Annotation Guidelines for the STEM-ECR Dataset V1.0

A Reference Dataset for Scientific Entity Extraction, Classification, and Resolution in Science, Technology, Engineering, and Mathematics

Abstract

The annotation task: Given a scientific article abstract consisting of a set of sentences, the goal is to mark the boundaries of *scientific entities* in each sentence. These annotated entities are then classified as one of a set of predefined *concept categories* (we consider four), and finally perform *entity linking* for the linkable entities to Wikipedia and Wiktionary for their encyclopedic and lexicographic enrichment in the real-world.

Purpose of this document: This document provides readers with basic knowledge that is required to successfully perform the annotation task defined above.

Document organisation: This document is organised in five parts. Sections 1 and 2 provide the introduction, and the task background along with basic definitions. After reading these sections, an annotator is assumed to understand our interpretation of scientific entities which form the core constituent in the annotation task facilitating later steps.
Section 3 provides some basic definitions regarding the annotation task. Section 4 provides information about the data used in the annotation task. Section 5 describes the annotation scheme intended to help annotators when identifying entities, marking term boundaries, assigning scientific concept classes, and linking entities with examples. Section 6 provides information about recommended tools for annotators to perform the task.

1. Introduction

These guidelines discuss the annotation of scientific entities in the context of the Open Research Knowledge Graph (ORKG)[1] project. The goal of this work is to produce an annotated corpus to facilitate the evaluation of natural language processing (NLP) techniques for automatically extracting entities from multidisciplinary scientific articles, and later classifying or linking and disambiguating them, in order to facilitate constructing a Scientific Knowledge Graph. We divide this annotation task into two parts: 1) scientific entity identification and classification; and 2) scientific entity resolution including entity linking (EL) to Wikipedia and entity word sense disambiguation (WSD) to Wiktionary.

For part 1, based on prior work,[2,3,4] these guidelines define four *generic* scientific concepts, viz. ***process***, ***method***, ***material***, and ***data***, for use by annotators in annotating multidisciplinary scientific entities.
For part 2, the guidelines describe a three-step annotation procedure to arrive at the EL and WSD annotations.

2. Background

Knowledge Graphs (KG) play a crucial role in many modern applications[5] as solutions to the information access and search problem. There have been several initiatives in the NLP[6,7] and the Semantic Web[8,9] 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.

These guidelines should guide users through the process of annotating a multidisciplinary corpus with entities geared towards supplying the needed signals at a multidisciplinary scale, specifically as the following two information units: 1) spans of scientific entities in at least the following 10 domains in Science (viz. Agriculture, Astronomy, Biology, Chemistry, Computer Science, Earth Science, Engineering, Materials Science, and Mathematics); and 2) entity linking annotations for the scientific entities to Wikipedia and disambiguated to Wiktionary.

---

[1] https://projects.tib.eu/orkg/

[2] QasemiZadeh, Behrang, and Anne-Kathrin Schumann. "The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods." *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. 2016.

[3] Augenstein, Isabelle, et al. "SemEval 2017 Task 10: ScienceIE-Extracting Keyphrases and Relations from Scientific Publications." *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 2017.

[4] Luan, Yi, et al. "Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.

[5] https://developers.google.com/knowledge-graph

[6] Ammar, Waleed, et al. "Construction of the Literature Graph in Semantic Scholar." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. 2018.

[7] Luan, Yi, et al. "Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction." *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018.

[8] Auer, Sören, et al. "Towards a knowledge graph for science." *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. ACM, 2018.

[9] Jaradeh, Mohamad Yaser et al. 2019. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. In Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19). ACM, New York, NY, USA, 243-246. DOI: https://doi.org/10.1145/3360901.3364435

3. Basic Definitions

This section provides verbose definitions for the annotators including: the four considered scientific concepts for classifying entities; and the linking of entities to the external knowledge sources.

Scientific Entities and Concepts: We allude to the concept formalism of Process, Method, Material, or Data as defined in Table 1 to define a scientific entity, where Process, Method, Material, or Data serve as the concept type for the annotated entity.

Concept	Definition	Examples
Process	Natural phenomenon, or independent/dependent activities.	growing (*Biology*), cured (*Materials Science*), flooding (*Earth Science*)
Method	A commonly used procedure that acts on entities.	powder X-ray (*Chemistry*), the PRAM analysis (*Computer Science*), magnetoencephalography (*Medicine*)
Material	A physical or digital entity used for scientific experiments.	soil (*Agriculture*), the moon (*Astronomy*), the set (*Mathematics*)
Data	The data themselves, or quantitative or qualitative characteristics of entities.	rotational energy (*Engineering*), tensile strength (*Material Sciences*), vascular risk (*Medicine*)

Table 1: Scientific Entity Concepts considered in this study

Further, the annotated scientific entities need to be linked to the real world concepts found in domain-agnostic knowledge sources, if possible. For scientific entities to be semantically machine-interpretable, we need the lexical knowledge present in collaboratively constructed knowledge sources such as Wikipedia and Wiktionary. To this end, we adopt BabelNet's[10] integrated view of a lexical item's complementary encyclopedic and lexicographic roles.

Linking Scientific Entities: Of the annotated scientific entities, the ones that are found in Wikipedia or Wiktionary are linked and annotated with their corresponding unique identifier from a specific time-stamped Wiki data release.

Linking entities to these knowledge sources makes the entity annotations locatable in the real world, without which the annotations risk becoming somewhat random since they are based on subjective decisions in the first step.

[10] Navigli, Roberto, and Simone Paolo Ponzetto. "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." *Artificial Intelligence* 193 (2012): 217-250.

4. Dataset

Of other data sources, 110 multidisciplinary scientific article abstracts are made available in the Open Access Corpus of Scientific, Technical, and Medical Content (OA-STM)[11] provided by Elsevier. The OA-STM corpus is a selection of 11 articles from 10 different STEM domains that are the most published. The domains are Agriculture, Astronomy, Biology, Chemistry, Computer Science, Earth Science, Engineering, Materials Science, Math, and Medicine. These guidelines have been developed based on an annotation task trials conducted on this data, hence all future references are made to domains defined in the OA-STM copus. Nevertheless, the annotation guidelines are generally portable to articles and domains other than the ones we consider.

5. The Annotation Scheme

5.1 Annotating Scientific Entities

Identification of scientific entities as Process, Method, Material, and Data concepts: The four concepts considered for the annotation of scientific entities are defined in Table 1.

Unlike most previous attempts in scientific entity annotations, annotations in STEM-ECR 1) are attempted at a consistent fine-grained detail for its considered entities; and 2) incorporate 10 different STEM (Science, Technology, Engineering and Mathematics) domains. These are described in some detail in the subsequent sections. The overall features characterizing our annotation task for scientific entities are summarized below:

1. To ensure consistent scientific entity spans, entities are annotated as definite noun phrases whenever possible. In later stages, the extraneous determiners and articles are dropped.
2. Coreferring lexical units for scientific entities in the context of a single abstract are annotated with the same concept type.
3. Quantifiable lexical units such as numbers (e.g., years 1999, measurements 4km) or even phrases (e.g., vascular risk) are annotated as *Data*.
4. Where possible, the most precise text reference (i.e. including qualifiers) regarding materials used in the experiment are annotated. For instance, “carbon atoms in graphene” will be annotated as one *Material* entity and not separately as “carbon atoms”, “graphene”.
5. Any confusion in classifying scientific entities as one of four types is resolved using the following concept precedence: Method > Process > Data > Material, where the concept appearing earlier in the list is preferred.

With the annotation conceptual framework now in place, the detailed annotation scheme is presented next for each scientific concept type. The annotation scheme is described with the help of examples for inclusion and exclusion.

5.1.1 Annotating Scientific Entities as *Process*

[11] https://github.com/elsevierlabs/OA-STM-Corpus

Heuristics for identifying *Process* candidates

a. Verbs (e.g., measured), verb phrases (e.g., integrating results), or noun phrases (e.g., an assessment, future changes, this transport process, the transfer) are scientific entity candidates for *Process*.
b. *Process* 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 a *Method* when the action is a named instance.

Examples

1) 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 transfer”, “a laboratory process”, and “the large scale production” are each annotated as *Process*

2) The transterminator ion flow in the Venusian ionosphere is observed at solar minimum for the first time.

“The transterminator ion flow” and “solar minimum” are annotated are *Process*

The verb “observed” is not annotated as *Process* since it doesn’t act upon another object.

3) It is suggested that this ion flow contributes to maintaining the nightside ionosphere.

“this ion flow” and “maintaining” are annotated as *Process*

4) Modified protocols were developed for the automated system.

“Modified protocols” is annotated as *Process*.

The verb “developed” is not annotated as *Process* since it does not act upon another object.

5) The management of cells aggregates (clumps) was identified as the critical step.

“The management” is annotated as *Process*

The verb “identified” is not annotated as *Process* since it doesn’t act upon another object.

6) Cellular morphology, pluripotency gene expression and differentiation into the three germ layers have been used compare the outcomes of manual and automated processes.

“pluripotency gene expression”, “differentiation”, “compare”, and “manual and automated processes” are each annotated as *Process*.

5.1.2 Annotating Scientific Entities as *Method*

Heuristics for identifying *Method* candidates

a. We annotate as *Method*, the phrases containing any of the following words: simulation, method, algorithm, scheme, technique, system, function, derivative, proportion, strategy, solver, experiment, test, computation, program.

Example

1) 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.

“finite-element modelling” is annotated as a *Method*

5.1.3 Annotating Scientific Entities as *Material*

Examples

1) Based on the results of the LUCAS topsoil survey we performed an assessment of plant available P status of European croplands.

“European croplands” is annotated as *Material*

2) 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.

“a manufacturing facility” and “cell-based therapy products” are annotated as *Material*

3) 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 three germ layers” is annotated as *Material*

5.1.4 Annotating Scientific Entities as *Data*

1) Based on the results of the LUCAS topsoil survey we performed an assessment of plant available P status of European croplands.

“the results” and “plant available P status” are annotated as *Data*

2) 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.

“a status of a baseline period”, “nutrient levels”, and “P” are annotated as *Data*

3) 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.

“asymmetry between the noon and midnight sectors”, “a velocity”, and “(2.5±1.5)kms-1” are annotated as *Data*

4) “We established [a P fertilizer need map] based on integrating results from the two systems.”

“a P fertilizer need map” is annotated as *Data* overriding “a P fertilizer” as *Material* by the tag precedence annotation guideline.

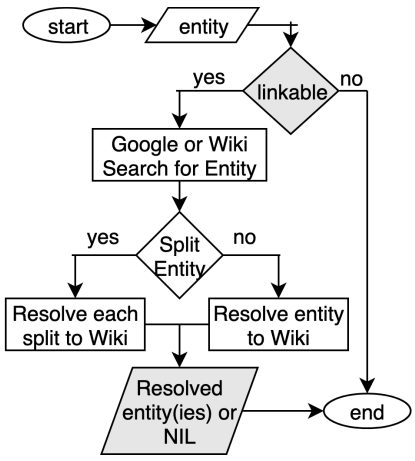
5.2 Scientific Entity Resolution

Resolution of the scientific entities: For the resolution of a set of scientific entities, the guidelines are elicited below.

In this step, annotators ground linkable scientific entities to a combination of knowledge from two different collaboratively-constructed lexical resources: (1) the Wikipedia encyclopedia whose articles are seen as individual real-world concepts; and (2) the lexicographic resource Wiktionary. Consequently, the scientific entities are enriched with *additional information* provided by resources (1) and (2). Formally, a scientific entity *e = (p, s)* where *p* is a Wikipedia page and *s* is the corresponding Wiktionary term sense. In the case that either *p* or *s* is not present for the entity *e*, the one present is used and the other is left empty, or both are empty for the linkable entities that can’t be found.

Note that in these guidelines, the definitions for scientific entities and linkable scientific entities are not equivalent. The latter set of entities is a subset of the former. In other words, given the set of scientific entities annotated in the preceding section (Section 5.1), not all are valid candidates for linking. The fine-grained nature of the scientific entity annotation task entailed annotating various generic phrases (e.g., measuring, documenting, increasing, etc.). In this step, such entities are not resolved.

The overall workflow for resolving scientific entities including is pictorially depicted below.



The salient components in the workflow are described next.

5.2.1 Linkable?

Only the scientific entity phrases that are self-contained units of scientific information targeting precisely those entities conveying domain-specific scientific jargon are deemed linkable.

In this step, annotators are assumed to be given as input entities from stage 1 and decide whether or not the entities are valid real-world candidates for linking. The following are types of entities that are not:

1. Scientific entities annotated as Data, specifically numbers apart from their units, are unlinkable; these entities form the easiest exclusion candidate for linkability.
2. Phrases that do not convey a precise scientific entity target are unlinkable. Consider “younger and more labile sources” where neither the phrasal unit as a whole, nor any parts of it are self-contained units of scientific information. It leaves the question "sources of what?" Therefore, it is unlinkable. As another example, the phrase “the development” isn’t what can be considered scientific jargon in any scientific domain, and as a generic phrase is unlinkable.
3. Generic verbs that take on their most common sense (e.g., document, increase, etc.) and do not convey any particular scientific information are unlinkable. These verbs should be found as Process.

We make a note here for annotators that linkability of an entity does not imply that the entity can be found in Wiki sources, but merely that is an eligible candidate for linking. This lets the corpus remain amenable to semantic annotations from ontologies other than Wikipedia or Wiktionary.

5.2.2 Google or Wiki Search for Entity

Given our set of linkable scientific entities, their longest meaningful span is preferred for linking. Annotations of the scientific entity phrases required capturing the most precise lexical unit (as in the “carbon atoms in graphene” as opposed to “carbon atoms” and “graphene” separately). Thus the annotated scientific entity phrases often include qualifiers or more than one noun phrase. However, such multiword expressions are not suitable candidates for linking.

Somewhat similar to BabelNet, longest linkable parts of the scientific entity phrase that constitute valid multi-word expressions interpreted as entities on the WWW are candidates for linking. But unlike BabelNet, iterative annotations for even smaller parts of the phrase are not performed since we are not attempting linking of scientific entities linguistically in a general sense but rather as semantic annotations for scientific jargon preserving intended scientific conceptual integrity. For instance, “plant biomass” means plant material used for energy production, whereas if split “plant” means an organism capable of photosynthesis and “biomass” means a plant or animal material used for energy production, where “plant biomass” is then a lost concept.

The next natural question is: how to determine valid multi-word expressions? Given an entity phrase, split decisions are determined based on annotator knowledge or Google or Wiki Search. For instance, if a Google Search returns the whole concept as a valid search term, the phrase is not further split irrespective of whether the phrase can be found in Wikipedia. The motivation behind this strategy is that multi-word expressions (MWE) over the course of time evolve into single concepts and Wikipedia/Google search engines being based on massive amounts of data are reliable sources to discover collocations.

As an example of Google Search: Microbial biomass-C is found via Google search as a Web entity http://www.soilquality.org.au/factsheets/microbial-biomass-carbon-nsw, therefore it is retained as is. Note, however, that Microbial biomass-C is not found in Wikipedia, so it is left unlinked in the subsequent phase or linked as a NIL concept. As an example of Wikipedia Search: the phrase “Cassini spacecraft” is a redirect link in Wiki to the page “Cassini,” so we do not split the phrase for linking as “Cassini” and “spacecraft,” separately, but rather link “Cassini spacecraft” to the Wiki page “Cassini.”

5.2.3 Split Entity?

For scientific entities not found as valid MWE Web entities, the following two heuristics are used to determine splits: 1) can the phrase be rewritten as a prepositional phrase? If it can, then the phrase is split accordingly and each split is linked. E.g., “soil phosphorus” can be rewritten as “phosphorus in soil”, where “phosphorus” and “soil” are linked individually. Else, 2) can the phrase be rewritten as a relative clause indicating properties of a subject? E.g., “planted soils” can be rewritten as “soils that are planted,” such that “planted” is a property of the “soil”; “planted” and “soils” are then linked individually. Note, however, that for all phrase parts that qualify as properties we do not link them, particularly the ones that are common (considered secondary information). E.g. “The key topological properties” which is “topological properties” that are “key”, but we do not link “key.”
Note that while most valid MWEs can be rewritten as prepositional phrases or in a relative clause structure, the scientific entities are preferred to be linked in their original form. E.g., “organic carbon” is not split since it is a valid MWE.

5.2.4 Resolve to Wikipedia and Wiktionary

Depending on the split results, entities or parts of it are then linked to Wikipedia and Wiktionary. This resolution is carried out based on the criteria elicited below.

Conceptual and Linguistic Basis for Linking to Wiki

Entities are linked to Wikipedia pages and Wiktionary senses, separately. For Wikipedia, entity phrase variants are allowed as long as the meaning is not changed significantly. However, this means that their original part-of-speech information could change. For Wiktionary, the term is preserved as is with the exception of plural words which are reduced to their singular form.

For WSD to Wiktionary

WSD is carried out for the exact word as found in the text (considering all parts-of-speech or tenses) with the following four exceptions:

1. *Plurals* are reduced to singular for WSD;
2. *Abbreviations* are queried in their expanded form, but not symbols (such as units or chemical elements). Symbols are not considered as abbreviations;
3. *Alternative spellings* such as “clear cut” instead of “clearcut”; and
4. Case of words such as uppercase or lowercase, such as “laboratory” instead of sentence case “Laboratory.”
5. *Possessives* are reduced to normal form.

For Linking to Wikipedia

Cases where the linked phrase includes extensions, so that with or without the meaning is the same thing. In other words, the phrases are not split for the extension.

1. “the clearfelling operations” is linked as “clearfelling operations” to https://en.wikipedia.org/wiki/Clearcutting,. Note here “clearfelling” and “clearfelling operations” means

the same thing. As a counter example, consider “clearcut sites” which are linked as “clearcut,sites” since “clearcut” is an operation and “sites” is a location.

2. “upland conifer plantation sites” is linked as “upland,conifer,plantation sites” to https://en.wikipedia.org/wiki/Upland_and_lowland,ADJECTIVE_184784_1_1;https://en.wikipedia.org/wiki/Pinophyta,NOUN_191538_0_1;https://en.wikipedia.org/wiki/Plantation,. Consider “plantation sites” and “plantation” are the same entity. Only that “plantation sites” does not have a WSD result. As a counter example, consider “their selection sites” which is linked as “selection,sites” to https://en.wikipedia.org/wiki/Natural_selection,NOUN_54009_0_1;,NOUN_62598_1_7 since “selection” is a process and “sites” is a different concept.
3. Similar to (2), “plant roots” is linked as is to https://en.wikipedia.org/wiki/Root, where “roots” and “plant roots” mean the same thing. As a counter example, consider “field soil” where not all soils are “field soil.” So it is linked as “field,soil” to https://en.wikipedia.org/wiki/Field_(agriculture),NOUN_8536_0_1;https://en.wikipedia.org/wiki/Soil,NOUN_5217_0_1
4. “Aspalathus species” is linked to https://en.wikipedia.org/wiki/Aspalathus. Similarly, “baseline level” is linked to https://en.wikipedia.org/wiki/Baseline where even simply “baseline” would’ve been linked to https://en.wikipedia.org/wiki/Baseline.
5. “unpruned timber trees” is linked as “unpruned,timber trees” to ,ADJECTIVE_1364314_0_1;https://en.wikipedia.org/wiki/Lumber,
6. “Cyclopia plants” is linked as “Cyclopia plants” to https://en.wikipedia.org/wiki/Cyclopia_(plant),

With this strategy, we collect all phrase variants. Note, however, that if “species” is explicitly meant, we do normalize just “species.”

Cases where extensions are dropped because they convey the most common sense

7. “a tilled environment” is linked as “tilled” to https://en.wikipedia.org/wiki/Till,ADJECTIVE_407524_1_1 where “environment” is dropped
8. “the physiological processes” is linked as “physiological” to https://en.wikipedia.org/wiki/Physiology,ADJECTIVE_78316_0_1 where “processes” is dropped
9. “local native seed sources” is linked as “local,native,seed” to ,ADJECTIVE_38364_0_1;,ADJECTIVE_49762_0_5;https://en.wikipedia.org/wiki/Seed,NOUN_39502_0_1 where “sources” is dropped
10. “multiple independent genes” is linked as “genes” to https://en.wikipedia.org/wiki/Gene,NOUN_41461_0_1 where “multiple independent” is dropped
11. “downy mildew incidence” is linked as “downy mildew” to https://en.wikipedia.org/wiki/Downy_mildew,NOUN_492286_0_2 where “incidence” is dropped
12. “higher versus moderate agronomic inputs” is linked as “agronomic,inputs” to https://en.wikipedia.org/wiki/Agricultural_economics,ADJECTIVE_310148_0_1;,NOUN_6965_0_2 with “higher versus moderate” dropped.
13. “SOC stocks” is linked as “SOC” to https://en.wikipedia.org/wiki/Soil_carbon, with “stocks” dropped as a generic phrase. Similarly, “SOC stock changes” is linked as “SOC” with “stock changes” dropped.
14. “P supply” is linked as “P” to https://en.wikipedia.org/wiki/Phosphorus,SYMBOL_145394_1_1
15. “bacterial community composition” is linked as “bacterial community” to https://en.wikipedia.org/wiki/Consorms, with “composition” dropped.
16. “the key responsible variables” is linked as “variables” to https://en.wikipedia.org/wiki/Variable_and_attribute_(research),NOUN_41481_1_3

17. “scale-free, small world and modular properties” Is split as “scale-free,small world,modular” where “properties” is dropped as a generic reference.
18. “integrated Striga control” is linked as “Striga” where “integrated” and “control” are dropped. Note, had “control” reflected a specialized sense such as in Medicine “control population”, it wouldn’t have been dropped.
19. “plasma absorbing interactions” is linked as “plasma” to https://en.wikipedia.org/wiki/Plasma_(physics),NOUN_8397_0_1 and “absorbing interactions” is dropped.

<u>Cases where extensions which seem generic are not dropped because they do not convey the most common sense</u>

We do retain generic phrases if they have a different sense from the general domain sense

20. “poorly correlated” is linked as “poorly,correlated” to ,ADVERB_78732_0_2;https://en.wikipedia.org/wiki/Correlation_and_dependence,ADJECTIVE_67229_1_1 where “poorly” means “weakly” or “indifferently”
21. “topsoil P survey” is linked as “topsoil,P,survey” to https://en.wikipedia.org/wiki/Topsoil,NOUN_465872_0_1;https://en.wikipedia.org/wiki/Phosphorus,SYMBOL_145394_1_1;https://en.wikipedia.org/wiki/Soil_survey,NOUN_94590_0_2 where “survey” is linked as “soil survey”
22. Also “the C5 selection cycle” is linked as “selection,cycle” to https://en.wikipedia.org/wiki/Natural_selection,NOUN_54009_0_1;,NOUN_12150_0_1 where “cycle” is linked to the “cyclic process” sense
23. “topological properties” is linked to https://en.wikipedia.org/wiki/Topological_property,NOUN_6386673_0_1

<u>Cases where extensions are partly subsumed because of conjunctions</u>

24. acid/alkaline phosphatase activity
    Is split as “acid,alkaline phosphatase activity” where “phosphatase activity” would’ve been subsumed within “acid.” However, since there is no entity for “acid phosphatase activity” nor for “alkaline phosphatase activity,” they are reduced for linking as “acid,alkaline phosphatase” where “phosphatase” is subsumed with “acid.” So we link the first part i.e. “acid” to https://en.wikipedia.org/wiki/Acid_phosphatase and the second part “alkaline phosphatase” to https://en.wikipedia.org/wiki/Alkaline_phosphatase. The word “activity” is dropped from consideration as a generic reference. Here, however, we do not drop “phosphatase” from the explicitly mentioned linking phrase.
    We do not do a WSD for such phrases, in particular, for “acid” since the linked meaning is “acid phosphatase”

25. Acid and alkaline activity
    Is split as “Acid,alkaline” where “activity” is dropped as a generic reference. And the linked solution is https://en.wikipedia.org/wiki/Acid,ADJECTIVE_1625_0_3;https://en.wikipedia.org/wiki/Alkali,ADJECTIVE_202709_1_1
    Note for WSD the query is exactly “acid,alkaline”

26. “Cyclopia and Aspalathus species” is linked as “Cyclopia,Aspalathus species” where “species” is considered subsumed in Cyclopia. The query phrase for WSD is “Cyclopia,Aspalathus species” as it is.

27. “Honeybush and Rooibos tea” is linked as “Honeybush,Rooibos tea” to https://en.wikipedia.org/wiki/Cyclopia_(plant),;https://en.wikipedia.org/wiki/Rooibos, where “tea” is considered subsumed in Honeybush.

28. “Laboratory and field studies” is linked as “Laboratory,field studies” to https://en.wikipedia.org/wiki/Laboratory,;https://en.wikipedia.org/wiki/Field_research, where “studies” is subsumed in Laboratory. Note we do not do the WSD for “Laboratory.”

Cases where extensions are not subsumed since they are a valid separate entity

29. “solar energy, water, and mineral nutrients” is linked as “solar energy,water,mineral,nutrients” to

Handling generic phrases

30. “an assessment” is not linked.
31. “multiple or changing environments” is left unlinked
32. “the environment” is left unlinked
33. “predict” is left unlinked
34. “rapid transport” is left unlinked.

Handling formulaic entities

35. “mass loss rate” is linked as is and is not treated linguistically, such as “rate of the loss in mass.”

Abbreviations

All author-madeup abbreviations in the writing of the paper are not normalized and dropped from consideration. E.g., “Two-hundred full-sib families (FS).” Here “FS” is not a candidate for normalization

Symbols

1. “soil P management” is linked as “soil,P” to https://en.wikipedia.org/wiki/Soil,NOUN_5217_0_1;https://en.wikipedia.org/wiki/Phosphorus,SYMBOL_145394_1_1 where “management” is dropped from consideration and the chemical element “P” is queried as is for WSD
2. “altitudes ranging from 120m to 380m” is linked as “altitudes,m,m” to https://en.wikipedia.org/wiki/Altitude,NOUN_7950_0_1;https://en.wikipedia.org/wiki/Metre,SYMBOL_8697_1_1;https://en.wikipedia.org/wiki/Metre,SYMBOL_8697_1_1

Hyphenations

1. Negations: “non-rhizosphere bulk soils” is linked as “rhizosphere,bulk soils” where “non” is dropped.
2. “on-farm” is linked as “farm”
3. “PGRN-dependent pathogenic mechanisms” is linked as “PGRN,pathogenic” with the hyphenation extension “dependent” dropped.

Handling long phrase resolutions

1. “planted (with or without mycorrhizal fungi) and in unplanted macrocosms” is linked as “planted, mycorrhizal fungi, unplanted, macrocosms” to https://en.wikipedia.org/wiki/Sowing,VERB_191403_0_1;https://en.wikipedia.org/wiki/Mycorrhiza,NOUN_489043_0_1

6. Annotation Tools

6.1 Annotating Scientific Entities

To perform the annotation task and insert markups for scientific entity spans and their corresponding concept type, we suggest annotators use the brat rapid annotation tool[12] (for both Windows and Unix system users).

The interface of the tool presents sentence-split text of the abstracts of the scientific articles. In this view of the data, annotators can select a span of text that they identify as a scientific entity, and further annotate it with a scientific entity type. The four scientific entity types used in this work can be separately configured for the annotation project within the tool. Scientific entity span annotations can be created simply by selecting text using the mouse, as in most text editors and similar software, selecting a type from the NEW ANNOTATION DIALOG that pops up after selection, and pressing OK. Thus annotators can annotate scientific entities with their types with just 3 mouse clicks.

Annotations created in brat for each document or a collection of documents can be exported with a few clicks from the interface in a simple standoff format that can be easily analysed, processed, and converted into other formats.

6.2 Linking Scientific Entities

To perform the linking task, we suggest annotators use Google Excel Sheets[13] (OS-independent). They are flexible enough to incorporate various linking task definitions. Each column in the excel sheet can be a specific data type.

In our task, we create the following eight columns: Domain, Filename, Entity, Linkable, Split, Split Terms, Wiki IDs, and Notes. The values under column “Domain” correspond to the data domain under consideration; the values under “Filename” correspond to the name of the file from the OA-STM corpus in the specified domain; under “Entity” is the original annotated scientific entity phrase; under “Linkable” is the annotator decision about whether the scientific entity is linkable or not--values can be 1 or 0, or yes or no; under “Split” is the annotator decision only for Linkable entities about whether they should be split or not--values can be 1 or 0, or yes or no; under “Split Terms” is the original entity text split by the annotator if the “Split” decision was 1 or yes; under “Wiki IDs” annotations are performed as below:

Based on the formalism we adopt, if the phrase is not split or is a unigram, then Wiki IDs value is p;s, where p is the Wikipedia page title and s is the Wiktionary sense identifier (e.g. NOUN_0_1 meaning that the entity is a NOUN, and the NOUN is the 0th entry in the Wiktionary page, and is the 1st sense gloss corresponding to particular entry).
Otherwise, for a split phrase e = $e_a;e_b;e_c;e_d;...$ it’s normalized phrase = $p_a,s_a;p_b,s_b;p_c,s_c;p_d,s_d;...$ where p and s correspond to the Wikipedia and Wiktionary identifiers as before.

[12] https://brat.nlplab.org/
[13] https://www.google.com/sheets/about/

If the entity or a split part of the split entity cannot be linked to either Wikipedia or Wiktionary or both, the corresponding space is left blank, however the separators are still included. Consider as in the following three examples.

1) the experimentally measured fracture energies → experimentally,measured,fracture energies → ,ADVERB_94599_0_1;,ADJECTIVE_227450_0_1;
where for “experimentally” and “measured”, their corresponding Wikipedia pages are absent and for “fracture energies” both the Wikipedia and Wiktionary pages are absent.

2) a fracture energy of 481 J/m2 → fracture energy,J,m → ,;https://en.wikipedia.org/wiki/Joule,NOUN_36288_0_1;https://en.wikipedia.org/wiki/Metre,NOUN_28923_0_1

3) this very low test temperature → low,test,temperature → ,ADJECTIVE_8384_0_9;,NOUN_27637_0_4;https://en.wikipedia.org/wiki/Temperature,NOUN_4531_0_3

The final component to the linking annotation setup includes local installations of specific time-stamped Wikipedia and Wiktionary dumps to enable future persistent references for entities since they undergo active revisions. We use the DKPro JWPL[14] for querying a static Wikipedia dump as an optimized database enabling efficient search. Correspondingly, we use the DKPro JWKTL[15] for querying a static Wiktionary dump as an optimized database similarly enabling efficient search.

Annotations created in the excel sheets can be easily saved in either csv or tsv format enabling clear separation between the different data columns for later analysis, processing, and storage.

[14] https://dkpro.github.io/dkpro-jwpl/JWPLCore_GettingStarted/
[15] https://dkpro.github.io/dkpro-jwktl/documentation/getting-started/

7. Supplementary Material: Text Graphs from Annotations of Scientific Entities

In this section, are provided example text graphs for the stage 1 created scientific concept annotations from one abstract per domain. These figures can be employed as reference structures for annotators attempting the annotation task at stage 1. In all graphs, nodes are color-coded by their concept type: orange corresponds to PROCESS, green corresponds to MATERIAL, blue for DATA, and purple for METHOD.

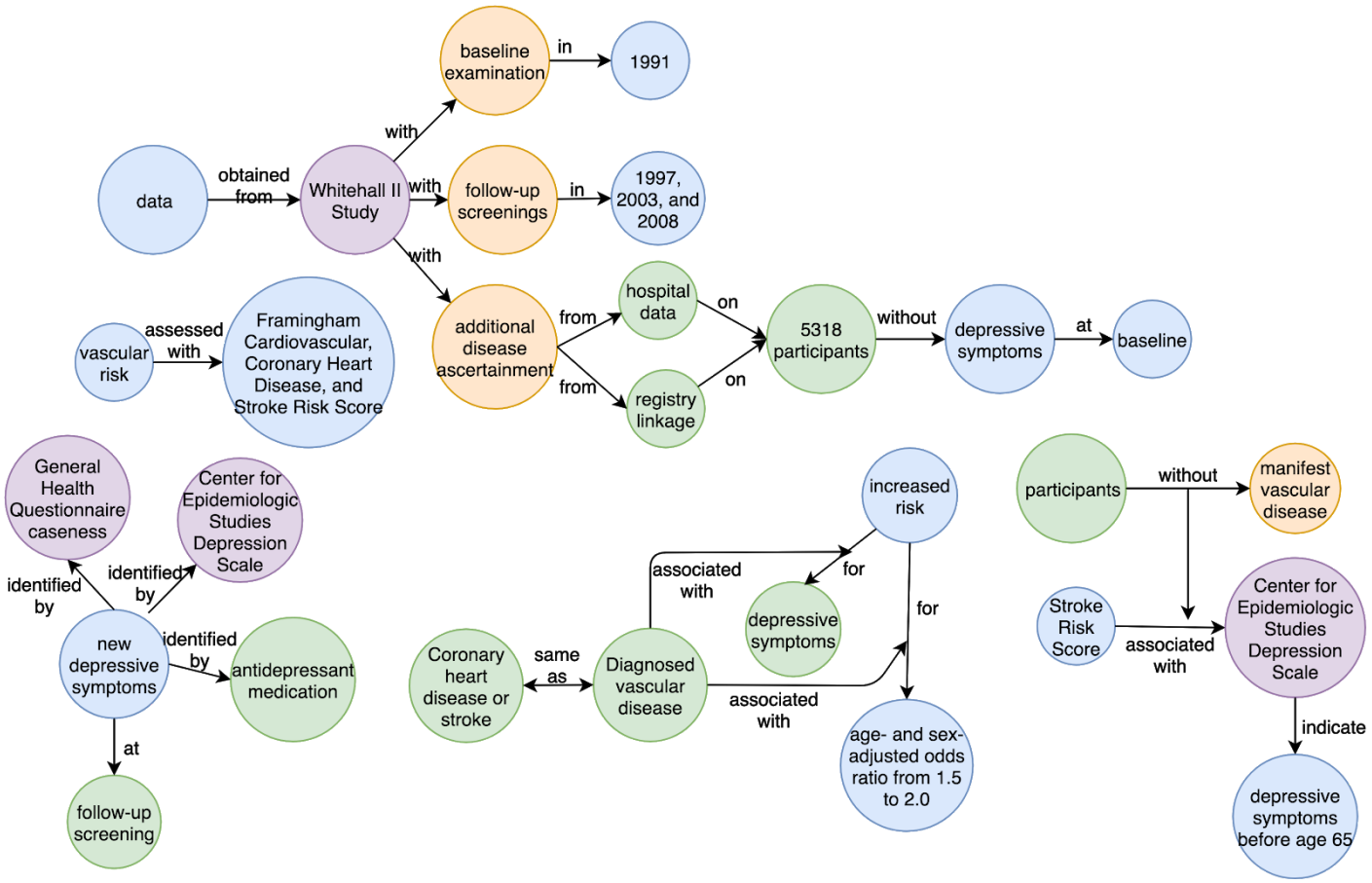


Figure 1: Knowledge Graph representation of the Abstract of the Elsevier article "*Soil structural responses to alterations in soil microbiota induced by the dilution method and mycorrhizal fungal inoculation*"[16] in the Agriculture domain.

[16] Martin, Sarah L., et al. "Soil structural responses to alterations in soil microbiota induced by the dilution method and mycorrhizal fungal inoculation." *Pedobiologia* 55.5 (2012): 271-281.

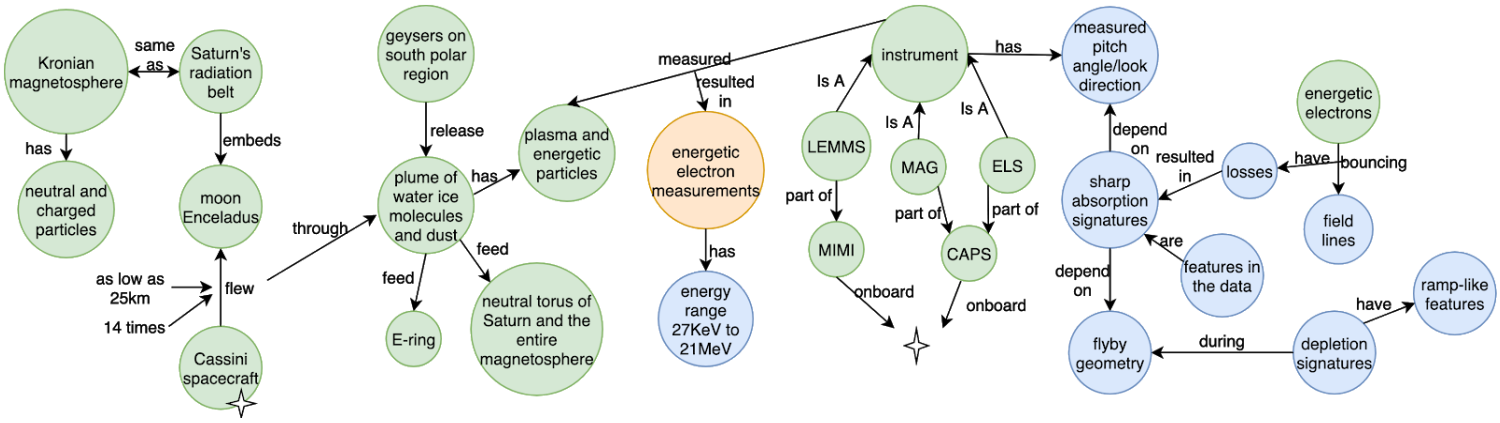


Figure 2: Knowledge Graph representation of the Abstract of the Elsevier article “The Cassini Enceladus encounters 2005–2010 in the view of energetic electron measurements”[17] in the Astronomy domain

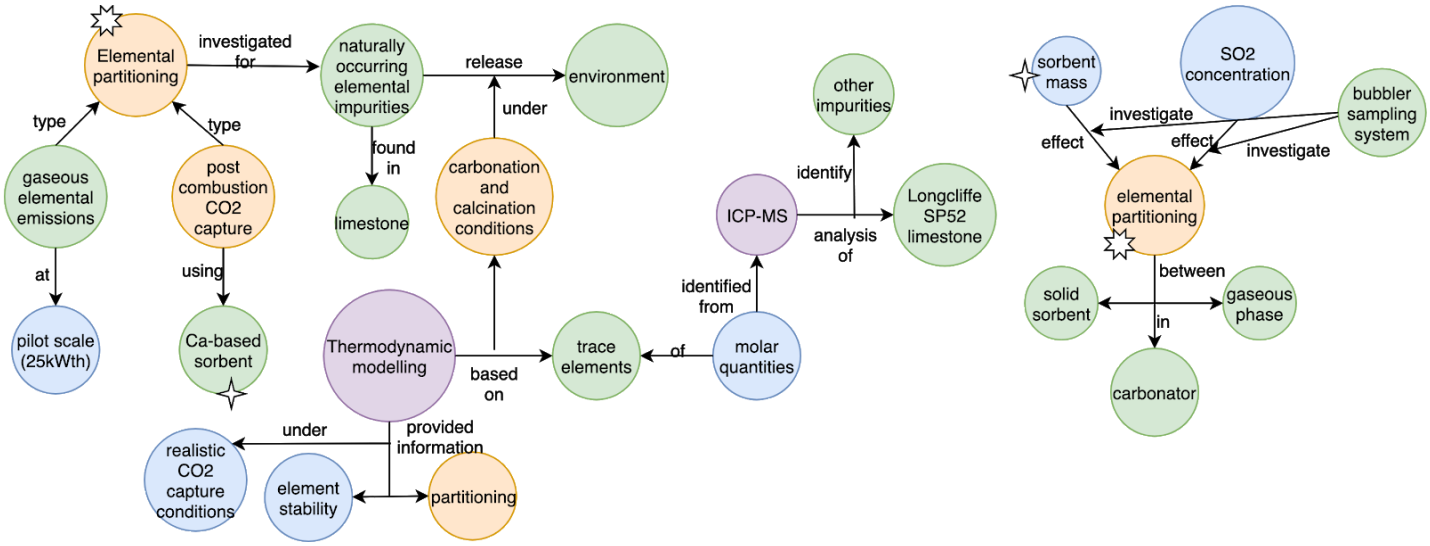


Figure 3: Knowledge Graph representation of the Abstract of the Elsevier article “Minor and trace element emissions from post-combustion CO2 capture from coal: Experimental and equilibrium calculations ”[18] in the Chemistry domain

[17] Krupp, N., et al. "The Cassini Enceladus encounters 2005–2010 in the view of energetic electron measurements." *Icarus* 218.1 (2012): 433-447.
[18]Cotton, Alissa, Kumar Patchigolla, and John E. Oakey. "Minor and trace element emissions from post-combustion CO2 capture from coal: Experimental and equilibrium calculations." *Fuel* 117 (2014): 391-407.

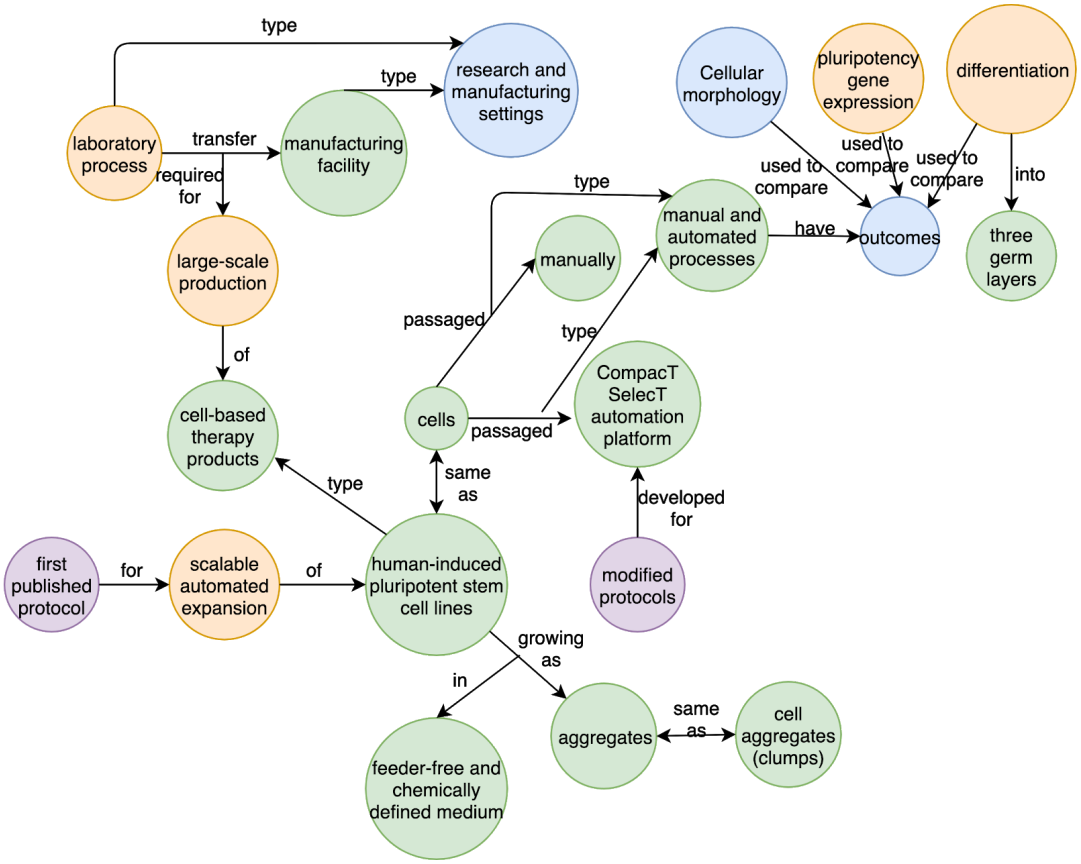


Figure 4: Knowledge Graph representation of the Abstract of the Elsevier article “Investigating the feasibility of scale up and automation of human induced pluripotent stem cells cultured in aggregates in feeder free conditions”[19] in the Biology domain

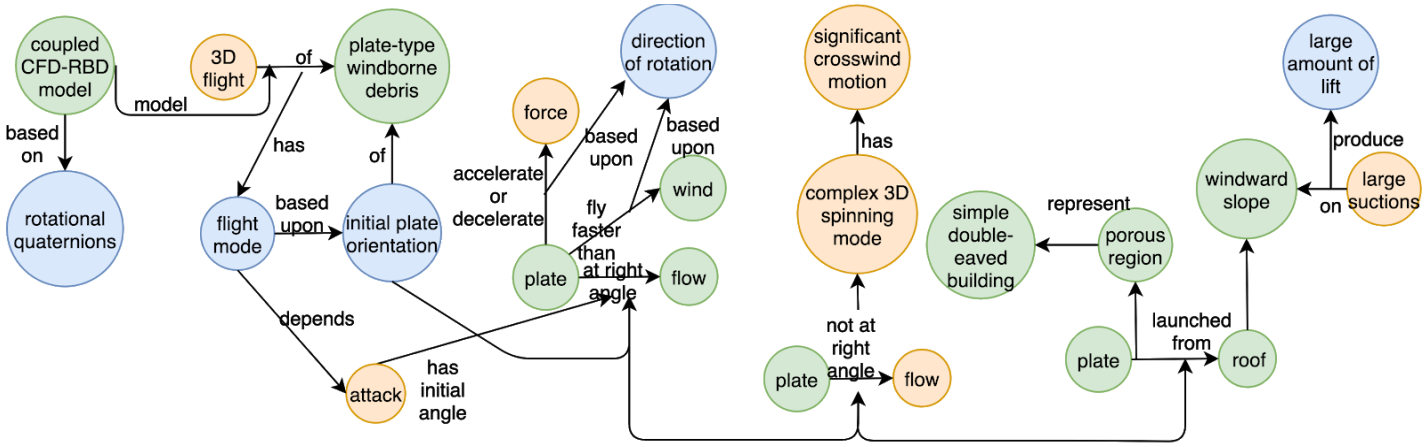


Figure 5: Knowledge Graph representation of the Abstract of the Elsevier article “An investigation of plate-type windborne debris flight using coupled CFD-RBD models”[20] in the Engineering domain

[19]Soares, Filipa AC, et al. "Investigating the feasibility of scale up and automation of human induced pluripotent stem cells cultured in aggregates in feeder free conditions." *Journal of biotechnology* 173 (2014): 53-58.

[20]Kakimpa, B., D. M. Hargreaves, and J. S. Owen. "An investigation of plate-type windborne debris flight using coupled CFD–RBD models. Part II: Free and constrained flight." *Journal of Wind Engineering and Industrial Aerodynamics* 111 (2012): 104-116.

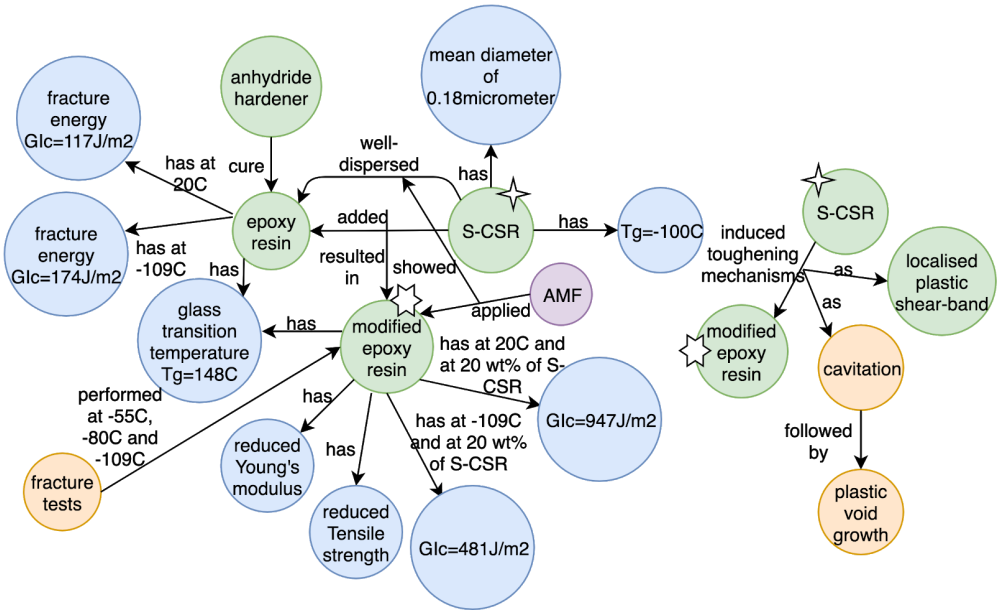


Figure 6: Knowledge Graph representation of the Abstract of the Elsevier article “The mechanical properties and toughening mechanisms of an epoxy polymer modified with polysiloxane-based core-shell particles”[21] in the Materials Science domain

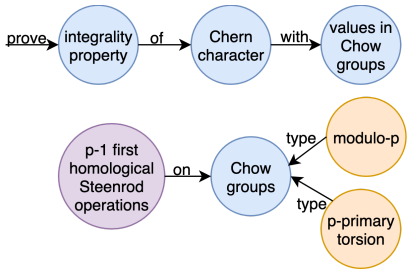


Figure 7: Knowledge Graph representation of the Abstract of the Elsevier article “Integrality of the Chern character in small codimension”[22] in the Mathematics domain

[21]Kakimpa, B., D. M. Hargreaves, and J. S. Owen. "An investigation of plate-type windborne debris flight using coupled CFD–RBD models. Part II: Free and constrained flight." *Journal of Wind Engineering and Industrial Aerodynamics* 111 (2012): 104-116.
[22]Haution, Olivier. "Integrality of the Chern character in small codimension." *Advances in Mathematics* 231.2 (2012): 855-878.

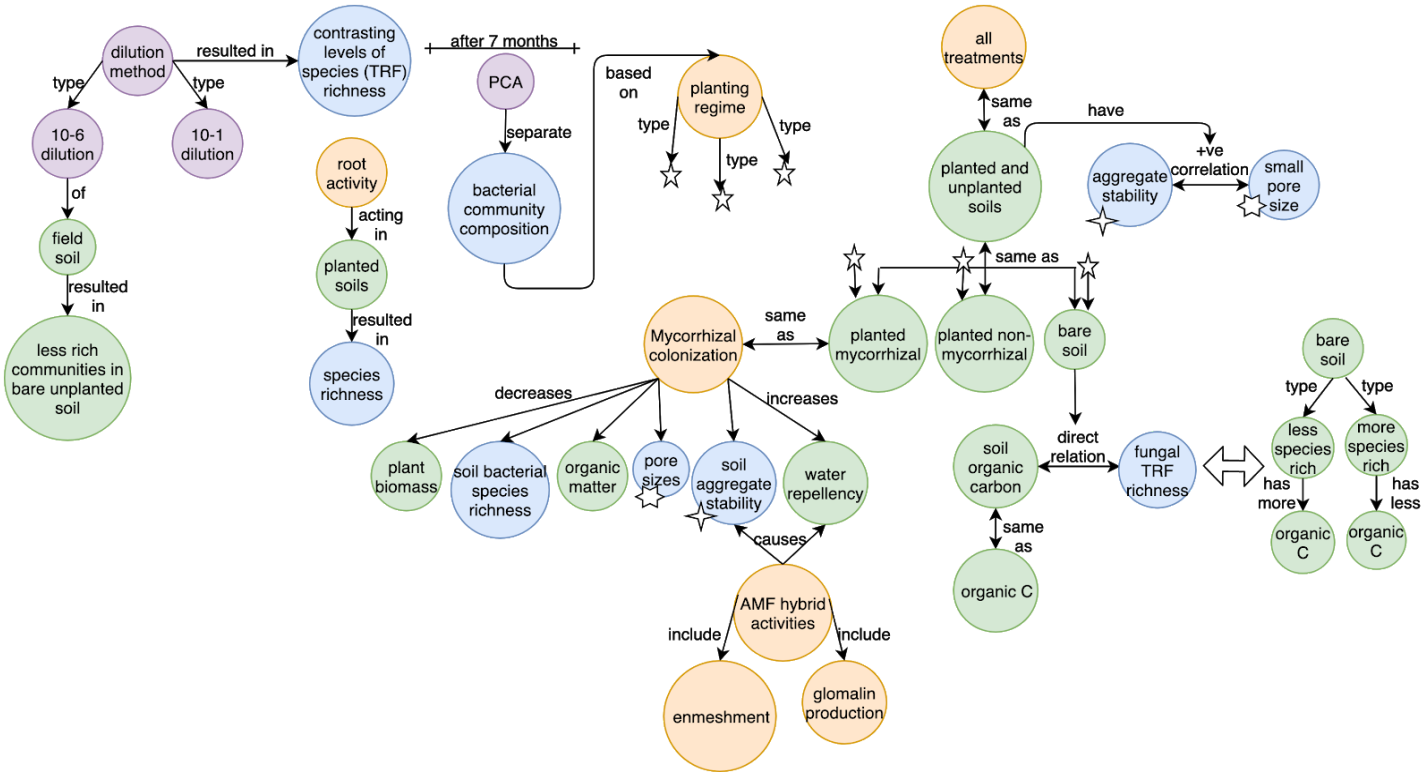


Figure 8: Knowledge Graph representation of the Abstract of the Elsevier article “Soil structural responses to alterations in soil microbiota induced by the dilution method and mycorrhizal fungal inoculation”[23] in the Agriculture domain

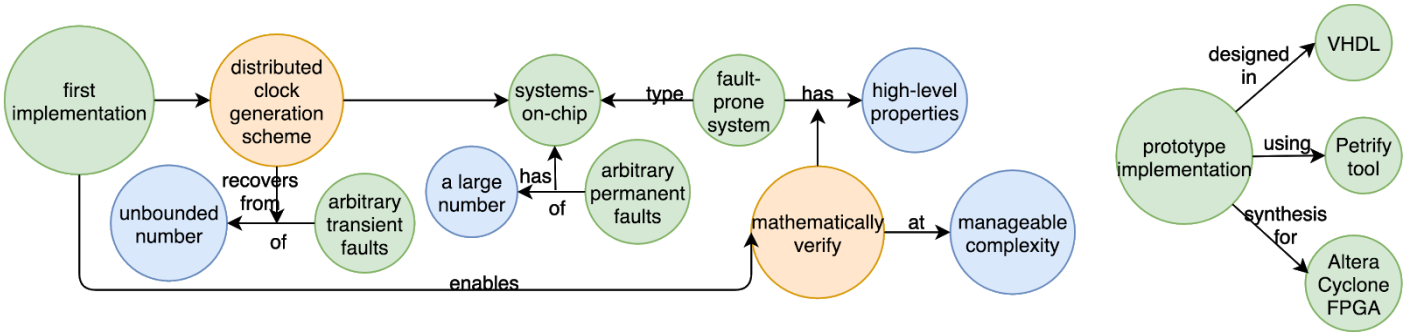


Figure 9: Knowledge Graph representation of the Abstract of the Elsevier article “Rigorously modeling self-stabilizing fault-tolerant circuits: An ultra-robust clocking scheme for systems-on-chip.”[24] in the Computer Science domain

[23] Martin, Sarah L., et al. "Soil structural responses to alterations in soil microbiota induced by the dilution method and mycorrhizal fungal inoculation." Pedobiologia 55.5 (2012): 271-281.

[24] Dolev, Danny, et al. "Rigorously modeling self-stabilizing fault-tolerant circuits: An ultra-robust clocking scheme for systems-on-chip." *Journal of computer and system sciences* 80.4 (2014): 860-900.

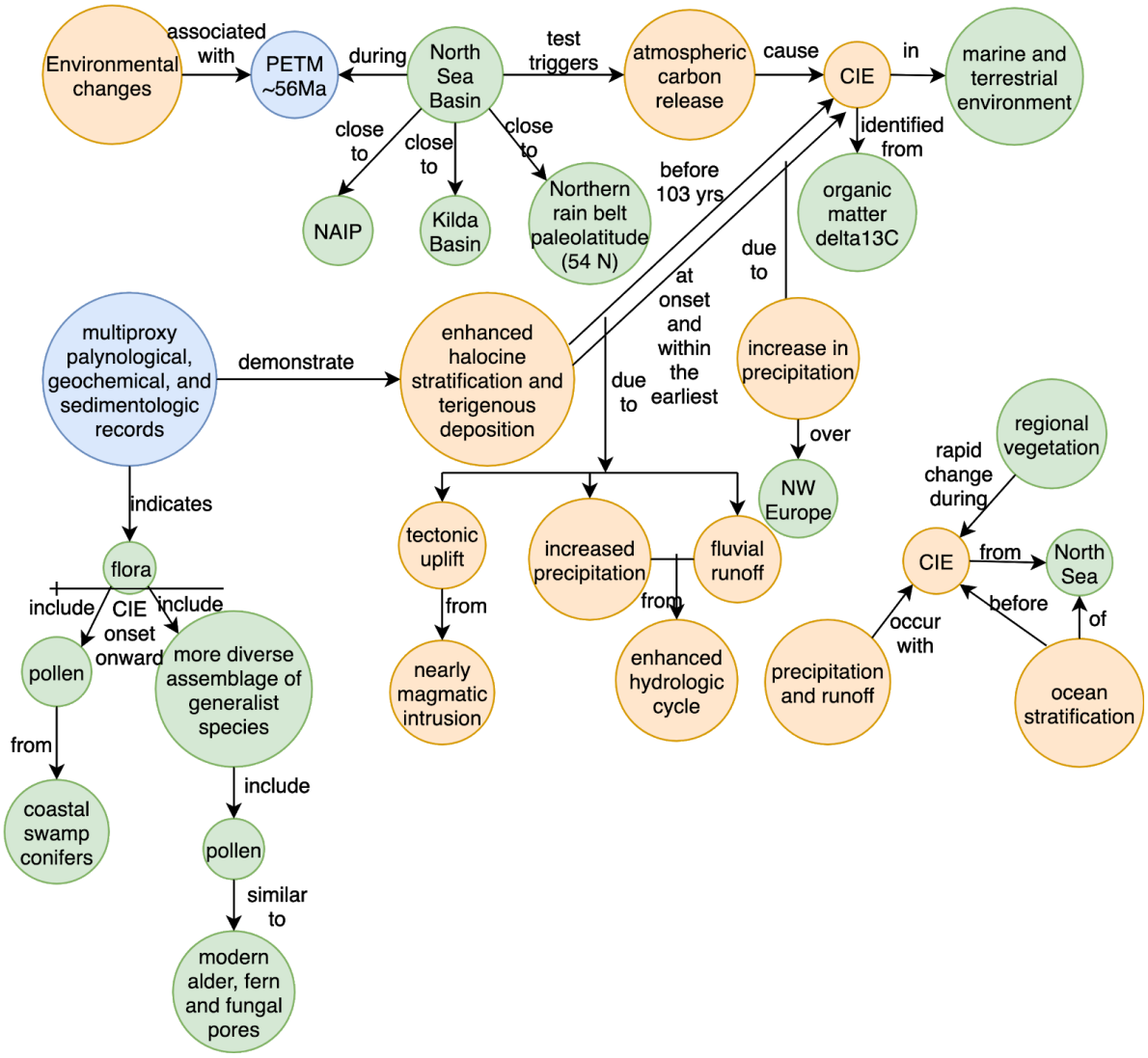


Figure 10: Knowledge Graph representation of the Abstract of the Elsevier article “Marine and terrestrial environmental changes in NW Europe preceding carbon release at the Paleocene–Eocene transition”[25] in the Earth Science domain

[25]Kender, Sev, et al. "Marine and terrestrial environmental changes in NW Europe preceding carbon release at the Paleocene–Eocene transition." *Earth and Planetary Science Letters* 353 (2012): 108-120.