WikiMed-DE: Constructing a Silver-Standard Dataset for German Biomedical Entity Linking using Wikipedia and Wikidata

Yi Wang[1], Corina Dima[1] and Steffen Staab[1,2]

[1]University of Stuttgart, Germany
[2]University of Southampton, UK

Abstract
This paper introduces WikiMed-DE, a silver-standard, automatically annotated biomedical entity linking dataset for the German language. WikiMed-DE encompasses a substantial collection of 53,981 articles from the German Wikipedia annotated with 1,951,081 mentions corresponding to 317,010 unique mention URLs. The hyperlinks of Wikipedia articles are used to connect concept mentions to Wikidata and transitively to three biomedical concept IDs: the Concept Unique Identifier from the Unified Medical Language System, the MeSH ID from Medical Subject Headings hierarchy, and the DOID from the Disease Ontology. A curated subset, WikiMed-DE-BEL, is released as a ready-to-use benchmark for biomedical entity linking in German. It features the same number of articles as WikiMed-DE, but only the highest-quality information is retained: 413,913 mentions corresponding to 35,012 unique concepts. Both resources are available at: https://doi.org/10.5281/zenodo.8188966.

1. Introduction

Biomedical entity linking (BEL) is an important task for automatically processing text from the medical domain. It enables the disambiguation of entities in text to unique identifiers in ontologies like the Unified Medical Language System (UMLS)[1] [1]. In the example *MRI technology has revolutionized medical imaging*, the term *MRI* could stand for `Magnetic Resonance Imaging (C0024485)`, `Multidrug Resistance Induction (C1513738)`, or `Most Recent Inpatient (C1546460)`. The goal of biomedical entity linking is to identify that the text span *MRI* in the example should be mapped to the concept `Magnetic Resonance Imaging`, which has the identifier `C0024485` in the UMLS. In this paper we refer to text span *MRI* as a *mention* of the concept `Magnetic Resonance Imaging`, and to the sentence in which this text span appears as the *context* of the mention.

Many datasets have been created in order to foster the development of reliable BEL systems, e.g. the NCBI Disease dataset [2], which annotates mentions of diseases in PubMed[2] abstracts; MedMentions [3], a collection of PubMed abstracts annotated with concepts from the UMLS;

Wikidata'23: Wikidata workshop at ISWC 2023
yi.wang@ipvs.uni-stuttgart.de (Y. Wang); corina.dima@ipvs.uni-stuttgart.de (C. Dima); steffen.staab@ipvs.uni-stuttgart.de; s.r.staab@soton.ac.uk (S. Staab)

[1] UMLS: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html
[2] PubMed: https://pubmed.ncbi.nlm.nih.gov/

BC5CDR [4], a dataset that focuses on extracting and linking chemical compounds and diseases from PubMed articles, where every entity was manually annotated by a team of Medical Subject Headings (MeSH) [3] indexers; COMETA [5], a dataset consisting of 20,000 biomedical entity mentions from Reddit, annotated by experts with links to SNOMED CT[4] or RegEl [6], which maps the manual annotation of regulatory DNA elements within PubMed abstracts to various ontologies - e.g. the tissue entities are mapped to Brenda Tissue Ontology (BTO) [7], while diseases are mapped to MONDO [8].

These datasets were manually annotated by domain experts and focus on entities of interest from various domains - e.g. diseases, genes, tissues, chemical compounds, etc. And while many of these datasets are not very large, they do provide invaluable information for training machine learning models for the automatic disambiguation of biomedical entities.

However, the vast majority of biomedical datasets provide annotations for English texts. For other languages, like German, datasets annotated with biomedical concepts are extremely scarce due to the resource-intensive nature of the manual creation process, which requires trained professionals to perform the annotation process.

This paper addresses this issue by introducing **WikiMed-DE**, a silver-standard dataset for biomedical entity linking for the German language. The automatic annotation process makes use of the links connecting the text of the German Wikipedia[5] pages with the structured information available in the Wikidata [9] knowledge base and in three knowledge sources from the biomedical domain: the Unified Medical Language System (UMLS) [1], the Medical Subject Headings (MeSH) hierarchy [10] and the Disease Ontology (DO) [11].

Our contributions are the following:

1. We build upon and extend the procedure introduced by Vashishth et al. [12] for creating the English WikiMed dataset and construct a German dataset for biomedical entity linking called WikiMed-DE, which we make publicly available[6]; WikiMed-DE is annotated with a wide range of concepts from the UMLS; a subset of this dataset, WikiMed-DE-BEL, can be readily used for training broad-coverage biomedical entity linking systems for German.
2. We provide an extensive description of the steps and resources required to create the dataset, as well as a public code repository[7]; this makes it straightforward to apply the procedure for creating similar datasets for other languages, or for updating the dataset once new information is available in Wikipedia and Wikidata.

The remainder of this article is organized as follows: Section 2 provides an overview of the related work, Section 3 introduces the knowledge sources used to create WikiMed-DE, Section 4 describes the methodology used to construct WikiMed-DE, Section 5 presents the statistics of the dataset and asseses its quality, Section 6 discusses the limitations of WikiMed-DE and concludes the paper.

[3] MeSH: https://www.ncbi.nlm.nih.gov/mesh/

[4] SNOMED-CT: https://www.snomed.org/

[5] German Wikipedia: https://de.wikipedia.org/wiki/Wikipedia:Hauptseite

[6] WikiMed-DE dataset: https://doi.org/10.5281/zenodo.8188966

[7] WikiMed-DE code repo:https://github.com/AI4MedCode/wikimed-de

2. Related Work

Datasets in the BEL domain typically focus on annotating biomedical entities from existing ontologies. A point in case is the BC6BioID [13] dataset published for the BioCreative challenges, which focuses on identifying genes or chemicals in English text and maps them to ontologies like the Gene ontology[8]. It consists of 17,883 documents and 133,033 mentions referring to 7,652 unique concepts.

Medical Concept Normalization (MCN) [14] is a dataset that focuses primarily on annotating entities of clinical utility, such as disorders, problems, tests, and treatments in discharge summaries written in English. These entities are mapped to widely adopted medical terminologies, such as the UMLS and the International Classification of Diseases (ICD)[9]. MCN consists of 100 discharge summaries and provides normalization for a total of 10,919 concept mentions corresponding to 3,792 unique concepts.

The NCBI disease corpus [2] provides annotations of disease mentions along with their corresponding concepts, which are represented using either MeSH or Online Mendelian Inheritance in Man (OMIM) [10] identifiers. The corpus contains 6,892 disease mentions mapped to 790 unique concepts for a collection of 793 PubMed abstracts written in English.

MedMentions [3] is a biomedical dataset containing 4,392 PubMed abstracts annotated with 203,282 mentions. Mentions are linked to UMLS concepts as well as to UMLS semantic types.

Vashishth et al. [12] introduce the WikiMed and PubMedDS datasets to facilitate research in biomedical natural language processing. In constructing the WikiMed dataset, they select the English Wikipedia as a comprehensive source of articles, while Wikidata[11] [9] and Freebase [15] are used to establish mappings between the Wikipedia articles and UMLS concepts.

The BRONCO [16] dataset is a valuable German-language resource for BEL and healthcare research. It consists of 200 manually de-identified discharge summaries of cancer patients. The discharge summaries were meticulously annotated with various medical terminologies, including diagnoses, treatments, and medications, and further mapped to the German Modification of the International Classification of Diseases (ICD-10-GM) [12]. Because of its limited size, the BRONCO dataset is mainly used for the evaluation of biomedical named entity recognition/entity linking models (e.g. [17]).

In an effort to support the biomedical entity linking in languages other than English, Liu et al. [18] propose a cross-lingual biomedical entity linking evaluation benchmark, XL-BEL, for evaluating BEL in 10 typologically diverse languages, including German. However, while they make use of the WikiMed process introduced by Vashishth et al. [12] for creating this resource, the benchmark contains only 1000 annotated sentences for each language. While this amount of annotations is adequate for evaluation purposes, it does not suffice for training a good quality biomedical entity linker.

[8] GeneOntology: http://geneontology.org/

[9] ICD-10-CM: https://www.cdc.gov/nchs/icd/icd-10-cm.htm

[10] OMIM: https://www.omim.org/

[11] Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page

[12] ICD-10-GM: www.dimdi.de/dynamic/de/klassifikationen/icd/icd-10-gm/

3. Structured Knowledge Repositories

3.1. Wikidata

Wikidata [9] is a collaborative knowledge base that serves as a data source for numerous projects in the Wikimedia sphere [19], such as Wikipedia. The main objective of Wikidata is to ensure consistent and high-quality data across the multiple language versions of Wikipedia. At the moment of writing Wikidata contains more than 100M items[13]. Wikidata has garnered significant attention from researchers across diverse fields of study [20], for example, in the biomedical field, Mitraka et al. [21] used Wikidata as a knowledge base to collect biomedical concepts such as NCBI Gene and map them to Wikipedia articles.

3.2. Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) [1] is a repository of biomedical vocabularies developed by the US National Library of Medicine. The main component of UMLS is the Metathesaurus, which integrates more than 100 vocabularies from different subdomains, some in several languages. These include the NCBI taxonomy [22], the MeSH hierarchy [10] in multiple languages, the Gene Ontology, the International Classification of Diseases (ICD) 9 and 10 in multiple languages, DrugBank, the Logical Observation Identifiers Names and Codes (LOINC) in several languages, the Medical Dictionary for Regilatory Activities (MeDRA) and the SNOMED-CT terminology, to name a few. When a concept is added to the Metathesaurus, it receives a unique identifier entitled the Concept Unique Identifier (CUI). The CUI is used to connect all the concepts from different source vocabularies that refer to the same meaning. For example, the entry *Carcinoma of breast* from the SNOMEDCT_US [14] terminology and the entry *Carcinomas, Breast* from the MeSH vocabulary are associated with the same UMLS CUI, `C0678222`. The UMLS Metathesaurus is released twice a year. The current release, 2023AA[15], contains ~3.31 million concepts and 15.7 million unique concept names from 185 source vocabularies.

3.3. Medical Subject Headings (MeSH)

Medical Subject Headings (MeSH) [10] is a comprehensive controlled vocabulary used for indexing, cataloging, and searching for biomedical and health-related information and documents. MeSH consists of terms or descriptors representing various aspects of biomedical concepts, such as diseases, anatomy, drugs, and medical procedures. Each MeSH term is assigned a unique identifier called a MeSH ID or MeSH Heading (MH). Here is an example of an entry from MeSH listing a term, its MeSH ID and its description:

- `MeSH Term`: *Hypertension*
- `MeSH ID: D005260`
- `Description`: *A condition characterized by elevated blood pressure persistently exceeding 140 mm Hg systolic or 90 mm Hg diastolic.*

[13] Wikidata statistics: https://www.wikidata.org/wiki/Wikidata:Statistics

[14] SNOMED-CT, US edition: https://www.nlm.nih.gov/healthit/snomedct/us_edition.html

[15] UMLS release: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/notes.html

3.4. Disease Ontology (DO)

The Disease Ontology (DO) [11] is a publicly accessible ontological representation of human diseases, designed to establish unambiguous disease definitions based on etiological classifications. Its primary objective is to ensure standardized utilization and incorporation of disease information in biomedical data annotation. The latest version of the Disease Ontology was released in June 2023 and contains 11,349 disease terms [16]. Each disease in the Disease Ontology is assigned a unique identifier called a Disease Ontology ID (DOID). For example, `DOID:162` indicates the disease *Breast Cancer*. The Disease Ontology undergoes biannual updates to its vocabulary mappings by extracting CUIs from the UMLS `MRCONSO.RRF` file. In the current release 7075 of the DO terms are mapped to corresponding UMLS CUIs.

4. Constructing the WikiMed-DE Dataset

The WikiMed-DE dataset consists of a selection of articles from the German Wikipedia. The core of the construction process is the fact that each Wikipedia article has a unique Wikipedia page ID and that the majority of the Wikipedia page IDs have a corresponding Wikidata ID, the *QID*. The QID is a unique identifier of a data item in Wikidata, consisting of the letter `Q` followed by a sequence of digits. The QID connects Wikipedia pages written in various languages to the language-independent Wikidata items and transitively to all the structured knowledge already linked to each Wikidata item.

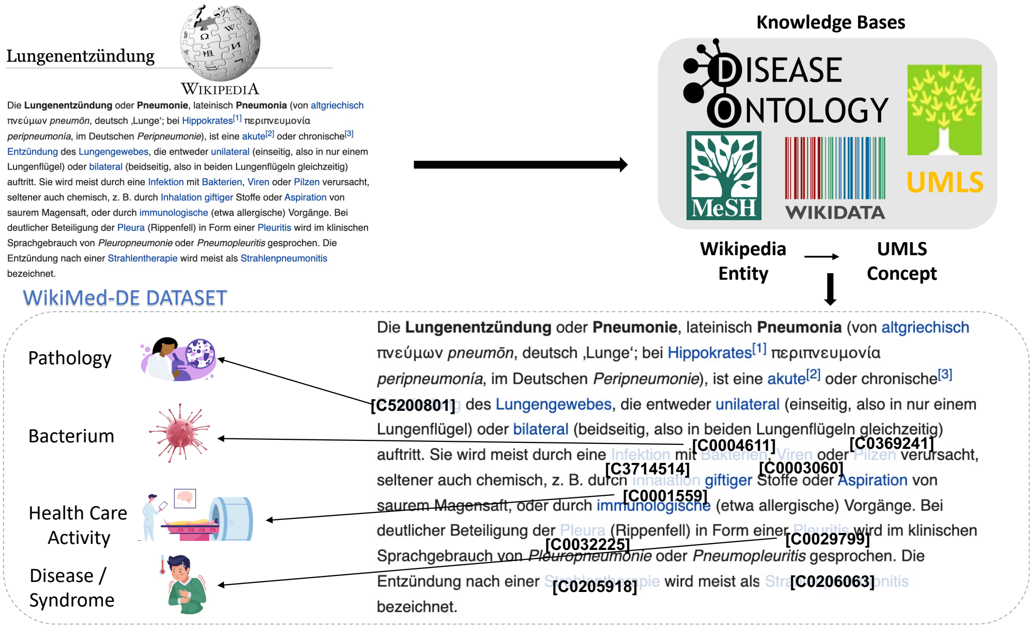
In the case of WikiMed-DE the QID is used to map each German Wikipedia article to three types of biomedical concept IDs: the *Concept Unique Identifier (CUI)* from the UMLS, the *MeSH ID* from the MeSH hierarchy, and the *Disease Ontology ID (DOID)* from the Disease Ontology.

The UMLS provides extensive coverage for biomedical concepts and encompasses several medical terminologies from different biomedical vocabularies. Annotating mentions with UMLS CUIs ensures that WikiMed-DE covers a diverse range of medical concepts. The MeSH hierarchy is used for the semantic indexing of PubMed. Because the MeSH hierarchy is integrated into the UMLS, each MeSH ID can be mapped to UMLS CUIs, thereby increasing the number of mentions in WikiMed-DE. The DOID is widely employed in biomedical research to annotate and analyze disease-related data. Similar to the MeSH ID, DOIDs can also be mapped to UMLS CUIs, thus enabling the integration of more disease-specific information into WikiMed-DE. By incorporating mentions linked to the UMLS, MeSH, and Disease Ontology WikiMed-DE gains access to an extensive array of interconnected biomedical concepts, resulting in a rich and diverse collection of biomedical-specific information.

Following WikiMed [12], the WikiMed-DE mentions annotated with UMLS CUIs are further mapped to UMLS semantic types. The UMLS semantic types serve as broad categories or classes that group medical concepts within UMLS. Each semantic type is identified by a unique identifier called the Type Unique Identifier (TUI), which is composed of the letter T followed by three digits. Semantic types in the UMLS encompass various categories such as *sign or symptom* (T184), *cell component* (T026), *immunologic factor* (T129), and others. There are a total of 127 semantic types connected via 54 relations in the Semantic Network of the UMLS.

[16] Disease Ontology release: https://github.com/DiseaseOntology/HumanDiseaseOntology/releases

Figure 1 illustrates the process of constructing the WikiMed-DE dataset: Wikidata is used to map German Wikipedia articles to biomedical concepts such as the UMLS CUI. Both the articles and the mentions (i.e. the links) in the articles are mapped to biomedical concepts.



Lungenentzündung

Die **Lungenentzündung** oder **Pneumonie**, lateinisch **Pneumonia** (von altgriechisch πνεύμων *pneumōn*, deutsch ‚Lunge'; bei Hippokrates[1] περιπνευμονία *peripneumonía*, im Deutschen *Peripneumonie*), ist eine akute[2] oder chronische[3] Entzündung des Lungengewebes, die entweder unilateral (einseitig, also in nur einem Lungenflügel) oder bilateral (beidseitig, also in beiden Lungenflügeln gleichzeitig) auftritt. Sie wird meist durch eine Infektion mit Bakterien, Viren oder Pilzen verursacht, seltener auch chemisch, z. B. durch Inhalation giftiger Stoffe oder Aspiration von saurem Magensaft, oder durch immunologische (etwa allergische) Vorgänge. Bei deutlicher Beteiligung der Pleura (Rippenfell) in Form einer Pleuritis wird im klinischen Sprachgebrauch von *Pleuropneumonie* oder *Pneumopleuritis* gesprochen. Die Entzündung nach einer Strahlentherapie wird meist als Strahlenpneumonitis bezeichnet.

Figure 1: Constructing WikiMed-DE: links in a German Wikipedia article are mapped to corresponding CUIs using Wikidata information. Figure layout inspired by Fig. 3 from Vashishth et al. [12].

4.1. Obtaining German Wikipedia Articles

The WikiMed-DE dataset is based on a recent database dump of the German Wikipedia from 20.06.2023[17]. `WikiExtractor`[18] was used to extract the article title, page id, URL and text from the archive *dewiki-20230620-pages-articles-multistream.xml.bz2* and save them in JSON format. The text of each article contains HTML-encoded hyperlink tags to create clickable links, which are retained for the annotation step. The WikiExtractor outputs roughly ~10,000 files from this archive, each containing on average 450 Wikipedia articles. These files are combined in a post-processing step into a single JSON file containing 4,579,135 Wikipedia articles.

4.2. Mapping Wikipedia Articles to Wikidata

We map the Wikipedia articles and their mentions to Wikidata QIDs using another file from the German Wikipedia database dump, namely *dewiki-20230620-page_props.sql.gz*. This file relates a page ID from the German Wikipedia to the corresponding Wikidata QID. For instance, the

[17] German Wikipedia dump, 20.06.2023: https://dumps.wikimedia.org/dewiki/20230620/

[18] `WikiExtractor`: https://github.com/attardi/wikiextractor

first entry `(1, 'wikibase_item', 'Q734916', NULL)` in this file indicates that the Wikidata QID associated with German Wikipedia page ID `1` is `Q734916`.

The result of this mapping step is a CSV file containing the QID and the Wikipedia page ID for each entry. Among the 4,579,135 Wikipedia articles from the previous step, a total of 1,754,551 page IDs lack a corresponding QID. This is due to the fact that not all the page IDs in the archive *dewiki-20230620-pages-articles-multistream.xml.bz2* appear in the archive *dewiki-20230620-page_props.sql.gz*. In the next steps we focus on the 2,824,584 articles that have a corresponding QID.

4.3. Mapping Wikidata QIDs to Biomedical Concept IDs

The official Wikidata SPARQL endpoint[19] was used to generate a mapping from QIDs to biomedical concept IDs. Three properties from Wikidata were targeted:

- `P2892:UMLS CUI`[20], which maps a Wikidata item to its UMLS CUI, if one is available
- `P486:MeSH descriptor ID`[21], which maps a Wikidata item to the Medical Subject Headings identifier, if it has one and
- `P699:Disease Ontology ID`[22], which connects a Wikidata item to its ID in the Disease Ontology, if such a mapping exists.

All three properties feature a *single-value constraint*[23] in Wikidata, which states that this property generally contains a single value per item. However, as we will show in Section 4.6, these constraints are not enforced in Wikidata, making it possible for a Wikidata item to have, for example, multiple CUIs associated to it. We obtained a mapping from QIDs to 763,859 UMLS CUIs, 38,607 MeSH IDs and 10,609 DOIDs.

4.4. Filtering Wikipedia Articles

WikiMed-DE is meant to serve as a training material for BEL models. It is therefore important to filter the German Wikipedia articles and retain only those related to the biomedical domain. The articles are filtered based on the mapping of QIDs to the three biomedical concepts of interest described in Section 4.3. We retain only those articles where the QID is associated with at least one of these three biomedical IDs, resulting in 54,514 articles. However, in some cases, an article will have a title, a QID and a valid mapping, but no text - such articles are also filtered out. At the end of the filtering step there are 53,981 German Wikipedia articles left.

4.5. Mapping Mentions to Wikidata

In order to fulfill its goal as a training resource for BEL, the hyperlinked words or phrases in the WikiMed-DE articles need to be identified and annotated with biomedical concepts. These

[19] Official Wikidata SPARQL endpoint: https://query.wikidata.org/

[20] `P2892:UMLS CUI`: https://www.wikidata.org/wiki/Property:P2892

[21] `P486:MeSH descriptor ID`: https://www.wikidata.org/wiki/Property:P486

[22] `P699:Disease Ontology ID`: https://www.wikidata.org/wiki/Property:P699

[23] Wikidata item for *single-value constraint*: https://www.wikidata.org/wiki/Q19474404

are extracted using regular expressions, by identifying HTML-encoded tags in the article text and decoding them to generate URLs and clean text.

In this step we store, for each article, it's title, text, URL, and a list of mentions corresponding to the hyperlinked words or phrases in the text. For each mention we record the text of the mention and the corresponding URL. We then save a list containing 317,010 unique mention URLs, which we need to map to Wikipedia page IDs. However, the URLs that are originally associated with the mentions are different from the URLs needed to obtain the page ID for each mention. For example, for the page with the title *Getreide* we need to map from the original mention URL[24] to the corresponding Wikipedia page information URL[25].

To obtain this mapping we do a request for each Wikipedia page information URL and retrieve the corresponding Wikipedia page ID. 259,250 mention URLs are successfully matched to a corresponding page ID in this way. The rest of the mention URLs (57,760 links) yield no page information because some links do not exist in the German Wikipedia. For example, in Figure 2, the URL corresponding to the red mention *Aktin-bindenden Proteinen* lacks a corresponding page ID because it is a link to a page that does not yet exist in the German Wikipedia. Such cases are relatively frequent, as Wikipedia editors routinely add links to pages that will be created only in a subsequent step.

The next step is to map the page IDs corresponding to each mention to QIDs. We use the previously generated CSV file (from Section 4.2), which contains a mapping from Wikipedia page IDs to QIDs. Leveraging this resource 206,549 page IDs out of the 259,250 are uniquely mapped to a QID. The mapping is, however, incomplete, with 52,701 page IDs lacking corresponding QIDs. We examined a small sample of these URLs and discovered two issues: (i) some hyperlinks point to sections within the same article, and therefore cannot be mapped to a separate QID and (ii) some of the URLs extracted from the hyperlink tags are redirects. For example, the page *Hydrophil*[26] redirects to *Hydrophilie*[27]. The page ID information is not typically stored on the redirect page, but only on the target URL.

After the two previous steps 110,461 mention URLs are still not successfully mapped to their corresponding QIDs: 57,760 mention URLs lack a page ID and 52,701 mention URLs have a page ID but no QID. To address the redirection problem, we used the `wikipedia`[28] Python package to obtain a page ID for the URLs that were still missing a QID. This package allows one to look for a Wikipedia page given the title of the page and the language code of the wiki (`'de'` in our case). By specifying the flag `redirect=True` one can also find the page IDs for redirects. We applied `wikipedia`'s `page` function to these 110,461 mention URLs and obtained an extra 46,541 correct mappings to QIDs.

4.6. Integrating the Mappings to Biomedical Concept IDs

The methodology described in Section 4.3 is used to map the QIDs to three biomedical concepts: the UMLS CUI, the MeSH ID, and the DOID. When a particular Wikidata item contains state-

[24] Mention URL for *Getreide*: https://de.wikipedia.org/wiki/Getreide

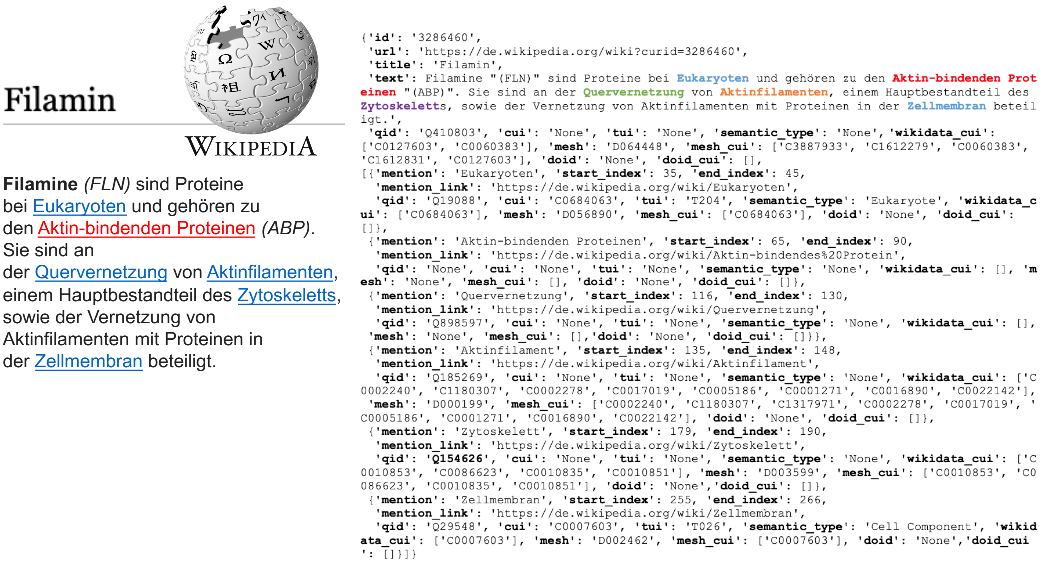
[25] Wikipedia page info URL for *Getreide*: https://de.wikipedia.org/w/index.php?title=Getreide&action=info

[26] *Hydrophil*: https://de.wikipedia.org/wiki/Hydrophil

[27] *Hydrophilie*: https://de.wikipedia.org/wiki/Hydrophilie

[28] `wikipedia` Python package: https://pypi.org/project/wikipedia/

ments about biomedical IDs, it will, most frequently, contain a statement about the associated UMLS CUI of that item. However, in some cases, the Wikidata items contain statements about their MeSH IDs or their DOIDs, but does not include statements recording their UMLS CUIs. Consequently, to increase the number of linked biomedical concepts, we extract UMLS CUIs, MeSH IDs and DOIDs for each article. In the WikiMed-DE dataset, we use the tags `wikidata_cui`, `mesh` and `doid` to label this information. Figure 2 shows a sample of WikiMed-DE: an article together with its meta-information and its annotations.



Filamine *(FLN)* sind Proteine bei Eukaryoten und gehören zu den Aktin-bindenden Proteinen *(ABP)*. Sie sind an der Quervernetzung von Aktinfilamenten, einem Hauptbestandteil des Zytoskeletts, sowie der Vernetzung von Aktinfilamenten mit Proteinen in der Zellmembran beteiligt.

```
{'id': '3286460',
 'url': 'https://de.wikipedia.org/wiki?curid=3286460',
 'title': 'Filamin',
 'text': Filamine "(FLN)" sind Proteine bei Eukaryoten und gehören zu den Aktin-bindenden Prot
einen "(ABP)". Sie sind an der Quervernetzung von Aktinfilamenten, einem Hauptbestandteil des
Zytoskeletts, sowie der Vernetzung von Aktinfilamenten mit Proteinen in der Zellmembran beteil
igt.',
 'qid': 'Q410803', 'cui': 'None', 'tui': 'None', 'semantic_type': 'None','wikidata_cui':
['C0127603', 'C0060383'], 'mesh': 'D064448', 'mesh_cui': ['C3887933', 'C1612279', 'C0060383',
'C1612831', 'C0127603'], 'doid': 'None', 'doid_cui': [],
[{'mention': 'Eukaryoten', 'start_index': 35, 'end_index': 45,
  'mention_link': 'https://de.wikipedia.org/wiki/Eukaryoten',
  'qid': 'Q19088', 'cui': 'C0684063', 'tui': 'T204', 'semantic_type': 'Eukaryote', 'wikidata_c
ui': ['C0684063'], 'mesh': 'D056890', 'mesh_cui': ['C0684063'], 'doid': 'None', 'doid_cui':
[]},
 {'mention': 'Aktin-bindenden Proteinen', 'start_index': 65, 'end_index': 90,
  'mention_link': 'https://de.wikipedia.org/wiki/Aktin-bindendes%20Protein',
  'qid': 'None', 'cui': 'None', 'tui': 'None', 'semantic_type': 'None', 'wikidata_cui': [], 'm
esh': 'None', 'mesh_cui': [], 'doid': 'None', 'doid_cui': []},
 {'mention': 'Quervernetzung', 'start_index': 116, 'end_index': 130,
  'mention_link': 'https://de.wikipedia.org/wiki/Quervernetzung',
  'qid': 'Q898597', 'cui': 'None', 'tui': 'None', 'semantic_type': 'None', 'wikidata_cui': [],
 'mesh': 'None', 'mesh_cui': [],'doid': 'None', 'doid_cui': []}},
 {'mention': 'Aktinfilament', 'start_index': 135, 'end_index': 148,
  'mention_link': 'https://de.wikipedia.org/wiki/Aktinfilament',
  'qid': 'Q185269', 'cui': 'None', 'tui': 'None', 'semantic_type': 'None', 'wikidata_cui': ['C
0002240', 'C1180307', 'C0002278', 'C0017019', 'C0005186', 'C0001271', 'C0016890', 'C0022142'],
 'mesh': 'D000199', 'mesh_cui': ['C0002240', 'C1180307', 'C1317971', 'C0002278', 'C0017019', '
C0005186', 'C0001271', 'C0016890', 'C0022142'], 'doid': 'None', 'doid_cui': []},
 {'mention': 'Zytoskelett', 'start_index': 179, 'end_index': 190,
  'mention_link': 'https://de.wikipedia.org/wiki/Zytoskelett',
  'qid': 'Q154626', 'cui': 'None', 'tui': 'None', 'semantic_type': 'None', 'wikidata_cui': ['C
0010853', 'C0086623', 'C0010835', 'C0010851'], 'mesh': 'D003599', 'mesh_cui': ['C0010853', 'C0
086623', 'C0010835', 'C0010851'], 'doid': 'None','doid_cui': []},
 {'mention': 'Zellmembran', 'start_index': 255, 'end_index': 266,
  'mention_link': 'https://de.wikipedia.org/wiki/Zellmembran',
  'qid': 'Q29548', 'cui': 'C0007603', 'tui': 'T026', 'semantic_type': 'Cell Component', 'wikid
ata_cui': ['C0007603'], 'mesh': 'D002462', 'mesh_cui': ['C0007603'], 'doid': 'None','doid_cui
': []}]}
```

Figure 2: A sample of WikiMed-DE. In this snippet, there are five kinds of mentions. Mentions in blue have a valid URL, a QID, and a unique CUI. The mention in red links to a German Wikipedia article that does not yet exist, so it has no QID. The mention in green has a valid URL and QID, but no CUI. The mention in orange has multiple CUIs. For the mention in lila, the HTML tags do not coincide with a full token of the underlying text - *Zytoskelett* vs *Zytoskeletts*.

The MeSH hierarchy is integrated into the UMLS. Therefore we can use the file `MRCONSO.RRF` from the UMLS release to map the MeSH IDs to UMLS CUIs, resulting in a mapping that is not always unique. In WikiMed-DE, the CUIs mapped from MeSH IDs are saved under the tag `mesh_cui`. A similar mapping is also performed for the DOIDs, using the file `doid.json` from Disease Ontology's current release. In WikiMed-DE the CUIs mapped from the DOIDs are saved as `doid_cui`.

Finally, we consolidate the CUI information saved under the tags `wikidata_cui`, `mesh_cui` and `doid_cui`. If, under all these tags, there is only a single CUI for a mention or an article, this unique CUI is saved under the tag `cui` in WikiMed-DE. This unique CUI is further mapped to one or more TUIs using the file `MRSTY.RRF` from the UMLS release. The list of TUIs and the corresponding semantic type labels are saved under `tui` and `semantic_type`, respectively.

A part of the mentions will not have a single CUI, but several possible CUIs. This can have multiple reasons: either the Wikidata item already maps to multiple CUIs, despite the fact that the property P2892 has a single-value constraint; or the mapping from MeSH ID to CUIs resulted in several CUIs; or the consolidation step lead to a list of CUIs rather than a single CUI. In any case, we have no automatic method to choose a single correct CUI among the given ones, so we will typically just record all this information.

4.7. Combining All the Information to Create WikiMed-DE

To obtain the final version of WikiMed-DE, we reprocess each of the articles and add the information we extracted in the previous steps for each article. The text within each Wikipedia page is decoded, removing any HTML tags and producing clean text and mention URLs. The start and end indices for each mention are recorded, thus enabling precise identification of the mention's position. In some cases, the extracted start and end positions will not overlap with natural token boundaries (see Figure 2 for an example). Each mention is associated with its corresponding QID and all the extracted biomedical information.

At the end of the extraction process, WikiMed-DE consists of a list of German Wikipedia articles. For each article we save the article's title, text, URL, QID, the biomedical indices (CUI, TUIs, semantic type labels, Wikidata CUI, MeSH ID, MeSH-derived CUI, DOID and DOID-derived CUI) and mention list, where the mentions correspond to the hyperlinked words or phrases in the text. For each mention we record the start and end indices, URL, QID and the biomedical indices (same as for the article).

5. WikiMed-DE Dataset Statistics

WikiMed-DE is a dataset consisting of 53,981 German Wikipedia articles, each containing multiple mentions. The filtering step described in Section 4.4 ensures that all the articles in the dataset describe biomedical concepts mentioned either in the UMLS, in the MeSH hierarchy or in the Disease Ontology. The WikiMed-DE articles and the mentions therein are mapped to QIDs and to biomedical concept IDs. We can therefore analyze the dataset at two levels: at the article level and at the mention level.

5.1. Statistics at the Mention Level

WikiMed-DE contains a total of 1,951,081 mentions corresponding to 317,010 unique mention URLs. As shown in Table 1, 95.79% of these mentions have an assigned QID, with the rest having either missing links or missing QID information. CUI information was assigned for 29.59% of the mentions. Note that both directly through Wikidata and through the MeSH ID we obtained a larger amount of assigned CUIs (38.30% and 35.46%, respectively). However, part of these mentions have multiple CUIs assigned, and are therefore ambiguous from the point of view of biomedical entity linking. The disambiguation step is non-trivial and cannot be done without the help of experts, so we decided to just keep all the information in the dataset. Disease Ontology information is available only for a small percentage of the mentions.

From 317,010 unique mention URLs, 79.82% have an assigned QID and 14.94% have an assigned CUI. WikiMed-DE contains therefore links to 47,380 unique biomedical concepts. The number

Table 1
The distribution of mentions and unique entities one biomedical concepts

	qid	cui	wikidata_cui	mesh	mesh_cui	doid	doid_cui
Number of Mentions	1,867,554	577,547	746,022	691,383	691,084	44,557	41,595
Percentage	95.79%	29.59%	38.30%	35.46%	35.44%	2.28%	2.13%
Number of Unique Mention URLs	253,090	47,380	57,751	31,901	31,866	4,039	3,534
Percentage	79.82%	14.94%	18.21%	10.06%	10.05%	1.27%	1.11%
Total Mentions				1,951,081			
Total Unique Mention URLs				317,010			

of unique CUIs is much smaller than the number of QIDs. However, we believe that a large number of the items that have a QID in this dataset will, at some point, be connected to the UMLS and assigned a CUI. This is because many of the linked entities are still biomedical entities that are just not yet marked as such in Wikidata, or are marked using other biomedical IDs (e.g. `P351: Entrez Gene ID`). By including the QID information in the dataset, we give the research community the possibility to customize the dataset annotations to include other relevant biomedical information.

5.2. Statistics at the Article Level

Table 2 displays the percentage of WikiMed-DE articles associated with various identifiers. All articles have an associated QID and are mapped to at least one biomedical concept ID. 88.39% of articles have a unique CUI associated to them, and are annotated with the three biomedical concept IDs of interest in various degrees - 98.34% are annotated with one or multiple CUIs based on Wikidata information, 29.15% are annotated with a MeSH ID and 4.40% are annotated with DOIDs. WikiMed-DE contains a total of 198,356 unique QIDs, 66,955 unique UMLS CUIs (including the MeSH-derived CUIs and DOID-derived CUIs), 15,915 MeSH IDs, 2,400 DOID annotations, and 125 TUIs.

Table 2
The distribution biomedical concepts connected to the WikiMed-DE articles

	qid	cui	wikidata_cui	mesh	mesh_cui	doid	doid_cui
Number of articles	53,981	47,729	53,085	15,727	15,696	2,377	1,914
Percentage	100%	88.39%	98.34%	29.15%	29.09%	4.40%	3.55%
Total Articles				53,981			

For the convenience of the researchers interested only in a biomedical entity linking benchmark we provide a curated subset, called *WikiMed-DE-BEL*, which focuses exclusively on the mentions that have an unique UMLS CUI associated to them. WikiMed-DE-BEL contains 413,913 mentions corresponding to 35,012 unique mention URLs. All mentions are automatically annotated with a single CUI. Mentions annotated with multiple CUIs are discarded. We also discard any mentions where the mention start or end index does not coincide with the start or end of a token. The dataset was divided into train, test and development splits using an 80-10-10 ratio — leading to 43,184 train articles, 5,399 test articles and 5,398 dev articles. The dataset portions contain 330,233 (train), 41,120 (test) and 42,560 (dev) mentions, annotated with 222,247 (train), 9,123 (test) and 9,149 (dev) unique UMLS CUIs, respectively. 833 (9.12%) of the concepts in the test set do not occur in training.

5.3. WikiMed-DE-BEL quality

To assess the data quality of WikiMed-DE-BEL, a sample of 50 mentions annotated with a single CUI was randomly selected from the dataset. One of the authors checked the automatically annotated CUI for each mention, comparing it to the information available in the UMLS, using the UMLS Metathesaurus Browser[29]. The information was also compared to the context available in the Wikipedia article. 100% of the mentions were found to link to the correct concept and to match their context accurately. This shows that the strict filtering of problematic instances (e.g. mentions with multiple CUIs, or with missing links or QIDs) lead to the creation of a high-quality dataset.

6. Limitations and Conclusion

Accuracy of Information: WikiMed-DE's annotation process heavily relies on the accuracy and completeness of information available in Wikipedia and Wikidata. However, these sources are not immune to errors, inconsistencies, or vandalism. As a result, inaccuracies or outdated information present in the source material can propagate into the annotations in WikiMed-DE.
Noise and Ambiguity: Automated annotation processes can introduce noise and ambiguity in the dataset. The automated methods used to match Wikipedia articles with biomedical concepts may still encounter challenges in fully disambiguating mentions - for example, we cannot systematically choose a single CUI for an entity if multiple CUIs are annotated in Wikidata. We try to limit the noise and ambiguity by enforcing stricter constraints - e.g., we only consider as valid annotations the mentions with a unique CUI mapping. However, this makes the dataset a silver standard dataset, since not all the proposed annotations were manually verified by domain experts.
Link coverage: Not all entities mentioned in a Wikipedia page are exhaustively marked with hyperlinks, meaning that many possible mentions will not be annotated. Furthermore, because we focus on the quality of annotations, we also end up discarding a portion of the marked hyperlinks. This leads to a dataset that has a lower mention coverage than a typical biomedical dataset. WikiMed-DE is therefore less suited for biomedical named entity recognition tasks. However, we believe that it is a useful resource for training BEL systems, and it has great potential to be further developed as new information is added to Wikidata.

This paper presented a new resource for disambiguating biomedical entities in German, WikiMed-DE, and a benchmark dataset for biomedical entity linking in German, WikiMed-DE-BEL, thus supporting biomedical entity linking research focusing on the German language.

[29] UMLS Metathesaurus Browser: https://uts.nlm.nih.gov/uts/umls/home

Acknowledgments

Yi Wang and Corina Dima were supported by the Ministry for Economics, Labour and Tourism from Baden-Württemberg, Germany via grant agreement number BW1_1456 (AI4MedCode).

References

[1] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) 267–270. URL: https://doi.org/10.1093/nar/gkh061. doi:10.1093/nar/gkh061.

[2] R. I. Dogan, R. Leaman, Z. Lu, NCBI disease corpus: A resource for disease name recognition and concept normalization, J. Biomed. Informatics 47 (2014) 1–10. URL: https://doi.org/10.1016/j.jbi.2013.12.006. doi:10.1016/j.jbi.2013.12.006.

[3] S. Mohan, D. Li, Medmentions: A large biomedical corpus annotated with UMLS concepts, in: 1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019, 2019. URL: https://doi.org/10.24432/C5G59C. doi:10.24432/C5G59C.

[4] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, Z. Lu, Biocreative V CDR task corpus: a resource for chemical disease relation extraction, Database J. Biol. Databases Curation 2016 (2016). URL: https://doi.org/10.1093/database/baw068. doi:10.1093/database/baw068.

[5] M. Basaldella, F. Liu, E. Shareghi, N. Collier, COMETA: A corpus for medical entity linking in the social media, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 3122–3137. URL: https://doi.org/10.18653/v1/2020.emnlp-main.253. doi:10.18653/v1/2020.emnlp-main.253.

[6] S. Garda, F. Lenihan-Geels, S. Proft, S. Hochmuth, M. Schuelke, D. Seelow, U. Leser, Regel corpus: identifying DNA regulatory elements in the scientific literature, Database J. Biol. Databases Curation 2022 (2022). URL: https://doi.org/10.1093/database/baac043. doi:10.1093/database/baac043.

[7] M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, D. Schomburg, The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources, Nucleic Acids Res. 39 (2011) 507–513. URL: https://doi.org/10.1093/nar/gkq968. doi:10.1093/nar/gkq968.

[8] N. A. Vasilevsky, S. Essaid, N. Matentzoglu, N. L. Harris, M. A. Haendel, P. N. Robinson, C. J. Mungall, Mondo disease ontology: Harmonizing disease concepts across the world (short paper), in: J. Hastings, F. Loebe (Eds.), Proceedings of the 11th International Conference on Biomedical Ontologies (ICBO) joint with the 10th Workshop on Ontologies and Data in Life Sciences (ODLS) and part of the Bolzano Summer of Knowledge (BoSK 2020), Virtual conference hosted in Bolzano, Italy, September 17, 2020, volume 2807 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 1–2. URL: https://ceur-ws.org/Vol-2807/abstractY.pdf.

[9] D. Vrandecic, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[10] C. E. Lipscomb1, Medical Subject Headings (MeSH), in: Bulletin of the Medical Library Association, volume 88(3) of *CEUR Workshop Proceedings*, 2000, pp. 265–266. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC35238/, pMID: 10928714.

[11] L. M. Schriml, C. Arze, S. Nadendla, Y. W. Chang, M. Mazaitis, V. Felix, G. Feng, W. A. Kibbe, Disease ontology: a backbone for disease semantic integration, Nucleic Acids Res. 40 (2012) 940–946. URL: https://doi.org/10.1093/nar/gkr972. doi:10.1093/nar/gkr972.

[12] S. Vashishth, D. Newman-Griffis, R. Joshi, R. Dutt, C. P. Rosé, Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets, J. Biomed. Informatics 121 (2021) 103880. URL: https://doi.org/10.1016/j.jbi.2021.103880. doi:10.1016/j.jbi.2021.103880.

[13] C. Arighi, L. Hirschman, T. Lemberger, S. Bayer, R. Liecht, D. Comeau, C. Wu1, Bio-ID Track Overview , BioCreative Workshop 482 (2017) 376. URL: https://biocreative.bioinformatics.udel.edu/media/store/files/2018/BC6_track1_1.pdf.

[14] Y. Luo, W. Sun, A. Rumshisky, MCN: A comprehensive corpus for medical concept normalization, J. Biomed. Informatics 92 (2019). URL: https://doi.org/10.1016/j.jbi.2019.103132. doi:10.1016/j.jbi.2019.103132.

[15] K. D. Bollacker, C. Evans, P. K. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: J. T. Wang (Ed.), Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008, ACM, 2008, pp. 1247–1250. URL: https://doi.org/10.1145/1376616.1376746. doi:10.1145/1376616.1376746.

[16] M. Kittner, M. Lamping, D. T. Rieke, J. Götze, B. Bajwa, I. Jelas, G. Rüter, H. Hautow, M. Sänger, M. Habibi, M. Zettwitz, T. de Bortoli, L. Ostermann, J. Ševa, J. Starlinger, O. Kohlbacher, N. P. Malek, U. Keilholz, U. Leser, Annotation and initial evaluation of a large annotated german oncological corpus, JAMIA Open 4 (2021) ooab025.

[17] S. Liang, M. Hartmann, D. Sonntag, Cross-domain German medical named entity recognition using a pre-trained language model and unified medical semantic types, in: Proceedings of the 5th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 259–271. URL: https://aclanthology.org/2023.clinicalnlp-1.31.

[18] F. Liu, I. Vulic, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021, Association for Computational Linguistics, 2021, pp. 565–574. URL: https://doi.org/10.18653/v1/2021.acl-short.72. doi:10.18653/v1/2021.acl-short.72.

[19] C. Müller-Birn, B. Karran, J. Lehmann, M. Luczak-Rösch, Peer-production system or collaborative ontology engineering effort: what is wikidata?, in: D. Riehle (Ed.), Proceedings of the 11th International Symposium on Open Collaboration, San Francisco, CA, USA, August 19-21, 2015, ACM, 2015, pp. 20:1–20:10. URL: https://doi.org/10.1145/2788993.2789836. doi:10.1145/2788993.2789836.

[20] M. Farda-Sarbas, C. Müller-Birn, Wikidata from a research perspective - A systematic mapping study of wikidata, CoRR abs/1908.11153 (2019). URL: http://arxiv.org/abs/1908.

11153. arXiv:1908.11153.

[21] E. Mitraka, A. Waagmeester, S. Burgstaller-Muehlbacher, L. M. Schriml, A. I. Su, B. M. Good, Wikidata: A platform for data integration and dissemination for the life sciences and beyond, in: J. Malone, R. Stevens, K. Forsberg, A. Splendiani (Eds.), Proceedings of the 8th Semantic Web Applications and Tools for Life Sciences International Conference, Cambridge UK, December 7-10, 2015, volume 1546 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015, pp. 69–73. URL: https://ceur-ws.org/Vol-1546/paper_38.pdf.

[22] S. Federhen, The NCBI taxonomy database, Nucleic Acids Res. 40 (2012) 136–143. URL: https://doi.org/10.1093/nar/gkr1178. doi:10.1093/nar/gkr1178.