knowledge

MDPI

Article

Overview of STEM Science as Process, Method, Material, and Data Named Entities

Jennifer D'Souza

TIB Leibniz Information Centre for Science and Technology, 30167 Hannover, Germany; jennifer.dsouza@tib.eu

Abstract: We are faced with an unprecedented production in scholarly publications worldwide. Stakeholders in the digital libraries posit that the document-based publishing paradigm has reached the limits of adequacy. Instead, structured, machine-interpretable, fine-grained scholarly knowledge publishing as Knowledge Graphs (KG) is strongly advocated. In this work, we develop and analyze a large-scale structured dataset of STEM articles across 10 different disciplines, viz. *Agriculture*, *Astronomy*, *Biology*, *Chemistry*, *Computer Science*, *Earth Science*, *Engineering*, *Material Science*, *Mathematics*, and *Medicine*. Our analysis is defined over a large-scale corpus comprising 60K abstracts structured as four scientific entities process, method, material, and data. Thus, our study presents, for the first time, an analysis of a large-scale multidisciplinary corpus under the construct of four named entity labels that are specifically defined and selected to be domain-independent as opposed to domain-specific. The work is then inadvertently a feasibility test of characterizing multidisciplinary science with domain-independent concepts. Further, to summarize the distinct facets of scientific knowledge per concept per discipline, a set of word cloud visualizations are offered. The STEM-NER-60k corpus, created in this work, comprises over 1 M extracted entities from 60k STEM articles obtained from a major publishing platform and is publicly released.

Keywords: named entity recognition; information extraction; scholarly knowledge graphs; STEM science; open research knowledge graph

Citation: D'Souza, J. Overview of STEM Science as Named Entities. *Knowledge* **2022**, *2*, 735–754. https://doi.org/10.3390/knowledge2040042

Academic Editor: Gwanggil Jeon

Received: 14 October 2022
Accepted: 13 December 2022
Published: 19 December 2022

1. Introduction

In the quest for knowledge [1], significant progress has been made toward the automated understanding of the meaning of text in the commonsense domain. Some state-of-the-art systems that power commonsense machine interpretability or readability are Babelfy [2], DBpedia Spotlight [3], NELL [4], and FRED [5], to name a few. In contrast, scholarly literature remains relatively understudied for its intelligible machine interpretability. Consequently, fine-grained scholarly knowledge remains largely inaccessible for machine reading. In terms of data standards, particularly the FAIR guiding principles for scientific data creation [6], this implies a wider unexplored scope for obtaining scientific resources that are findable, accessible, interpretable, and reusable.

There are a multitude of recent emerging large-scale initiatives to build specialised Scholarly Knowledge Graphs (SKGs) capable of serving specific user needs. Consider Google Scholar, Web of Science [7], Microsoft Academic Graph [8], Open Research Knowledge Graph [9,10], Semantic Scholar [11], etc. These initiatives already create a practical systems need for automated graph encoding methods (e.g., to associate IRIs as URIs or URLs with graph nodes for their RDFization—a W3C labeled graph standard).

We position this paper within the broader aim of supporting the machine reading of scientific terms multidisciplinarily as dereferencable Web resources. Specifically, in this work, we release a large-scale, multidisciplinary, structured dataset of scientific named entities called STEM-NER-60k comprising over a 1M extracted entities from 60k articles across the 10 most prolific STEM disciplines on Elsevier, viz. Agriculture (*agr*), Astronomy (*ast*), Biology (*bio*), Chemistry (*chem*), Computer Science (*cs*), Earth Science (*es*), Engineering (*eng*),

Materials Science (*ms*), and Mathematics (*mat*). The NER extraction objective was based on the STEM-ECR corpus in our prior work [12,13] which consisted of 110 abstracts from open access publications with each of the aforementioned 10 domains equally represented with 11 abstracts per domain. Each abstract was structured in terms of the scientific entities, where the entities were classified within a four-class formalism (viz. process, method, material, and data) and were resolved to Wikipedia and Wiktionary respectively. The 4-class formalism had the following objective: *to be restricted within the single broad domain of Science, while still bridging terms multidisciplinarily*, thus, in a sense, facilitating the validation of semantic adaptability of the concepts. To this end, while our prior gold-standard STEM-ECR corpus [13] initiated the development of semantically adaptable systems, the silver-standard STEM-NER-60k corpus of this work (https://github.com/jd-coderepos/stem-ner-60k, accessed on 12 December 2022) fosters the development of not just semantically adaptable, but also scalable solutions.

2. Background

Scholarly domain NER is not entirely new: the flagship Semantic Evaluation (SemEval) series has so far seen five related tasks organized [14–18]—however, *none of this work has been done so far in the broad multidisciplinary setting of Science*. Extending our prior work on this theme [12,13], we explore scholarly NER on a large-scale, silver-standard corpus of structured STEM articles which has *a wide-ranging application scope in the emerging field of the creation and discovery of SKGs that strive toward representing knowledge from scholarly articles in machine-interpretable form.* The characteristics of our corpus are unique and noteworthy. (1) The linguistic phenomenon of interdisciplinary word-sense switching is pervasive. For example, consider the term "the Cloud" which in *cs* takes on the meaning of a technological solution for hosting software, versus in *ast* where it takes the common interpretation of the mass of water vapor we see in the sky. (2) There is a seemingly evident shift of the sense interpretation of terms to take on our corpus domain-specific scientific word senses as opposed to their common sense interpretations which may be more widely known. For example, "power" in *mat* refers to exponentiation, which, otherwise in a common sense, takes on a human social interpretation. Thus, our work on multidisciplinary NER with semantically bridging concepts with regard to both our prior gold-standard STEM-ECR corpus (https://data.uni-hannover.de/dataset/stem-ecr-v1-0, accessed on 12 December 2022) and the STEM-NER-60k corpus, discussed in this paper, facilitates designing novel solutions attempting multidisciplinary NER within a generic four-class formalism capable of bridging semantic concepts multidisciplinarily.

The paper is structured in two main parts. First, related work in terms of existing class formalisms for annotating scientific entities is discussed. Second, insights into the STEM-NER-60k corpus released in this work (https://github.com/jd-coderepos/stem-ner-60k, accessed on 12 December 2022) are given.

3. Related Work: Scientific Named Entity Recognition (NER) Formalisms

The structuring of unstructured articles as an NER task has been taken up at a wide-scale in three scientific disciplines: Computer Science (CS), Biomedicine (Bio), and Chemistry (Chem).

In this section, we discuss the different NER conceptual formalisms defined in the three domains.

3.1. Computer Science NER (CS NER)

CS NER corpora can be compared along five dimensions: (1) domain, (2) annotation coverage, (3) semantic concepts, (4) size, and (5) annotation method. Most of the corpora consist of relatively short documents. The shortest is the CL-Titles corpus [19] with only paper titles. The longer ones have sentences from full-text articles, viz. ScienceIE [16], NLP-TDMS [20], SciREX [21], and ORKG-TDM [22]. We see that the corpora have had from one [18] to atmost seven NER concepts [23]. Each corpora' concepts purposefully

informs an overarching knowledge extraction objective. For example, the concepts focus, technique, and domain in the FTD corpus [24] helped examine the influence between research communities; ACL-RD-TEC [23] made possible a broader trends analysis with seven concepts. Eventually, corpora began to shed light on a novel scientific community research direction toward representing the entities as knowledge graphs [9] with hierarchical relation annotations such as synonymy [16] or semantic relations such 'Method *Used-for* a Task' [25]; otherwise, concepts were combined within full-fledged semantic constructs as Leaderboards with between three and four concepts [20–22,26], viz. research problem, dataset, method, metric, and score; or were in extraction objectives with solely contributions-focused entities of a paper [19,27,28].

3.2. Biomedical NER (BioNER)

BioNER predates CS NER. It was one of the earliest domains taken up for text mining of fine-grained entities from scholarly publications to enhance search engine performance in health applications. It aims to recognize concepts in bioscience and medicine. For example, protein, gene, disease, drug, tissue, body part and location of activity such as cell or organism. The most frequently used corpora are GENETAG (full-text articles annotated with protein/gene entities) [29], JNLPBA (~2400 abstracts annotated with DNA, RNA, protein, cell type and cell line concepts) [30], GENIA (~200 Medline abstracts annotated with 36 different concepts from the Genia ontology and several levels of linguistic/semantic features) [31], NCBI disease corpus (793 abstracts annotated with diseases in the MeSH taxonomy) [32], CRAFT (the second largest corpus with 97 full-text papers annotated with over 4000 corpus) [33] linking to the NCBI Taxonomy, the Protein, Gene, Cell, Sequence ontologies etc. Finally, the MedMentions corpus [34] as the largest dataset with ~4000 abstracts with ~34,724 concepts from the UMLS ontology. By leveraging ontologies such as the Gene Ontology [35], UMLS [36], MESH, or the NCBI Taxonomy [37], for the semantic concepts, these corpora build on years of careful knowledge representation work and are semantically consistent with a wide variety of other efforts that exploit these community resources. This differs from CS NER which is evolving toward standardized concepts.

Structured knowledge as knowledge bases (KB) was initially seen as necessary in organizing biomedical scientific findings. Biomedical NER was applied to build such KBs. For example, protein–protein interaction (PPI) databases as MINT [38] and IntAct [39] or the more detailed KBs as pathway [40] or Gene Ontology Annotation [41]. Community challenges help curate these KBs via text mining at a large-scale. For example, BioCreative for PPI [42,43], protein-mutation associations [44], and gene–disease relations [45]; or BioNLP [46] for complex n-ary bio events. CS NER is also been addressed in equivalent series such as SemEval [16–18] which is promising to foster rapid task progress.

3.3. Chemistry NER (ChemNER)

BioNER in part fosters Chemistry NER. Text mining for drug and chemical compound entities [47,48] is indispensable to mining chemical disease relations [49], and drug and chemical–protein interactions [50]. Obtaining this structured knowledge has implications in precision medicine, drug discovery as well as basic biomedical research. Corpora for ChemNER are Corbett et al.'s [51] dataset (42 full-text papers with ~7000 chemical entities), ChemDNER (10,000 PubMed abstracts with 84,355 chemical entities) [48], and NLM-Chem (150 full-text papers with 38,342 chemical entities normalized to 2064 MeSH identifiers) [52].

The high-level identification of entities in text is a staple of most modern NLP pipelines over commonsense knowledge. This, in the context of the scientific entities formalisms presented above is pertinent to scholarly knowledge as well. While being structurally domain-wise related, our work has a unique objective: *to extract entities across STEM Science which follow a generic four-entity conceptual formalism.* Thus apart from having an impact in the emerging field of the discovery of science graphs, the STEM-NER-60k corpus can have specific applications in higher-level NLP tasks including information extraction [53],

factuality ascertainment of statements in knowledge base population [54], and question answering over linked data [55].

4. Materials and Methods

4.1. Our STEM-NER-60k Corpus

We now describe our corpus in the following terms: (1) definitions of the four scientific concepts [13], viz. process, method, material, and data, are given; (2) the process by which the silver-standard STEM-NER-60k corpus is created is explained; and (3) corpus insights are offered specifically in terms of the multidisciplinary entities annotated under the formalism of four concepts bridging the 10 domains.

4.1.1. Concept Definitions

Following an iterative process of concept refinement [13] via a process that involved expert adjudication of which scientific concepts were multidisciplinarily semantically meaningful, the following four concepts were agreed upon to be relevant for entities across STEM, specifically across the following 10 domains, viz. *agri*, *ast*, *bio*, *chem*, *cs*, *es*, *eng*, *ms*, and *math*.

- Process. Natural phenomenon, or independent/dependent activities. For example, growing (*bio*), cured (*ms*), flooding (*es*).
- Method. A commonly used procedure that acts on entities. For example, powder X-ray (*chem*), the PRAM analysis (*cs*), magnetoencephalography (*med*).
- Material. A physical or digital entity used for scientific experiments. For example, soil (*agri*), the moon (*ast*), the set (*math*).
- Data. The data themselves, or quantitative or qualitative characteristics of entities. For example, rotational energy (*eng*), tensile strength (*ms*), vascular risk (*med*).

4.1.2. Corpus Creation

The silver-standard STEM-NER-60k corpus was created as follows. Roughly 60,000 articles in text format and restricted only to the articles with the CC-BY redistributable license on Elsevier were first downloaded https://tinyurl.com/60k-raw-dataset, accessed on 12 December 2022. Next, our aim was to obtain the four-concept entity annotations for the Abstracts in this corpus of publications. We leveraged our prior-developed state-of-the-art NER system for this purpose [12]. The Brack et al. [12] system was based on Beltagy et al.'s [56] SciBERT which in turn for the specific NER configuration makes use of the original BERT NER [57,58] prediction architecture. The Brack et al. [12] system, however, was pretrained on a smaller gold-standard STEM corpus [13] expert-annotated with the four generic scientific entities described in Section 4.1.1. Applying this system on the Abstracts on the newly downloaded 60 k articles produced the silver-standard STEM-NER-60k corpus of this work. The resulting silver-standard corpus statistics are shown in Table 1.

Table 1. STEM-NER-60k corpus statistics comprising 59,984 articles structured in terms of process, method, material, and data concepts.

	Articles	**Process**	**Method**	**Material**	**Data**
agriculture (agri)	4944	20,532	3252	62,043	33,608
astronomy (ast)	15,003	31,104	10,423	55,753	97,011
biology (bio)	9038	54,029	6777	100,454	43,418
chemistry (chem)	5232	18,185	6044	48,779	30,596
computer science (cs)	5389	17,014	13,650	35,554	33,575
earth science (es)	4363	28,432	5129	56,571	50,211
engineering (eng)	2441	12,705	3293	24,633	24,238
material science (ms)	2144	10,102	2437	23,054	16,981
mathematics (math)	1765	8002	1941	11,381	15,631
medicine (med)	15,054	89,637	19,580	148,059	134,249

Our corpus is publicly released (https://github.com/jd-coderepos/stem-ner-60k, accessed on 12 December 2022) to support related R&D endeavors and for other researchers interested in further investigating the task of scholarly NER.

4.1.3. Corpus Insights

Here, details of the STEM-NER-60k corpus are offered.

First, given our large-scale multidisciplinary scientific corpus, we take the opportunity to briefly examine the difference between scientific writing and non-scientific writing, if any, with the help of entropy formulations for each of our data domains. The entropies obtained for our corpus per domain is as follows: *med* (4.58) > *chem* (4.58) > *bio* (4.56) > *ast* (4.53) > *agri* (4.5) > *ms* (4.5) > *es* (4.48) > *eng* (4.42) > *math* (4.4). Intriguingly, these numbers are close to non-scientific English (4.11 bits [59]). Thus, given our corpus scientific English, one can rule out any atypical English usage syntax other than domain-specific jargon vocabulary.

Next, briefly, we address *which general scientific entity annotation patterns can one anticipate in our silver-standard corpus?* (1) Entity annotations can be expected as definite noun phrases whenever found. (2) Coreferring lexical units for scientific entities in the context of a single Abstract can be expected to be annotated with the same concept type. (3) Quantifiable lexical units such as numbers (e.g., years 1999, measurements 4 km) or even phrases (e.g., vascular risk) should be data. (4) Where found, the most precise text reference (i.e., including qualifiers) regarding materials used in the experiment should be annotated. For example, the term "carbon atoms in graphene" was annotated as a single material entity and not separately as "carbon atoms", "graphene". (5) The precedence of annotation of the scientific concepts in the original corpus in case of any confusion in classifying the four classes of scientific entities was resolved as follows: method > process > data > material, where the concept appearing earlier in the list was selected as the preferred class.

STEM scientific terms as process

Verbs (e.g., measured), verb phrases (e.g., integrating results), or noun phrases (e.g., an assessment, future changes, this transport process, the transfer) can be expected to be scientific entity candidates for process. Generally, it can be one of two things: an occurrence natural to the state/properties of the entity or an action performed by the investigators. In the latter case, however, it is better aptly expected as a method entity when the action is a named instance.

Some examples are offered for scientific entities as process candidates. (1) In the sentence, "The transfer of a laboratory process into a manufacturing facility is one of the most critical steps required for the large scale production of cell-based therapy products", the terms "The transfer", "a laboratory process", and "the large scale production" all are of type process. (2) In "The transterminator ion flow in the Venusian ionosphere is observed at solar minimum for the first time.", the terms "The transterminator ion flow" and "solar minimum" process entities. The verb "observed", however, is not considered a process since it doesn't act upon another object. (3) On the other hand, in "It is suggested that this ion flow contributes to maintaining the nightside ionosphere.", the terms "this ion flow" and "maintaining" are both considered valid process candidates. Finally, (4) in the sentence "Cellular morphology, pluripotency gene expression and differentiation into the three germ layers have been used compare the outcomes of manual and automated processes" the terms "pluripotency gene expression", "differentiation", "compare", and "manual and automated processes" are each annotated as process.

STEM scientific terms as method

Phrases which can be expected to be annotated as a method entity are those which contain the following trigger words: simulation, method, algorithm, scheme, technique, system, function, derivative, proportion, strategy, solver, experiment, test, computation, program. As an example, consider the sentence "Here finite-element modelling has demonstrated that once one silica nanoparticle debonds then its nearest neighbours are shielded

from the applied stress field, and hence may not debond." In this sentence, the term "finite-element modelling" is annotated as a method.

STEM scientific terms as material

This concept is expounded merely with the following three examples. (1) In "Based on the results of the LUCAS topsoil survey we performed an assessment of plant available P status of European croplands." the term "European croplands" should be material. (2) In "The transfer of a laboratory process into a manufacturing facility is one of the most critical steps required for the large scale production of cell-based therapy products." there are two material terms, viz. "a manufacturing facility" and "cell-based therapy products". Finally, (3) in the sentence "Cellular morphology, pluripotency gene expression and differentiation into the three germ layers have been used to compare the outcomes of manual and automated processes." the phrase "the three germ layers" is a material.

STEM scientific terms as data

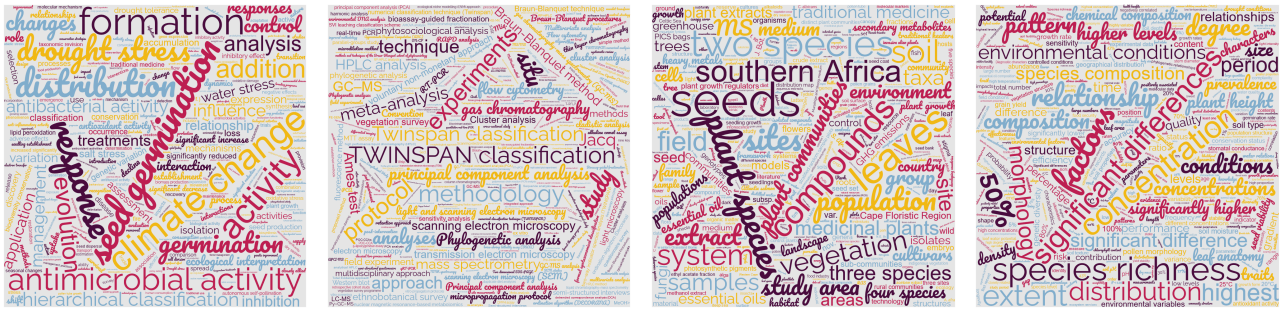
Phrases satisfying the patterns in the following examples can be expected to be data. (1) In "Based on the results of the LUCAS topsoil survey we performed an assessment of plant available P status of European croplands.", the phrases "the results" and "plant available P status" are considered as data in the original annotation scheme. (2) In "Our analysis shows a status of a baseline period of the years 2009 and 2012, while a repeated LUCAS topsoil survey can be a useful tool to monitor future changes of nutrient levels, including P in soils of the EU." the phrases: "a status of a baseline period", "nutrient levels", and "P" are data items. (3) Further, in "Observations near the terminator of the energies of ions of ionospheric origin showed asymmetry between the noon and midnight sectors, which indicated an antisunward ion flow with a velocity of (2.5 ± 1.5) kms^{-1}." the terms "asymmetry between the noon and midnight sectors", "a velocity", and "(2.5 ± 1.5) kms^{-1}" are data. Finally, (4) in "We established a P fertilizer need map based on integrating results from the two systems." the phrase "a P fertilizer need map" is data which should override the selection of "a P fertilizer" as material by the concept precedence stated earlier.

Discovered STEM scientific research trends based on the terminology of our corpus

A distinct characteristic of our corpus is that it enables uncovering the predominant research trends of the 10 considered research domains with respect to our four annotated concepts, i.e., process, method, material, and data. Answers to questions such as *which are the frequently applied processes, methods, materials, and data entities in any given research domain?* can be automatically discovered. Thus our four-concept information extraction task applied over 60,000 scholarly abstracts presents a new perspective on analyzing the dynamics of a research community [24] in terms of the most common research trends found across the descriptions in scholarly papers. The methodology to uncovering research trends that we adopt is straightforward. The STEM-NER-60k corpus is organized, for each of the 10 domains, as four lists of the extracted entities per concept and then sorted in descending order of their frequency of occurrence. Our claim is that the most common entities that surface to the top of the list reflect the predominant process, method, material, and data entities for the underlying domain. To our knowledge, thus far such insights have not been discussed in any other research endeavor. This subsection is devoted to a detailed discussion offering for the first time insights about the predominant research trends across 10 STEM disciplines with the help of the visual device of word clouds. Specifically, in Figures 1–10, word clouds of the top 100 entities for the 10 STEM disciplines are shown which form the subject of the subsequent discussions. Note, however, our annotations of the abstracts included original as well as coreferential entities as generic terms. The subsequent discussion is focused only on the original noun phrase mentions of the entities.

Agriculture domain

Starting with Agriculture (Figure 1), some highly researched process entities are seed germination, climate change, antimicrobial activity, or drought stress. TWINSPAN classification, ames test, phylogenetic analysis, or gas chromatography are commonly applied methods. Plant species, plant communities, leaves, or medicinal plants are common materials. data seems to be expressed in terms of percentages, species richness, time period, or size.



(**a**) process (**b**) method (**c**) material (**d**) data

Figure 1. Agriculture domain word clouds.

Astronomy domain

For Astronomy (Figure 2), some highly researched processes are proton collisions, string theory, neutrino oscillations, or first order phrase transition. Commonly applied methods are clinical perturbation theory, Lagrangian approach, quantum field theory, or LHCb experiment. Dark matter, Higgs boson, scalar field, or black hole are highly researched or used materials. data is commonly expressed as neutrino masses, cosmological constant, dark energy, or sensitivity.

(**a**) process (**b**) method (**c**) material (**d**) data

Figure 2. Astronomy domain word clouds.

Biology domain

In the Biology research domain (Figure 3), molecular mechanisms, oxidative stress, DNA replication, or cell death appear as highly researched processes. Commonly applied methods are in vitro, in vivo, flow cytometry, or mass spectrometry. Nucleus, stem cells, plasma membrane, or cancer cells are the highly researched or used materials. Frequent expressions of data are as increased risk, genome stability, therapeutic potential, or phenotype.

(**a**) process (**b**) method (**c**) material (**d**) data

Figure 3. Biology domain word clouds.

Chemistry domain

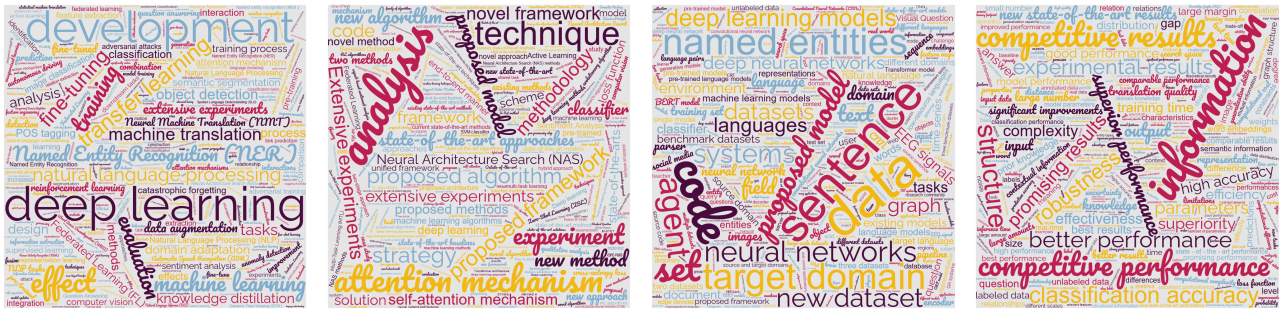
For the fourth Chemistry research domain (Figure 4), there appears some overlap with the Biology domain in terms of the entities. Some of the highly researched processes are oxidative stress, gene expression, "expression" generally, or relationship. Commonly applied methods are scanning electron microscopy, gas chromatography, cross-sectional study, or regression analysis. Frequently researched or used materials are proteins, drinking water, catalyst, or aqueous solution. Common data entities are temperature, concentrations, room temperature, or parameters.

(**a**) process (**b**) method (**c**) material (**d**) data

Figure 4. Chemistry domain word clouds.

Computer Science domain

Now the words clouds for the fifth STEM domain considered viz. Computer Science (Figure 5) are examined. Here we see notably different research trends for Computer Science versus the previous four STEM domains discussed viz. Astronomy, Agriculture, Biology, and Chemistry. In this domain, deep learning, fine-tuning, machine translation, or named entity recognition (ner) are some of the highly researched processes. Commonly applied methods are attention mechanism, neural architecture search (nas), loss function, or self-attention mechanism. Frequently researched or used materials are code, neural networks, sentence, or deep learning models. Common data entities are structure, information, classification accuracy, or translation quality.

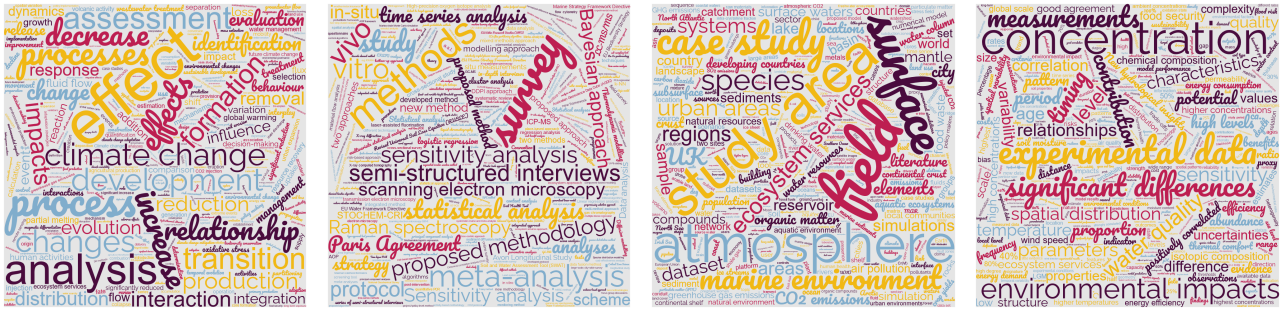


(**a**) process (**b**) method (**c**) material (**d**) data

Figure 5. Computer Science domain word clouds.

Earth Science domain

In our sixth STEM research domain, i.e., Earth Science (Figure 6), some of the highly researched processes are transition, evolution, formation, or reduction. Commonly applied methods are Paris agreement, survey, sensitivity analysis, or semi-structured interviews. Common researched or used materials are atmosphere, field, marine environment, or CO2 emissions. Common data entities are expressed as environmental impacts, spatial distribution, water quality, or energy efficiency.

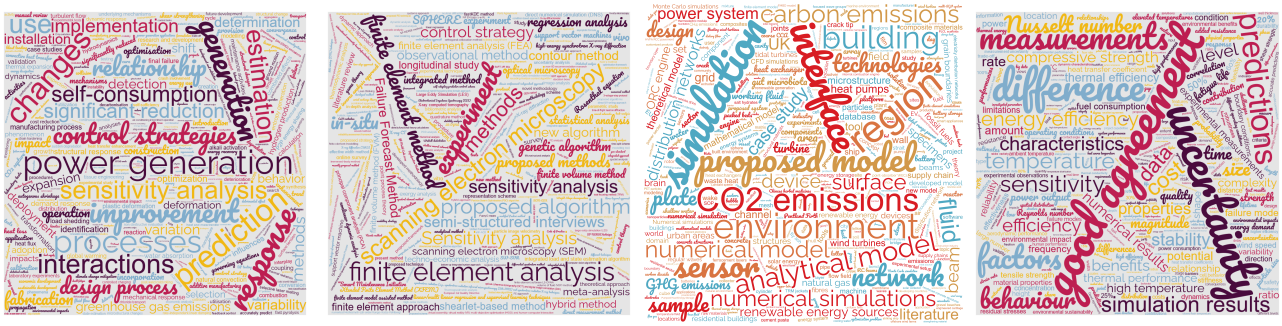


(**a**) process (**b**) method (**c**) material (**d**) data

Figure 6. Earth Science domain word clouds.

Engineering domain

In our seventh STEM research domain, i.e., Engineering (Figure 7), research trends show power generation, control strategies, design processes, or manufacturing process as predominant investigated processes. The frequently applied methods are finite element method, genetic algorithm, finite volume method, or finite element analysis. The highly researched or used materials are sensor, CO2 emissions, numerical simulations, or plate. Common data entities are expressed as good agreement, simulation results, uncertainty, or Nusselt number.

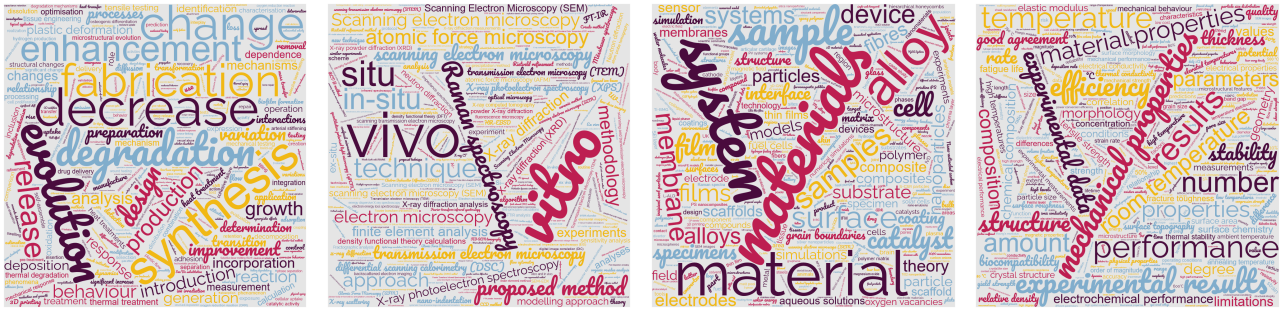


(**a**) process (**b**) method (**c**) material (**d**) data

Figure 7. Engineering domain word clouds.

Material Science domain

For the eighth STEM research domain, i.e., Material Science (Figure 8), research trends show fabrication, plastic deformation, thermal treatment, or heat treatment as predominant investigated processes. The frequently applied methods are Raman spectroscopy, in-situ, X-ray photoelectron spectroscopy, or atomic force microscopy. The highly researched or used materials are alloys, films, membranes, or particles. Common data entities are expressed as mechanical properties, material properties, electrochemical performance, or crystal structure.

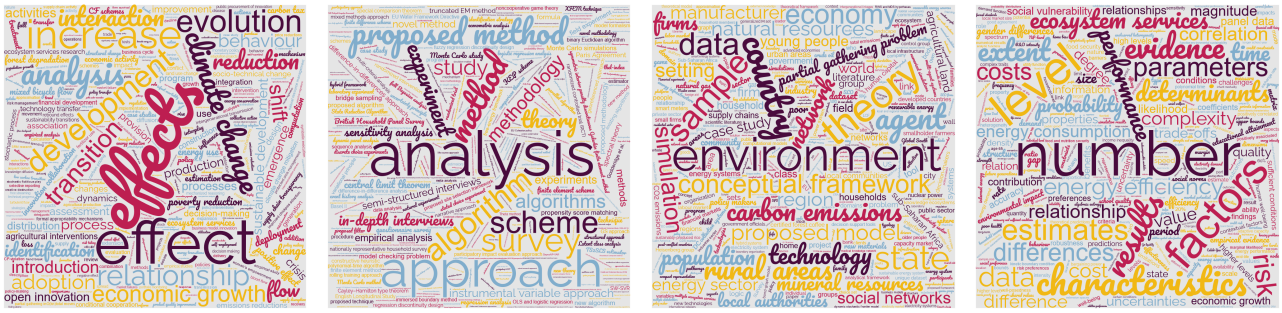


(**a**) process (**b**) method (**c**) material (**d**) data

Figure 8. Material Science domain word clouds.

Mathematics domain

In the ninth STEM research domain, i.e., Mathematics (Figure 9), economic growth, flow, open innovation, or poverty reduction are among the highly researched processes. The frequently applied methods are in-depth interviews, instrumental variable approach, Monte Carlo study, or full Bayesian analysis. Among the highly researched or used materials are mineral resources, social networks, firms, or energy sector. Common data entities are expressed as probability, complexity, estimates, or number.



(**a**) process (**b**) method (**c**) material (**d**) data

Figure 9. Mathematics domain word clouds.

Medicine domain

Finally, for the tenth STEM research domain considered in this work, i.e., Medicine, predominantly researched process entities are Alzheimer's disease, Parkinson's disease, disease progression, or cardiovascular disease. Common method entities included cross-sectional study, functional magnetic resonance imaging, vitro, or vivo. Frequent material entities that were either used or researched included patients, control group, brain, or general population. Common data entities are expressed as primary outcome, incidence, positively associated, or quality of life.

(**a**) process (**b**) method (**c**) material (**d**) data

Figure 10. Medicine domain word clouds.

5. Results

In this section, we offer some directions for practical applications developed based on STEM-ECR NER. With this we hope to offer inspiration for similar or related usages of the STEM-NER-60k corpus developed in this work.

STEM Entities Recommendation Service in the Open Research Knowledge Graph

The STEM-ECR annotation project (https://data.uni-hannover.de/dataset/stem-ecr-v1-0, accessed on 12 December 2022) was initiated to support the population of structured scientific concepts in the Open Research Knowledge Graph (ORKG). Figure 11 demonstrates the integration of our prior developed machine learning model of STEM entities [12] in the ORKG frontend. The model takes as input the Abstract of a new incoming publication and structures the Abstract per the four concepts. The user can then flexibly select and deselect annotations based on their prediction confidences or the user preference to automatically structure the contribution data of the work.

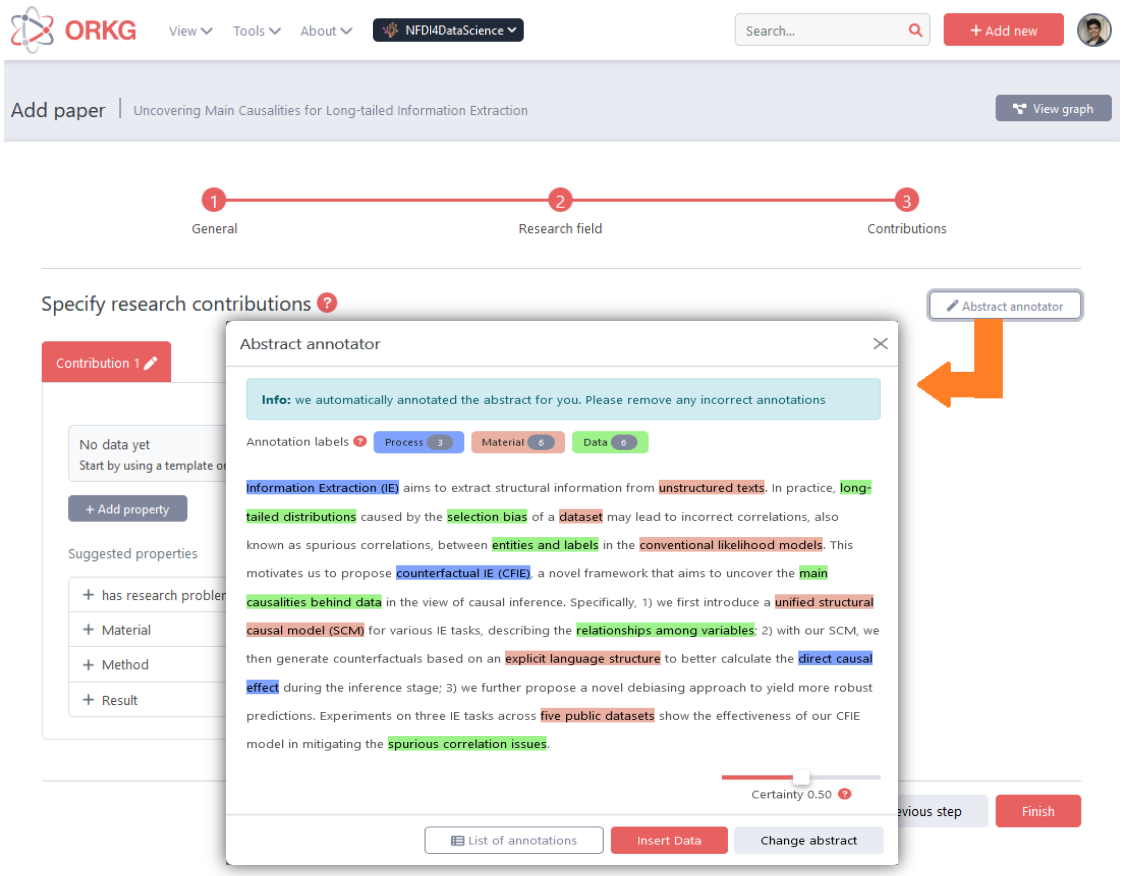


Figure 11. STEM Entities-based Abstract Annotator Recommendation Service for process, method, material, and data entities in the Add-paper wizard of the next-generation Open Research Knowledge Graph (ORKG) digital library front-end service.

6. Discussion

In this section, we discuss the potential of our entities corpus to be extended as a knowledge graph.

Knowledge Graph Construction for Fine-Grained Structured Search

Knowledge Graphs (KG) play a crucial role in many modern applications (https://developers.google.com/knowledge-graph, accessed on 12 December 2022) as solutions to the information access and search problem. There have been several initiatives in the NLP [18,25,60,61], and the Semantic Web [62,63], communities suggesting an increasing trend toward adoption of KGs for scientific articles. The automatic construction of KGs from text is a challenging problem, more so owing to the multidisciplinary nature of Science at large. While machines can better handle the volume of scientific literature, they need supervisory signals to determine which elements of the text have value. The STEM-NER-60k corpus can be leveraged to construct knowledge graphs underlying fine-grained search over publications. Figure 12 demonstrates an example KG that was manually annotated with relations between the entities. In the figure, the entity nodes are color-coded by their concept type: orange corresponds to process, purple for method, green corresponds to material, and blue for data.

As a practical illustration of the relation triples studied in this work, we build a knowledge graph from the annotations in the Combined corpus. This is depicted in Figure 12. Looking at the corpus-level graph (the right graph), we observe that generic scientific terms such as "method," "approach," and "system" are the most densely connected nodes, as expected since generic terms are found across research areas. In the zoomed-in ego-network of the term "machine_translation" (the left graph), Hyponym-

Of is meaningfully highlighted by its role linking "machine_translation" and its sibling nodes as the research tasks "speech_recognition," and "natural_language_generation" to the parent node "NLP_problems." The term "lexicon" is related by Usage to "machine_translation" and "operational_foreign_language." The Conjunction link joins the term "machine_translation" and "speech_recognization", both of which aim at translating information from one source to the other one. In summary, this knowledge graph can represent the relationships between scientific terms either at macro-level in terms of the whole corpus or at micro-level with respect to the ego-network of a specific concept.

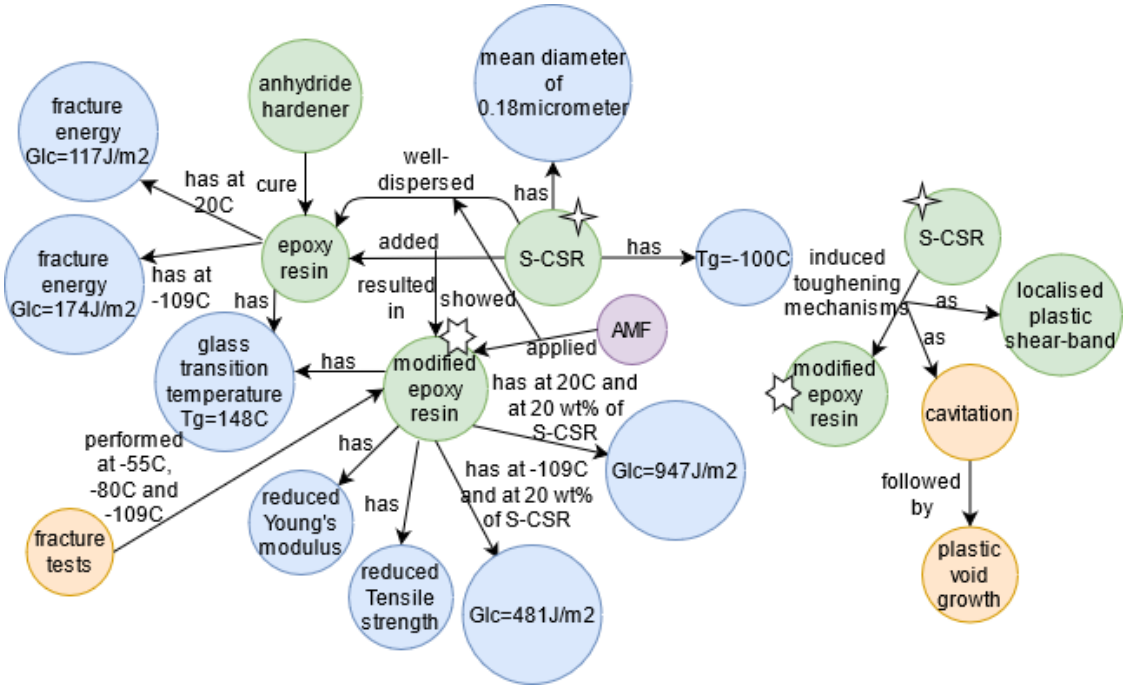


Figure 12. Structured Knowledge Graph (KG) representation of a *Material Science* domain publication Abstract [64] as process, method, material, and data typed entities. For KGs in the remaining 9 STEM domains we consider, see Appendix A.

7. Conclusions

In this paper, we have systematically introduced a large-scale multidisciplinary STEM corpus spanning 10 disciplines with structured abstracts in terms of generic process, method, material, and data entities. The corpus called STEM-NER-60k is publicly released for facilitating future research https://github.com/jd-coderepos/stem-ner-60k, accessed on 12 December 2022. Based on the application presented in Section 5 and the proposed knowledge graph extension in Section 6, we envision two main future research directions: (1) extending the set of generic concepts beyond the four proposed in this work. For example, our prior work [12,13] proposed concepts such as task, objective, and result. Additionally, (2) for automatic KG construction multidisciplinarily, annotating semantic relations between the entities.

Funding: Supported by TIB Leibniz Information Centre for Science and Technology, the EU H2020 ERC project ScienceGraph (GA ID: 819536) and the BMBF project SCINEXT (GA ID: 01lS22070).

Data Availability Statement: The dataset developed for this study can be found on the Github platform at https://github.com/jd-coderepos/stem-ner-60k.

Acknowledgments: The author would like to acknowledge the ORKG Team for support with implementing the ORKG frontend interfaces of the Abstract Annotator service.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Supplementary to the Knowledge Graph in the Material Science domain presented in Section 6, in Figures A1–A9, we demonstrate nine examples of knowledge graphs for the respective nine remaining STEM domains we consider. In all graphs, nodes are color-coded by their concept type: orange corresponds to process, purple for method, green corresponds to material, and blue for data.

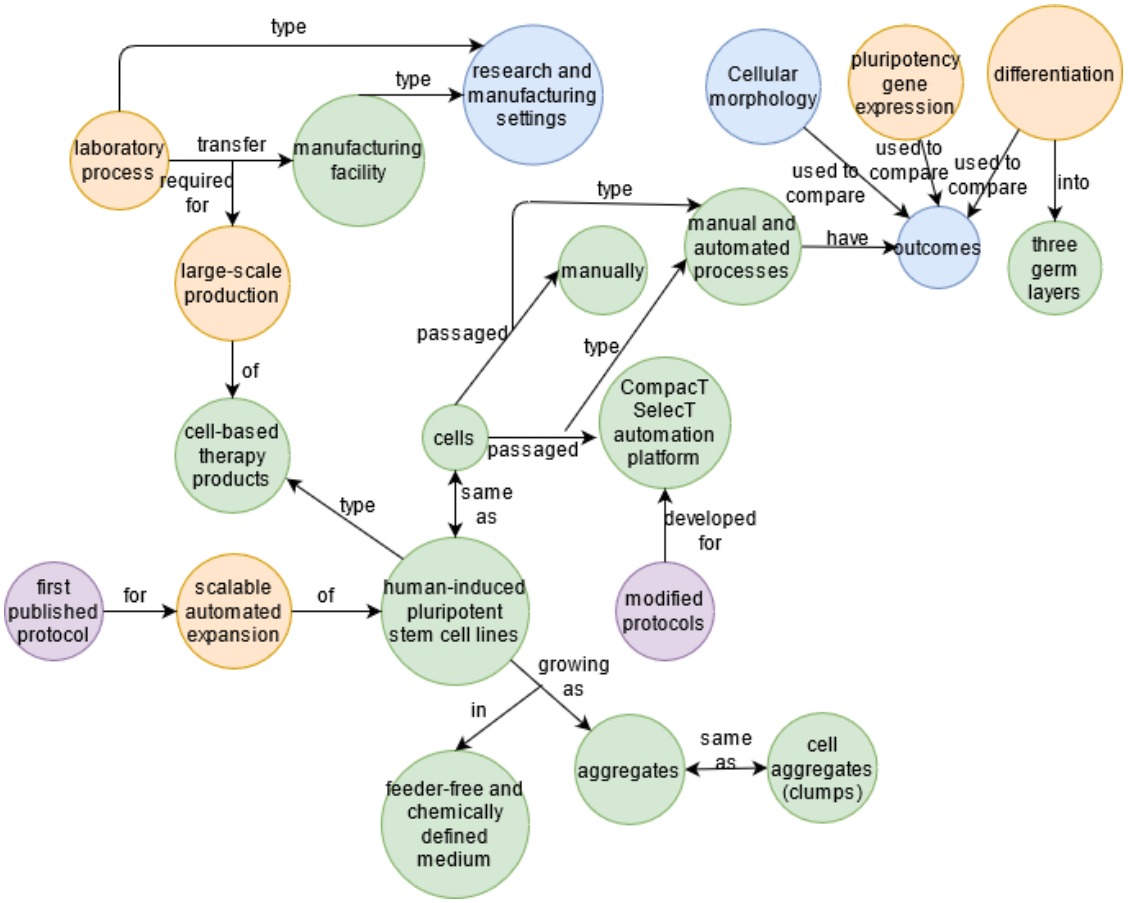


Figure A1. Structured KG representation of a Biology domain publication Abstract [65] as process, method, material, and data entities.

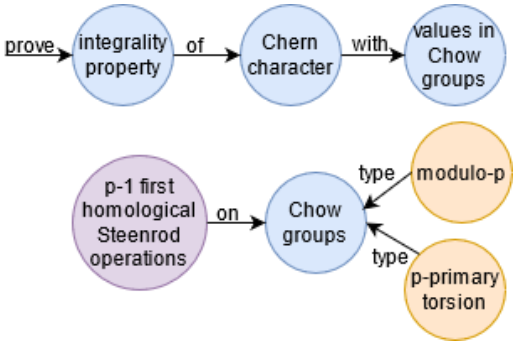


Figure A2. Structured KG representation of a Mathematics domain publication Abstract [66] as process, method, material, and data entities.

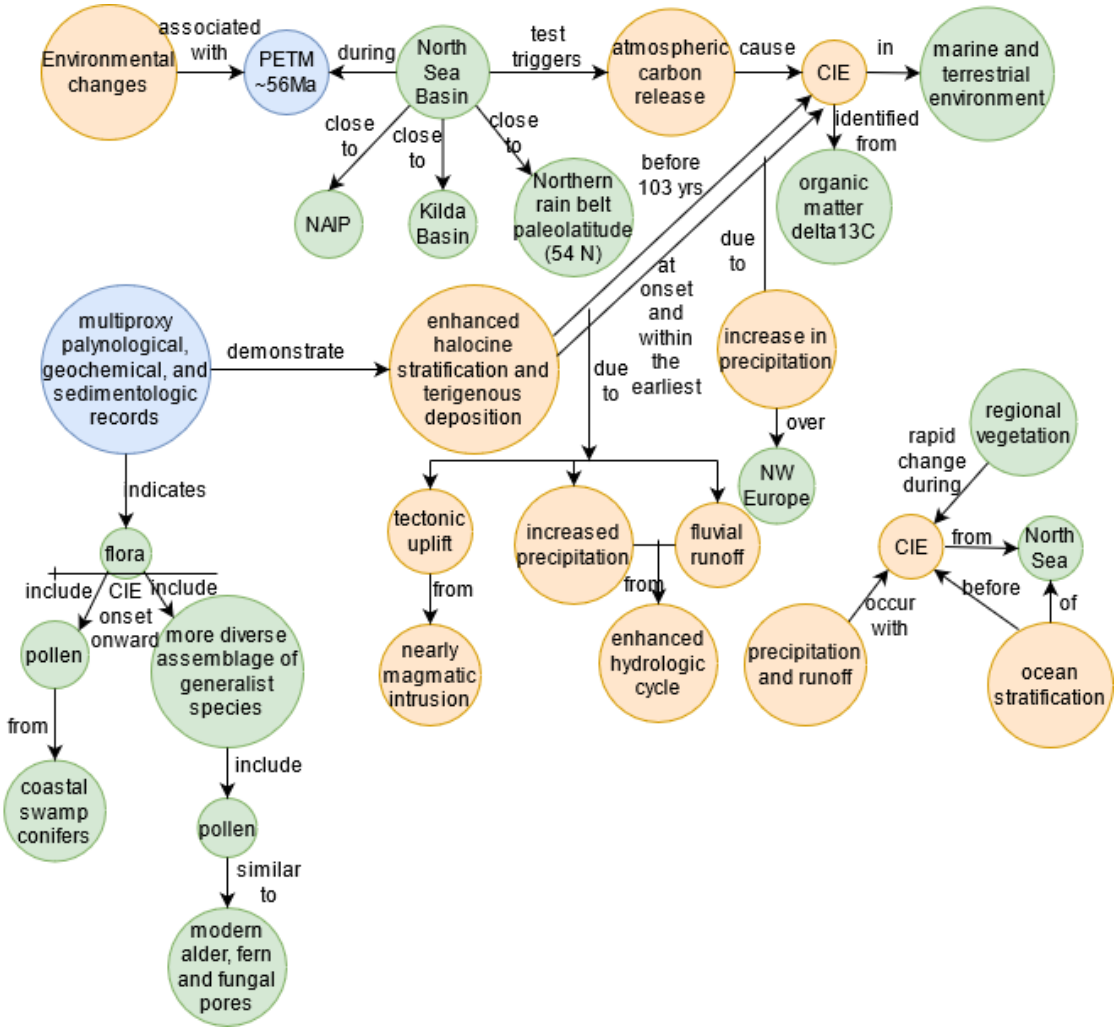


Figure A3. Structured KG representation of a Earth Science domain publication Abstract [67] as process, method, material, and data typed entities.

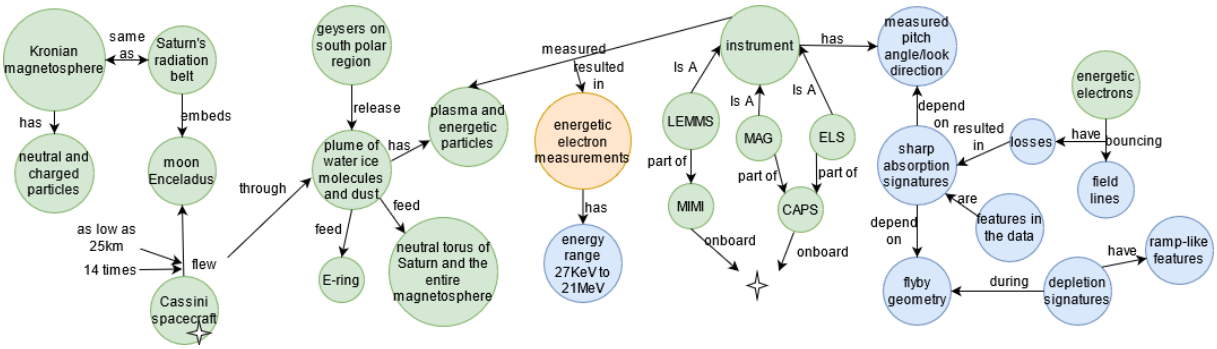


Figure A4. Structured KG representation of an Astronomy domain publication Abstract [68] as process, method, material, and data typed entities.

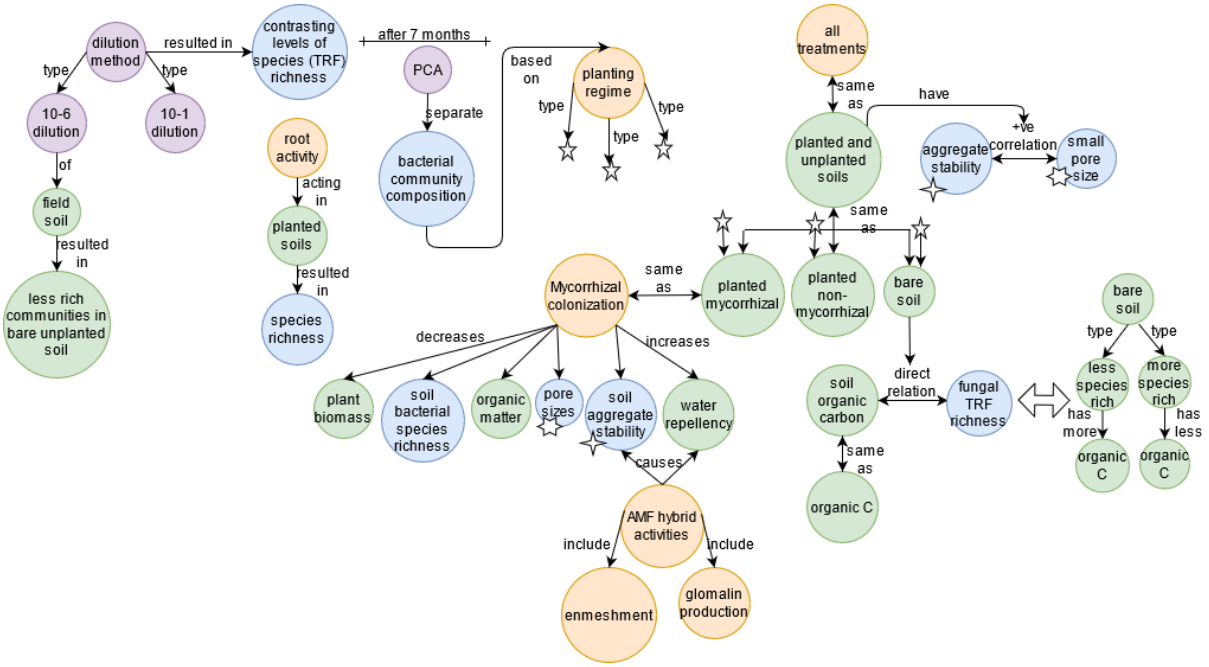


Figure A5. Structured KG representation of an Agriculture domain publication Abstract [69] as process, method, material, and data typed entities.

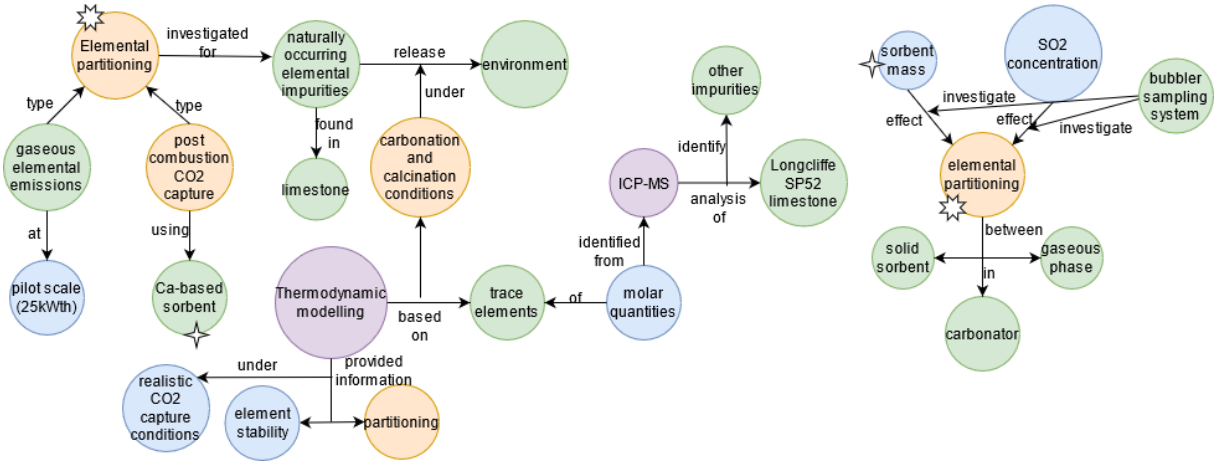


Figure A6. Structured KG representation of a Chemistry domain publication Abstract [70] as process, method, material, and data typed entities.

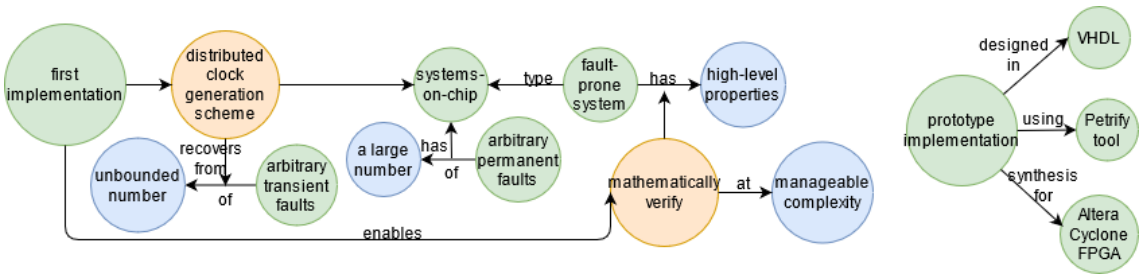


Figure A7. Structured KG representation of a Computer Science domain publication Abstract [71] as process, method, material, and data typed entities.

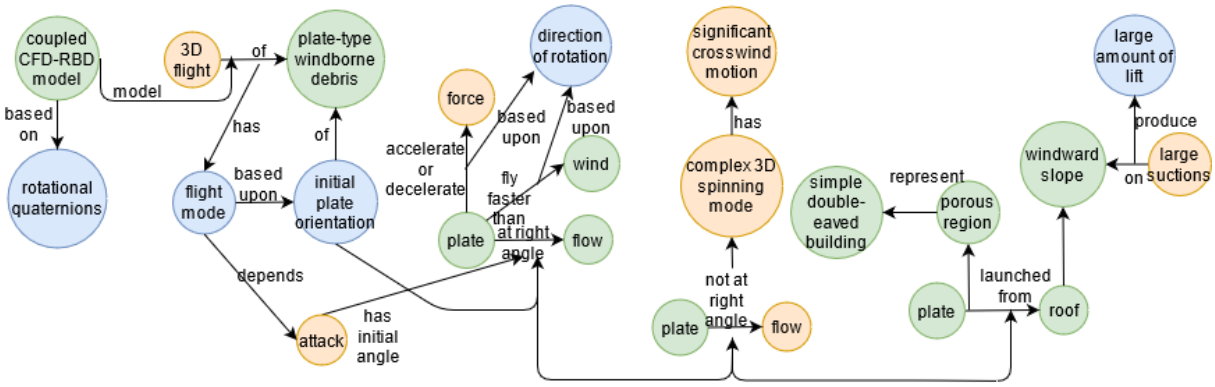


Figure A8. Structured KG representation of an Engineering domain publication Abstract [72] as process, method, material, and data typed entities.

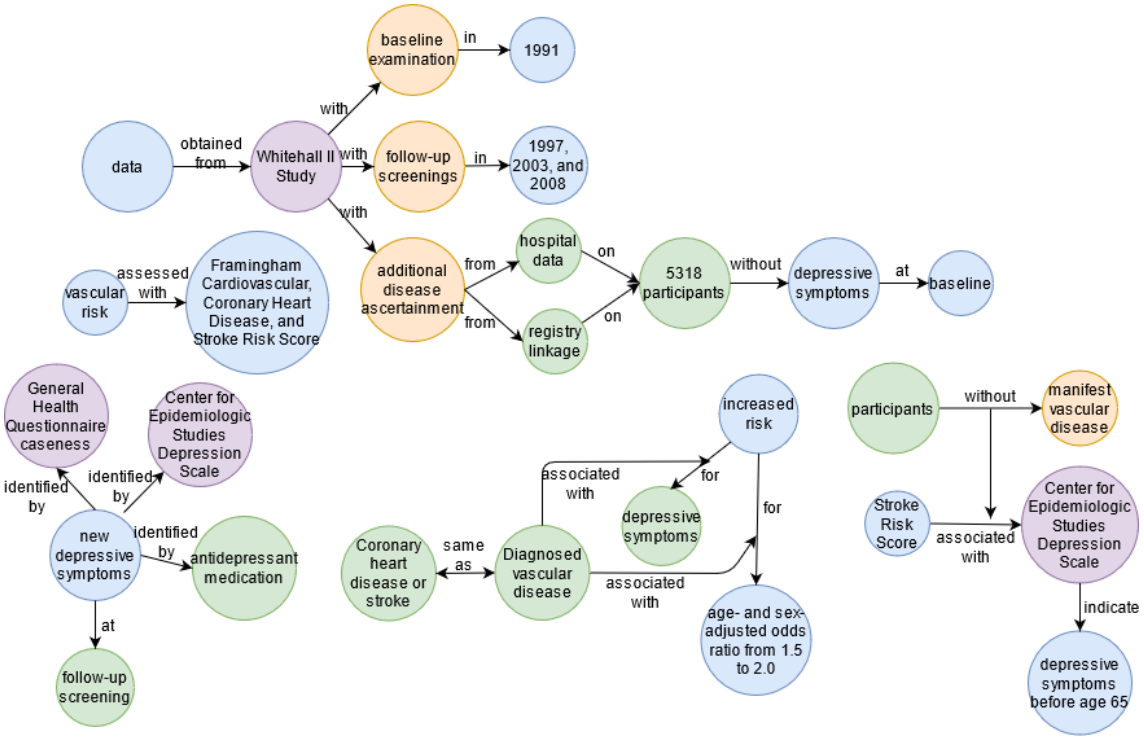


Figure A9. Structured KG representation of a Medical domain publication Abstract [73] as process, method, material, and data typed entities.

References

1. Schubert, L. Turing's dream and the knowledge challenge. In Proceedings of the National Conference on Artificial Intelligence, Boston, MA, USA, 16–20 June 2006; Volume 21, p. 1534.
2. Moro, A.; Cecconi, F.; Navigli, R. Multilingual Word Sense Disambiguation and Entity Linking for Everybody. In Proceedings of the ISWC, Downtown Seattle, WA, USA, 13–17 September 2014; pp. 25–28.
3. Mendes, P.N.; Jakob, M.; García-Silva, A.; Bizer, C. DBpedia spotlight: Shedding light on the web of documents. In Proceedings of the 7th International Conference on Semantic Systems, Graz, Austria, 7–9 September 2011; ACM: New York, NY, USA, 2011; pp. 1–8.
4. Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Yang, B.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; et al. Never-ending learning. *Commun. ACM* **2018**, *61*, 103–115. [CrossRef]
5. Gangemi, A.; Presutti, V.; Recupero, D.R.; Nuzzolese, A.G.; Draicchio, F.; Mongiovi, M. Semantic Web Machine Reading with FRED. *Semant. Web* **2017**, *8*, 873–893. [CrossRef]
6. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]

7. Birkle, C.; Pendlebury, D.A.; Schnell, J.; Adams, J. Web of Science as a data source for research on scientific and scholarly activity. *Quant. Sci. Stud.* **2020**, *1*, 363–376. [CrossRef]
8. Wang, K.; Shen, Z.; Huang, C.; Wu, C.H.; Dong, Y.; Kanakia, A. Microsoft academic graph: When experts are not enough. *Quant. Sci. Stud.* **2020**, *1*, 396–413. [CrossRef]
9. Auer, S. Towards an Open Research Knowledge Graph. *Ser. Libr.* **2018**, *76*, 35–41. [CrossRef]
10. Auer, S.; Oelen, A.; Haris, M.; Stocker, M.; D'Souza, J.; Farfar, K.E.; Vogt, L.; Prinz, M.; Wiens, V.; Jaradeh, M.Y. Improving access to scientific literature with knowledge graphs. *Bibl. Forsch. Und Prax.* **2020**, *44*, 516–529. [CrossRef]
11. Fricke, S. Semantic scholar. *J. Med Libr. Assoc.* **2018**, *106*, 145. [CrossRef]
12. Brack, A.; D'Souza, J.; Hoppe, A.; Auer, S.; Ewerth, R. Domain-independent extraction of scientific concepts from research articles. In Proceedings of the European Conference on Information Retrieval, Lisbon, Portugal, 14–17 April 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 251–266.
13. D'Souza, J.; Hoppe, A.; Brack, A.; Jaradeh, M.Y.; Auer, S.; Ewerth, R. The STEM-ECR Dataset: Grounding Scientific Entity References in STEM Scholarly Content to Authoritative Encyclopedic and Lexicographic Sources. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 13–15 May 2020; pp. 2192–2203.
14. Kim, S.N.; Medelyan, O.; Kan, M.Y.; Baldwin, T. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, 15–17 July 2010; pp. 21–26.
15. Moro, A.; Navigli, R. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, 4–5 June 2015; pp. 288–297.
16. Augenstein, I.; Das, M.; Riedel, S.; Vikraman, L.; McCallum, A. SemEval 2017 Task 10: ScienceIE—Extracting Keyphrases and Relations from Scientific Publications. In Proceedings of the SemEval@ACL, Vancouver, BC, Canada, 3–4 August 2017.
17. Gábor, K.; Buscaldi, D.; Schumann, A.K.; QasemiZadeh, B.; Zargayouna, H.; Charnois, T. Semeval-2018 Task 7: Semantic relation extraction and classification in scientific papers. In Proceedings of the SemEval, New Orleans, LA, USA, 5–6 June 2018; pp. 679–688.
18. D'Souza, J.; Auer, S.; Pedersen, T. SemEval-2021 Task 11: NLPContributionGraph-Structuring Scholarly NLP Contributions for a Research Knowledge Graph. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Bangkok, Thailand, 5–6 August 2021; pp. 364–376.
19. D'Souza, J.; Auer, S. Pattern-based acquisition of scientific entities from scholarly article titles. In Proceedings of the International Conference on Asian Digital Libraries, Virtual Event, 1–3 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 401–410.
20. Hou, Y.; Jochim, C.; Gleize, M.; Bonin, F.; Ganguly, D. Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 5203–5213. [CrossRef]
21. Jain, S.; van Zuylen, M.; Hajishirzi, H.; Beltagy, I. SciREX: A Challenge Dataset for Document-Level Information Extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7506–7516. [CrossRef]
22. Kabongo, S.; D'Souza, J.; Auer, S. Automated mining of leaderboards for empirical ai research. In Proceedings of the International Conference on Asian Digital Libraries, Virtual Event, 1–3 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 453–470.
23. QasemiZadeh, B.; Schumann, A.K. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoroz, Slovenia, 23–28 May 2016; pp. 1862–1868.
24. Gupta, S.; Manning, C. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, 8–23 November 2011; Asian Federation of Natural Language Processing: Chiang Mai, Thailand, 2011; pp. 1–9.
25. Luan, Y.; He, L.; Ostendorf, M.; Hajishirzi, H. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3219–3232.
26. Mondal, I.; Hou, Y.; Jochim, C. End-to-End Construction of NLP Knowledge Graph. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 1885–1895. [CrossRef]
27. Färber, M.; Albers, A.; Schüber, F. Identifying used methods and datasets in scientific publications. In Proceedings of the Workshop on Scientific Document Understanding: Co-located with 35th AAAI Conference on Artificial Inteligence (AAAI 2021), Online, 9 February 2021 .
28. D'Souza, J.; Auer, S. Computer Science Named Entity Recognition in the Open Research Knowledge Graph. In Proceedings of the International Conference on Asian Digital Libraries, Hybrid Event, 30 November–2 December 2022; Springer: Berlin/Heidelberg, Germany, 2021; pp. 35–45. ._3. [CrossRef]
29. Tanabe, L.; Xie, N.; Thom, L.H.; Matten, W.; Wilbur, W.J. GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinform.* **2005**, *6*, S3. [CrossRef]

30. Collier, N.; Kim, J.D. Introduction to the Bio-entity Recognition Task at JNLPBA. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), Geneva, Switzerland, 28–29 August 2004; pp. 73–78.
31. Kim, J.D.; Ohta, T.; Tateisi, Y.; Tsujii, J. GENIA corpus—A semantically annotated corpus for bio-textmining. *Bioinformatics* **2003**, *19*, i180–i182. [CrossRef]
32. Doğan, R.I.; Leaman, R.; Lu, Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **2014**, *47*, 1–10. [CrossRef]
33. Bada, M.; Eckert, M.; Evans, D.; Garcia, K.; Shipley, K.; Sitnikov, D.; Baumgartner, W.A.; Cohen, K.B.; Verspoor, K.; Blake, J.A.; et al. Concept annotation in the CRAFT corpus. *BMC Bioinform.* **2012**, *13*, 161. [CrossRef] [PubMed]
34. Mohan, S.; Li, D. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In Proceedings of the Automated Knowledge Base Construction (AKBC), San Diego, CA, USA, 17 June 2016.
35. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
36. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267. [CrossRef] [PubMed]
37. Schoch, C.L.; Ciufo, S.; Domrachev, M.; Hotton, C.L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; et al. NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database* **2020**, *2020*, baaa062. Available online: https://academic.oup.com/database/article-pdf/doi/10.1093/database/baaa062/33570620/baaa062.pdf (accessed on 12 December 2022). [CrossRef]
38. Chatr-Aryamontri, A.; Ceol, A.; Palazzi, L.M.; Nardelli, G.; Schneider, M.V.; Castagnoli, L.; Cesareni, G. MINT: The Molecular INTeraction database. *Nucleic Acids Res.* **2007**, *35*, D572–D574. [CrossRef] [PubMed]
39. Kerrien, S.; Alam-Faruque, Y.; Aranda, B.; Bancarz, I.; Bridge, A.; Derow, C.; Dimmer, E.; Feuermann, M.; Friedrichsen, A.; Huntley, R.; et al. IntAct—Open source resource for molecular interaction data. *Nucleic Acids Res.* **2007**, *35*, D561–D565. [CrossRef]
40. Bader, G.D.; Cary, M.P.; Sander, C. Pathguide: A pathway resource list. *Nucleic Acids Res.* **2006**, *34*, D504–D506. [CrossRef]
41. Camon, E.; Magrane, M.; Barrell, D.; Lee, V.; Dimmer, E.; Maslen, J.; Binns, D.; Harte, N.; Lopez, R.; Apweiler, R. The Gene Ontology Annotation (GOA) Database: Sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* **2004**, *32*, D262. [CrossRef]
42. Krallinger, M.; Leitner, F.; Rodriguez-Penagos, C.; Valencia, A. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol.* **2008**, *9*, S4. [CrossRef]
43. Krallinger, M.; Vazquez, M.; Leitner, F.; Salgado, D.; Chatr-Aryamontri, A.; Winter, A.; Perfetto, L.; Briganti, L.; Licata, L.; Iannuccelli, M.; et al. The Protein-Protein Interaction tasks of BioCreative III: Classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinform.* **2011**, *12*, S3. [CrossRef]
44. Krallinger, M.; Izarzugaza, J.M.; Rodriguez-Penagos, C.; Valencia, A. Extraction of human kinase mutations from literature, databases and genotyping studies. *BMC Bioinform.* **2009**, *10*, S1. [CrossRef] [PubMed]
45. Krallinger, M.; Leitner, F.; Valencia, A. Analysis of biological processes and diseases using text mining approaches. *Bioinform. Methods Clin. Res.* **2010**, 341–382.
46. Kim, J.D.; Ohta, T.; Pyysalo, S.; Kano, Y.; Tsujii, J. Extracting bio-molecular events from literature—The BIONLP'09 shared task. *Comput. Intell.* **2011**, *27*, 513–540. [CrossRef]
47. Herrero-Zazo, M.; Segura-Bedmar, I.; Martínez, P.; Declerck, T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *J. Biomed. Inform.* **2013**, *46*, 914–920. [CrossRef]
48. Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D.M.; et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminform.* **2015**, *7*, S2. [CrossRef]
49. Li, J.; Sun, Y.; Johnson, R.J.; Sciaky, D.; Wei, C.H.; Leaman, R.; Davis, A.P.; Mattingly, C.J.; Wiegers, T.C.; Lu, Z. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database* **2016**, *2016*, baw068. [CrossRef]
50. Krallinger, M.; Miranda, A.; Mehryary, F.; Luoma, J.; Pyysalo, S.; Valencia, A. DrugProt Shared Task (BioCreative VII track 1-2021) Text Mining Drug-Protein/Gene Interactions (DrugProt) Shared Task. 2021. Available online: https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-1/ (accessed on 13 October 2022).
51. Corbett, P.; Batchelor, C.; Teufel, S. Annotation of chemical named entities. In Proceedings of the Biological, Translational, and Clinical Language Processing, Prague, Czech Republic, 7–10 June 2007; pp. 57–64.
52. Islamaj, R.; Leaman, R.; Kim, S.; Kwon, D.; Wei, C.H.; Comeau, D.C.; Peng, Y.; Cissel, D.; Coss, C.; Fisher, C.; et al. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci. Data* **2021**, *8*, 91. [CrossRef]
53. Shah, P.K.; Perez-Iratxeta, C.; Bork, P.; Andrade, M.A. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinform.* **2003**, *4*, 20. [CrossRef]
54. Adel, H. Deep Learning Methods for Knowledge Base Population. Ph.D. Thesis, LMU Munchen, Munchen, Germany, 2018.
55. Unger, C.; Forascu, C.; Lopez, V.; Ngomo, A.C.N.; Cabrio, E.; Cimiano, P.; Walter, S. Question Answering over Linked Data (QALD-4). 2014. Available online: https://pub.uni-bielefeld.de/record/2763516 (accessed on 13 October 2022).
56. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A pretrained language model for scientific text. In Proceedings of the EMNLP-IJCNLP, Hong Kong, China, 3–7 November 2019; pp. 3606–3611.

57. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
58. Ma, X.; Hovy, E.H. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *arXiv* **2016**, arXiv:1603.01354.
59. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
60. Ammar, W.; Groeneveld, D.; Bhagavatula, C.; Beltagy, I.; Crawford, M.; Downey, D.; Dunkelberger, J.; Elgohary, A.; Feldman, S.; Ha, V.; et al. Construction of the Literature Graph in Semantic Scholar. In Proceedings of the NAACL-HLT (3), New Orleans, LA, USA, 1–6 June 2018.
61. D'Souza, J.; Auer, S. Sentence, phrase, and triple annotations to build a knowledge graph of natural language processing contributions—A trial dataset. *J. Data Inf. Sci.* **2021**, *6*, 6–34. [CrossRef]
62. Auer, S.; Kovtun, V.; Prinz, M.; Kasprzik, A.; Stocker, M.; Vidal, M.E. Towards a knowledge graph for science. In Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, Novi Sad, Serbia, 25–27 June 2018; pp. 1–6.
63. Jaradeh, M.Y.; Oelen, A.; Farfar, K.E.; Prinz, M.; D'Souza, J.; Kismihók, G.; Stocker, M.; Auer, S. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. In Proceedings of the 10th International Conference on Knowledge Capture, New York, NY, USA, 19–21 November 2019; pp. 243–246. [CrossRef]
64. Chen, J.; Kinloch, A.; Sprenger, S.; Taylor, A. The mechanical properties and toughening mechanisms of an epoxy polymer modified with polysiloxane-based core-shell particles. *Polymer* **2013**, *54*, 4276–4289. [CrossRef]
65. Soares, F.A.; Chandra, A.; Thomas, R.J.; Pedersen, R.A.; Vallier, L.; Williams, D.J. Investigating the feasibility of scale up and automation of human induced pluripotent stem cells cultured in aggregates in feeder free conditions. *J. Biotechnol.* **2014**, *173*, 53–58. [CrossRef] [PubMed]
66. Haution, O. Integrality of the Chern character in small codimension. *Adv. Math.* **2012**, *231*, 855–878. [CrossRef]
67. Kender, S.; Stephenson, M.H.; Riding, J.B.; Leng, M.J.; Knox, R.W.; Peck, V.L.; Kendrick, C.P.; Ellis, M.A.; Vane, C.H.; Jamieson, R. Marine and terrestrial environmental changes in NW Europe preceding carbon release at the Paleocene–Eocene transition. *Earth Planet. Sci. Lett.* **2012**, *353*, 108–120. [CrossRef]
68. Krupp, N.; Roussos, E.; Kollmann, P.; Paranicas, C.; Mitchell, D.; Krimigis, S.; Rymer, A.; Jones, G.; Arridge, C.; Armstrong, T.; et al. The Cassini Enceladus encounters 2005–2010 in the view of energetic electron measurements. *Icarus* **2012**, *218*, 433–447. [CrossRef]
69. Martin, S.L.; Mooney, S.J.; Dickinson, M.J.; West, H.M. Soil structural responses to alterations in soil microbiota induced by the dilution method and mycorrhizal fungal inoculation. *Pedobiologia* **2012**, *55*, 271–281. [CrossRef]
70. Cotton, A.; Patchigolla, K.; Oakey, J.E. Minor and trace element emissions from post-combustion CO2 capture from coal: Experimental and equilibrium calculations. *Fuel* **2014**, *117*, 391–407. [CrossRef]
71. Dolev, D.; Függer, M.; Posch, M.; Schmid, U.; Steininger, A.; Lenzen, C. Rigorously modeling self-stabilizing fault-tolerant circuits: An ultra-robust clocking scheme for systems-on-chip. *J. Comput. Syst. Sci.* **2014**, *80*, 860–900. [CrossRef]
72. Kakimpa, B.; Hargreaves, D.; Owen, J. An investigation of plate-type windborne debris flight using coupled CFD–RBD models. Part II: Free and constrained flight. *J. Wind. Eng. Ind. Aerodyn.* **2012**, *111*, 104–116. [CrossRef]
73. Kivimäki, M.; Shipley, M.J.; Allan, C.L.; Sexton, C.E.; Jokela, M.; Virtanen, M.; Tiemeier, H.; Ebmeier, K.P.; Singh-Manoux, A. Vascular risk status as a predictor of later-life depressive symptoms: A cohort study. *Biol. Psychiatry* **2012**, *72*, 324–330. [CrossRef] [PubMed]