting More Out Of Biomedical Documents With GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology*, 9(2):e1002854–e1002854.

[Curators Herbarium B, 2020] Curators Herbarium B (2020). Curators Herbarium B (2020). Digital specimen images at the Herbarium Berolinense. [Dataset]. Version: 2020-10-07;. Data Publisher: Botanic Garden and Botanical Museum Berlin. https://data.bgbm.org/dataset/gfbio/0001/.

[Dakka and Ipeirotis, 2008] Dakka, W. and Ipeirotis, P. G. (2008). Automatic extraction of useful facet hierarchies from text databases. In *2008 IEEE 24th International Conference on Data Engineering*, pages 466–475.

[Data Citation Synthesis Group, 2014] Data Citation Synthesis Group (2014). Joint Declaration of Data Citation Principles, Martone M. (ed.) San Diego CA: FORCE11;.

[DataONE, 2022] DataONE (2022). Make your data fair: Evaluate your metadata with community established fair principles. `https://www.dataone.org/fair/`. accessed on 15th May 2022.

[Deléger et al., 2016] Deléger, L., Bossy, R., Chaix, E., Ba, M., Ferre, A., Bessieres, P., and Nedellec, C. (2016). Overview Of The Bacteria Biotope Task At BioNLP Shared Task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 12–22.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Diefenbach et al., 2020] Diefenbach, D., Both, A., Singh, K., and Maret, P. (2020). Towards a Question Answering System over the Semantic Web. *Semantic Web*, 11(3):421–439.

[Diehl et al., 2016] Diehl, A. D., Meehan, T. F., Bradford, Y. M., Brush, M. H., Dahdul, W. M., Dougall, D. S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., Van Slyke, C. E., Vasilevsky, N. A., Haendel, M. A., Blake, J. A., and Mungall, C. J. (2016). The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *Journal of biomedical semantics*, 7:44.

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

[Dietze and Schroeder, 2009] Dietze, H. and Schroeder, M. (2009). GoWeb: a semantic search engine for the life science web. *BMC Bioinformatics*, 10(10):S7.

[Dillman et al., 2009] Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., and Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, 38(1):1–18.

[Dimitrova et al., 2020] Dimitrova, M., Poelen, J., Zhelezov, G., Georgiev, T., Agosti, D., and Penev, L. (2020). Semantic Publishing Enables Text Mining of Biotic Interactions. *Biodiversity Information Science and Standards*, 4:e59036.

[Dimou et al., 2014] Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., and Van de Walle, R. (2014). RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In Bizer, C., Heath, T., Auer, S., and Berners-Lee, T., editors, *Proceedings of the 7th Workshop on Linked Data on the Web*, volume 1184 of *CEUR Workshop Proceedings*.

[Dixit et al., 2017] Dixit, R., Rogith, D., Narayana, V., Salimi, M., Gururaj, A., Ohno-Machado, L., Xu, H., and Johnson, T. R. (2017). User needs analysis and usability assessment of DataMed – a biomedical data discovery index. *Journal of the American Medical Informatics Association*, 25(3):337–344.

[Duerst and Suignard, 2005] Duerst, M. and Suignard, M. (2005). Internationalized Resource Identifiers (IRI). `https://www.ietf.org/rfc/rfc3987.txt`, accessed on 17.08.2023.

[Dumais, 2005] Dumais, S. (2005). *The Interactive TREC Track: Putting the User Into Search*. MIT Press.

[Eaton et al., 2021] Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., Horne, R., Whiteaker, T., Blodgett, D., Zender, C., Lee, D., Hassell, D., Snow, A. D., Kölling, T., Allured, D., Jelenak, A., Soerensen, A. M., Gaultier, L., and Herlédan, S. (2021). NetCDF Climate and Forecast (CF) Metadata Conventions. `http://cfconventions.org/`, accessed on 13th July 2022.

[Elbedweihy et al., 2012] Elbedweihy, K., Wrigley, S. N., and Ciravegna, F. (2012). Evaluating semantic search query approaches with expert and casual users. In Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J. X., Hendler, J., Schreiber, G., Bernstein, A., and Blomqvist, E., editors, *The Semantic Web – ISWC 2012*, pages 274–286, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Elbedweihy et al., 2015] Elbedweihy, K. M., Wrigley, S. N., Clough, P., and Ciravegna, F. (2015). An overview of semantic search evaluation initiatives. *Journal of Web Semantics*, 30:82–105. Semantic Search.

[Eliason et al., 2019] Eliason, C. M., Edwards, S. V., and Clarke, J. A. (2019). phenotools: An R package for visualizing and analyzing phenomic datasets. *Methods in Ecology and Evolution*.

[Ernst et al., 2016] Ernst, P., Siu, A., Milchevski, D., Hoffart, J., and Weikum, G. (2016). DeepLife: An entity-aware search, analytics and exploration platform for health and life sciences. In *Proceedings of ACL-2016 System Demonstrations*, pages 19–24, Berlin, Germany. Association for Computational Linguistics.

[Ernst et al., 2019] Ernst, P., Terolli, E., and Weikum, G. (2019). LongLife: a Platform for Personalized Searchfor Health and Life Sciences. In Suárez-Figueroa, M., Cheng, G., Gentile, A., Guéret, C., Keet, M., and Bernstein, A., editors, *Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas) co-located with 18th International Semantic Web Conference (ISWC 2019) (pp. 237-240)*.

[Faessler and Hahn, 2017] Faessler, E. and Hahn, U. (2017). Semedico: A Comprehensive Semantic Search Engine for the Life Sciences. In *Proceedings of ACL 2017, System Demonstrations*, pages 91–96, Vancouver, Canada. Association for Computational Linguistics.

[FAIRsharing.org: QUDT, 2011] FAIRsharing.org: QUDT (2011). Quantities, Units, Dimensions and Types. `https://doi.org/10.25504/FAIRsharing.d3pqw7`. accessed on 14.04.2022.

[Faith et al., 2013] Faith, D., Collen, B., Arturo, A., Koleff, P., Guinotte, J., Kerr, J., and Chavan, V. (2013). Bridging the biodiversity data gaps: Recommendations to meet users' data needs. *Biodiversity Informatics*, 8(2).

[FDA, 2022] FDA (2022). FDA: U.S. Food and Drug Administration. `http://www.fda.gov/`. accessed on 22.04.2022.

[Feddoul et al., 2019] Feddoul, L., Schindler, S., and Löffler, F. (2019). Automatic facet generation and selection over knowledge graphs. In Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., and Sure-Vetter, Y., editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 310–325, Cham. Springer International Publishing.

[Fenner et al., 2019] Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., and Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data*, 6(1):28.

[Ferrucci, 2012] Ferrucci, D. A. (2012). Introduction to "This is Watson". *IBM Journal of Research and Development*, 56(3.4):1:1–1:15.

[Fleiss, 1971] Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382.

[Frenzel et al., 2016] Frenzel, M., Dussl, F., Höhne, R., Nickels, V., and Creutzburg, F. (2016). Dataset: Wild bee monitoring in six agriculturally dominated landscapes of Saxony-Anhalt (Germany) in 2014, `https://doi.org/10.1594/PANGAEA.865100`. PANGAEA. In: Frenzel, Mark; Preiser, Christine; Dussl, Franz; Höhne, René; Nickels, Volker; Creutzburg, Frank (2016): TERENO (Terrestrial Environmental Observatories) wild bee monitoring in six agriculturally dominated landscapes of Saxony-Anhalt (Germany). Helmholtz Centre for Environmental Research - UFZ, PANGAEA, https://doi.org/10.1594/PANGAEA.864908.

[Gaiji et al., 2013] Gaiji, S., Chavan, V., Ariño, A. H., Otegui, J., Hobern, D., Sood, R., and Robles, E. (2013). Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodiversity Informatics*, 8(2).

[Gaizauskas and Wilks, 1998] Gaizauskas, R. and Wilks, Y. (1998). Information Extraction: Beyond Document Retrieval. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 3, Number 2, August 1998*, pages 17–60.

[GBIF, 2018] GBIF (2018). GBIF Science Review 2018. Technical report, `https://doi.org/10.15468/VA9B-3048`, accessed on 20.02.2019.

[GBIF, 2020] GBIF (2020). Global Biodiversity Information Facility. `https://www.gbif.org/`, accessed on 12.11.2022.

[Geonames, 2022] Geonames (2022). The Geonames geographical databases. `https://www.geonames.org/`. accessed on 22.04.2022.

[Germany and Erfmeier, 2019] Germany, M. and Erfmeier, A. (2019). Dataset: Main Experiment: Seedling addition experiment - growth and biomass data. (accessed through URL: `http://china.befdata.biow.uni-leipzig.de/datasets/577`).

[Gerner et al., 2010] Gerner, M., Nenadic, G., and Bergman, C. M. (2010). LINNAEUS: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1):85.

[Gkoutos, G. et al., 2022] Gkoutos, G. et al. (2022). Phenotype And Trait Ontology: An ontology of phenotypic qualities (properties, attributes or characteristics). `https://obofoundry.org/ontology/pato.html`. accessed on 16.06.2022.

[GND, 2022] GND (2022). GND: The Integrated Authority File. `https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html`. accessed on 22.04.2022.

[GO, 2021] GO (2021). Gene Ontology Consortium: The Gene Ontology resource: enriching a GOld mine. *Nucleic acids research*, 49:D325–D334.

[Goossen et al., 2011] Goossen, F., IJntema, W., Frasincar, F., Hogenboom, F., and Kaymak, U. (2011). News Personalization Using the CF-IDF Semantic Recommender. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 10:1–10:12, New York, NY, USA. ACM.

[Gregory et al., 2020] Gregory, K., Groth, P., Scharnhorst, A., and Wyatt, S. (2020). Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/gw3r97ht.

[Gremse et al., 2011] Gremse, M., Chang, A., Schomburg, I., Grote, A., Scheer, M., Ebeling, C., and Schomburg, D. (2011). The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research*, 39:D507–13.

[Guo et al., 2009] Guo, J., Xu, G., Cheng, X., and Li, H. (2009). Named Entity Recognition in Query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 267–274, New York, NY, USA. Association for Computing Machinery.

[Gurulingappa et al., 2010] Gurulingappa, H., Klinger, R., Hofmann-Apitius, M., and Fluck, J. (2010). An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. In *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining, LREC, Valetta, Malta*.

[Gwet, 2008] Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

[Gwinn and Rinaldo, 2009] Gwinn, N. E. and Rinaldo, C. (2009). The Biodiversity Heritage Library: sharing biodiversity literature with the world. *IFLA Journal*, 35(1):25–34.

[Güntsch et al., 2012] Güntsch, A., Fichtmueller, D., Kirchhoff, A., and Berendsohn, W. (2012). Efficient rescue of threatened biodiversity data using rebind-workflows. *Plant Biosystems*, 146:752–755.

[Harispe et al., 2015] Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2015). Semantic Similarity from Natural Language and Ontology Analysis. *Synthesis Lectures on Human Language Technologies*, 8:1–254.

[Hasibi et al., 2015] Hasibi, F., Balog, K., and Bratsberg, S. E. (2015). Entity Linking in Queries: Tasks and Evaluation. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ICTIR '15, page 171–180, New York, NY, USA. Association for Computing Machinery.

[Hastings et al., 2016] Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1):D1214—9.

[Hawizy et al., 2011] Hawizy, L., Jessop, D. M., Adams, N., and Murray-Rust, P. (2011). ChemicalTagger: A tool for semantic text-mining in chemistry. *Journal of Cheminformatics*, 3(1):17.

[He et al., 2014] He, Y., Liu, Y., and Zhao, B. (2014). OGG: a biological ontology for representing genes and genomes in specific organisms. In Hogan, W. R., Arabandi, S., and Brochhausen, M., editors, *Proceedings of the 5th International Conference on Biomedical Ontologies (ICBO)*, pages 13–20, Houston, Texas, USA.

[He et al., 2020] He, Y., Zhu, Z., Zhang, Y., Chen, Q., and Caverlee, J. (2020). Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online. Association for Computational Linguistics.

[Hearst, 2006] Hearst, M. A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM*, 49(4):59–61.

[Hearst, 2011] Hearst, M. A. (2011). *Modern Information Retrieval*, chapter User Interfaces and Visualization, pages 257–340. Addison-Wesley Publishing Company, USA, 2nd edition.

[Heath and Bizer, 2011] Heath, T. and Bizer, C. (2011). Linked Data: Evolving The Web Into A Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136.

[Herrera et al., 2019] Herrera, J.-M., Hogan, A., and Käfer, T. (2019). Btc-2019: The 2019 billion triple challenge dataset. In Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., and Gandon, F., editors, *The Semantic Web – ISWC 2019*, pages 163–180, Cham. Springer International Publishing.

[Hersh and Voorhees, 2009] Hersh, W. and Voorhees, E. (2009). Trec genomics special issue overview. *Information Retrieval*, 12(1):1–15.

[Hey et al., 2009] Hey, T., Tansley, S., and Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

[HGNC, 2022] HGNC (2022). HUGO Gene Nomenclature Committee: The resource for approved human gene nomenclature. `https://www.genenames.org/`. accessed on 22.04.2022.

[Hildebrand et al., 2006] Hildebrand, M., van Ossenbruggen, J., and Hardman, L. (2006). /facet: A Browser for Heterogeneous Semantic Web Repositories. In Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., and Aroyo, L. M., editors, *The Semantic Web - ISWC 2006*, pages 272–285, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Hitzler et al., 2009] Hitzler, P., Krötzsch, M., and Rudolph, S. (2009). *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC.

[Hoehndorf et al., 2016] Hoehndorf, R., Alshahrani, M., Gkoutos, G. V., Gosline, G., Groom, Q., Hamann, T., Kattge, J., de Oliveira, S. M., Schmidt, M., Sierra, S., Smets, E., Vos, R. A., and Weiland, C. (2016). The flora phenotype ontology (FLOPO): tool for integrating morphological traits and phenotypes of vascular plants. *Journal of Biomedical Semantics*, 7(1):65.

[Hoehndorf et al., 2015] Hoehndorf, R., Slater, L., Schofield, P. N., and Gkoutos, G. V. (2015). Aber-OWL: a framework for ontology-based data access in biology. *BMC Bioinformatics*, 16(1):26.

[Horton et al., 2022] Horton, T., Kroh, A., Ahyong, S., Bailly, N., Bieler, R., Boyko, C., Brandão, S., Gofas, S., Hooper, J., Hernandez, F., Mees, J., Molodtsova, T., Paulay, G., Bouirig, K., Decock, W., Dekeyzer, S., Vandepitte, L., Vanhoorne, B., Adlard, R., Agatha, S., Ahn, K., Akkari, N., Alvarez, B., Amorim, V., Anderberg, A., Anderson, G., Andrés, S., Ang, Y., Antic, D., Antonietto, L., Arango, C., Artois, T., Atkinson,

S., Auffenberg, K., Baldwin, B., Bank, R., Barber, A., Barbosa, J., Bartsch, I., Bellan-Santini, D., Bergh, N., Bernot, J., Berta, A., Bezerra, T., Blanco, S., Blasco-Costa, I., Blazewicz, M., Bock, P., Bonifacino de León, M., Böttger-Schnack, R., Bouchet, P., Boury-Esnault, N., Bouzan, R., Boxshall, G., Bray, R., Bruce, N., Bruneau, A., Bueno, V., Bueno-Villegas, J., Cairns, S., Calvo Casas, J., Carballo, J., Cárdenas, P., Carstens, E., Chan, B., Chan, T., Cheng, L., Christenhusz, M., Churchill, M., Coleman, C., Collins, A., Collins, G., Corbari, L., Cordeiro, R., Cornils, A., Costa Corgosinho, P., Coste, M., Costello, M., Crandall, K., Cremonte, F., Cribb, T., Cutmore, S., da Silva Pereira, J., Dahdouh-Guebas, F., Daly, M., Daneliya, M., Dauvin, J., Davie, P., De Broyer, C., De Grave, S., de Lima Ferreira, P., de Mazancourt, V., de Voogd, N., Decker, P., Defaye, D., Dekker, H., d'Hondt, J., Dippenaar, S., Dohrmann, M., Dolan, J., Domning, D., Downey, R., Dreyer, N., Ector, L., Eisendle, U., Eitel, M., Encarnação, S., Enghoff, H., Epler, J., Evenhuis, N., Ewers-Saucedo, C., Faber, M., Figueroa, D.and Fišer, C., Fordyce, E., Foster, W., Frank, J., Fransen, C., Freire, S., Furuya, H., Galbany, M., Gale, A., Galea, H., Gao, T., Garcia-Alvarez, O., Garcia-Jacas, N., Garic, R., Garnett, S., Gasca, R., Gaviria-Melo, S., Gerken, S., Gibson, D., Gibson, R., Gil, J., Gittenberger, A., Glasby, C., Glenner, H., Glover, A., Gómez-Noguera, S., Gondim, A., González-Solís, D., Goodwin, C., Gostel, M., Grabowski, M., Gravili, C., Grossi, M., Guerra-García, J., Guerrero, J., Guidetti, R., Guimarães, S., Guiry, M., Gutierrez, D., Hadfield, K., Hajdu, E., Hallermann, J., Hayward, B., Hegna, T., Heiden, G., Hendrycks, E., Herbert, D., Herrera Bachiller, A., Hirsch, H., Ho, J., Hodda, M., Høeg, J., Hoeksema, B., Holovachov, O., Houart, R., Hughes, L., Hyžný, M., Iniesta, L., Iseto, T., Ivanenko, V., Iwataki, M., Janssen, R., Jaume, D., Jazdzewski, K., Jersabek, C., Jóźwiak, P., Kabat, A., Kantor, Y., Karanovic, I., Karthick, B., Kathirithamby, J., Katinas, L., Kim, Y., King, R., Kirk, P., Klautau, M., Kociolek, J., Köhler, F., Kolb, J., Konowalik, K., Kotov, A., Kovács, Z., Kremenetskaia, A.and Kristensen, R., Kulikovskiy, M., Kullander, S., Kupriyanova, E., Lamaro, A., Lambert, G., Lazarus, D., Le Coze, F., Le Roux, M., LeCroy, S., Leduc, D., Lefkowitz, E., Lemaitre, R., Lichter-Marck, I., Lim, S., Lindsay, D., Liu, Y., Loeuille, B., Lörz, A., Ludwig, T., Lundholm, N., Macpherson, E., Madin, L., Mah, C., Mamo, B., Mamos, T., Manconi, R., Mapstone, G., Marek, P., Marshall, B., Marshall, D., Martin, P., McFadden, C., McInnes, S., McKenzie, R., Means, J., Mejía-Madrid, H., Meland, K., Merrin, K., Messing, C., Miller, J., Mills, C., Moestrup, ö., Mokievsky, V., Monniot, F., Mooi, R., Morandini, A., Moreira da Rocha, R., Morrow, C., Mortelmans, J., Mortimer, J., Muñoz Gallego, A., Musco, L., Nery, D., Nesom, G., Neubauer, T., Neubert, E., Neuhaus, B., Ng, P., Nguyen, A., Nielsen, C., Nishikawa, T., Norenburg, J., O'Hara, T., Opresko, D., Osawa, M., Osigus, H., Ota, Y., Páll-Gergely, B., Panero, J., Pasini, E., Patterson, D., Paxton, H., Pedram, M., Pelser, P., Peña Santiago, R., Perez-Losada, M., Petrescu, I., Pfingstl, T., Pica, D., Picton,

B., Pilger, J., Pisera, A., Polhemus, D., Poore, G., Potapova, M., Půža, V., Read, G., Reich, M., Reimer, J., Reip, H., Reuscher, M., Reynolds, J., Richling, I., Rimet, F., Ríos, P., Rius, M., Rodríguez, E., Rogers, D., Roque, N., Rosenberg, G., Rützler, K., Saavedra, M., Sabbe, K., Saiz-Salinas, J., Sala, S., Santagata, S., Santos, S., Sar, E., Satoh, A., Saucède, T., Schärer, L., Schierwater, B., Schilling, E., Schmidt-Lebuhn, A., Schmidt-Rhaesa, A., Schneider, S., Schönberg, C., Schuchert, P., Senna, A., Sennikov, A., Serejo, C., Shaik, S., Shamsi, S., Sharma, J., Shear, W., Shenkar, N., Short, M., Sicinski, J., Sierwald, P., Simmons, E., Sinniger, F., Sinou, C., Sivell, D., Sket, B., Smit, H., Smit, N., Smol, N., Souza-Filho, J., Spelda, J., Sterrer, W., Stienen, E., Stoev, P., Stöhr, S., Strand, M., Suárez-Morales, E., Susanna, A., Suttle, C., Swalla, B., Taiti, S., Tanaka, M., Tandberg, A., Tang, D., Tasker, M., Taylor, J., Taylor, J., Taylor, K., Tchesunov, A., Temereva, E., ten Hove, H., ter Poorten, J., Thomas, J., Thuesen, E., Thurston, M., Thuy, B., Timi, J., Timm, T., Todaro, A., Turon, X., Uetz, P., Urbatsch, L., Uribe-Palomino, J., Urtubey, E., Utevsky, S., Vacelet, J., Vachard, D., Vader, W., Väinölä, R., Van de Vijver, B., van der Meij, S., van Haaren, T., van Soest, R., Vanreusel, A., Venekey, V., Vinarski, M., Vonk, R., Vos, C., Vouilloud, A., Walker-Smith, G., Walter, T., Watling, L., Wayland, M., Wesener, T., Wetzel, C., Whipps, C., White, K., Wieneke, U., Williams, D., Williams, G., Wilson, R., Witkowski, J., Wyatt, N., Wylezich, C., Xu, K., Zanol, J., Zeidler, W., Zhao, Z., and Zullini, A. (2022). World Register of Marine Species (WoRMS), `https://www.marinespecies.org`. `https://www.marinespecies.org`. accessed on 01.08.2022.

[Hripcsak and Rothschild, 2005] Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298.

[Ibáñez and Simperl, 2022] Ibáñez, L.-D. and Simperl, E. (2022). A comparison of dataset search behaviour of internal versus search engine referred sessions. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, page 158–168, New York, NY, USA. Association for Computing Machinery.

[Ide et al., 2007] Ide, N. C., Loane, R. F., and Demner-Fushman, D. (2007). Essie: A Concept-based Search Engine for Structured Biomedical Text. *Journal of the American Medical Informatics Association*, 14(3):253–263.

[idiv, 2019] idiv (2019). German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. `https://www.idiv.de`, accessed on 11.04.2019.

[Ikkala et al., 2022] Ikkala, E., Hyvönen, E., Rantala, H., and Koho, M. (2022). Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. *Semantic Web*, 13(1):69–84.

[Islamaj Dogan et al., 2009] Islamaj Dogan, R., Murray, G. C., Névéol, A., and Lu, Z. (2009). Understanding PubMed® user search behavior through log analysis. *Database*, 2009.

[ISO 9241-11, 2018] ISO 9241-11 (2018). ISO 9241-11:2018, Ergonomics of human-system interaction — part 11: Usability: Definitions and concepts.

[ISO 9241-210, 2019] ISO 9241-210 (2019). ISO 9241-210:2019, Ergonomics of human-system interaction — part 210: Human-centred design for interactive systems.

[Janowicz et al., 2014] Janowicz, K., Hitzler, P., Adams, B., Kolas, D., and Vardeman II, C. (2014). Five stars of linked data vocabulary use. *Semantic Web*, 5(3):173–176.

[Järvelin and Kekäläinen, 2002] Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

[Jones et al., 2019] Jones, M., O'Brien, M., Mecum, B., Boettiger, C., Schildhauer, M., Maier, M., Whiteaker, T., Earl, S., and Chong, S. (2019). Ecological Metadata Language version 2.2.0.

[Jovanović and Bagheri, 2017] Jovanović, J. and Bagheri, E. (2017). Semantic annotation in biomedicine: the current landscape. *Journal of Biomedical Semantics*, 8(1):44.

[Jurafsky and Martin, 2008] Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

[Kacprzak et al., 2018] Kacprzak, E., Koesten, L., Ibáñez, L.-D., Blount, T., Tennison, J., and Simperl, E. (2018). Characterising dataset search - An analysis of search logs and data requests. *Journal of Web Semantics*.

[Kalantari et al., 2021] Kalantari, M., Syahrudin, S., Rajabifard, A., and Hubbard, H. (2021). Synchronising spatial metadata records and interfaces to improve the usability of metadata systems. *ISPRS International Journal of Geo-Information*, 10(6).

[Kalantari et al., 2020] Kalantari, M., Syahrudin, S., Rajabifard, A., Subagyo, H., and Hubbard, H. (2020). Spatial metadata usability evaluation. *ISPRS International Journal of Geo-Information*, 9(7).

[Kamdar et al., 2014] Kamdar, M. R., Zeginis, D., Hasnain, A., Decker, S., and Deus, H. F. (2014). ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. *Journal of Biomedical Informatics*, 47:112–130.

[Kapanipathi et al., 2014] Kapanipathi, P., Jain, P., Venkataramani, C., and Sheth, A. (2014). User interests identification on twitter using a hierarchical knowledge base. In Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., and Tordai, A., editors, *The Semantic Web: Trends and Challenges*, pages 99–113, Cham. Springer International Publishing.

[Karam et al., 2016] Karam, N., Müller-Birn, C., Gleisberg, M., Fichtmüller, D., Tolksdorf, R., and Güntsch, A. (2016). A Terminology Service Supporting Semantic Annotation, Integration, Discovery and Analysis of Interdisciplinary Research Data. *Datenbank-Spektrum*, 16(3):195–205.

[Kaufmann and Bernstein, 2007] Kaufmann, E. and Bernstein, A. (2007). How Useful Are Natural Language Interfaces to the Semantic Web for Casual End-Users? In Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., and Cudré-Mauroux, P., editors, *The Semantic Web*, pages 281–294, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Kelly, 2009] Kelly, D. (2009). Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends Inf. Retr.*, 3(1—2):1–224.

[Khalili et al., 2016] Khalili, A., Loizou, A., and van Harmelen, F. (2016). Adaptive Linked Data-Driven Web Components: Building Flexible and Reusable Semantic Web Interfaces. In Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S. P., and Lange, C., editors, *The Semantic Web. Latest Advances and New Domains*, pages 677–692, Cham. Springer International Publishing.

[Khalili et al., 2018] Khalili, A., van den Besselaar, P., and de Graaf, K. A. (2018). FERASAT: A Serendipity-Fostering Faceted Browser for Linked Data. In Gangemi, A., Navigli, R., Vidal, M.-E., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., and Alam, M., editors, *The Semantic Web*, pages 351–366, Cham. Springer International Publishing.

[Khalsa et al., 2018] Khalsa, S., Cotroneo, P., and Wu, M. (2018). A survey of current practices in data search services. Technical report, Research Data Alliance Data (RDA) Discovery Paradigms Interest Group.

[Kibbe et al., 2015] Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C. J., Binder, J. X., Malone, J., Vasant, D., Parkinson, H., and Schriml, L. M. (2015). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research*, 43:D1071–8.

[Kilicoglu et al., 2018] Kilicoglu, H., Ben Abacha, A., Mrabet, Y., Shooshan, S. E., Rodriguez, L., Masterton, K., and Demner-Fushman, D. (2018). Semantic annotation of consumer health questions. *BMC Bioinformatics*, 19(1):34.

[Kim et al., 2019] Kim, D., Lee, J., So, C. H., Jeon, H., Jeong, M., Choi, Y., Yoon, W., Sung, M., and Kang, J. (2019). A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. *IEEE Access*, 7:73729–73740.

[Kim et al., 2003] Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:i180–i182.

[Koho et al., 2016] Koho, M., Heino, E., and Hyvönen, E. (2016). SPARQL Faceter—Client-side Faceted Search Based on SPARQL. In Troncy, R., Verborgh, R., Nixon, L., Kurz, T., Schlegel, K., and Vander Sande, M., editors, *Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshopco-located with the 13th Extended Semantic Web Conference ESWC 2016, Heraklion, Crete, Greece*.

[Kors et al., 2015] Kors, J. A., Clematide, S., Akhondi, S. A., van Mulligen, E. M., and Rebholz-Schuhmann, D. (2015). A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association : JAMIA*, 22:948–56.

[Krallinger et al., 2015] Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., Leaman, R., Lu, Y., Ji, D., Lowe, D. M., Sayle, R. A., Batista-Navarro, R. T., Rak, R., Huber, T., Rocktäschel, T., Matos, S., Campos, D., Tang, B., Xu, H., Munkhdalai, T., Ryu, K. H., Ramanan, S., Nathan, S., Žitnik, S., Bajec, M., Weber, L., Irmer, M., Akhondi, S. A., Kors, J. A., Xu, S., An, X., Sikdar, U. K., Ekbal, A., Yoshioka, M., Dieb, T. M., Choi, M., Verspoor, K., Khabsa, M., Giles, C. L., Liu, H., Ravikumar, K. E., Lamurias, A., Couto, F. M., Dai, H.-J., Tsai, R. T.-H., Ata, C., Can, T., Usié, A., Alves, R., Segura-Bedmar, I., Martínez, P., Oyarzabal, J., and Valencia, A. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(1):S2.

[Krug, 2013] Krug, S. (2013). *Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability*. New Riders.

[Köhler et al., 2019] Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N. L., Matentzoglu, N., McMurry, J. A., Osumi-Sutherland, D., Cipriani, V., Balhoff, J. P., Conlin, T., Blau, H., Baynam, G., Palmer, R., Gratian, D., Dawkins, H., Segal, M., Jansen, A. C., Muaz, A., Chang, W. H., Bergerson, J., Laulederkind, S. J. F., Yüksel, Z., Beltran, S., Freeman, A. F.,

Sergouniotis, P. I., Durkin, D., Storm, A. L., Hanauer, M., Brudno, M., Bello, S. M., Sincan, M., Rageth, K., Wheeler, M. T., Oegema, R., Lourghi, H., Della Rocca, M. G., Thompson, R., Castellanos, F., Priest, J., Cunningham-Rundles, C., Hegde, A., Lovering, R. C., Hajek, C., Olry, A., Notarangelo, L., Similuk, M., Zhang, X. A., Gómez-Andrés, D., Lochmüller, H., Dollfus, H., Rosenzweig, S., Marwaha, S., Rath, A., Sullivan, K., Smith, C., Milner, J. D., Leroux, D., Boerkoel, C. F., Klion, A., Carter, M. C., Groza, T., Smedley, D., Haendel, M. A., Mungall, C., and Robinson, P. N. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic acids research*, 47:D1018–D1027.

[Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

[Lebo et al., 2013] Lebo, T., Sahoo, S., McGuinness, D., Belhaj-jame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2013). PROV-O: The PROV Ontology. `https://www.w3.org/TR/prov-o/`, accessed on 23th August 2023.

[Lee et al., 2019] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

[Lee et al., 2016] Lee, S., Kim, D., Lee, K., Choi, J., Kim, S., Jeon, M., Lim, S., Choi, D., Kim, S., Tan, A.-C., and Kang, J. (2016). BEST: Next-Generation Biomedical Entity Search Tool for Knowledge Discovery from Biomedical Literature. *PLOS ONE*, 11(10):1–16.

[Lewis, 1995] Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *Int. J. Hum. Comput. Interact.*, 7:57–78.

[Li et al., 2019] Li, Z., Xu, G., Liang, X., Li, F., Wang, L., and Zhang, D. (2019). Exploring the Importance of Entities in Semantic Ranking. *Information*, 10(2).

[Likert, 1932] Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22(140):55.

[Liu et al., 2015] Liu, Y., Liang, Y., and Wishart, D. (2015). PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. *Nucleic acids research*, 43:W535–42.

[Liu et al., 2017] Liu, Y.-H., Thomas, P., Bacic, M., Gedeon, T., and Li, X. (2017). Natural Search User Interfaces for Complex Biomedical Search: An Eye Tracking Study. *Journal of the Australian Library and Information Association*, 66(4):364–381.

[Löffler et al., 2020] Löffler, F., Abdelmageed, N., Babalou, S., Kaur, P., and König-Ries, B. (2020). Tag me if you can! semantic annotation of biodiversity metadata with the QEMP corpus and the BiodivTagger. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4557–4564, Marseille, France. European Language Resources Association.

[Löffler and Klan, 2016] Löffler, F. and Klan, F. (2016). Does Term Expansion Matter for the Retrieval of Biodiversity Data? In Martin, M., Cuquet, M., and Folmer, E., editors, *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16), co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016)*. CEUR Workshop Proceedings.

[Löffler et al., 2017] Löffler, F., Opasjumruskit, K., Karam, N., Fichtmüller, D., Schindler, U., Klan, F., Müller-Birn, C., and Diepenbroek, M. (2017). Honey Bee Versus Apis Mellifera: A Semantic Search for Biological Data. In Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., and Hartig, O., editors, *The Semantic Web: ESWC 2017 Satellite Events: Portorož, Slovenia*, pages 98–103. Springer International Publishing.

[Löffler et al., 2023] Löffler, F., Shafiei, F., Witte, R., König-Ries, B., and Klan, F. (2023). Semantic Search for Biological Datasets: A Usability Study on Modes of Querying and Explaining Search Results. In König-Ries, B., Scherzinger, S., Lehner, W., and Vossen, G., editors, *BTW 2023*, pages 851–864. Gesellschaft für Informatik e.V., Bonn.

[Löffler et al., 2020] Löffler, F., Wesp, V., Babalou, S., Kahn, P., Lachmann, R., Sateli, B., Witte, R., and König-Ries, B. (2020). ScholarLensViz: A Visualization Framework for Transparency in Semantic User Profiles. In Taylor, K., Gonçalves, R., Lecue, F., and Yan, J., editors, *Proceedings of the ISWC 2020 Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 19th International Semantic Web Conference (ISWC 2020), Globally online, November 1-6, 2020 (UTC).*

[Loreau, 2010] Loreau, M. (2010). *Excellence in Ecology: Book 17, The Challenges of Biodiversity Science*. International Ecology Institute, Oldendorf, Germany.

[Löffler and Klan, 2022] Löffler, F. and Klan, F. (2022). Dataset: Comparative evaluation of a keyword based search and semantic search in a data portal for biodiversity research., `https://doi.org/10.5281/zenodo.7374398`.

[Löffler et al., 2022a] Löffler, F., Klan, F., and König-Ries, B. (2022a). Dataset: Survey for ranking preferences in search for biological datasets, `https://doi.org/10.5281/zenodo.7396827`.

[Löffler et al., 2022b] Löffler, F., König-Ries, B., and Klan, F. (2022b). Dataset: Entity based dataset retrieval - Evaluation results, `https://doi.org/10.5281/zenodo.7396781`.

[Löffler and Opasjumruskit, 2022] Löffler, F. and Opasjumruskit, K. (2022). Dataset: Relevance and usability evaluation in a data portal for biodiversity research, `https://doi.org/10.5281/zenodo.7374443`.

[Löffler et al., 2021] Löffler, F., Schuldt, A., König-Ries, B., Bruelheide, H., and Klan, F. (2021). A test collection for dataset retrieval in biodiversity research. *Research Ideas and Outcomes*, 7:e67887.

[Löffler et al., 2022c] Löffler, F., Shafiei, F., Witte, R., König-Ries, B., and Klan, F. (2022c). Dataset: Supplementary material for a usability evaluation of a semantic search for biological datasets, `https://doi.org/10.5281/zenodo.7388037`.

[Löffler et al., 2021] Löffler, F., Wesp, V., König-Ries, B., and Klan, F. (2021). Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs? *PLOS ONE*, 16(3):1–36.

[Madin et al., 2007] Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279–296. Meta-information systems and ontologies. A Special Feature from the 5th International Conference on Ecological Informatics ISEI5, Santa Barbara, CA, Dec. 4–7, 2006.

[Manning et al., 2014] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Maynard et al., 2017] Maynard, D., Bontcheva, K., and Augenstein, I. (2017). *Natural Language Processing for the Semantic Web*. Morgan & Claypool.

[Maynard et al., 2006] Maynard, D., Peters, W., and Li, Y. (2006). Metrics for Evaluation of Ontology-based Information Extraction. In *Workshop on Evaluation of Ontologies for the Web, WWW 2006, Edinburgh, UK*.

[McCowan et al., 2004] McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., and Bourlard, H. (2004). On the use of information retrieval measures for speech recognition evaluation. Technical report.

[McCrae et al., 2020] McCrae, J. P., Abele, A., Buitelaar, P., Cyganiak, R., Jentzsch, A., Andryushechkin, V., and Debattista, J. (2020). The Linked Open Data Cloud. `https://www.lod-cloud.net/`.

[Megler and Maier, 2013] Megler, V. M. and Maier, D. (2013). Data Near Here: Bringing Relevant Data Closer to Scientists. *Computing in Science Engineering*, 15(3):44–53.

[Megler and Maier, 2015] Megler, V. M. and Maier, D. (2015). Are Data Sets Like Documents?: Evaluating Similarity-Based Ranked Searchover Scientific Data. *TKDE: Transactions on Knowledge and Data Engineering*, 27(1).

[Meij et al., 2011] Meij, E., Bron, M., Hollink, L., Huurnink, B., and de Rijke, M. (2011). Mapping queries to the Linking Open Data cloud: A case study using DBpedia. *Journal of Web Semantics*, 9(4):418–433. JWS special issue on Semantic Search.

[MeSH, 2022] MeSH (2022). The National Center for Biotechnology Information, U.S. National Library of Medicine: MeSH (Medical Subject Headings). `https://www.ncbi.nlm.nih.gov/mesh`. accessed on 08.04.2022.

[Moreno-Vega and Hogan, 2018] Moreno-Vega, J. and Hogan, A. (2018). GraFa: Scalable Faceted Browsing for RDF Graphs. In Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M. C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.-A., and Simperl, E., editors, *The Semantic Web – ISWC 2018*, pages 301–317, Cham. Springer International Publishing.

[Mungall et al., 2012] Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E., and Haendel, M. A. (2012). Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13:R5.

[Müller et al., 2017] Müller, B., Poley, C., Pössel, J., Hagelstein, A., and Gübitz, T. (2017). LIVIVO - the Vertical Search Engine for Life Sciences. *Datenbank-Spektrum*, 17(1):29–34.

[Müller et al., 2008] Müller, H.-M., Rangarajan, A., Teal, T. K., and Sternberg, P. W. (2008). Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers. *Neuroinformatics*, 6(18949581):195–204.

[Naderi et al., 2011] Naderi, N., Kappler, T., Baker, C. J. O., and Witte, R. (2011). OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, 27(19):2721–2729.

[Natale et al., 2017] Natale, D. A., Arighi, C. N., Blake, J. A., Bona, J., Chen, C., Chen, S.-C., Christie, K. R., Cowart, J., D'Eustachio, P., Diehl, A. D., Drabkin, H. J., Duncan, W. D., Huang, H., Ren, J., Ross, K., Ruttenberg, A., Shamovsky, V., Smith, B., Wang, Q., Zhang, J., El-Sayed, A., and Wu, C. H. (2017). Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. *Nucleic acids research*, 45:D339–D346.

[Nature, 2018] Nature (2018). Scientific Data, Recommended Data Repositories. `https://www.nature.com/sdata/policies/repositories`, access date: 18.12.2022.

[NCBIGene, 2022] NCBIGene (2022). NCBI Gene: National Library of Medicine, National Center of Biotechnology Information. `https://www.ncbi.nlm.nih.gov/gene`. accessed on 22.04.2022.

[NCBITaxon, 2022] NCBITaxon (2022). NCBI organismal classification: An ontology representation of the NCBI organismal taxonomy. `https://obofoundry.org/ontology/ncbitaxon.html`. accessed on 22.11.2022.

[NCIT, 2022] NCIT (2022). Enterprise Vocabulary Services group of the Center for Biomedical Informatics and Information Technology, National Cancer Institute, NCI Thesaurus OBO Edition. `https://obofoundry.org/ontology/ncit.html`. accessed on 22.04.2022.

[Nentidis et al., 2017] Nentidis, A., Bougiatiotis, K., Krithara, A., Paliouras, G., and Kakadiaris, I. (2017). Results of the fifth edition of the BioASQ Challenge. In *BioNLP 2017*, pages 48–57, Vancouver, Canada,. Association for Computational Linguistics.

[Nguyen et al., 2019] Nguyen, N. T., Gabud, R. S., and Ananiadou, S. m. A. (2019). COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal*, (7).

[Nielsen, 1993] Nielsen, J. (1993). Chapter 6 - Usability Testing. In NIELSEN, J., editor, *Usability Engineering*, pages 165–206. Morgan Kaufmann, San Diego.

[Nielsen, 1999] Nielsen, J. (1999). *Designing Web Usability: The Practice of Simplicity*.

[Nielsen and Norman, 2022] Nielsen, J. and Norman, D. (2022). Nielsen Norman Group (NNG): World Leaders in Research-Based User Experience. `https://www.nngroup.com/`. accessed on 14.06.2022.

[Nishioka and Scherp, 2016] Nishioka, C. and Scherp, A. (2016). Profiling vs. time vs. content: What does matter for top-k publication recommendation based on twitter profiles? In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, pages 171–180.

[NKGCF, 2008] NKGCF (2008). German National Committee on Global Change Research: Global Change Research in Germany 2008. `http://ccsl.iccip.net/gcrg2008.pdf`.

[OBOFoundry, 2022] OBOFoundry (2022). Physico-chemical process: An ontology of physico-chemical processes, i.e. physico-chemical changes occurring in course of time. `https://obofoundry.org/ontology/rex.html`. accessed on 01.08.2022.

[Olsen and Stevens, 2010] Olsen, L. M. and Stevens, T. (2010). Updates to the directory interchange format (dif) standard. `https://www.earthdata.nasa.gov/esdis/esco/standards-and-references/directory-interchange-format-dif-standard`, accessed on 13th July 2022.

[Osgood et al., 1975] Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1975). *The measurement of meaning*. Urbana : University of Illinois Press.

[Pachzelt et al., 2021] Pachzelt, A., Kasperek, G., Lücking, A., Abrami, G., and Driller, C. (2021). Semantic Search in Legacy Biodiversity Literature: Integrating data from different data infrastructures. *Biodiversity Information Science and Standards*, 5:e74251.

[Pafilis et al., 2013] Pafilis, E., Frankild, S. P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., Arvanitidis, C., and Jensen, L. J. (2013). The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLOS ONE*, 8(6):1–6.

[Page, 2019] Page, R. D. (2019). Ozymandias: a biodiversity knowledge graph. *PeerJ*, 7:e6739.

[Parker et al., 2016] Parker, T. H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J. D., Chee, Y. E., Kelly, C. D., Gurevitch, J., and Nakagawa, S. (2016). Transparency in Ecology and Evolution: Real Problems, Real Solutions. *Trends in Ecology & Evolution*, 31(9):711 – 719.

[Parr et al., 2014] Parr, C. S., Wilson, N., Leary, P., S. Schulz, K., Lans, K., Walley, L., A. Hammock, J., Goddard, A., Rice, J., Studer, M., G. Holmes, J. T., and Robert J. Corrigan, J. (2014). The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodiversity Data Journal*, 2:e1079.

[Pazzani and Billsus, 2007] Pazzani, M. J. and Billsus, D. (2007). *Content-Based Recommendation Systems*, pages 325–341. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Peng et al., 2018] Peng, Y., Rios, A., Kavuluru, R., and Lu, Z. (2018). Extracting chemical–protein relations with ensembles of SVM and deep learning models. *Database*, 2018. bay073.

[Pérez-Luque et al., 2020] Pérez-Luque, A. J., Gea-Izquierdo, G., and Zamora, R. (2020). Resilience to drought of relict Mediterranean Quercus pyrenaica populations in the southern Iberian (Sierra Nevada, Spain).

[Plepi et al., 2021] Plepi, J., Kacupaj, E., Singh, K., Thakkar, H., and Lehmann, J. (2021). Context Transformer with Stacked Pointer Networks for Conversational Question Answering over Knowledge Graphs. In Verborgh, R., Hose, K., Paulheim, H., Champin, P.-A., Maleshkova, M., Corcho, O., Ristoski, P., and Alam, M., editors, *The Semantic Web*, pages 356–371, Cham. Springer International Publishing.

[Polychronopoulos et al., 2013] Polychronopoulos, D., Almirantis, Y., Krithara, A., and Paliouras, G. (2013). Expert Team of the BioASQ project, `http://www.bioasq.org/sites/default/files/PublicDocuments/BioASQ_D3.1-ExpertTeam_final_0.pdf`. Project deliverable D3.1.

[Pyysalo and Ananiadou, 2013] Pyysalo, S. and Ananiadou, S. (2013). Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

[Qi et al., 2020] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

[Qu et al., 2019] Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., and Iyyer, M. (2019). BERT with History Answer Embedding for Conversational Question Answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1133–1136, New York, NY, USA. Association for Computing Machinery.

[Quarfoot and Levine, 2016] Quarfoot, D. and Levine, R. A. (2016). How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution? *The American Statistician*, 70(4):373–384.

[Ramakers et al., 2018] Ramakers, J., Culina, A., Visser, M., and Gienapp, P. (2018). Environmental coupling of heritability and selection is rare and of minor evolutionary significance in wild populations. *Nature Ecology & Evolution*, 2.

[re3data, 2018] re3data (2018). https://https://www.re3data.org, accessed on 21.11.2022.

[Research Data Alliance, 2020] Research Data Alliance (2020). RDA Metadata Standards Catalog, version 2. https://rdamsc.bath.ac.uk/, accessed on 15.11.2022.

[Resnik, 1999] Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-based Measure and Its Application to Problems of Ambiguity in Natural Language. *J. Artif. Int. Res.*, 11(1):95–130.

[Roberts et al., 2017] Roberts, K., Gururaj, A. E., Chen, X., Pournejati, S., Hersh, W. R., Demner-Fushman, D., Ohno-Machado, L., Cohen, T., and Xu, H. (2017). Information retrieval for biomedical datasets: the 2016 bioCADDIE dataset retrieval challenge. *Database*, 2017.

[Russel and Norvig, 2021] Russel, S. J. and Norvig, P. (2021). *Artificial Intelligence - A Modern Approach, 4th edition*. Pearson.

[Saggion et al., 2007] Saggion, H., Funk, A., Maynard, D., and Bontcheva, K. (2007). Ontology-Based Information Extraction for Business Intelligence. In Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., and Cudré-Mauroux, P., editors, *The Semantic Web*, pages 843–856, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620.

[Sateli et al., 2017] Sateli, B., Löffler, F., König-Ries, B., and Witte, R. (2017). ScholarLens: extracting competences from research publications for the automatic generation of semantic user profiles. *PeerJ Computer Science*, 3:e121.

[Sateli and Witte, 2015] Sateli, B. and Witte, R. (2015). Semantic representation of scientific literature: bringing claims, contributions and named entities onto the Linked Open Data cloud. *PeerJ Computer Science*, 1:e37.

[Schieber, 1987] Schieber, P. (1987). The Wit and Wisdom of Grace Hopper (OCLC Newsletter, March/April, No. 167). https://cs.yale.edu/homes/tap/Files/hopper-wit.html, accessed on 7th August 2023.

[Schriml et al., 2021] Schriml, L. M., Munro, J. B., Schor, M., Olley, D., McCracken, C., Felix, V., Baron, J., Jackson, R., Bello, S., Bearer, C., Lichenstein, R., Bisordi, K., Dialo, N. C., Giglio, M., and Greene, C. (2021). The Human Disease Ontology 2022 update. *Nucleic Acids Research*, 50(D1):D1255–D1261.

[Schweiger et al., 2014] Schweiger, D., Trajanoski, Z., and Pabinger, S. (2014). SPARQLGraph: a web-based platform for graphically querying biological Semantic Web databases. *BMC bioinformatics*, 15:279.

[Semantic Web Education and Outreach Interest Group, 2009] Semantic Web Education and Outreach Interest Group (2009). `https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`.

[Shafiei et al., 2021] Shafiei, F., Löffler, F., Thiel, S., Opasjumruskit, K., Grabiger, D., Rauh, P., and König-Ries, B. (2021). [Dai:Si] - A Modular Dataset Retrieval Framework with a Semantic Search for Biological Data. In Sanfilippo, E. M., Kutz, O., Troquard, N., Hahmann, T., Masolo, C., Hoehndorf, R., Vita, R., Algergawy, A., Karam, N., Klan, F., Michel, F., and Rosati, I., editors, *S4BioDiv 2021: 3rd International Workshop on Semantics for Biodiversity, held at JOWO 2021: Episode VII The BolzanoSummer of Knowledge, September 11–18, 2021, Bolzano, Italy*.

[Shen et al., 2019] Shen, T., Geng, X., Qin, T., Guo, D., Tang, D., Duan, N., Long, G., and Jiang, D. (2019). Multi-Task Learning for Conversational Question Answering over a Large-Scale Knowledge Base. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2442–2451, Hong Kong, China. Association for Computational Linguistics.

[Shimoyama et al., 2012] Shimoyama, M., Nigam, R., McIntosh, L. S., Nagarajan, R., Rice, T., Rao, D. C., and Dwinell, M. R. (2012). Three ontologies to define phenotype measurement data. *Frontiers in genetics*, 3:87.

[Shneiderman et al., 2016] Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., and Diakopoulos, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction (6th edition)*. Pearson.

[Silva, 2009] Silva, D. (2009). Internet has only just begun, say founders, *Phys.org*. `https://phys.org/news/2009-04-internet-begun-founders.html`, accessed on 17th August, 2023.

[Smith et al., 2005] Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A. L., and Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biology*, 6(R46).

[Smith et al., 2015] Smith et al. (2015). Basic Formal Ontology 2.0 Specification and Users's Guide, `https://github.com/bfo-ontology/bfo/raw/master/docs/bfo2-reference/bfo2-reference.pdf`, `https://raw.githubusercontent.com/bfo-ontology/bfo/master/releases/2.0/bfo.owl`. Technical report.

[Sommerville, 2021] Sommerville, I. (2021). *Software Engineering (Tenth Edition)*. PEARSON.

[Soto et al., 2018] Soto, A. J., Przybyła, P., and Ananiadou, S. (2018). Thalia: semantic search engine for biomedical abstracts. *Bioinformatics*, 35(10):1799–1801.

[Soylu et al., 2013] Soylu, A., Giese, M., Jimenez-Ruiz, E., Kharlamov, E., Zheleznyakov, D., and Horrocks, I. (2013). OptiqueVQS: Towards an Ontology-Based Visual Query System for Big Data. In *Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems*, MEDES '13, page 119–126, New York, NY, USA. Association for Computing Machinery.

[Sparck Jones et al., 2000] Sparck Jones, K., Walker, S., and Robertson, S. (2000). A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management*, 36(6):779–808.

[Stoyanovich et al., 2011] Stoyanovich, J., Lodha, M., Mee, W., and Ross, K. A. (2011). SkylineSearch: Semantic Ranking and Result Visualization for Pubmed. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

[Suchanek et al., 2008] Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). YAGO: A Large Ontology from Wikipedia and WordNet. *Journal of Web Semantics*, 6(3):203–217. World Wide Web Conference 2007Semantic Web Track.

[Sy et al., 2012] Sy, M. F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., and Ranwez, V. (2012). User centered and ontology based information retrieval system for life sciences. *BMC Bioinformatics*.

[Tang and Mooney, 2001] Tang, L. R. and Mooney, R. J. (2001). Using Multiple Clause Constructors in Inductive Logic Programming for Semantic Parsing. In De Raedt, L. and Flach, P., editors, *Machine Learning: ECML 2001*, pages 466–477, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Tchechmedjiev et al., 2018] Tchechmedjiev, A., Abdaoui, A., Emonet, V., Melzi, S., Jonnagaddala, J., and Jonquet, C. (2018). Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator+. *Bioinformatics*, 34(11):1962–1965.

[Terolli et al., 2020] Terolli, E., Ernst, P., and Weikum, G. (2020). Focused Query Expansion with Entity Cores for Patient-Centric Health Search. In Pan, J. Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., and Kagal, L., editors, *The Semantic Web – ISWC 2020*, pages 547–564, Cham. Springer International Publishing.

[Thessen et al., 2015] Thessen, A. E., Bunker, D. E., Buttigieg, P. L., Cooper, L. D., Dahdul, W. M., Domisch, S., Franz, N. M., Jaiswal, P., Lawrence-Dill, C. J., Midford, P. E., Mungall, C. J., Ramírez, M. J., Specht, C. D., Vogt, L., Vos, R. A., Walls, R. L., White, J. W., Zhang, G., Deans, A. R., Huala, E., Lewis, S. E., and Mabee, P. M. (2015). Emerging semantics to link phenotype and environment. *PeerJ*, 3:e1470.

[Thessen et al., 2012] Thessen, A. E., Cui, H., and Mozzherin, D. (2012). Applications of Natural Language Processing in Biodiversity Science. *Advances in Bioinformatics*, 2012(Article ID 391574):17 pages.

[Thomas et al., 2012] Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S., and Leser, U. (2012). GeneView: a comprehensive semantic search engine for PubMed. *Nucleic acids research*, 40(22693219):W585–W591.

[Thornton and Shiri, 2021] Thornton, G. M. and Shiri, A. (2021). Challenges with organization, discoverability and access in Canadian open health data repositories. *Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada*, 42(1).

[Tsatsaronis et al., 2015] Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artiéres, T., Ngomo, A.-C. N., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., and Paliouras, G. (2015). An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

[Tsatsaronis and Panagiotopoulou, 2009] Tsatsaronis, G. and Panagiotopoulou, V. (2009). A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL '09, pages 70–78, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Tullis and Albert, 2013a] Tullis, T. and Albert, B. (2013a). Chapter 3 - Planning. In Tullis, T. and Albert, B., editors, *Measuring the User Experience (Second Edition)*, Interactive Technologies, pages 41–62. Morgan Kaufmann, Boston, second edition edition.

[Tullis and Albert, 2013b] Tullis, T. and Albert, B. (2013b). Chapter 4 - Performance Metrics. In Tullis, T. and Albert, B., editors, *Measuring the User Experience (Second Edition)*, Interactive Technologies, pages 63–97. Morgan Kaufmann, Boston, second edition edition.

[Tullis and Albert, 2013c] Tullis, T. and Albert, B. (2013c). Chapter 5 - Issue-Based Metrics. In Tullis, T. and Albert, B., editors, *Measuring the User Experience (Second Edition)*, Interactive Technologies, pages 99–120. Morgan Kaufmann, Boston, second edition edition.

[Tullis and Albert, 2013d] Tullis, T. and Albert, B. (2013d). Chapter 6 - Self-Reported Metrics. In Tullis, T. and Albert, B., editors, *Measuring the User Experience (Second Edition)*, Interactive Technologies, pages 121–161. Morgan Kaufmann, Boston, second edition edition.

[Tullis and Stetson, 2004] Tullis, T. S. and Stetson, J. N. (2004). A Comparison of Questionnaires for Assessing Website Usability. In *Usability Professionals Association (UPA) 2004 conference*, Minneapolis, MN.

[UMLS, 2022] UMLS (2022). The National Center for Biotechnology Information, U.S. National Library of Medicine: Unified Medical Language System (UMLS). `https://www.nlm.nih.gov/research/umls/`. accessed on 08.04.2022.

[Unger et al., 2014] Unger, C., Freitas, A., and Cimiano, P. (2014). *An Introduction to Question Answering over Linked Data*, pages 100–140. Springer International Publishing, Cham.

[van Mulligen et al., 2012] van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J. A., and Furlong, L. I. (2012). The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884. Text Mining and Natural Language Processing in Pharmacogenomics.

[Vega-Gorgojo et al., 2016a] Vega-Gorgojo, G., Slaughter, L., Giese, M., Heggestøyl, S., Klüwer, J. W., and Waaler, A. (2016a). PepeSearch: Easy to Use and Easy to Install Semantic Data Search. In Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., and Lange, C., editors, *The Semantic Web*, pages 146–150, Cham. Springer International Publishing.

[Vega-Gorgojo et al., 2016b] Vega-Gorgojo, G., Slaughter, L., Giese, M., Heggestøyl, S., Soylu, A., and Waaler, A. (2016b). Visual query interfaces for semantic datasets: An evaluation study. *Journal of Web Semantics*, 39:81–96.

[VIVO,] VIVO. VIVO Core Ontology. `http://vivoweb.org/ontology/core`. accessed on 01.09.2023.

[Volentine et al., 2015] Volentine, R., Owens, A., Tenopir, C., and Frame, M. (2015). Usability Testing to Improve Research Data Services. *Qualitative and Quantitative Methods in Libraries*, 4(1):59–68.

[W3C, 2014] W3C (2014). The RDF Data Cube Vocabulary. https://www.w3.org/TR/vocab-data-cube/, accessed on 12th July 2022.

[W3C, 2020] W3C (2020). Data Catalog Vocabulary (DCAT). https://www.w3.org/TR/vocab-dcat/, accessed on 13th July 2022.

[Waitelonis et al., 2015] Waitelonis, J., Exeler, C., and Sack, H. (2015). Enabled Generalized Vector Space Model to Improve Document Retrieval. In *Proceedings of the Third NLP & DBpedia Workshop (NLP & DBpedia 2015) co-located with the 14th International Semantic Web Conference 2015 (ISWC 2015) in Bethlehem, Pennsylvania, USA, October 11, 2015.*

[Walls et al., 2014] Walls, R. L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., and et al. (2014). Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. *PLoS ONE 9(3): e89606*.

[Walls, R. et al., 2022] Walls, R. et al. (2022). Plant Phenology Ontology: An ontology for describing the phenology of individual plants and populations of plants, and for integrating plant phenological data across sources and scales. https://obofoundry.org/ontology/ppo.html. accessed on Jan 4, 2022.

[Wang et al., 2019] Wang, Z., Ng, P., Ma, X., Nallapati, R., and Xiang, B. (2019). Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China. Association for Computational Linguistics.

[Wei et al., 2019] Wei, C.-H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1):W587–W593.

[Wei et al., 2013] Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41:W518–22.

[Weinberger, 2008] Weinberger, D. (2008). Knowledge at the End of the Information Age, Bertha Bassum lecture at the Faculty of Information at the University of Toronto. https://archive.org/details/KnowledgeAtTheEndOfTheInformationAge, accessed on 06 August 2023.

[Wentzel et al., 2023] Wentzel, B., Kirstein, F., Jastrow, T., Sturm, R., Peters, M., and Schimmler, S. (2023). An extensive methodology and framework for quality assessment of dcat-ap datasets. In Lindgren, I., Csáki, C., Kalampokis, E., Janssen, M.,

Viale Pereira, G., Virkar, S., Tambouris, E., and Zuiderwijk, A., editors, *Electronic Government*, pages 262–278, Cham. Springer Nature Switzerland.

[Whetzel et al., 2011] Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(suppl 2):W541–W545.

[Wikidata, 2022] Wikidata (2022). Wikidata. `https://www.wikidata.org`. accessed on April 22, 2022.

[Wikipedia, 2022] Wikipedia (2022). Wikipedia. `https://www.wikipedia.org/`. accessed on April 22, 2022.

[Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*, (160018).

[Wilson, 1988] Wilson, E. O. e. a. (1988). *Commission on Life Sciences, Division on Earth and Life Studies, National Academy of Sciences/Smithsonian Institution: Biodiversity*. National Academies Press.

[Wishart et al., 2018] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46:D1074–D1082.

[Wishart et al., 2007] Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., Macinnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D.,

Vogel, H. J., and Querengesser, L. (2007). HMDB: the Human Metabolome Database. *Nucleic acids research*, 35:D521–6.

[Witte and Gitzinger, 2008] Witte, R. and Gitzinger, T. (2008). Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients. In Domingue, J. and Anutariya, C., editors, *The Semantic Web*, pages 360–374, Berlin, Heidelberg. Springer Berlin Heidelberg.

[WordNet, 2022] WordNet (2022). WordNet: A Lexical Database for English. `https://wordnet.princeton.edu/`. accessed on 22.04.2022.

[Wrigley et al., 2010] Wrigley, S. N., Elbedweihy, K., Reinhard, D., Bernstein, A., and Ciravegna, F. (2010). Evaluating Semantic Search Tools using the SEALS platform. In *Proceedings of the International Workshop on Evaluation of Semantic Technologies (IWEST 2010)Workshop at the 9th International Semantic Web Conference (ISWC2010) - ISWC 2010 Workshops Volume IIIShanghai, China, November 8, 2010.*

[Wu et al., 2019] Wu, M., Psomopoulos, F., Khalsa, S. J., and de Waard, A. (2019). Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. *Data Science Journal*, 18(1):3.

[Xiong et al., 2016] Xiong, C., Callan, J., and Liu, T.-Y. (2016). Bag-of-Entities Representation for Ranking. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 181–184, New York, NY, USA. Association for Computing Machinery.

[Xiong et al., 2017a] Xiong, C., Callan, J., and Liu, T.-Y. (2017a). Word-Entity Duet Representations for Document Ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 763–772, New York, NY, USA. Association for Computing Machinery.

[Xiong et al., 2017b] Xiong, C., Liu, Z., Callan, J., and Hovy, E. (2017b). Jointsem: Combining query entity linking and entity based document ranking. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 2391–2394, New York, NY, USA. Association for Computing Machinery.

[Xiong et al., 2017c] Xiong, C., Power, R., and Callan, J. (2017c). Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1271–1279, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[Xu and Zhuge, 2014] Xu, B. and Zhuge, H. (2014). Automatic faceted navigation. *Future Generation Computer Systems*, 32:187 – 197. Special Section: The Management of Cloud Systems, Special Section: Cyber-Physical Society and Special Section: Special Issue on Exploiting Semantic Technologies with Particularization on Linked Data over Grid and Cloud Architectures.

[Xu et al., 2020] Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., Kang, J., Rousseau, J. F., Li, X., Xu, W., Torvik, V. I., Bu, Y., Chen, C., Ebeid, I. A., Li, D., and Ding, Y. (2020). Building a PubMed knowledge graph. *Scientific data*, 7:205.

[Yates et al., 2021] Yates, A., Nogueira, R., and Lin, J. (2021). *Pretrained Transformers for Text Ranking: BERT and Beyond*, page 1154–1156. Association for Computing Machinery, New York, NY, USA.

[Yilmaz et al., 2008] Yilmaz, E., Kanoulas, E., and Aslam, J. A. (2008). A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 603–610, New York, NY, USA. Association for Computing Machinery.

[Zahiri et al., 2021] Zahiri, R., Tarmann, G., Efetov, K. A., Rajaei, H., Fatahi, M., Seidel, M., Jaenicke, B., Dalsgaard, T., Sikora, M., and Husemann, M. (2021). Supplementary material 1 from: Zahiri R, Tarmann G, Efetov KA, Rajaei H, Fatahi M, Seidel M, Jaenicke B, Dalsgaard T, Sikora M, Husemann M (2021) An illustrated catalogue of the type specimens of Lepidoptera (Insecta) housed in the Zoological Museum Hamburg (ZMH): Part I. superfamilies Hepialoidea, Cossoidea, and Zygaenoidea. Evolutionary Systematics 5(1): 39-70. `https://doi.org/10.3897/evolsyst.5.62003`.

[Zaki and Tennakoon, 2017] Zaki, N. and Tennakoon, C. (2017). BioCarian: search engine for exploratory searches in heterogeneous biological databases. *BMC Bioinformatics*, 18(1):435.

[Özgür et al., 2016] Özgür, A., Hur, J., and He, Y. (2016). The Interaction Network Ontology-supported modeling and mining of complex interactions represented with multiple keywords in biomedical literature. *BioData Mining*, 9(1):41.