Tag Me If You Can! Semantic Annotation of Biodiversity Metadata with the QEMP Corpus and the BiodivTagger

Felicitas Löffler1, Nora Abdelmageed1 2, Samira Babalou1, Pawandeep Kaur1, Birgitta König-Ries1 2 3
1Heinz Nixdorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Germany
2Michael Stifel Center for Data-Driven and Simulation Science, Jena, Germany
3German Center for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Germany
{felicitas.loeffler, nora.abdelmageed, samira.babalou, pawandeep.kaur, birgitta.koenig-ries}@uni-jena.de

Abstract

Dataset Retrieval is gaining importance due to a large amount of research data and the great demand for reusing scientific data. Dataset Retrieval is mostly based on metadata, structured information about the primary data. Enriching these metadata with semantic annotations based on Linked Open Data (LOD) enables datasets, publications and authors to be connected and expands the search on semantically related terms. In this work, we introduce the BiodivTagger, an ontology-based Information Extraction pipeline, developed for metadata from biodiversity research. The system recognizes biological, physical and chemical processes, environmental terms, data parameters and phenotypes as well as materials and chemical compounds and links them to concepts in dedicated ontologies. To evaluate our pipeline, we created a gold standard of 50 metadata files (QEMP corpus) selected from five different data repositories in biodiversity research. To the best of our knowledge, this is the first annotated metadata corpus for biodiversity research data. The results reveal a mixed picture. While materials and data parameters are properly matched to ontological concepts in most cases, some ontological issues occurred for processes and environmental terms.

Keywords: Semantic Annotation, Ontology-Based Information Extraction, Gold Standard, Metadata, Biodiversity Research, Life Sciences

1. Introduction

Metadata are an important source in Dataset Retrieval (Sirjodha Khalsa, 2018) as they contain compressed information about data collection objects, geographic location, author and temporal expressions in a structured and machine-readable form. Driven by research fields such as Semantic Publishing (Shadbolt et al., 2006) and initiatives such as the FAIR principles (Wilkinson et al., 2016), metadata are increasingly semantically enriched with Uniform Resource Identifiers (URI) based on Linked Open Data (LOD) (Heath and Bizer, 2011) to foster interoperability of information about authors, papers and datasets. Scholars also benefit from semantically enriched metadata in the retrieval process as the search results can be expanded on related information such as synonyms (Löffler et al., 2017).

One domain that requires such improved retrieval techniques and ontological enhancement is biodiversity research, a discipline dealing with the variety of species, the genetic diversity and the diversity of functions, interactions and ecosystems. Numerous text mining applications have been developed to detect named entities such as species, persons and geographic locations (Naderi et al., 2011), (Cunningham et al., 2013). However, based on our previous research (Löffler et al., 2020), we figured that further entity types are important to be considered, when searching for datasets in biodiversity research. Besides organisms and geographic locations, scholars are interested in habitats and environmental information where the organisms occur. Furthermore, biological, chemical and physical processes influencing these environments or organisms, materials and chemical compounds as well as data parameters are further important entity types in biodiversity research. The introduction of the Essential Annotation Schema for Ecology (Pfaff et al., 2017), a new metadata schema that was particularly developed for search, also confirms that interests in biodiversity research go beyond species observation. Our literature review (Section 2) reveals that there are very limited Information Extraction (IE) approaches that are based on biodiversity metadata and that provide access to the LOD cloud.

In this paper, we introduce the BiodivTagger, a text mining pipeline using domain knowledge from selected ontologies to identify main entities in biodiversity research metadata. To evaluate our pipeline, we manually annotated 50 metadata files selected from five different biodiversity research data repositories and projects. We annotated four main entity types that are highly relevant for Dataset Retrieval tasks in this field, namely, phenotypic qualities and characteristics that can be measured or observed (QUALITY), environmental terms (ENVIRONMENT), materials and chemicals (MATERIAL) and biological, chemical and physical processes (PROCESS). The pipeline and the gold standard, called QEMP (Quality, Environment, Material, Process) corpus, are publicly available in our github repository.

The contribution of this paper is twofold:

1. The QEMP gold standard corpus which we believe to be the very first gold standard corpus created from biodiversity metadata.
2. The BiodivTagger, a text mining pipeline based on
ontological information extraction that detects ENVIRONMENT, PROCESS, MATERIAL and QUALITY entities.

The structure of the paper is as follows: Related work is presented in Section 2, we describe our approach in Section 3, followed by the evaluation in Section 4. The conclusion and future work are presented in Section 5.

2. Related Work

The advancement of Natural Language Processing (NLP) techniques leads to the development of many different Information Extraction tools for biological research. However, due to a specialized language in scientific communication in the Life Sciences, diverse content, imprecise and inconsistent naming (Thessen et al., 2012; Ananiadou et al., 2004), the extraction of biological entities remains a challenge.

Named Entity Recognition (NER) is the first step in an IE task, as it provides the backbone for the subsequent semantic text interpretation. In the biomedical domain, numerous tools have already been developed to extract entities related to biomedicine such as diseases, chemical compounds, genes, enzymes or proteins, e.g., (Cunningham et al., 2013). Due to the increasing semantic domain knowledge in the LOD cloud, a variety of approaches have also been introduced that link detected entities to ontological concepts (Jovanović and Bagheri, 2017). For instance, Bioportal (Jonquet et al., 2009) provides a graphical interface and API to match terms and phrases to entries in biomedical and biological ontologies. However, disambiguation or determination of entity types are not provided yet. The suitability of ontologies for NER tasks has already been studied by (Gurulingappa et al., 2010). Their outcome reveals that the usage of several ontologies lead to a good match of ontological concepts in biomedical literature.

Despite of numerous resources available for the biomedical domain, unfortunately, very few studies focus on the special needs for biodiversity research. Mainly, existing tools concentrate on the extraction of taxonomic information and species names, e.g., (Naderi et al., 2011), (Wood et al., 2004) and TaxonFinder. The detection of morphological characters or phylogenetic attributes was studied by (Balhoff et al., 2010; Elaison et al., 2019). Other relevant entity types such as environmental terms or processes are not investigated.

In order to evaluate IE tools, manually created gold standard corpora are required. Two of those which are closely related to our work are both based on biodiversity literature: 2

1. The Bacteria Biotope (BB) (Deleger et al., 2016) corpus is a result of the subtask of BioNLP task which was first introduced in 2011 with the ambition to use IE from scientific documents at a large scale in order to automatically fill knowledge bases (Bossy et al., 2012). The BB task consists of the extraction of bacteria and their locations (habitats or geographical places) from scientific literature, their categorization according to the NCBI taxonomy and OntoBiotope ontology and the linking of bacteria to their locations through localization events.

2. The COPIOUS corpus (Nguyen et al., 2019) is another gold standard corpus in which geographical locations, habitats, temporal expressions and person name entities from species occurrence records are annotated in 200 scientific documents.

To the best of our knowledge, there is neither an annotated corpus for metadata of biodiversity research data, nor tools available to detect further related entities such as materials, processes and environmental terms. Furthermore, corpora from biodiversity literature are not suitable for our overall research goal of using the pipeline for improved Dataset Retrieval as it is primarily based on metadata. Therefore, it is also important to evaluate our pipeline on a corpus created from biodiversity metadata only.

Moreover, publications and metadata differ in format and size, and the complexity can vary greatly from basic entries such as author, title and citation to detailed information on measured data parameters and research methods used. In addition, more research is needed in terms of the suitability of ontologies for NER and IE tasks for applied domains.

3. Methodology

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 or the GFBio Terminology Service. Initiatives such as the OBO Foundry aim to provide terminologies that are interlinked and that adhere to several principles including open use and strictly-tailored content. This well-structured domain knowledge can be used as ontological gazetteers in IE tasks. In the following, we introduce the selected entity types, explain what ontologies are used and provide an overview of the architecture.

3.1. Important Entity Types in Biodiversity Dataset Search

In our previous research (Löffler et al., 2020), we explored what information biodiversity researchers are interested in when searching for datasets. The outcome reveals that species, environmental terms, processes, chemical compounds and materials, geographic locations, data types and data parameters (quality) are main categories in biodiversity dataset search. As a variety of taggers and approaches exist to determine species or geographic locations in textual resources, we did not consider these entity types. For data types, very few ontologies exist. Therefore, we also left them out in our development. Table 1 introduces the identified entity types that are considered in the following

1 In March 2019, one-third of the registered datasets in the LOD cloud were terminologies from the biological and biomedical domain: https://www.lod-cloud.net/ http://taxonfinder.org/
approach. We concentrate 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).

3.2. Ontology Selection
Figure 1 presents the ontologies used. In order to form ontological gazetteers, we carefully selected terminologies or parts of them from OBO Foundry and assigned them to the identified entity types. That diminishes the risk of too broad terms and also ensures a light-weight disambiguation. In order to determine environmental terms, we utilize parts of the ENVO ontology (Buttigieg et al., 2013). As ENVO has a broad scope and comprises also processes, materials and quality concepts, we fetch only concepts starting from nodes that correspond to our definition of Environment. The same applies for PATO to detect phenotypic qualities. PATO also contains concepts that are too broad for our purpose, and therefore, we utilize only concepts from specific nodes. Process entities are obtained from several ontologies, e.g., ENVO, Gene Ontology (GO) and UBERON. For detecting materials, we use the Chemical Entities of Biological Interest (CheBI) and parts of ENVO describing environmental materials.

3.3. Architecture
Our approach is based on the widely used text mining framework GATE (Cunningham et al., 2013) that already provides basic text mining functions and offers various plugins for the Life Sciences. Figure 2 presents the overall workflow. At first, the metadata documents go through a text extraction phase in which the XML structure is removed. Afterwards, in a pre-processing phase, syntactical steps such as tokenization, sentence splitting and Part-Of-Speech (POS) tagging are executed. The results are token annotations including the identification of noun entities, verbs and adjectives. In order to use all inflected forms of nouns (singular vs. plural), we lemmatize the document’s text. In this syntactical phase, we mainly use GATE’s in-build processing steps and the ANNIE pipeline that in addition also extracts general named entities such as Person, Location, Date and Time. Our own contribution is represented in the following semantic analysis. Each entity type is formed by large ontological gazetteer lists. GATE’s Large Knowledge Base (LKB) Gazetteer plugin offers an easy access to any remote knowledge base with a SPARQL interface. However, in order not to be dependent on external providers, we downloaded and host all ontologies in our own GraphDB triple store. The SPARQL queries, an excerpt is provided in Listing 1, are stored in text files. The LKB plugin takes the stored queries, sends them to the SPARQL interface and receives a list of ontology concepts with URIs and labels.

Listing 1: Excerpt from a SPARQL query to retrieve all concepts and subconcepts for entity type Environment

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX oboInOwl: <http://purl.obolibrary.org/obo/owl#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX uberon: <http://uberon.org/>
PREFIX chebi: <http://www.ebi.ac.uk/chebi/>
PREFIX envo: <http://gfbio.github.io/envo/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/>
PREFIX gfbio: <http://gfbio.github.io/gfbio/

As synonyms have the same meaning as the given term, we also consider semantic relations such as hasRelatedSynonym, hasExactSynonym. Once the ontological lists are received, per entity type, transducers match the document’s tokens against the list entries, link them to their corresponding resource URI and create a semantic annotation. The look up is performed using case-insensitive, word order-sensitive and all possible string matches as constraints. We use GATE’s Chemi-
Figure 1: Ontologies and their concepts used, from which all sub-nodes and concepts with labels and synonyms are retrieved.

Figure 2: This figure presents the overall flow. Subsequently, the processing resources are executed.

Metadaten und ihre Konzepte, aus denen alle Sub-Nodes und Konzepte mit Labeln und Synonymen extrahiert werden.

4. Evaluation & Results

Due to the lack of semantically annotated metadata corpora in biodiversity research, we created our own gold standard (Subsection 4.1.) to evaluate our text mining pipeline. We compared the generated annotations of the pipeline with the manually labeled annotation set (Subsection 4.2.) and discuss the results in Subsection 4.3.

4.1. Dataset

Metadata descriptions can greatly vary from sparse content mentioning only author, title and identifier to very detailed information on data structure, units and methodology. This diversity becomes even stronger by the variety of metadata standards across different research disciplines. In order to reflect this heterogeneity and diversity of biological datasets, we selected 50 publicly available metadata from five different data repositories and project databases with various metadata formats. The corpus contains 10 files each from Dryad [16], a generic data repository in Dublin Core [17] format, PANGAEA [18] (a data archive for environmental data) in an extended Dublin Core format called PanMDI [19] and three project-related portals and databases such as BEFChina [20] (a joint Chinese-German-Swiss biodiversity research project) in EML [21] for...

---

[16] https://v1.datadryad.org/
[17] https://dublincore.org/
[18] https://www.pangaea.de/
[19] https://wiki.pangaea.de/wiki/Metadata
[20] http://china.befdata.biow.uni-leipzig.de/
[21] https://knb.ecoinformatics.org/tools/eml
mat, iDiv (The German Centre for Integrative Biodiversity Research Halle-Jena-Leipzig) and the Biodiversity Exploratories (a large, long-running functional biodiversity research project in Germany), both providing an own XML based metadata schema. Listing 2 presents an excerpt of a metadata file from BEFChina in EML format.

Annotation Guidelines: We considered only noun entities and adjectives as they mainly correspond to biodiversity terms. In case of compound nouns using the indications described below, we determined whether they need to stay together and are specifically relevant to biodiversity research such as “climate change”, “oxygen uptake” and “carbon cycling”, or whether they are too specific such as “benthic oxygen uptake”. In this case, not only the whole phrase was labeled but also the nested entities, e.g., benthic [ENVIRONMENT], oxygen uptake [PROCESS]. However, as oxygen itself is a chemical element, in addition, it was labeled with [MATERIAL]. In order to decide whether a term is domain-specific, we allowed the annotators to use various sources such as BioPortal and BiodiversityA-Z. In addition, annotators took reference from a result sheet of our previous research (Löffler et al., 2020), which contains around 600 biodiversity terms annotated from 9 biodiversity researchers. We left out abbreviations and units but permitted several annotations per entity type on the same phrase or term. For instance, “biomass” was annotated with MATERIAL and QUALITY as both meanings do apply. However, “water” was only annotated in its first, simple meaning MATERIAL. As “water” is also a habitat for species, we considered this case only if concrete habitats such as ocean, sea or river were mentioned in the datasets.

Listing 2: Excerpt from a BEFChina metadata file in EML metadata format (Germany and Erfmeier, 2019)

Annotation Process: Four authors of this work performed the manual labeling. All annotators are PhD students in computer science, and two of them have additional experience in Biodiversity Informatics. In addition, a PostDoc in biology gave advice with respect to the definitions of the entity types and also guided the annotators on the complex terms. GATE was used for the gold standard creation as it provides various functions for manual annotation tasks as well as support in merging different annotation sets into one final set.

The greatest challenge was to train the group properly to get a common understanding of the entity types. Hence, training took place in several phases over a period of two months:

- **Trial round:** At first, the overall goal and the entity types were presented to the annotators. They got familiar with GATE, the metadata files and their structure. Each annotator labeled 5 files individually and the results were discussed within the group. Afterwards, the initial annotation guidelines were refined.

- **Pilot phase:** In this phase, the annotators were grouped in teams of two. Four files were double-annotated per annotator team. Afterwards, the results again were discussed to finalize the annotation guidelines.

- **Main phase:** In the main annotation phase, each annotator pair received 25 metadata files. At first, the
files were annotated separately. Afterwards, they were swapped and labeled by the second annotator. The Inter-Annotator-Agreement measures are reported in our github repository. The F-score values are low for some files because, despite a thorough training, biological entities remain fuzzy and difficult to annotate. In case of disagreements, we collected the terms in a list and discussed them with a biodiversity expert who took the final annotation decision. Afterwards, the biodiversity expert’s decisions were incorporated into the files.

**Corpus Statistics:** In total, 5357 tokens were annotated. Table 2 presents the overall statistics of the QEMP corpus. It points out that PROCESS annotations are rare (4.6% of the annotated tokens). The other three entity type categories vary between 26.22% for ENVIRONMENT, 31.9% for MATERIAL to 37.1% for QUALITY. Figure 4 illustrates the distribution of the entity types over all tokens per repository. The picture confirms the diversity of the content. For instance, materials occur more often in BEFChina and PANGAEA than in the other repositories.

![Figure 4: Distribution of entity types over all tokens per data repository.](image)

**Table 2: QEMP corpus statistics:** the total number of annotated tokens per entity type and data source.

|          | Dryad | Panja | BEFChina | Biodiv. Explo. | IDiv | Total |
|----------|-------|-------|----------|----------------|------|-------|
| All Token| 3942  | 14069 | 47778    | 22817          | 1640 | 92857 |
| ENVIRONMENT| 74    | 1134  | 641      | 3041           | 166  | 1405 |
| MATERIAL | 17    | 283   | 1132     | 354            | 27   | 1715 |
| PROCESS  | 50    | 74    | 45       | 41             | 39   | 253  |
| QUALITY  | 24    | 207   | 1359     | 232            | 125  | 1896 |
| Total    | 210   | 694   | 3179     | 919            | 355  | 5387 |

**4.2. Results**

A screenshot from the results in GATE’s graphical interface is presented in Figure 4. The right panel contains the created annotations. The URI concepts are displayed in the bottom panel. We ran the pipeline on the QEMP corpus and compared the pipeline results with the manually created annotations.

**Measurements:** Precision, Recall and F-score are common metrics in Information Retrieval tasks (Manning et al., 2008). As introduced by (Maynard et al., 2006), they can be adapted for evaluations of ontology-based IE tasks. In this case, statistics are calculated in terms of ‘Correct’ (exact match), ‘Missing’ (no ontological concept found), ‘Spurious’ (additional match in the ontology but not labeled in the gold standard) and ‘Partial’ (ontological concepts cover named entities only partially) matches. Thus, Precision, Recall and F-Measure in ontology-based tasks can be defined as follows:

\[
\text{Precision} = \frac{\text{Correct} + \frac{1}{2}\text{Partial}}{\text{Correct} + \text{Spurious} + \text{Partial}}
\]

\[
\text{Recall} = \frac{\text{Correct} + \frac{1}{2}\text{Partial}}{\text{Correct} + \text{Missing} + \text{Partial}}
\]

\[
F\text{-Measure} = \frac{(\beta^2 + 1) P + R}{(\beta^2 P) + R}
\]

\(\beta\) denotes the weighting of Precision versus Recall. If Precision and Recall should be weighted equally, \(\beta\) is 1. In order to put more emphasis on the precision, \(\beta\) is set to 0.5. If it is set to 2, the recall is twice as weighted as the precision.

We also computed Macro and Micro measurement as global metrics to give an overview of the whole performance. Macro measurement denotes a single value by averaging the desired metric, giving all categories an equal weight. However, for unbalanced datasets micro measurement is preferred as it treats the corpus as a very large document (McCowan et al., 2004).

We wrote a Python script to calculate the above presented metrics and to process the corpus in a batch mode. A correct match was counted if an annotation is labeled with the correct entity type and possess at least one URI concept. In a partial match, the span of the ontology concept only partially covered the originally labeled term but the entity type was correctly assigned. If no URI concept could be found, it was counted as ‘Missing’. All concepts that were additionally labeled by the pipeline were considered as ‘Spurious’ in the metrics. The script is publicly available in our github repository.

**Outcomes:** The Precision denotes how many items the system detects are correct, whereas the Recall measures how many of the items that should have been returned were actually returned. In contrast to Information Retrieval tasks, where usually the Precision is considered to be the more important value, in annotation tasks, the Recall has at least the same impact. Table 3 presents the results over all documents per entity type. The values vary between 0.423 and 0.589 for the Precision and 0.38 and 0.74 for the Recall. The Precision values are not that high as numerous additional terms were annotated by the pipeline. In particular, for processes we received twice the number of spurious terms as correct matches. Furthermore, the Recall is low, too, as many process terms could not be matched to an entry in an ontology. Despite the fact that the ontological gazetteer for PROCESS already contain five ontologies, this picture reveals that this entity type is currently not well covered by the selected ontologies. In terms of the Recall value for environmental terms, the value is low as some...
Our results basically confirm our assumption. For biodiver-
sities, that are strictly tailored to a specific scope and that are
in ontology management. Hence, we only considered ter-
regular maintenance and interoperability play a crucial role
should be covered by several domain ontologies. However,
ready contains numerous biological terminologies, we as-
Entity Recognition tasks. As the LOD cloud nowadays al-
The aim of this research was to examine if biological on-
terminology of biodiversity research data.

4.3. Discussion
The aim of this research was to examine if biological ontologies are suitable for Information Extraction and Named Entity Recognition tasks. As the LOD cloud nowadays already contains numerous biological terminologies, we assumed that the domain knowledge of a certain research field should be covered by several domain ontologies. However, regular maintenance and interoperability play a crucial role in ontology management. Hence, we only considered terminologies that are maintained by large research communities, that are strictly tailored to a specific scope and that are interlinked among each other.

Our results basically confirm our assumption. For biodiver-
sity research, most entity types already reached good Recall values which denotes an overall good coverage of biodiver-
sity terms in ontologies. This also holds for the results across the different data repositories and projects. However, our approach is highly dependent on a certain ontology ver-
sion. If concepts and sub-concepts are moved, removed or renamed, all SELECT statements in the SPARQL queries need to be updated. Therefore, all used ontologies are cur-
rently cached and are provided locally in the pipeline.

Nevertheless, the Recall values leave room for improve-
ment. Currently, there are still a large amount of missing terms. In particular, for biological, chemical and physical processes the ontological coverage is low. Here, the bio-
diversity research community should put more effort into extending existing terminologies. Furthermore, for some terms we noticed wrong classifications which also lowered the Recall. The Precision values are also not that high as numerous additional terms and phrases were returned from the ontological gazetteers. In most cases, these additional annotations are too broad terms such as “position” or “content”. In the SPARQL statements, we already excluded spe-
cific nodes that contain too broad terms. Obviously, this needs further revision. For instance, machine learning ap-
proaches could be applied to remove unimportant terms and to gain higher precision values. However, that would re-
quire a much larger corpus of manually labeled datasets for training.

5. Conclusion
In this work, we introduced the BiodivTagger, a text mining pipeline that annotates main entities in metadata of biodiver-
sity research data, namely, environmental terms, bi-
ological, chemical and physical processes, materials and chemicals, phenotypic qualities and characteristics that can be measured. We evaluated the pipeline with a manually created gold standard, the QEMP corpus, which is the first annotated biological metadata corpus. Our results confirm our assumption that several domain ontologies from a valid source for the representation of domain knowledge can be used for IE tasks. However, a few ontological issues as well as pipeline issues remain. Numerous broad terms were an-
notated and for some environmental terms we discovered a
wrong classification. In particular, processes received very
less matches in ontologies. This also holds for the results
across the different data repositories and projects. However, our approach is highly dependent on a certain ontology ver-
sion. If concepts and sub-concepts are moved, removed or renamed, all SELECT statements in the SPARQL queries need to be updated. Therefore, all used ontologies are cur-
rently cached and are provided locally in the pipeline.

Nevertheless, the Recall values leave room for improve-
ment. Currently, there are still a large amount of missing terms. In particular, for biological, chemical and physical processes the ontological coverage is low. Here, the bio-
diversity research community should put more effort into extending existing terminologies. Furthermore, for some terms we noticed wrong classifications which also lowered the Recall. The Precision values are also not that high as numerous additional terms and phrases were returned from the ontological gazetteers. In most cases, these additional annotations are too broad terms such as “position” or “content”. In the SPARQL statements, we already excluded spe-
cific nodes that contain too broad terms. Obviously, this needs further revision. For instance, machine learning ap-
proaches could be applied to remove unimportant terms and to gain higher precision values. However, that would re-
quire a much larger corpus of manually labeled datasets for training.

5. Conclusion
In this work, we introduced the BiodivTagger, a text mining pipeline that annotates main entities in metadata of biodiver-
sity research data, namely, environmental terms, bi-
iological, chemical and physical processes, materials and chemicals, phenotypic qualities and characteristics that can be measured. We evaluated the pipeline with a manually created gold standard, the QEMP corpus, which is the first annotated biological metadata corpus. Our results confirm our assumption that several domain ontologies from a valid source for the representation of domain knowledge can be used for IE tasks. However, a few ontological issues as well as pipeline issues remain. Numerous broad terms were an-
notated and for some environmental terms we discovered a
wrong classification. In particular, processes received very
less matches in ontologies. Here, we need to further ana-
lyze whether the integration of additional ontologies would improve the results.

In a next step, we will add more datasets to the corpus to apply machine learning techniques that will support the re-
moval of the spurious annotations.

6. Acknowledgements
S. Babalou is supported by a scholarship of the German Academic Exchange Service (DAAD). P. Kaur is funded by the Deutsche Forschungsgemeinschaft (DFG) Priority Program 1374 “Infrastructure-Biodiversity-Exploratories” (KO 2209 / 12-2). N. Abdelmageed is supported by the Carl-Zeiss-Stiftung and F. Löffler is funded by the DFG within the scope of the GFBio project.
7. Bibliographical References

Ananiadou, S., Friedman, C., and Tsujii, J. (2004). Introduction: Named entity recognition in biomedicine. *Journal of Biomedical Informatics*, 37(6):393 – 395.

Balhoff, J. P., Dahdul, W. M., Kothari, C. R., Lapp, H., Lundberg, J. G., Mabée, P., Midford, P. E., Westerfield, M., and Vision, T. J. (2010). Phenex: Ontological Annotation Of Phenotypic Diversity. *PLOS ONE*, 5(5):1–10.

Bossy, R., Jourde, J., Manine, A.-P., Veber, P., Alphonse, E., Van De Guchte, M., Bessières, P., and Nédéllec, C. (2012). BioNLP Shared Task - The Bacteria Track. In *BMC Bioinformatics*, volume 13, page S3.

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., and the ENVO Consortium. (2013). The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, 4(1):43, December.

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.

Deléger, L., Bossy, R., Chaix, E., Ba, M., Ferre, A., Bessières, 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.

Eliaison, C. M., Edwards, S. V., and Clarke, J. A. (2019). phenootools: An R package for visualizing and analyzing phenomic datasets. *Methods in Ecology and Evolution*.

Germany, M. and Erfmeier, A. (2019). Main experiment: Seedling addition experiment - growth and biomass data. (accessed through URL: https://china.befdata.biol.uni-leipzig.de/datasets/57).

Gurulingappa, H., Klinger, R., Hofmann-Aptius, 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.

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.

Jonquet, C., Shah, N. H., and Musen, M. A. (2009). The Open Biomedical Annotator. In *Summit on Translational Bioinformatics*, 2009:56–60, March.

Jovanović, J. and Bagheri, E. (2017). Semantic annotation in biomedicine: the current landscape. *Journal of Biomedical Semantics*, 8(1):44, September.

Löffler, F., Opasjumruskit, K., Karam, N., Fichtmüller, D., Schindler, U., Klan, F., Müller-Birn, C., and Diepenbroek, M. (2017). Honey Bee VersusApis Mellifera: A Semantic Search For Biological Data. In Eva Blomqvist, et al., editors, *The Semantic Web: ESWC 2017 Satellite Events: Portorož, Slovenia*, pages 98–103. Springer International Publishing.

Löffler, F., Wesp, V., König-Ries, B., and Klan, F. (2020). Dataset Search In Biodiversity Research: Do Meta-data In Data Repositories Reflect Scholarly Information Needs? https://arxiv.org/abs/2002.12021

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

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, 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, IDIAP.

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, 08.

Nguyen, N. T., Gabud, R. S., and Ananiadou, S. (2019). COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity Data Journal*, (7).

Pfaff, C.-T., Eichenberg, D., Liebergesell, M., König-Ries, B., and Wirth, C. (2017). Essential Annotation Schema For Ecology (EASE) - A framework supporting the efficient data annotation and faceted navigation in ecology. *PLOS ONE*, 12(10):1–13, 10.

Shadbolt, N., Berners-Lee, T., and Hall, W. (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, Jan.

Sirirjodha Khalsa, Peter Cotroneo, M. W. (2018). A survey of current practices in data search services. Technical report, Research Data Alliance Data (RDA) Discovery Paradigms Interest Group.

Thessen, A. E., Cui, H., and Mozherin, D. (2012). Applications of natural language processing in biodiversity science. *Advances in Bioinformatics*, 2012(Article ID 391574):17 pages.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blommers, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crossas, 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 Hoorn, 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 Leij, 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).

Wood, M. M., Lydon, S. J., Tablan, V., Maynard, D., and Cunningham, H. (2004). Populating A Database From Parallel Texts Using Ontology-Based Information Extraction. In Farid Meziane et al., editors, *Natural Language Processing and Information Systems*, pages 254–264, Berlin, Heidelberg. Springer Berlin Heidelberg.