Wikidata Hierarchy for Named Entity Type Discovery in the Climate Change Domain

Andrija Poleksić[1,2,*], Sanda Martinčić-Ipšić[1,2]

[1]Faculty of Informatics and Digital Technologies (University of Rijeka), Radmile Matejčić 2, Rijeka, 51000, Croatia
[2]Center for Artificial Intelligence and Cybersecurity

Abstract

Named Entity Recognition (NER) is a fundamental task in information extraction, yet general-purpose NER categories often fail to capture the specificity required for specialized domains such as climate change research. This paper presents a methodology for the automatic construction of a domain-specific NER type set with minimal supervision, leveraging a schema-based bottom-up approach to knowledge graph construction. The process begins with the identification of 655 core climate change-related terms, sourced from authoritative domain-specific resources. These terms are then semi-automatically aligned with Wikidata using SPARQL queries to take advantage of its hierarchical structure. A neighbourhood graph is constructed based on *instance of* (P31) and *subclass of* (P279) properties, forming the basis for community detection via the weighted Louvain algorithm. The resulting 59 communities are manually analyzed to derive a final set of 21 NER types, including *Ecosystem, Energy Source, Natural Disaster, Meteorological Phenomenon,* and *Chemical.* Validation against existing ontologies and terminological knowledge base (SWEET, ENVO, and EcoLexicon) reveals that the SWEET ontology provides the highest coverage, containing 57.25% of core terms and 65.38% of the proposed NER types. The findings demonstrate that integrating knowledge graphs, NLP-based information extraction, and community detection provides an effective approach for domain-specific NER schema construction.

Keywords

Named Entity Recognition, Information Extraction, Climate Change, Wikidata, Knowledge Graphs, Community Detection

1. Introduction

Climate change is a global threat that affects various sectors and poses serious risks to sustainability [1]. The agricultural sector is facing declining food production due to unpredictable weather patterns, endangering food security, especially in economies that depend on agriculture [2]. Shifts in temperature ranges threaten biodiversity and accelerate species extinction and ecosystem degradation. Climate change is also increasing the spread of foodborne, waterborne and vector-borne diseases, with rising antimicrobial resistance compounding the health crisis. Additionally, extreme weather events and changing environmental conditions have increased in frequency and intensity [3]. Addressing these challenges requires urgent mitigation and adaptation efforts to prevent further economic, social and environmental consequences.

Climate change research, like other areas of scholar interest, has seen a significant increase in research literature. Motivated by this growing body of literature, many research domains [4, 5, 6, 7, 8] have turned to natural language processing (NLP) methods, particularly tasks surrounding information extraction, to levarage structuring capabilities of these methods on a large amount of unstructured textual data. A well established solution to represent textual information in a structured, machine-interpretable manner is a knowledge graph (KG). Knowledge graphs can be formally defined as a directed graph (G), where G = (V , E) [9]. V refers to the vertices (V) or nodes that represent the real-world entities. E refers to the edges (E) or links between the nodes that represent the relations between the entities [10]. Pairs of entities (e ∈ V), together with an edge that describes their relation form a triple in the KG. Core schema of the knowledge graph is defined as an ontology or taxonomy, depending on the use of

LLM-TEXT2KG 2025: 4th International Workshop on LLM-Integrated Knowledge Graph Generation from Text (Text2KG), June 1 - June 5, 2025, co-located with the Extended Semantic Web Conference (ESWC 2025), Portorož, Slovenia
*Corresponding author.
andrija.poleksic@uniri.hr (A. Poleksić); smarti@uniri.hr (S. Martinčić-Ipšić)

the knowledge graph itself [11, 12]. When building a KG, it is desirable to define classes or types of (named) entities and relations. For example, *cumulonimbus* and *stratocumulus* could be combined with the class *clouds*, further, *clouds* and the entity *El Niño Southern Oscillation* could be defined as elements of the class *meteorological phenomenon* - [METP]. With regard to the newly defined class, it is possible to set specific restrictions for individual relations, e.g. for the relation "*causes*" a restriction ([METP], *causes*, [METP]) can be set.

Named entity recognition (NER) is an information extraction (IE) component that plays a fundamental role in the automated analysis of scientific literature [13, 14]. Traditionally framed as a sequence labeling task, NER aims to assign predefined entity types - such as location, organization, and person - to text spans. However, such coarse-grained categories are often insufficient to capture the domain-specific nuances required for specialized domains such as climate change research. To address this issue, this work focuses on refining the NER for the automatic construction of KGs from textual data in the climate change domain. We utilize existing resources (i.e. climate change terminology dictionaries) to develop a domain-specific set of NER types that are consistent with the Wikidata types terminology [15]. Our approach grounds derived entity types in a corpus of scientific publications in the climate change domain curated by [16] to ensure consistency with real-world climate change research discourse.

Specifically, the contributions are:

- NER types discovery methodology for a selected domain (e.g. climate change) with minimal supervision;
- Derived set of NER types for the climate change domain;
- An alignment of derived entity types with Wikidata supported by coverage in existing climate change domain ontologies.

The paper is structured as follows. Section 2 discusses the principles of KG construction with a focus on the construction of domain-specific KGs and problems. Section 3 covers related work discussing the use of existing resources (dictionaries and KGs) for various information extraction tasks with a emphasis on NER. Section 4 discusses existing NLP resources in climate change domain that can be utilized. Section 5 follows with entity type discovery methodology, in particular the creation of a core entity set for climate change and the use of the Wikidata hierarchy for (named) entity type discovery. In Sections 6 the results are presented. We conclude with Section 7 and discuss the limitations and future work in Section 8.

2. Knowledge Graph Construction

The creation of general, comprehensive, encyclopedic knowledge graphs is a long-term and continuous process that requires a large amount of resources, and traditionally relies on the scientific research results and projects based on community collaboration. Examples of such knowledge graphs are DBpedia [17] (2007), YAGO [18] (2007), BabelNet [19] (2012), and Wikidata [15] (2014) as the currently largest knowledge graph with 114,097,305 nodes and 24,190 active users[1].

In the work of Abu-Salih [9], the creation of a knowledge graph is divided into a schema-based, a schema-free and a hybrid approach, of which the first approach is applicable for the aims of this research. In addition, the schema-based approach can be realized based on two strategies: bottom-up and top-down [10, 20]. The top-down approach implies the initial construction of an ontology/schema or the use of an existing schema and the extraction of knowledge based on a given schema. An example of this approach is the YAGO knowledge graph with strictly defined, non-redundant types of entities and relations and logical constraints on them. In the bottom-up approach, the focus of creation is on the content itself, i.e. the data. Potential entities and relations are first extracted, and the initial knowledge graph schema or ontology is created based on the extracted data. Tamašauskaitė and Groth [10] in a systematic review of 57 scientific papers on the process of creating knowledge graphs, find that 70% of

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

the papers describe a bottom-up approach, an approach that corresponds to the current data-centric trend that we follow in our research as well.

So far, only encyclopedic, (i.e. cross-domain) knowledge graphs (e.g. Wikidata, DBpedia and YAGO) have been mentioned, but there are also increasingly popular domain-specific knowledge graphs such as: KnowLife [4], PaintKG [21] and CS-KG [5] in the fields of health, art and computer science respectively. The creation of knowledge graphs for the selected domain encounters domain-specific challenges in addition to the general problems of building knowledge graphs:

- **Complexity of domain terminology** - a specific domain usually has a specialized vocabulary and technical terms that are not correctly represented in multi-domain (general) knowledge bases;
- **The need for expert domain knowledge** - for the evaluation and validation of knowledge graphs, it is necessary to ensure a domain expert evaluation, and expertise is also required when creating the schema/ontology of the knowledge graph itself;
- **Limitations of existing models for information extraction** - specific domains have their specific entities and relations, which general models fail to extract (i.e. they have not learned the domain-specific relations and entities and are not capable of distinguishing nuanced meanings of domain phrases);
- **Lack of domain ontology** - usually, in a specific domain, there is no clearly defined ontology, which makes it difficult to structure and organize knowledge graph schema. Without an established domain ontology, it becomes difficult to define relations between entities, while ensuring consistency and enabling coherent integration of new information.

To overcome these challenges, automation of the domain knowledge graphs construction, in terms of developing NLP (natural language processing) methods in information extraction, plays a central role.

3. Related Work

The automation of knowledge graph construction is based on unsupervised and/or semi-supervised information extraction procedures, reducing the need for time-consuming and expensive manual data labeling. When building a domain knowledge graph, it is necessary to utilize existing (digital) resources to automate the process and reduce the amount of manual labeling.

Thus, Cai et al. [22] use an existing, more general (coarse), medical domain knowledge graph to create a specific (fine) knowledge graph for the oncology domain. The authors address three types of triples: overlapping triples, where both the coarse and fine domain KGs contain certain triples; triples of new relations but overlapping entities, where the fine domain KG includes both entities but lacks the relation between them; and triples of new entities, where at least one entity does not exist in the coarse KG. To tackle coarse-to-fine KG domain adaptation, they propose an end-to-end KG domain adaptation (KGDA) framework using distant supervision. This framework enables the construction of a KG from fully unlabeled raw text data under the guidance of an existing KG. While this system provides promising results, it relies on the assumption that both KGs have the same types of entities and relations.

Wang et al. [6] use a dictionary and classification of terminology from the geology and mineral resources domain and create a directed graph based on the frequency of bigrams and the order of words in the sentence.

Yuan et al. [7] argue that most existing knowledge graph construction methods are based on large knowledge graphs or existing extensive ontologies/taxonomies, and therefore use the available UMLS thesaurus [23], based on which they recognize domain entities. High-frequency pairs of entities in sentences become potential facts (i.e. triples: entity - relation - entity) for which latent groups (clusters) of relation types are obtained using contextualized embeddings. The clusters of potential relation types are then manually labeled. This significantly reduces annotation cost without loss of quality (instead of labeling each instance of relations, the entire group or all instances of a type are labeled simultaneously).

Frei and Kramer [24] integrate Wikipedia[2] and Wikidata to systematically extract text data and annotation information for Named Entity Recognition (NER). Their approach utilizes the graph relations (properties) of Wikidata to derive NER types. In particular, they use properties such as P2176 (*drug or therapy used for treatment*) to identify entities - e.g. diseases with known treatments - and assign them the NER type *TREATABLE_HEALTH_ISSUE*. This method shows how structured knowledge graphs can be effectively used to generate domain-specific NER categories and improve the annotation of entities in specialized corpora.

Lippolis et al. [25] introduce two approaches for entity alignment between ArtGraph and Wikidata. The first method, Wikidata Entity Search (WES), uses simple SPARQL queries to establish entity correspondences. The second approach, pArtLink, leverages the generative capabilities of large language models in conjunction with established entity-linking techniques such as GENRE [26] and Wikimapper[3] to increase alignment accuracy. ArtGraph, a domain-specific knowledge graph created from WikiArt and DBpedia, encapsulates structured representations of concepts related to works of art.

Nie et al. [13] present the Know-Adapter framework for few-shot NER. The authors emphasize the benefits of incorporating explicit knowledge from external sources, such as knowledge graphs, while addressing the heterogeneity between knowledge graph entity types and NER types. Specifically, for a given mention in a sentence, they build a retriever to find its closest match in Wikidata. They then construct a 3-hop subgraph around the matched entity by traversing Wikidata properties (relations). This approach creates a structured mapping from multiple Wikidata entities that differ in specificity to a single NER type and utilizes the Wikidata hierarchy to improve entity type classification. In contrast to their approach, which expands entity types to improve the few-shot entity classification, our research focuses on the compression and standardization of entity types. By refining a broad and diverse set of entities into a finite set of well-defined NER types. Specifically, we aim to create a structured and domain-relevant taxonomy of the climate change research that ensures consistency and usability in automated knowledge graph construction.

Inspired by these lines of research, we use existing resources such as dictionaries [6, 7], which presumably contain domain entities of different granularity, and combine them with a more general knowledge graph (Wikidata) [22, 25] to construct a hierarchy [13] to produce a final set of NER types for the climate change research domain.

4. Existing Resources

As discussed in Section 3, knowledge-intensive research benefits from available resources. In this sense, this section looks at existing sources that have been used directly or as a reference point in this research, especially existing domain dictionaries, terminologies and ontologies.

Full Weather Glossary[4] from National Oceanic and Atmospheric Administration (NOAA) - National Weather Service (NWS) contains a total of 355 terms with definitions. There is also an extension of this glossary with more than 2000 terms, phrases and abbreviations used by the NWS[5]. Glossary of Meteorology[6] from American Meteorology Society (AMS) is the authoritative source for definitions of meteorological terms. From the AMS and NWS glossaries we have extracted a total of 9,511 climate-change related terms and corresponding definitions.

Webersinke et al. [27] expand the vocabulary when pretraining their models, they add a list of 255 terms[7] (tokens) with the highest frequency in their climate-change related pretraining corpus to the original DistilRoBERTa$_{BASE}$ [28] vocabulary. We add these 255 terms to our dictionary of climate-change related terms.

[2]https://www.wikipedia.org/
[3]https://github.com/jcklie/wikimapper
[4]https://www.weather.gov/otx/Full_Weather_Glossary
[5]https://forecast.weather.gov/glossary.php?
[6]https://glossary.ametsoc.org/wiki/Welcome
[7]https://huggingface.co/climatebert/distilroberta-base-climate-f

Reimerink et al. [8] construct a new multilingual terminological knowledge base (TKB) on the environment science - EcoLexicon[8]. The construction of EcoLexicon began in 2003 with a core list of 794 environmental terms in Spanish and English. For each term, definitions were elaborated, reflecting the level of generality or specificity of the concept as well as its relations with other concepts within the same knowledge domain. The original list of terms was enriched by the addition of new terms as well as by its transformation into a conceptual network. Currently, EcoLexicon contains 4,654 concepts of environmental science and 24,968 terms in eight languages (English, Spanish, German, French, Dutch, Modern Greek, Russian and Arabic) [29]. The EcoLexicon data includes concepts, terms, and semantic relations organized within a frame-like structure called the Environmental Event.

The Environment Ontology (ENVO)[9] is a community-driven ontology that supports the representation of environments beyond the biological and biomedical domains [30, 31]. ENVO consists of classes (terms) that refer to the main types of environments and can facilitate the retrieval and integration of a wide range of biological data. The authors follow the principles of the Open Biomedical and Biological Ontologies (OBO) Foundry and align their ontology with the Basic Formal Ontology (BFO) [32]. ENVO consists of 7,030 classes (terms), such as ENVO's biome, environmental feature, and environmental material hierarchies – the ontology's most developed branches and of the greatest interest to annotators. Recently, when adapting to BFO, some of the hierarchies were revised and made obsolete, such as environmental features.

Semantic Web for Earth and Environmental Terminology (SWEET)[10] [33] is a highly modular ontology suite with 10,239[11] concepts (classes) in 200 separate ontologies covering Earth system science. SWEET is a mid-level ontology and consists of nine top-level concepts that can be used as a foundation for deriving domain-specific ontologies that start from extending these top-level SWEET components.

In [16] we elaborate upon our climate research corpus, consisting of research papers from renowned journals on climate change, that we use in this work. We showed an exploratory prestudy in which we applied a readily available NER model and a POS tagger from flair[12] on a sample of 10,000 research papers (~ 5% of the corpus). With the insights gained from this preliminary experiment, we have decided to experiment with LLM-assisted annotation; in particular, using Phi-3-mini-4k-instruct[13] deployed locally for sentence-level triple extraction task.

5. Entity Discovery

5.1. Core Entity Set

Building upon authoritative sources, including the Full Weather Glossary, the Glossary of Meteorology, Wikipedia glossaries and term expansions in ClimateBERT (*dictionary*), as well as our prior research [16], which includes NER results (*NER*), exploratory LLM-based annotations (*Phi3*) and extracted keywords (*keywords*), we systematically construct a core entity set for the climate change domain. This selection process is based on a majority overlap criterion that requires an exact match of at least three out of four sources. In the initial experiments, we include POS tagging results (*POS*), treating noun phrases as candidate entity terms. However, this approach resulted in a noisy set of instances, which did not contribute to the expansion of the core set, therefore POS-derived votes are excluded. In refinement steps, we experimented with different overlap ratios and case sensitivity. Ultimately, with a majority (three out of four) votes, we settled on a case-sensitive overlap strategy that balances corpus-driven entity selection (*NER, Phi3* and *keywords*) with the integration of terminologies from authoritative sources (*dictionary*).

This process results in a set of 818 core terms, which subsequently undergo cleaning and deduplication.

[8] https://ecolexicon.ugr.es/en/index.htm
[9] https://sites.google.com/site/environmentontology/
[10] https://github.com/ESIPFed/sweet
[11] https://bioportal.bioontology.org/ontologies/SWEET
[12] https://github.com/flairNLP/flair
[13] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

After removing duplicates, 766 unique terms remain. These terms are then validated against entire corpus [16] by computing the occurrence frequency. Terms that occur less than 10 times are excluded from further analysis. This process corresponds to entity detection in phase one of building a knowledge graph, corresponding to the discovery section proposed in [34].

Next, inspired by [25], we perform an automatic alignment of the core terms with Wikidata using three SPARQL queries: exact match, case-invariant match, and a substring-based ("contains") query (see Appendix A). This automated process yields preliminary results, which are then manually curated. During curation, the results are categorized into four distinct groups: (1) Out of scope: 4 terms; (2) Requires disambiguation: 144 terms; (3) Manually corrected (fixed item): 255 terms; (4) Good match: 363 terms. We successfully matched 47.39% of the terms with Wikidata using a simple automatic comparison. The subsequent manual alignment corrects an additional 33.29%, bringing the total number of aligned terms to 618 (80.68%). For the ambiguous group, we align relevant climate-change related terms from Wikidata that are similar to the ambiguous entries and add 37 more terms to the set. As a result, we obtained a final set of 655 core terms aligned with Wikidata items. An example of the alignment is in Table 1, with some terms that have an inherent domain-specific contextualization. For instance, the term *Barber*, which is conventionally associated with an occupational role, is instead categorized within the meteorological domain as a specific type of wind.

5.2. Wikidata Subgraph

Wikidata incorporates several hierarchical (vertical) relations, referred to as properties, such as *instance of* (P31) and *subclass of* (P279). Using the core terms aligned with Wikidata items and these two relations, we construct a neighbourhood graph. In this graph, for each core term, we identify $(n, -m)$-hop neighbours in each direction, where $n, m \in \mathbb{N}$, with n representing height and m representing depth. Height refers to the number of hops in the abstraction direction (towards top), while depth refers to the number of hops in the concretization direction (towards bottom). Specifically, for each core term, we recursively search for items that are instances of or subclasses of the given term. Conversely, we also search for items that the given term is an *instance of* or a *subclass of*, based on the P31 and P279 relations. This process enables us to capture the hierarchical structure and the relationships between

Table 1
Core entity set examples: An exemplary list of core entities (terms) is compiled, including the corresponding Wikidata item, label and description which are automatically assigned. During manual curation, the Wikidata description is systematically compared with dictionary definitions from relevant glossaries to ensure accuracy and consistency. Based on the comparison, a category (cat.) is assigned and, if possible, necessary corrections are made.

Term	Wikidata Item	Wikidata Label	Dictionary Definition	Wikidata Description	Correction	Cat.
anticyclone	Q177414	anticyclone	A region of relatively high atmospheric *pressure, also known as a high. On a *synoptic chart, it appears as a set of closed, approximately circular or elliptical ...	opposite to a cyclone		4
carbon cycle	Q167751	carbon cycle	The set of processes by which carbon is exchanged between the various global reservoirs: sedimentary rocks, the *atmosphere, *...	biogeochemical cycle by which carbon is exchanged among the biosphere		4
cloud	Q113100	Cloud	A visible accumulation of minute water droplets or ice crystals (or both) suspended in the atmosphere, created by the condensation or freezing of ...	2005 indie puzzle video game	Q8074	3
frost heave	Q125822121	Frost heave	The disturbance of the surface of the ground when water, freezing in the form of ice lenses, expands with consequent movement of the soil. The mechanism is involved in the formation of polygonal ground (regular patterns of stones) in Arctic and ...	scientific article published in 2010	Q1432833	3
Barber	Q107198	barber	A wind that is carrying *sleet, *snow, or spray, when the air temperature is close to freezing. Named for the ...	person whose occupation is mainly to cut, dress, groom, style and shave males' hair	Q47209908	2
2	Q200	2	3	natural number		1

the terms within the graph.

Figure 1 illustrates a neighbourhood graph for five terms - *mistral, jet stream, sea breeze, westerlies* and *katabatic wind* - with height $n = 2$ and depth $m = 1$. In this graph, the *instance of* (P31) relations are represented by solid lines, while the *subclass of* (P279) relations are shown with dashed lines. In this case, the concretization direction is not relevant, as the starting terms (i.e. at level 0) are already sufficiently specific. However, moving in the direction of abstraction (i.e. towards the top) reveals a wealth of valuable instances. In particular, the level 2 instance *wind* serves as a direct abstraction for two starting terms (*sea breeze* and *westerlies*), while indirectly encompassing the remaining three terms (*jet stream* via *thermal wind*, *katabatic wind* via *fall wind* and *air current*, and *mistral* via *katabatic wind*). The *wind* effectively encapsulates the meaning of all starting terms in this context, suggesting that it could serve as a representative entity type. A further step in the abstraction can be a viable solution in the form of *meteorological phenomenon*. In this way, we proceed to identify potential Named Entity Recognition (NER) types for identified core entity set (i.e. 655 detected core terms) by utilising the

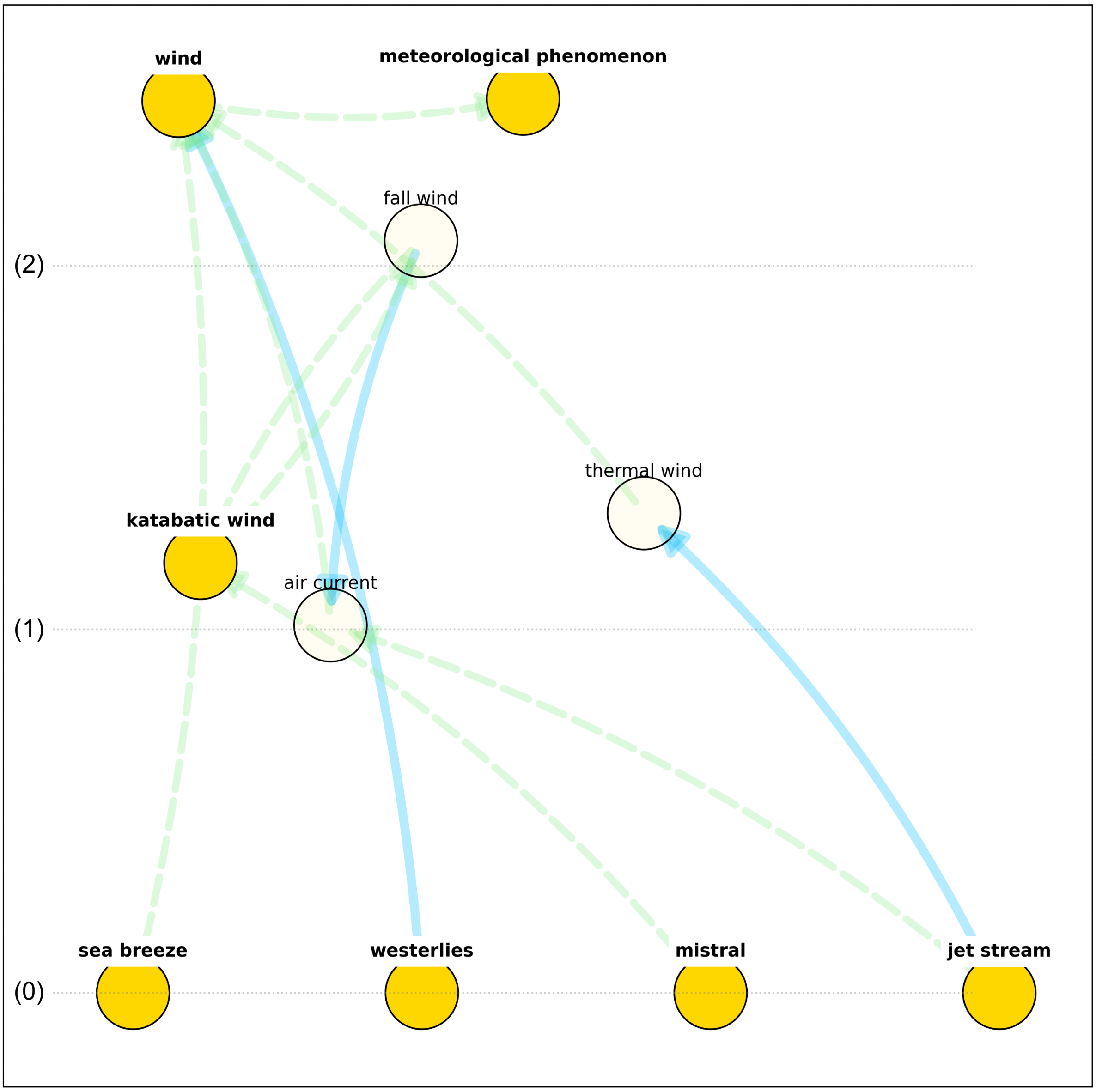

Figure 1: Neighbourhood Graph: A simplified preview of the neighbourhood graph for five terms - *mistral, jet stream, sea breeze, westerlies* and *katabatic wind* - with height $n = 2$ and depth $m = 1$. The *instance of* (P31) relations are represented by solid (blue) lines, and the *subclass* of (P279) relations are shown with dashed (green) lines.

hierarchical structure of the Wikidata graph that guides the discovery of relevant entity categories. Note that Figure 1 is a simplification of the original structure that would be created based on five terms used, a full preview is in the Appendix E.

The hierarchical structure of the Wikidata subgraph is rich and valuable. Still, it contains a large number of nodes and edges, making it difficult to manually navigate and identify an optimal representative node (i.e. a Wikidata item) for NER classification. To overcome this challenge, we utilized Graphia[14], an open-source visual analytics application designed to facilitate the interpretation of large and complex datasets. By leveraging Graphia's graph analysis and transformation capabilities, we refine the subgraph to improve its interpretability. To achieve this, we apply the following preprocessing steps:

- Removal of leaf nodes - not candidates for NER types;
- Filtering based on node height- removing all nodes with a height of $n \geq 4$ - height value indicates a term that is too abstract, e.g. *metaclass* (Q19478619);
- Removal of nodes with in-degree ≤ 1- terms do not contribute to the abstraction.

The height of the node is determined depending on its position to the initial core term. Specifically, for each term, we compute its outgoing n-hop neighbourhood using the *instance of* (P31) and the *subclass of* (P279) relations, as well as its incoming m-hop neighbourhood. Each term that appears in the neighbourhood is assigned a value based on the number of hops from the initial term. These assigned values are then averaged across occurrences to obtain a measure of overall height, which quantifies the level of abstraction of a given term (see Appendix C).

After these preprocessing steps, we perform a weighted Louvain algorithm [35] with a granularity parameter set to 1, using edge weights to reflect relation importance. We argue that the *instance of* (P31) should be considered more significant than the *subclass of* (P279) relation and, therefore assign it weights of 1.0 and 0.5, respectively. This weighting ensures that the communities formed by the Louvain algorithm better reflect meaningful entity groupings for NER classification. In this way, we obtain 59 components (i.e. communities) that are potential NER types for the climate change research domain. After manual inspection of each community we identified a central node (i.e. the node that has a high in-degree centrality), with many connected terms abstracting to it. We also favor nodes with a lower height value whenever possible, as this provides an optimal balance between over-abstraction and over-specificity. This ensures that the selected node serves as a well-generalized yet meaningful representative term within its community, making it a suitable candidate for NER type determination. Examples with the five highest in-degree values in three communities are in Table 2.

After acquiring 59 community or cluster representatives, we conducted a manual inspection to refine the selection. First, we merge similar classes, such as *mathematical expression* and *mathematical concept*. Additionally, we eliminate community representatives that are either overly abstract or unrelated to the field, including *metaclass*, *telecommunication network* and *second-order class* (refer to central row Table 2). Finally, we review and remove the majority of communities containing only a single instance, as they do not contribute to the overall classification structure. After this step, we retain 26 representative terms as potential NER types (classes). In the results Section (6), we ensure alignment with existing domain-specific classifications by manually comparing the extracted terms with established ontologies and terminological knowledge bases, including EcoLexicon, ENVO, and SWEET (see Section 4). Further, we compute the number of terms occurring in each domain-related KG, and we validate the NER types by counting the number of instances under each category. Finally, we preview Louvain clustering results with community statistics.

6. Results

As mentioned in Section 5.1, we calculated the frequency of occurrence for 766 unique terms (including the final 655 core entity terms) in the entire corpus. The top 10 most frequently occurring terms

[14]https://graphia.app/

Table 2
Weighted Louvain algorithm results: An exemplary overview of the weighted Louvain algorithm results, showcasing the top five nodes by in-degree within three detected communities (separated by horizontal lines). The bolded Wikidata label indicates the selected community representative; if none is bolded, the community was discarded during manual postprocessing. The table also provides additional node metrics, including height, total degree, in-degree, out-degree, and the overall size of the respective cluster.

Wikidata ID	Wikidata label	height	node degree	node in-degree	node out-degree	community size
Q107715	**physical quantity**	2.79999995231628	15	13	2	28
Q71758646	general quantity	3.39393949508667	10	10	0	28
Q181175	scalar quantity	3.21951222419739	10	8	2	28
Q71550118	individual quantity	3.484375	6	5	1	28
Q110653654	kind of quantity	3.28571438789368	4	4	0	28
Q24017414	second-order class	3.57894730567932	13	13	0	16
Q21871294	group or class of organisms	3	2	1	1	16
Q67015883	group or class of enzymes	2	2	1	1	16
Q108149	nuclide	3	2	1	1	16
Q112965645	symptom or sign	2.5	2	1	1	16
Q2041172	**measuring instrument**	3.6538462638855	4	4	0	9
Q3099911	scientific instrument	2.95000004768372	3	2	1	9
Q850281	radiometer	2.5	2	1	1	9
Q3743695	meteorological instrument	2.29999995231628	2	1	1	9
Q115797427	camera and optics product	3.75	1	1	0	9

are *water, model, Time, temperature, analysis, precipitation, climate, low, soil* and *level*. The bottom 10 are *Advanced Weather Interactive Processing System, dry line, red beds, pseudoboehmite, Tramontana, geomagnetism, North Greenland Ice Core Project, Advanced Baseline Imager, small hail* and *pressure jump*. The full list is reported in Table 5 (Appendix B).

Further, we perform a case-insensitive match of identified 655 core terms to other ontologies. In particular, we search for the core term in two available ontologies SWEET and ENVO, excluding EcoLexicon as it is not accessible via the API and can not be used locally. For the SWEET ontology, we find a match for 375 core terms (57.25 %), and for ENVO we find a match for 117 (17.86 %). Of the 117 terms that match in ENVO, 105 (89.74 %) are in the SWEET ontology. This limited alignment indicates that the SWEET ontology is a better candidate for future development, as in [22], where a coarse domain knowledge graph (i.e. SWEET) could be used to construct a more specific fine domain KG (i.e. KG for climate change research domain).

As elaborated in Section 5.2, we apply the Louvain algorithm for community detection, yielding a total of 59 communities. For each identified community, we designate a representative node as a potential NER type. The community size distribution is as follows: four large communities contain more than 20 nodes, 19 medium-sized communities have between 10 and 20 nodes, and 34 small communities consist of fewer than 10 nodes. Notably, half of the smallest communities are singleton nodes, that are omitted for further processing. Details are listed in Table 6 (Apendix D). Next, we compare the selected 26 communities (i.e. their representative terms) with SWEET, ENVO and EcoLexicon. The comparison results are shown together with the final selected class names (i.e. NER types) in Table 3. This process was carried out by manual examination of two ontologies (SWEET and ENVO) as well as a terminological knowledge base (EcoLexicon). SWEET and EcoLexicon have a better coverage of 26 representative terms (17 out of 26). Based on the occurrence of representative terms in other knowledge bases, we retain terms that occur at least once, with the exception of *Natural Phenomena*, which we believe is important for the climate change domain. We also merge several similar classes; in particular, *geographic region*, *geographic location* and *geographic entity* are merged into a single class *Location*. In this way, we create a final set of 21 NER types with the following classes: *Ecosystem, Energy Source, Natural Disaster, Meteorological Phenomenon, Quantity, Astronomical Object, Body of Water, Disease, Location, Measurement Unit, Physical Phenomenon, Chemical, Time Period, Organization, Natural Phenomenon, Field of Study, Mathematical Expression, Measuring Device, Geographical Feature, System* and *Satellite*.

Table 3: **Representative Wikidata item alignment**: Comparison of selected community representative Wikidata items (terms) with climate change-related structured sources (SWEET, EcoLexicon, and ENVO). The table includes the final selected entity types (last column) and the occurrence of each Wikidata item in other sources (penultimate column).

Wikidata	SWEET	EcoLexicon	ENVO	#	final entity type
ecosystem	ecosystem	Landscape	ecosystem	3	Ecosystem
energy source	energy source	Energy	oil; nuclear energy; fuel; solar panel array	3	Energy Source
natural disaster			flood; tsunami; vulcanic eruption; earthquake; wildfire	1	Natural Disaster
type of meteorological phenomenon	meteorological phenomena	Atmospheric phenomenon	atmospheric storm; gaseous astronomical body part; electrostatic discharge process; atmospheric aerosol	3	Meteorological Phenomenon
physical quantity	physical quantity	Measure	size	3	Quantity
astronomical object type	astronomical body	Fluid celestial body; Solid celestial body	astronomical object	3	Astronomical Object
body of water	body of water	Artificial body of water; Natural body of water	water body	3	Body of Water
class of disease	disease	Disease		2	Disease
geographic region			geographic feature	1	Location
physical system				0	-
SI unit	unit	Unit		2	Measurement Unit
physical phenomenon	physical process			1	Physical Phenomenon
structural class of chemical entities	chemical	Chemical substance	chemical entity	3	Chemical
time interval	time range	Period	temporal region	3	Time Period
organization	organization	Institution		2	Organization
natural phenomenon				0	Natural Phenomenon
academic discipline	knowledge domain	Discipline		2	Field of Study
mathematical expression	mathematical process	Mathematical expression		2	Mathematical Expression
geographic location		Area		1	Location
geographic entity				0	Location
measuring instrument	device	Measuring instrument		2	Measuring Device
geographical feature		Land		1	Geographical Feature
system	system	System		2	System
product category				0	-
social system				0	-
artificial satellite	satellite		artificial satellite	2	Satellite

For each NER type, we calculate the number of core entity terms that have a path in the Wikidata subgraph (Section 5.2) to Wikidata items corresponding to that NER type. The results are presented in Table 4. Note that we allow each term to have paths to multiple representative Wikidata items (NER types). In this way, we also gain insight into possible redundant classes. The top five class pairs in terms of overlap are: *Geographical Feature - Location* (77), *Field of Study - Quantity* (71), *Meteorological Phenomenon - Natural Phenomenon* (65), *Natural Phenomenon - Physical Phenomenon* (45) and *Field of Study - Physical Phenomenon* (37). On the other hand, we can also observe the terms with the largest number of classes to which they belong. The top five are: *typhoon* and *tropical cyclone* with six and *upwelling*, *cyclone* and *polar vortex*, all of which have five classes (types) to which they correspond.

7. Conclusion

This paper proposes a methodology for discovery of Named Entity Recognition (NER) types tailored to the climate change domain with minimal supervision, leveraging a schema-based bottom-up approach to knowledge graph construction. We use existing resources such as dictionaries [6, 7], which presumably contain domain entities of different granularity, and combine them with a more general knowledge graph (Wikidata) [22, 25] to construct a hierarchy [13] to produce a final set of NER types for the climate change research domain. This process begins with the identification of 655 core climate-change related terms, sourced from authoritative domain-specific resources. These terms are then semi-automatically aligned with Wikidata to fertilize from its hierarchical structure. The weighted Louvain algorithm is engaged for the community detection on a neighbourhood graph constructed from *instance of* (P31) and *subclass of* (P279) Wikidata properties. The resulting 59 communities are manually analyzed to derive a final set of 21 NER types in the climate change domain, including *Ecosystem, Energy Source, Natural Disaster, Meteorological Phenomenon,* and *Chemical*.

Validation against existing ontologies and terminological knowledge base (SWEET, ENVO, and EcoLexicon) reveals that the SWEET ontology provides the highest coverage, containing 57.25% of core terms. Similarly, SWEET also demonstrates strong alignment with the candidate NER types, covering 17 out of 26 types (65.38%). The final set of 21 NER types for the climate change research domain includes: *Ecosystem, Energy Source, Natural Disaster, Meteorological Phenomenon, Quantity, Astronomical Object, Body of Water, Disease, Location, Measurement Unit, Physical Phenomenon, Chemical, Time Period, Organization, Natural Phenomenon, Field of Study, Mathematical Expression, Measuring Device, Geographical Feature, System,* and *Satellite*. Finally, we report the occurrence frequency of core entities in the climate change research corpus. The cutoff threshold of 10 is an indicator that corpus will be well suited for downstream training of domain NER model. The findings demonstrate that refining a broad and diverse set of entities into a finite set of well-defined NER types can contribute to

Table 4
NER type core entity term frequency: Frequency of occurrence for each of the 21 NER types in 655 core terms, sorted descending.

NER label	#	NER label	#
Field of Study	181	Organization	21
Physical Phenomenon	126	Time Period	16
Natural Phenomenon	110	Satellite	13
Location	84	Body of Water	12
Geographical Feature	77	Natural Disaster	8
Quantity	71	Energy Source	6
Meteorological Phenomenon	65	Ecosystem	5
Chemical	46	Measurement Unit	3
System	44	Astronomical Object	2
Mathematical Expression	36	Disease	2
Measuring Device	26	TOTAL:	954

alignment with existing climate ontologies and subsequently to automated climate change knowledge graph construction.

8. Limitations and Future Work

As described in Section 5.2, we construct a neighbourhood graph based on two Wikidata properties - *instance of* (P31) and *subclass of* (P279). This construction is based on the assumption of Wikidata completeness, i.e. if information on these two relations is not available in the Wikidata knowledge graph, terms remain unused and thus potentially impact the overall quality of the results. Some exemplary terms from our core entity set that have neither P31 nor P279 properties are *absolute humidity, Action for climate empowerment, Shortwave radiation* and *pressure jump.* This problem can be tackled in two ways: firstly, by manually adding the missing Wikidata hierarchical properties (relations), thereby contributing to a valuable community-maintained resource, and secondly, by exploring other hierarchical relations such as *part of* (P361), *has part* (Q65964571), *facet of* (P1269) and *broader concept* (P4900). Incorporating these alternative properties could enhance the representation of hierarchical structures for a given domain.

Additionally, the results are potentially sensitive to parameter choices, such as the granularity parameter (set to 1) and the weighting of the *instance of* (1.0) and *subclass of* (0.5) relations in the weighted Louvain algorithm. Exploring alternative granularity values or different weighting schemes may lead to different community detection results and consequently to different NER types. The introduction of additional hierarchical relations further amplifies this sensitivity.

Finally, for future work, we plan to integrate the GLiNER model [36] with our generated NER types. This integration will facilitate the labeling of a larger corpus within the climate change research domain, further refining entity classification and improving automated knowledge extraction.

Acknowledgments

Croatian Science Foundation supports AP under the project DOK-2021-02.

Declaration on Generative AI

During the preparation of this work, the authors used InstaText to improve grammar, check spelling and reword. After using this tool, the authors have reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] K. Abbass, M. Z. Qasim, H. Song, M. Murshed, H. Mahmood, I. Younis, A review of the global climate change impacts, adaptation, and sustainable mitigation measures, Environmental Science and Pollution Research 29 (2022) 42539–42559. URL: https://doi.org/10.1007/s11356-022-19718-6. doi:10.1007/s11356-022-19718-6.

[2] A. Saleem, S. Anwar, T. Nawaz, S. Fahad, S. Saud, T. U. Rahman, M. N. R. Khan, T. Nawaz, Securing a sustainable future: the climate change threat to agriculture, food security, and sustainable development goals, Journal of Umm Al-Qura University for Applied Sciences (2024). URL: https://doi.org/10.1007/s43994-024-00177-3. doi:10.1007/s43994-024-00177-3.

[3] K. L. Ebi, J. Vanos, J. W. Baldwin, J. E. Bell, D. M. Hondula, N. A. Errett, K. Hayes, C. E. Reid, S. Saha, J. Spector, P. Berry, Extreme weather and climate change: Population health and health system implications, Annual Review of Public Health 42 (2021) 293–315. doi:10.1146/annurev-publhealth-012420-105026, epub 2021 Jan 6.

[4] P. Ernst, C. Meng, A. Siu, G. Weikum, Knowlife: A knowledge graph for health and life sciences, in: 2014 IEEE 30th International Conference on Data Engineering, 2014, pp. 1254–1257. doi:10.1109/ICDE.2014.6816754.

[5] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Scicero: A deep learning and nlp approach for generating scientific knowledge graphs in the computer science domain, Knowledge-Based Systems 258 (2022) 109945. URL: https://www.sciencedirect.com/science/article/pii/S0950705122010383. doi:https://doi.org/10.1016/j.knosys.2022.109945.

[6] C. Wang, X. Ma, J. Chen, J. Chen, Information extraction and knowledge graph construction from geoscience literature, Computers & Geosciences 112 (2018) 112–120. URL: https://www.sciencedirect.com/science/article/pii/S0098300417309020. doi:https://doi.org/10.1016/j.cageo.2017.12.007.

[7] J. Yuan, Z. Jin, H. Guo, H. Jin, X. Zhang, T. Smith, J. Luo, Constructing biomedical domain-specific knowledge graph with minimum supervision, Knowl. Inf. Syst. 62 (2020) 317–336. URL: https://doi.org/10.1007/s10115-019-01351-4. doi:10.1007/s10115-019-01351-4.

[8] A. Reimerink, P. L. Araúz, P. J. M. Redondo, EcoLexicon: An environmental TKB, in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010. URL: https://aclanthology.org/L10-1615/.

[9] B. Abu-Salih, Domain-specific knowledge graphs: A survey, Journal of Network and Computer Applications 185 (2021) 103076. URL: https://www.sciencedirect.com/science/article/pii/S1084804521000990. doi:https://doi.org/10.1016/j.jnca.2021.103076.

[10] G. Tamašauskaitė, P. Groth, Defining a knowledge graph development process through a systematic review, ACM Transactions on Software Engineering and Methodology 32 (2022) 1–40. URL: https://api.semanticscholar.org/CorpusID:248435579.

[11] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, Artificial Intelligence Review 56 (2023) 13071–13102. URL: https://doi.org/10.1007/s10462-023-10465-9. doi:10.1007/s10462-023-10465-9.

[12] L. Zhong, J. Wu, Q. Li, H. Peng, X. Wu, A comprehensive survey on automatic knowledge graph construction, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3618295. doi:10.1145/3618295.

[13] B. Nie, Y. Shao, Y. Wang, Know-adapter: Towards knowledge-aware parameter-efficient transfer learning for few-shot named entity recognition, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9777–9786. URL: https://aclanthology.org/2024.lrec-main.854/.

[14] X. Wang, V. Hu, X. Song, S. Garg, J. Xiao, J. Han, ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5227–5240. URL: https://aclanthology.org/2021.emnlp-main.424/. doi:10.18653/v1/2021.emnlp-main.424.

[15] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledge base, Communications of the ACM 57 (2014) 78–85. URL: http://cacm.acm.org/magazines/2014/10/178785-wikidata/fulltext.

[16] A. Poleksić, S. Martinčić-Ipšić, Towards dataset for extracting relations in the climate-change domain, in: S. Tiwari, N. Mihindukulasooriya, F. Osborne, D. Kontokostas, J. D'Souza, M. Kejriwal, M. A. Pellegrino, A. Rula, J. E. Labra Gayo, M. Cochez, M. Alam (Eds.), Joint Proceedings of the 3rd International Workshop on Knowledge Graph Generation from Text (TEXT2KG) and Data Quality Meets Machine Learning and Knowledge Graphs (DQMLKG), co-located with the Extended Semantic Web Conference (ESWC 2024), volume Vol-3747, CEUR-WS, 2024, pp. 9–15. URL: https://ceur-ws.org/Vol-3747/text2kg_paper9.pdf.

[17] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of

open data, in: K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux (Eds.), The Semantic Web, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 722–735.

[18] F. M. Suchanek, G. Kasneci, G. Weikum, Yago: A core of semantic knowledge, in: Proceedings of the 16th International Conference on World Wide Web, WWW '07, ACM, New York, NY, USA, 2007, pp. 697–706. URL: http://doi.acm.org/10.1145/1242572.1242667. doi:10.1145/1242572.1242667.

[19] R. Navigli, S. P. Ponzetto, Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artif. Intell. 193 (2012) 217–250. URL: https://api.semanticscholar.org/CorpusID:6063065.

[20] Z. Zhao, S.-K. Han, I.-M. So, Architecture of knowledge graph construction techniques, 2018. URL: https://api.semanticscholar.org/CorpusID:207900787.

[21] H. Wu, S. Y. Liu, W. Zheng, Y. Yang, H. Gao, Paintkg: the painting knowledge graph using bilstm-crf, in: 2020 International Conference on Information Science and Education (ICISE-IE), 2020, pp. 412–417. doi:10.1109/ICISE51755.2020.00094.

[22] H. Cai, W. Liao, Z. Liu, X. Huang, Y. Zhang, S. Ding, S. Li, Q. Li, T. Liu, X. Li, Coarse-to-fine knowledge graph domain adaptation based on distantly-supervised iterative training, 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2022) 1294–1299. URL: https://api.semanticscholar.org/CorpusID:253383970.

[23] A. Siu, Fast entity recognition in biomedical text, 2013. URL: https://api.semanticscholar.org/CorpusID:39345437.

[24] J. Frei, F. Kramer, Creating ontology-annotated corpora from Wikipedia for medical named-entity recognition, in: D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, J. Tsujii (Eds.), Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 570–579. URL: https://aclanthology.org/2024.bionlp-1.47/. doi:10.18653/v1/2024.bionlp-1.47.

[25] A. Lippolis, A. Klironomos, D. Milon-Flores, H. Zheng, A. Jouglar, E. Norouzi, A. Hogan, Enhancing entity alignment between wikidata and artgraph using llms, in: Semantic Web and Ontology Design for Cultural Heritage 2023, volume 3540 of *CEUR Workshop Proceedings*, CEUR-WS, 2023. Publisher Copyright: © 2023 Copyright for this paper by its authors.; 2023 International Workshop on Semantic Web and Ontology Design for Cultural Heritage, SWODCH 2023 ; Conference date: 07-11-2023.

[26] N. De Cao, G. Izacard, S. Riedel, F. Petroni, Autoregressive entity retrieval, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: https://openreview.net/forum?id=5k8F6UU39V.

[27] N. Webersinke, M. Kraus, J. A. Bingler, M. Leippold, Climatebert: A pretrained language model for climate-related text, 2022. arXiv:2110.12010.

[28] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[29] P. Faber, P. León-Araúz, R. Resi, P. ten Hacken, From specialized knowledge frames to linguistically based ontologies, Applied Ontology 19 (2024) 23–45. URL: https://doi.org/10.3233/AO-230033. doi:10.3233/AO-230033. arXiv:https://doi.org/10.3233/AO-230033.

[30] P. L. Buttigieg, N. Morrison, B. Smith, C. J. Mungall, S. E. Lewis, the ENVO Consortium, The environment ontology: contextualising biological and biomedical entities, Journal of Biomedical Semantics 4 (2013) 43. URL: https://doi.org/10.1186/2041-1480-4-43. doi:10.1186/2041-1480-4-43.

[31] P. L. Buttigieg, E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, C. J. Mungall, The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation, Journal of Biomedical Semantics 7 (2016) 57. URL: https://doi.org/10.1186/s13326-016-0097-6. doi:10.1186/s13326-016-0097-6.

[32] J. N. Otte, J. Beverley, A. Ruttenberg, Bfo: Basic formal ontology, Applied ontology 17 (2022) 17–43. doi:10.3233/ao-220262.

[33] R. G. Raskin, M. J. Pan, Knowledge representation in the semantic web for earth and environmental terminology (sweet), Computers & Geosciences 31 (2005) 1119–1125. URL: https://

www.sciencedirect.com/science/article/pii/S0098300405001020. doi:https://doi.org/10.1016/j.cageo.2004.12.004, application of XML in the Geosciences.
[34] G. Weikum, X. L. Dong, S. Razniewski, F. Suchanek, Machine knowledge: Creation and curation of comprehensive knowledge bases, Found. Trends Databases 10 (2021) 108–490. URL: https://doi.org/10.1561/1900000064. doi:10.1561/1900000064.
[35] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (2008) P10008. URL: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008. doi:10.1088/1742-5468/2008/10/p10008.
[36] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist model for named entity recognition using bidirectional transformer, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 5364–5376. URL: https://aclanthology.org/2024.naacl-long.300. doi:10.18653/v1/2024.naacl-long.300.

A. SPARQL queries

Inspired by the Wikidata Entity Search (WES) approach from [25] we construct three Wikidata SPARQL queries for automatic alignment of Wikidata items to our dictionary terms. For this task, we use the library SPARQLWrapper[15], which serves as a SPARQL endpoint interface to Python. Three queries - exact match, case-invariant match and substring-based ("contains") match - are each listed below.

Listing 1: **Exact Match**: Exact match SPARQL query used for automatic alignment.

```
SELECT ?item ?itemLabel ?itemDescription (GROUP_CONCAT(DISTINCT
    ?itemType; separator=", ") AS ?itemTypes) (GROUP_CONCAT(
    DISTINCT ?itemSubclass; separator=", ") AS ?itemSubclasses)
    WHERE {
     SERVICE wikibase:mwapi {
         bd:serviceParam wikibase:endpoint "www.wikidata.org";
                         wikibase:api "EntitySearch";
                         mwapi:search "{input_text}";
                         mwapi:language "en".
         ?item wikibase:apiOutputItem mwapi:item.
     }
     OPTIONAL { ?item wdt:P31 ?itemType. }  # Retrieve entity
         type (instance of)
     OPTIONAL { ?item wdt:P279 ?itemSubclass. }  # Retrieve
         subclass of
     OPTIONAL { ?item schema:description ?itemDescription. FILTER
          (lang(?itemDescription) = "en") }  # Retrieve
         description
     OPTIONAL { ?item rdfs:label ?itemLabel FILTER (lang(?
         itemLabel) = "en") }  # Retrieve labels
     FILTER (?itemLabel = "{input_text}")  # Ensure the label
         exactly matches the input term
}
GROUP BY ?item ?itemLabel ?itemDescription
LIMIT 10
```

[15] https://github.com/RDFLib/sparqlwrapper

Listing 2: **Case-Invariant Match**: Case-invariant match SPARQL query used for automatic alignment.

```
SELECT ?item ?itemLabel ?itemDescription (GROUP_CONCAT(DISTINCT
    ?itemTypeLabel; separator=", ") AS ?itemTypes) (
   GROUP_CONCAT(DISTINCT ?itemSubclassLabel; separator=", ") AS
    ?itemSubclasses) WHERE {
   SERVICE wikibase:mwapi {
       bd:serviceParam wikibase:endpoint "www.wikidata.org";
                       wikibase:api "EntitySearch";
                       mwapi:search "{input_text}";
                       mwapi:language "en".
       ?item wikibase:apiOutputItem mwapi:item.
   }
   OPTIONAL { ?item wdt:P31 ?itemType. ?itemType rdfs:label ?
      itemTypeLabel. FILTER (lang(?itemTypeLabel) = "en") }
   OPTIONAL { ?item wdt:P279 ?itemSubclass. ?itemSubclass rdfs:
      label ?itemSubclassLabel. FILTER (lang(?itemSubclassLabel
      ) = "en") }
   OPTIONAL { ?item schema:description ?itemDescription. FILTER
       (lang(?itemDescription) = "en") }
   OPTIONAL { ?item rdfs:label ?itemLabel FILTER (lang(?
      itemLabel) = "en") }
   FILTER (regex(?itemLabel, "^{input_text}$", "i"))
}
GROUP BY ?item ?itemLabel ?itemDescription
LIMIT 10
```

Listing 3: **Substring-Based ("contains") Match**: Substring-based ("contains") query match SPARQL query used for automatic alignment.

```
SELECT ?item ?itemLabel ?itemDescription (GROUP_CONCAT(DISTINCT
    ?itemType; separator=", ") AS ?itemTypes) (GROUP_CONCAT(
    DISTINCT ?itemSubclass; separator=", ") AS ?itemSubclasses)
    WHERE {
     SERVICE wikibase:mwapi {
         bd:serviceParam wikibase:endpoint "www.wikidata.org";
                         wikibase:api "EntitySearch";
                         mwapi:search "{input_text}";
                         mwapi:language "en".
         ?item wikibase:apiOutputItem mwapi:item.
     }
     OPTIONAL { ?item wdt:P31 ?itemType. } # Retrieve entity
        type (instance of)
     OPTIONAL { ?item wdt:P279 ?itemSubclass. } # Retrieve
        subclass of
     OPTIONAL { ?item schema:description ?itemDescription. FILTER
         (lang(?itemDescription) = "en") } # Retrieve
        description
     OPTIONAL { ?item rdfs:label ?itemLabel FILTER (lang(?
        itemLabel) = "en") } # Retrieve labels
     FILTER (CONTAINS(LCASE(?itemLabel), LCASE("{input_text}")))
         # Ensure the label contains the input term
}
```

```
GROUP BY ?item ?itemLabel ?itemDescription
LIMIT 10
```

B. Core Entity Terms

Table 5: **Core entity term corpus frequency**: Frequency of occurrence of each of the 655 core terms, sorted by highest occurrence (#) with corresponding Wikidata item ID,

Term	Item ID	#	Term	Item ID	#
water	Q283	1,862,744	confluence	Q723748	7,982
model	Q1979154	1,847,373	NAT	Q83320	7,837
Time	Q11471	1,345,502	ISCCP	Q6052840	7,753
temperature	Q11466	1,046,955	Kriging	Q225926	7,685
analysis	Q217602	1,009,806	biome	Q101998	7,660
precipitation	Q25257	772,605	tropical cyclone	Q8092	7,652
climate	Q7937	769,060	carbon cycle	Q167751	7,630
low	Q209190	757,646	lapse rate	Q66900467	7,620
soil	Q36133	741,357	diurnal variation	Q1469559	7,508
level	Q3686031	733,721	sunshine	Q193788	7,398
Energy	Q11379	686,232	dew	Q41097	7,234
period	Q2642727	597,095	diatoms	Q162678	7,085
Si	Q670	554,688	heat capacity	Q179388	7,061
Wind	Q8094	493,122	IMERG	Q121747699	7,032
rainfall	Q7925	423,591	thermodynamics	Q11473	7,013
SEA	Q11708	377,630	spectrophotometer	Q3492906	6,997
Power	Q25342	365,343	GPM	Q3108963	6,935
observations	Q193181	363,108	supersaturation	Q334104	6,844
Day	Q573	360,935	savanna	Q42320	6,817
Correlation	Q186290	358,478	water vapour	Q190120	6,765
frequency	Q11652	347,994	National Oceanic and Atmospheric Administration	Q214700	6,740
Current	Q5195029	341,183	transparency	Q487623	6,722
Ocean	Q9430	335,148	AVHRR	Q300146	6,684
cloud	Q8074	333,025	Acetonitrile	Q408047	6,630
Ice	Q23392	326,662	SLR	Q841083	6,617
Carbon	Q623	319,829	thunderstorm	Q2857578	6,567
pressure	Q39552	304,875	PSC	Q216417	6,463
resolution	Q3937033	302,930	CMA	Q906716	6,460
CO	Q2025	302,116	percolation	Q1367555	6,432
summer	Q1313	297,519	Prairie	Q194281	6,414
Sample	Q485146	289,046	microclimate	Q215108	6,302
Source	Q31464082	275,979	general circulation model	Q650994	6,282
index	Q1738991	272,002	Intergovernmental Panel on Climate Change	Q171183	6,276
variation	Q106645015	268,946	MISR	Q3867036	6,274
SST	Q1507383	268,895	Graupel	Q213202	6,077
Standard	Q367293	263,275	WMO	Q170424	6,071
Winter	Q1311	259,672	Walker Circulation	Q2142205	6,062
season	Q24384	242,303	steppe	Q123991	6,058
Ph	Q40936	237,393	subtropical high	Q972926	5,994

Term	Item ID	#	Term	Item ID	#
Stress	Q123414	222,831	APS	Q466113	5,950
temperatures	Q11466	216,572	Stratocumulus	Q40564	5,873
biomass	Q2945560	212,304	sea breeze	Q81242	5,765
Basin	Q813672	208,916	precipitable water	Q778526	5,742
aerosol	Q104541	206,552	MOC	Q4652675	5,718
drought	Q43059	195,902	Nevada	Q432381	5,690
Groundwater	Q161598	188,300	AMV	Q756835	5,624
Atmosphere	Q8104	187,837	internet	Q75	5,560
Sensitivity	Q521783	183,995	accretion	Q1402738	5,395
radiation	Q18335	178,900	p300	Q3136081	5,367
Age	Q568683	178,387	deuterium	Q102296	5,284
extreme	Q845060	173,728	brightness temperature	Q4538627	5,139
Channel	Q1210950	165,731	cloud amount	Q830457	5,121
thermal	Q752823	161,853	sublimation	Q131800	5,091
observation	Q193181	161,770	LLJ	Q11850562	5,069
probability	Q9492	158,667	European Centre for Medium-Range Weather Forecasts	Q1274195	5,037
spring	Q1312	155,658	trade winds	Q160603	5,035
evolution	Q1063	153,963	North Atlantic Oscillation	Q1137345	5,027
Accuracy	Q1298969	152,585	Alkanes	Q41581	5,018
CO2	Q1997	148,561	PMC	Q7209090	5,008
Runoff	Q66486514	147,941	tornado	Q8081	4,944
Snow	Q7561	147,532	storm surge	Q121742	4,941
measurement	Q12453	140,638	specific heat	Q487756	4,935
weather	Q11663	140,578	plankton	Q25367	4,897
ozone	Q36933	136,774	planetary boundary layer	Q1757268	4,886
ENSO	Q14524818	136,475	adenovirus	Q193447	4,769
Variance	Q175199	135,933	desertification	Q183481	4,737
Li	Q568	135,068	Kuroshio	Q53842	4,725
gradient	Q173582	131,432	CFC	Q23748224	4,630
Stability	Q2325497	130,248	Cretaceous	Q44626	4,629
threshold	Q29051774	130,137	power spectrum	Q1331626	4,611
dust	Q129129	128,726	glia	Q177105	4,581
Nitrogen	Q627	128,278	desiccation	Q903071	4,551
Vector	Q13471665	126,526	response time	Q578372	4,467
pollution	Q58734	122,509	GOME	Q1425042	4,426
accumulation	Q116844065	121,632	carbon monoxide	Q2025	4,317
Irrigation	Q21893647	120,674	Hadley Circulation	Q338589	4,315
Monsoon	Q42967	118,045	coalescence	Q2071902	4,292
assessment	Q123304503	117,212	treeline	Q207762	4,233
hypothesis	Q41719	115,864	Gulf Stream	Q130905	4,221
rain	Q7925	112,331	monsoon climate	Q122933063	4,157
force	Q11402	112,030	photochemistry	Q188651	4,152
humidity	Q180600	110,184	CGCM	Q650994	4,145
anomaly	Q567555	109,919	nitric oxide	Q207843	4,127
deposition	Q871279	109,532	Newton	Q12438	4,098
convection	Q160329	109,012	cyclogenesis	Q245472	3,963
amplitude	Q159190	105,901	drainage area	Q166620	3,933
elevation	Q2633778	105,434	SPCZ	Q5977788	3,920
Latitude	Q34027	105,131	AGL	Q323170	3,904

Term	Item ID	#	Term	Item ID	#
feedback	Q183635	104,950	radioactivity	Q11448	3,879
Oxygen	Q629	101,342	solar cycle	Q49385	3,868
Fluorescence	Q191807	98,686	solar activity	Q7297568	3,862
Validation	Q359176	96,631	planetary wave	Q1053589	3,860
Image	Q478798	94,192	lichen	Q43142	3,789
soil moisture	Q889507	92,774	MM5	Q1516983	3,789
forecast	Q748250	91,639	POP	Q1564294	3,763
equilibrium	Q11061286	89,713	Copernicus	Q1531636	3,750
storm	Q81054	88,557	Argon	Q696	3,743
theory	Q17737	86,270	volatile organic compounds	Q910267	3,724
altitude	Q190200	85,317	stratus	Q40526	3,702
Earth	Q2	83,341	Moon	Q405	3,641
aerosols	Q104541	82,000	refraction	Q72277	3,624
Spectrum	Q654182	81,830	eccentricity	Q208474	3,477
absorption	Q332828	80,604	overcast	Q1055865	3,472
diffusion	Q163214	80,602	SAF	Q7649638	3,464
evaporation	Q132814	78,605	IASI	Q1623073	3,459
hydrogen	Q556	77,660	helium	Q560	3,456
troposphere	Q40631	77,339	icing	Q12060664	3,442
sea ice	Q213926	76,405	MOPITT	Q1638480	3,385
Plasma	Q10251	71,480	occlusion	Q747330	3,383
fusion	Q106080	69,258	meridional circulation	Q463223	3,361
oxidation	Q1786087	67,928	atmospheric chemistry	Q287919	3,330
convergence	Q1783472	66,736	knot	Q128822	3,307
productivity	Q3289687	66,615	dew point	Q178828	3,244
jet	Q202325	65,669	anemometer	Q175029	3,207
adsorption	Q180254	63,735	MOS	Q1453537	3,185
watershed	Q166620	63,685	savannas	Q42320	3,174
salinity	Q179615	63,146	Intertropical Convergence Zone	Q753858	3,145
Albedo	Q101038	62,609	Rocky Mountains	Q5463	3,135
surface temperature	Q56297886	61,249	flash flood	Q860333	3,134
scattering	Q210028	60,496	nitrogen oxides	Q424418	3,101
Probe	Q96093522	59,052	critical point	Q111059	3,084
oscillation	Q170475	55,187	cold pool	Q104862831	3,066
p53	Q14818098	54,934	Firn	Q828861	3,054
autumn	Q1314	54,117	Headwaters	Q7376362	2,989
MJO	Q1170041	54,066	LIS	Q128405384	2,913
Nitrate	Q49916468	53,413	nitrous oxide	Q905750	2,871
Stratosphere	Q108376	52,329	avalanche	Q7935	2,838
NAO	Q1137345	51,719	tsunami	Q8070	2,836
boundary layer	Q752193	51,021	swell	Q185411	2,831
advection	Q379788	50,803	World Meteorological Organization	Q170424	2,827
El Niño	Q7939	49,885	phase change	Q185357	2,804
Divergence	Q85900110	49,567	Berg	Q8502	2,786
front	Q189796	48,813	sprite	Q904961	2,778
vortex	Q732722	48,788	Pliocene	Q76259	2,768
Streamflow	Q29425295	48,533	AOGCM	Q650994	2,749
climatology	Q52139	48,482	Pacific Decadal Oscillation	Q2033747	2,729

Term	Item ID	#	Term	Item ID	#
MODIS	Q676840	48,362	continental shelf	Q134851	2,708
sodium	Q658	47,313	SPC	Q751874	2,655
evapotranspiration	Q828158	47,288	aegypti	Q1148004	2,645
GCM	Q650994	47,127	ice shelf	Q46966	2,619
tropics	Q42530	47,009	Deconvolution	Q1183700	2,595
relative humidity	Q2499617	46,154	STP	Q102145	2,589
lidar	Q504027	45,011	SSI	Q81382741	2,587
tendency	Q55919789	44,877	Arctic Oscillation	Q674041	2,465
drop	Q185789	43,806	SEVIRI	Q117778573	2,465
eddy	Q994122	43,764	ocean acidification	Q855711	2,455
blocking	Q1540250	43,002	filopodia	Q14859810	2,396
Cd	Q83216	42,547	Jacobian	Q506041	2,287
turbulence	Q190132	40,596	ONI	Q117235275	2,264
NCEP	Q1966999	40,270	Paris Agreement	Q21707860	2,224
recombination	Q3373825	40,220	arid climate	Q190946	2,185
lightning	Q33741	39,991	GMS	Q2246672	2,175
Met	Q25261	39,857	greenhouse effect	Q41560	2,175
isotope	Q25276	39,605	stratopause	Q205397	2,147
nucleus	Q677070	39,474	TOGA	Q3540622	2,134
Methane	Q37129	38,440	hydrologic cycle	Q81041	2,129
aggregation	Q85248618	37,869	glomeruli	Q909882	2,118
Aspect	Q355730	37,612	NLDN	Q28458090	2,100
cyclone	Q79602	37,215	climate simulation	Q117829810	2,090
NOAA	Q214700	37,152	global radiation	Q1531731	2,090
Ir	Q11388	36,372	zonal flow	Q219838	2,087
Persistence	Q922395	36,162	photosynthetically active radiation	Q900892	2,060
reconstruction	Q116146313	36,009	tropical climate	Q135712	2,028
remote sensing	Q199687	35,667	inversion layer	Q25615856	2,026
Sun	Q525	34,997	low-level jet	Q11850562	2,008
Longitude	Q36477	34,765	synoptic scale	Q1233837	1,977
inversion	Q190096	34,714	thermohaline circulation	Q463223	1,964
global warming	Q7942	34,616	ODS	Q16607840	1,947
Forestry	Q38112	34,217	QuikSCAT	Q1734511	1,937
Nt	Q95976921	33,973	Meteosat	Q1429889	1,925
Equator	Q23538	33,730	Indian Ocean Dipole	Q1574518	1,901
instability	Q405372	32,678	laminar flow	Q189452	1,878
Wetlands	Q170321	31,762	AABW	Q3913650	1,815
nucleation	Q909022	31,459	continental climate	Q185005	1,807
latent heat	Q207721	30,008	levoglucosan	Q6535767	1,789
Seawater	Q184395	29,337	ozone hole	Q183140	1,789
dissociation	Q189673	29,180	carbon tax	Q288401	1,773
photosynthesis	Q11982	29,134	foehn	Q12314	1,753
desert	Q8514	28,743	melting point	Q15318	1,730
hydrolysis	Q103135	28,535	nitrogen dioxide	Q207895	1,717
tropopause	Q186433	28,013	ceilometer	Q1027486	1,659
phytoplankton	Q184755	27,616	convective available potential energy	Q1129355	1,591
dry season	Q146575	27,064	xenon	Q1106	1,586

Term	Item ID	#	Term	Item ID	#
eye	Q640404	26,844	POPS	Q912951	1,543
condensation	Q166583	26,827	UTCI	Q30347503	1,500
ECMWF	Q1274195	26,773	solar wind	Q79833	1,499
tracer	Q15835484	26,492	temperate zone	Q167466	1,495
glacier	Q35666	26,132	lithosphere	Q83296	1,468
Grass	Q643352	25,778	SMOS	Q280068	1,463
entropy	Q45003	25,384	long-wave radiation	Q82340792	1,458
ITCZ	Q753858	24,941	cryosphere	Q493109	1,443
deforestation	Q169940	24,806	geostrophic wind	Q929043	1,366
friction	Q82580	24,776	El Niño Southern Oscillation	Q14524818	1,352
IPCC	Q171183	24,750	National Weather Service	Q1066823	1,348
PDO	Q2033747	24,281	Atlantic Meridional Overturning Circulation	Q4652675	1,343
rotor	Q11998503	24,038	acid rain	Q40178	1,313
Ecology	Q7150	23,632	scatterometer	Q905295	1,309
radiative forcing	Q1463606	23,347	calving	Q868757	1,282
ammonia	Q4087	23,267	sintering	Q844613	1,278
AO	Q674041	23,079	Southern Oscillation Index	Q1550887	1,275
PG	Q2414143	22,305	photodissociation	Q16814	1,262
geopotential height	Q12432978	21,961	climate classification	Q267474	1,255
Pan	Q3342203	21,840	World Climate Research Programme	Q3407026	1,240
Autocorrelation	Q786970	21,576	SeaWiFS	Q2261857	1,231
greenhouse gas	Q167336	21,408	meteorite	Q60186	1,221
upwelling	Q215915	21,373	geomagnetic field	Q6500960	1,210
wind stress	Q8024052	21,099	zeaxanthin	Q169337	1,205
smoke	Q130768	20,878	Little Ice Age	Q190530	1,191
elastic	Q62932	20,620	megafauna	Q730371	1,161
TGF	Q1584373	20,588	orographic precipitation	Q11689358	1,155
diffraction	Q133900	20,533	gelsolin	Q18297560	1,147
depression	Q209190	20,465	Advanced Very High Resolution Radiometer	Q300146	1,143
CAPE	Q185113	20,294	ozone layer	Q79995	1,140
fog	Q37477	20,006	NHC	Q1329523	1,120
curvature	Q214881	19,949	MHS	Q17125174	1,115
hydrology	Q42250	19,853	acclimatization	Q419763	1,092
transpiration	Q167980	19,672	NEXRAD	Q3088597	1,090
La Niña	Q642867	19,552	GARP	Q16251355	1,084
attenuation	Q2357982	19,409	Kyoto Protocol	Q47359	1,073
intensification	Q38178665	19,332	bortezomib	Q419319	1,059
snowfall	Q7561	19,121	ODP	Q900522	1,049
PBL	Q1757268	19,113	land breeze	Q31374425	1,043
typhoon	Q140588	18,983	lamellipodia	Q3092607	1,028
reflection	Q165939	18,812	WRCC	Q30687889	1,027
TRMM	Q2001116	18,676	zonal circulation	Q3353804	1,025
AMOC	Q4652675	18,668	methane hydrate	Q389036	1,014
Permafrost	Q179918	18,554	Younger Dryas	Q944279	1,011
mixing ratio	Q171293	18,422	FAA	Q335357	979
FA	Q62008854	18,287	nitrogen cycle	Q82551	970
life cycle	Q67657988	17,931	Envisat	Q49692	950

Term	Item ID	#	Term	Item ID	#
Cirrus	Q185638	17,852	geophysics	Q46255	948
teleconnection	Q3982797	17,815	ultraviolet radiation	Q11391	923
phenology	Q272737	17,445	International Satellite Cloud Climatology Project	Q6052840	917
sensible heat	Q1480581	17,300	Western Pacific Warm Pool	Q7846140	900
peat	Q184624	17,278	Cyclohexane	Q211433	898
CAT	Q1101409	17,214	sea-surface temperature	Q1507383	882
Landsat	Q849791	17,019	cumulonimbus	Q182311	871
influenza	Q2840	16,872	freezing rain	Q11120024	863
GPS	Q18822	16,787	neon	Q654	853
entrainment	Q15733549	16,778	aldolase	Q421968	850
turbidity	Q898574	16,681	extratropical cyclone	Q1063457	848
rainy season	Q3117517	16,675	western boundary current	Q38178435	845
PAR	Q900892	16,651	absolute humidity	Q1048298	836
air mass	Q216823	16,640	meniscus	Q898732	828
surge	Q287381	16,550	synthetic aperture radar	Q740686	818
thermocline	Q849599	16,499	automatic weather station	Q846837	796
wet season	Q3117517	16,487	closed system	Q1468684	776
subsidence	Q2091656	16,480	EUMETSAT	Q692163	766
hurricane	Q34439356	16,426	Barber	Q47209908	752
soil temperature	Q889769	16,303	South Pacific Convergence Zone	Q5977788	739
carbon dioxide	Q1997	16,188	CCB	Q5133390	737
dissolution	Q3133701	16,031	Thermistor	Q175973	722
meteorology	Q25261	15,972	Somali Jet	Q122574051	706
GOES	Q976688	15,801	subtropical anticyclone	Q177414	685
ablation	Q322177	15,773	docetaxel	Q420436	670
AMO	Q756835	15,693	mean free path	Q756307	670
VOC	Q910267	15,396	wind rose	Q2336098	659
specific humidity	Q2253551	15,010	dendrochronology	Q80205	646
agarose	Q390697	15,000	California Current	Q281655	623
Isoprene	Q271943	14,764	anvil cloud	Q1358304	621
zebrafish	Q169444	14,745	ensemble forecasting	Q3433888	618
Holocene	Q25445	14,724	heat index	Q2141844	606
radiosonde	Q852817	14,589	Agulhas Current	Q398548	601
anticyclone	Q177414	14,479	Antarctic Circumpolar Current	Q55828	598
Sahel	Q66065	14,406	carbon capture and storage	Q41491	596
kinetic energy	Q46276	14,254	North Atlantic Current	Q211798	593
MCS	Q660968	14,093	hypothermia	Q1036696	587
frost	Q4590598	14,089	supercooling	Q213659	582
hydroxyl	Q104116	13,943	magnetosphere	Q6915	560
water table	Q3342272	13,843	North Atlantic Deep Water	Q921070	557
Cumulus	Q14189	13,821	Atlantic Niño	Q4816419	546
pandemic	Q12184	13,809	coupled general circulation model	Q650994	524
Radiance	Q1411145	13,733	speleothems	Q154507	504
termination	Q23582432	13,614	time-series analysis	Q11850042	498
Hf	Q15115271	13,575	planetary scale	Q124101881	493
visibility	Q654068	13,518	Mistral	Q193742	481

Term	Item ID	#	Term	Item ID	#
Haze	Q643546	13,436	AATSR	Q4649950	480
mass balance	Q121278173	13,375	mass balance model	Q121278173	472
wind shear	Q1027878	13,182	downburst	Q4847219	467
magnetic field	Q11408	12,951	frost heave	Q1432833	465
westerlies	Q12343832	12,947	Northern Annular Mode	Q674041	464
buoyancy	Q6497624	12,872	Maunder Minimum	Q827568	457
potential temperature	Q760765	12,727	katabatic wind	Q212903	441
loess	Q22723	12,663	mesoscale convective system	Q660968	409
ionization	Q190382	12,398	Antarctic Oscillation	Q3288815	395
eukaryotes	Q19088	12,167	sudden stratospheric warming	Q1583422	394
longwave radiation	Q82340792	12,152	bombykol	Q425845	378
BT	Q225561	11,921	gamma radiation	Q11523	366
shortwave radiation	Q7502259	11,745	olaparib	Q7083106	360
mercury	Q925	11,704	global dimming	Q211627	348
residence time	Q177453	11,642	Advanced Microwave Sounding Unit	Q4686237	345
ice sheet	Q12599	11,108	Nimbostratus	Q202278	326
Southern Oscillation	Q1423047	11,003	Oceanic Niño Index	Q117235275	325
subtropics	Q16305538	10,894	cut-off low	Q60967643	316
conduction	Q14946524	10,639	plate tectonics	Q7950	302
polar vortex	Q1197111	10,591	fibrillin-1	Q17927651	299
rain gauge	Q190052	10,432	Global Ozone Monitoring Experiment	Q1425042	296
carbon sequestration	Q15305550	10,417	Upper Atmosphere Research Satellite	Q534401	287
AGCM	Q650994	10,313	Loop Current	Q377116	275
ACE	Q30717004	10,252	National Lightning Detection Network	Q28458090	253
return period	Q2627230	10,221	CYGNSS	Q5198802	250
SAR	Q740686	10,196	Equatorial Undercurrent	Q1190478	248
Lf	Q17156810	10,041	Tropical Rainfall Measurement Mission	Q2001116	240
insolation	Q216973	9,972	mesocyclone	Q2002856	227
tundra	Q43262	9,943	dendroclimatology	Q2294113	215
cloudiness	Q830457	9,937	South Equatorial Current	Q1072306	202
adiabatic	Q182453	9,856	Benguela Current	Q59676	200
radon	Q1133	9,263	ketoconazole	Q407883	171
mantle	Q101949	9,252	synoptic meteorology	Q130221760	157
tilt	Q179745	9,179	pollen analysis	Q2737544	153
Skewness	Q9051521	9,156	Jason-1	Q1970012	150
CERES	Q1102659	9,127	COP26	Q7888355	141
gyre	Q1250263	8,881	Universal Thermal Climate Index	Q30347503	137
CCS	Q41491	8,802	glaciology	Q52120	126
NWP	Q837552	8,796	iridescence	Q957208	123
half-life	Q47270	8,794	turbidity current	Q1756774	120

Term	Item ID	#	Term	Item ID	#
biosphere	Q42762	8,632	International Polar Year	Q784374	114
Acetone	Q49546	8,596	pressure jump	Q7241727	108
Cal	Q26708069	8,522	small hail	Q3229952	104
Aqua	Q17397	8,445	Advanced Baseline Imager	Q110822048	94
black carbon	Q3233590	8,334	North Greenland Ice Core Project	Q9063437	90
hydrological cycle	Q81041	8,310	geomagnetism	Q114591	85
mass spectrometer	Q1327691	8,300	Tramontana	Q453122	75
hail	Q37602	8,264	pseudoboehmite	Q2115715	67
Terra	Q584697	8,204	red beds	Q2065586	63
harmonics	Q1148098	8,060	dry line	Q2742789	49
SOI	Q1550887	8,043	Advanced Weather Interactive Processing System	Q4686330	12
jet stream	Q202325	7,997	TOTAL:	36,516,003	

C. Node Depth and Node Height

Building upon the examples provided in this work, we consider five initial Wikidata terms: *mistral, katabatic wind, jet stream, sea breeze*, and *westerlies.* We perform a recursive search with a maximum height of $n = 2$ (two hops upward along *instance of* (P31) and *subclass of* (P279)) and a maximum depth of $m = 1$ (one hop downward along these relations).

For example, starting from *jet stream*, we identify *air current* as a one-hop neighbour. In turn, *wind* is a one-hop neighbour of *air current*, reaching the two-hop limit. Conversely, in the opposite direction (where *jet stream* is the object of P31 or P279 relations), we find *jet streak* as a direct neighbour. This procedure is applied to all starting terms, producing the following exemplary results:

a) (-1) jet streak	-> (0) **jet stream**	-> (1) air current	-> (2) wind
b) (-1) ______	-> (0) **mistral**	-> (1) katabatic wind	-> (2) fall wind
c) (-1) mistral	-> (0) **katabatic wind**	-> (1) fall wind	-> (2) air current
d) (-1) Sundowner	-> (0) **sea breeze**	-> (1) wind	-> (2) meteorological phenomena
e) (-1) Shrieking Sixties	-> (0) **westerlies**	-> (1) west wind	-> (2) wind

From this limited set of terms, we can compute each node's overall height as the average of all depths (or heights) at which it appears. For example, consider the node *katabatic wind*, which appears as a starting term at height 0 (example a) and as a one-hop neighbour at height 1 (example b). Its overall height is thus calculated as: $\frac{0+1}{2} = 0.5$.

D. Louvain Algorithm Results

Table 6

Louvain cluster results: Results of the Louvain algorithm on Wikidata subgraph with core entity terms. For each cluster/community the reported size is the number of nodes in the cluster and the selected representative node (the node with the highest in-degree centrality value, with several corrected nodes after manual inspection) is listed. The bolded node representatives are further used in the development of the final NER types, while the underlined nodes represent manually merged communities. Multiple representative terms separated by a semicolon indicate multiple potential NER types from a single community.

size	representative	size	representative
28	**physical quantity**	4	**artificial satellite**
25	**structural class of chemical entities**	4	**SI unit**
23	**mathematical expression**	4	**astronomical object type**
22	material	3	type of structure
19	**type of meteorological phenomenon**	2	statistic
19	**organization**	2	chronostratigraphic unit
18	**geographic location**	2	gene
17	**academic discipline**	2	production environment factor
16	process	2	radiation
16	second-order class	2	telecommunications network
14	metaclass	2	document
14	**body of water; geographical feature**	1	scientific model
14	philosophical concept	1	publishing company
14	**system; ecosystem;** physical system; social system; knowledge system	1	shell of an astronomical object
14	chemical element (structural class of chemical entities)	1	layer
13	**physical phenomenon**	1	computer simulation
13	legal concept	1	scientific law
13	**energy source**	1	circle
12	**geographic region**	1	geostationary satellite
11	result	1	differential operator
11	occurrence	1	beginning
11	field of study (academic discipline)	1	sense
11	**time interval**	1	s-block
10	**natural phenomenon**	1	observance
10	**class of disease; natural disaster;**	1	solution
9	**measuring instrument**	1	ecological unit
9	mathematical concept (mathematical expression)	1	mechanical wave
8	variable-order class	1	pigment
8	third-order class		
7	**product category**		
7	**geographic entity**		

E. Neighbourhood Graph

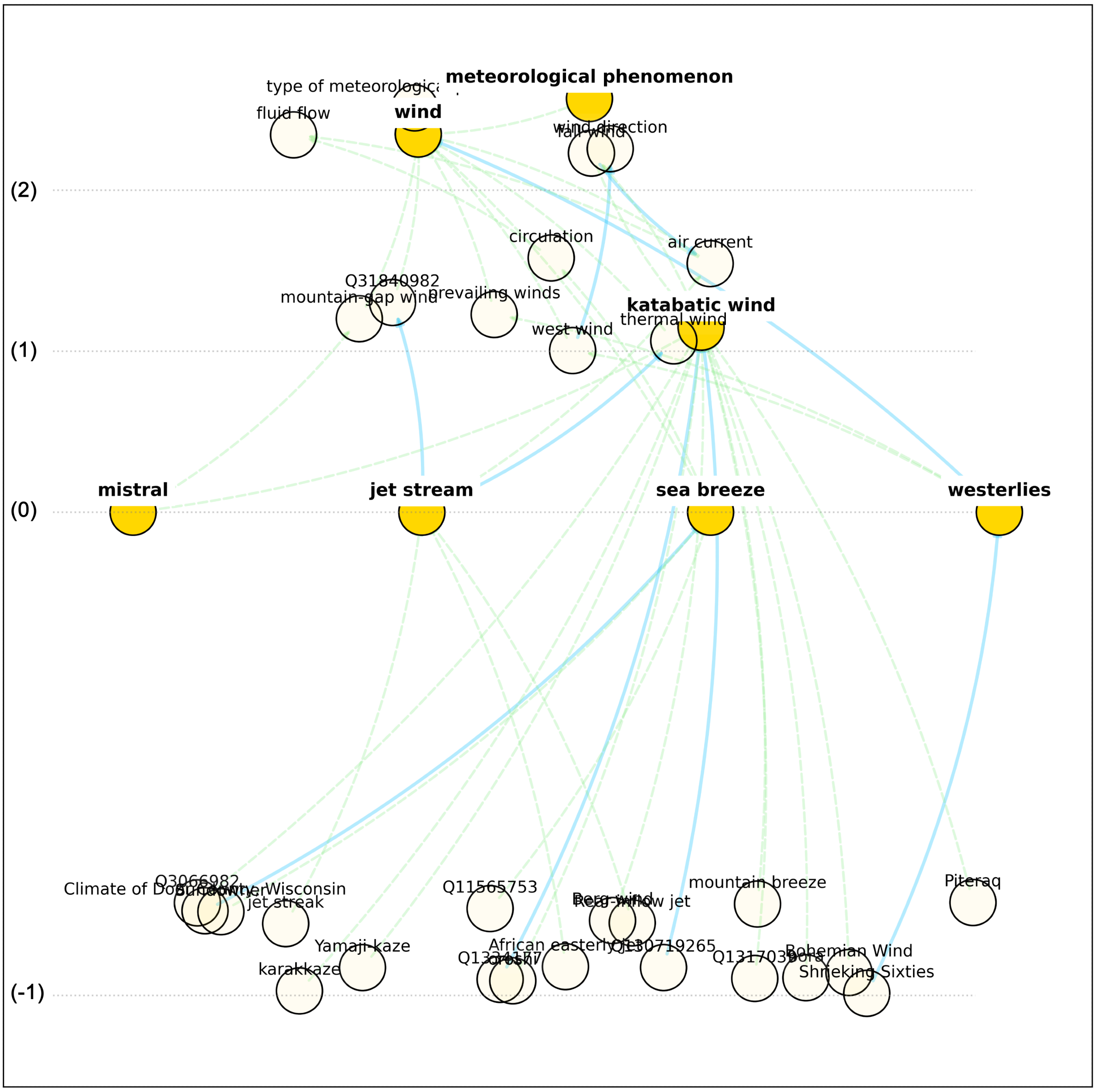

Figure 2: Neighbourhood Graph: A preview of the neighbourhood graph for five terms - *mistral, jet stream, sea breeze, westerlies* and *katabatic wind* - with height $n = 2$ and depth $m = 1$. The *instance of* (P31) relations are represented by solid (blue) lines, and the *subclass of* (P279) relations are shown with dashed (green) lines.