Analysing the Requirements for an Open Research Knowledge Graph: Use Cases, Quality Requirements and Construction Strategies

Arthur Brack¹² · Anett Hoppe · Markus Stocker¹ · Sören Auer¹² · Ralph Ewerth¹²

Received: date / Accepted: date

Abstract Current science communication has a number of drawbacks and bottlenecks which have been subject of discussion lately: Among others, the rising number of published articles makes it nearly impossible to get a full overview of the state of the art in a certain field, or reproducibility is hampered by fixed-length, document-based publications which normally cannot cover all details of a research work. Recently, several initiatives have proposed knowledge graphs (KG) for organising scientific information as a solution to many of the current issues. The focus of these proposals is, however, usually restricted to very specific use cases. In this paper, we aim to transcend this limited perspective and present a comprehensive analysis of requirements for an Open Research Knowledge Graph (ORKG) by (a) collecting and reviewing daily core tasks of a scientist, (b) establishing their consequential requirements for a KG-based system, (c) identifying overlaps and specificities, and their coverage in current solutions. As a result, we map necessary and desirable requirements for successful KG-based science communication, derive implications, and outline possible solutions.

Keywords scholarly communication · research knowledge graph · design science research · requirements analysis

1 Introduction

Today’s scholarly communication is a document-centred process and as such, rather inefficient. Scientists spend considerable time in finding, reading and reproducing research results from PDF files consisting of static text, tables, and figures. The explosion in the number of published articles [14] aggravates this situation further: It gets harder and harder to stay on top of current research, that is to find relevant works, compare and reproduce them and, later on, to make one’s own contribution known for its quality.

Some of the available infrastructures in the research ecosystem already use knowledge graphs (KG) to enhance their services. Academic search engines, for instance, such as Microsoft Academic Knowledge Graph [37] or Literature Graph [1] utilise metadata-based graph structures which link research articles based on citations, shared authors, venues and keywords.

Recently, initiatives have promoted the usage of KGs in science communication, but on a deeper, semantic level [3,49,55,72,77,83,112]. They envision the transformation of the dominant document-centred knowledge exchange to knowledge-based information flows by representing and expressing knowledge through semantically rich, interlinked KGs. Indeed, they argue that a shared structured representation of scientific knowledge has the potential to alleviate some of the science communication’s current issues: Relevant research could be easier to find, comparison tables automatically compiled, own insights rapidly placed in the current ecosystem. Such a powerful data structure could, more

¹ Acknowledging that knowledge graph is vaguely defined, we adopt the following definition: A knowledge graph (KG) consists of (1) an ontology describing a conceptual model (e.g. with classes and relation types), and (2) the corresponding instance data (e.g. objects, literals, and <subject, predicate, object>-triplets) following the constraints posed by the ontology (e.g. instance-of relations). The construction of a KG involves ontology design and population with instances.
than the current document-based system, also encourage the interconnection of research artefacts such as datasets and source code much more than current approaches (like Digital Object Identifier (DOI) references etc.); allowing for easier reproducibility and comparison. To come closer to the vision of knowledge-based information flows, research articles should be enriched and interconnected through machine-interpretable semantic content. The usage of Papers With Code [79] in the machine learning community and Jaradeh et al.'s study [55] indicate that authors are also willing to contribute structured descriptions of their research articles.

The work of a researcher is manifold, but current proposals usually focus on a specific use case (e.g. the aforementioned examples focus on enhancing academic search). In this paper, we present a detailed analysis of common literature-related tasks in a scientist’s daily life and analyse (a) how they could be supported by an ORKG, (b) what requirements result for the design of (b1) the KG and (b2) the surrounding system, (c) how different use cases overlap in their requirements and can benefit from each other. Our analysis is led by the following research questions:

1. Which use cases should be supported by an ORKG?
   (a) Which user interfaces are necessary?
   (b) Which machine interfaces are necessary?
2. What requirements can be defined for the underlying ontologies to support these use cases?
   (a) Which granularity of information is needed?
   (b) To what degree is domain specialisation needed?
3. What requirements can be defined for the instance data in context of the respective use cases?
   (a) Which completeness is sufficient for the instance data?
   (b) Which correctness is sufficient for the instance data?
   (c) Which approaches (human vs. machine) are suitable to populate the ORKG?

We follow the design science research (DSR) methodology [50]. In this study, we focus on the first phase of DSR and conduct a requirements analysis. The objective is to chart necessary (and desirable) requirements for successful KG-based science communication, and, consequently, provide a map for future research.

Compared to our paper at the 24th International Conference on Theory and Practice of Digital Libraries 2020 [16], this journal paper has been modified and extended as follows: The related work section is updated and extended with the new sections Quality of knowledge graphs and Systematic literature reviews. The new Appendix A contains comparative overviews of datasets for research knowledge graph population tasks such as sentence classification, relation extraction, and concept extraction. To be consistent with terminology in related work, we use the term “completeness” instead of “coverage” and “correctness” instead of “quality”. The requirements analysis in Section 3 is revised and contains more details with more justifications for the posed requirements and approaches.

The remainder of the paper is organised as follows. Section 2 summarises related work on research knowledge graphs, scientific ontologies, KG construction, data quality requirements, and systematic literature reviews. The requirements analysis is presented in Section 3, while Section 4 discusses implications and possible approaches for ORKG construction. Finally, Section 5 concludes the requirements analysis and outlines areas of future work. Appendix A contains comparative overviews for the tasks of sentence classification, relation extraction, and concept extraction.

2 Related work

This section gives a brief overview of (a) existing research KGs, (b) ontologies for scholarly knowledge, (c) approaches for KG construction, (d) quality dimensions of KGs, and (e) processes in systematic literature reviews.

2.1 Research knowledge graphs

Academic search engines (e.g. Google Scholar, Microsoft Academic, SemanticScholar) exploit graph structures such as the Microsoft Academic Knowledge Graph [37], SciGraph [110], the Literature Graph [1], or the Semantic Scholar Open Research Corpus (S2ORC) [69]. These graphs interlink research articles through metadata, e.g. citations, authors, affiliations, grants, journals, or keywords.

To help reproduce research results, initiatives such as Research Graph [2], Research Objects [71] and OpenAIRE [72] interlink research articles with research artefacts such as datasets, source code, software, and video presentations. Scholarly Link Exchange (Scholix) [20] aims to create a standardised ecosystem to collect and exchange links between research artefacts and literature.

Some approaches connect articles at a more semantic level: Papers With Code [79] is a community-driven effort to supplement machine learning articles with tasks, source code and evaluation results to construct leaderboards. Ammar et al. [11] link entity mentions in abstracts with DBpedia [65] and Unified Medical Language System (UMLS) [11], and Cohan et al. [23] extend the citation graph with citation intents (e.g. citation as background or used method).

Various scholarly applications benefit from semantic content representation, e.g. academic search engines by exploiting general-purpose KGs [109], and graph-based research paper recommendation systems [8] by utilising citation graphs and mentioned entities. However, the coverage of science-specific concepts in general-purpose KGs is rather low [11], e.g. the task “geolocation estimation of photos” from Com-
puter Vision is neither present in Wikipedia nor in the Computer Science Ontology (CSO) [94].

2.2 Scientific ontologies

Various ontologies have been proposed to model metadata such as bibliographic resources and citations [82]. Iniesta and Corcho [92] reviewed ontologies to describe scholarly articles. In the following, we describe some ontologies that conceptualise the semantic content in research articles.

Several ontologies focus on rhetorical [106,48,27] (e.g. Background, Methods, Results, Conclusion), argumentative [103,68] (e.g. claims, contrastive and comparative statements about other work) or activity-based structure [83] (e.g. sequence of research activities) of research articles. Others describe scholarly knowledge with linked entities such as problem, method, theory, statement [49,19], or focus on the main research findings and characteristics of research articles described in surveys with concepts such as problems, approaches, implementations, and evaluations [19,104].

Various domain-specific ontologies exist, for instance, mathematics [64] (e.g. definitions, assertions, proofs), machine learning [61,73] (e.g. dataset, metric, model, experiment), and physics [93] (e.g. formation, model, observation). The EXPeriments Ontology (EXPO) is a core ontology for scientific experiments that conceptualises experimental design, methodology, and results [97].

Taxonomies for domain-specific research areas support the characterisation and exploration of a research field. Salatino et al. [94] give an overview, e.g. Medical Subject Headings (MeSH), Physics Subject Headings (PhysSH), Computer Science Ontology (CSO), Gene Ontology [26] and Chemical Entities of Biological Interest (CheBi) [80] are KGs for genes and molecular entities.

2.3 Construction of knowledge graphs

Nickel et al. [76] classify KG construction methods into four groups: (1) curated approaches, i.e. triples created manually by a closed group of experts, (2) collaborative approaches, i.e. triples created manually by an open group of volunteers, (3) automated semi-structured approaches, i.e. triples extracted automatically from semi-structured text via handcrafted rules, and (4) automated unstructured approaches, i.e. triples are extracted automatically from unstructured text.

2.3.1 Manual approaches

WikiData [105] is one of the most popular KGs with semantically structured, encyclopaedic knowledge curated manually by a community. As of January 2021, WikiData comprises 92M entities curated by almost 27,000 active contributors. The community also maintains a taxonomy of categories and “infoboxes” which define common properties of certain entity types. Furthermore, Papers With Code [79] is a community-driven effort to link machine learning articles with tasks, source code and evaluation results. KGs such as Gene Ontology [26] or Wordnet [40] are curated by domain experts. Research article submission portals such as EasyChair (https://www.easychair.org/) enforce the authors to provide machine-readable metadata. Librarians and publishers tag new articles with keywords and subjects [110]. Virtual research environments enable the execution of data analysis on interoperable infrastructure and store the data and results in KGs [99].

2.3.2 Automated Approaches

Automatic KG construction from text: Petasis et al. [84] present a review on ontology learning, that is ontology creation from text, while Lubani et al. [71] review ontology population systems. Pajura and Singh [87] give an overview of the involved tasks for KG population: (a) information extraction to extract a graph from text with entity extraction and relation extraction, and (b) graph construction to clean and complete the extracted graph, as it is usually ambiguous, incomplete and inconsistent. Coreference resolution [17,70] clusters different mentions of the same entity in text and entity linking [62] maps mentions in text to entities in the KG. Entity resolution [102] identifies objects in the KG that refer to the same underlying entity. For taxonomy population, Salatino et al. [94] provide an overview of methods based on rule-based natural language processing (NLP), clustering and statistical methods.

The Computer Science Ontology (CSO) has been automatically populated from research articles [94]. The AI-KG was automatically generated from 333,000 research papers in the artificial intelligence (AI) domain [32]. It contains five entity types (tasks, methods, metrics, materials, others) linked by 27 relations types. Kannan et al. [57] create a multimodal KG for deep learning papers from text and images and the corresponding source code. Brack et al. [17] generate a KG for 10 different science domains with the concept types material, method, process, and data. Zhang et al. [112] suggest a rule-based approach to mine research problems and proposed solutions from research papers.

Information extraction from scientific text: Information extraction is the first step in the automatic KG population pipeline. Nasar et al. [74] survey methods on information extraction from scientific text. Beltagy et al. [9] present benchmarks for several scientific datasets and Peng et al. [81] especially for the biomedical domain. Appendix A presents comparative overviews of datasets for the tasks sentence classification, relation extraction, and concept extraction, respectively, in research papers.
There are datasets which are annotated at sentence level for several domains, e.g. biomedical [31,59], computer graphics [42], computer science [24], chemistry and computational linguistics [103], or algorithmic metadata [93]. They cover either only abstracts [31,59,24] or full articles [42,68,93,103]. The datasets differentiate between five and twelve concept classes (e.g. Background, Objective, Results). Machine learning approaches for datasets consisting of abstracts achieve an F1 score ranging from 66% to 92% and for datasets with full papers F1 scores range from 51% to 78% (see Table 2).

More recent corpora, annotated at phrasal level, aim at constructing a fine-grained KG from scholarly abstracts with the tasks of concept extraction [43,70,15,88], binary relation extraction [70,44,14], n-ary relation extraction [58,54,56], and coreference resolution [17,25,70]. They cover several domains, e.g. material sciences [43], computational linguistics [44,88], computer science, material sciences, and physics [43], machine learning [70], biomedicine [25,56,63], or a set of ten scientific, technical and medical domains [15,17,36]. The datasets differentiate between four to seven concept classes (like Task, Method, Tool) and between two to seven binary relation types (like used-for, part-of, evaluate-for). The extraction of n-ary relations involves extraction of relations among multiple concepts such as drug-gene-mutation interactions in medicine [50], experiments related to solid oxide fuel cells with involved material and measurement conditions in material sciences [43], or task-dataset-metric-score tuples for leaderboard construction for machine learning tasks [58].

Approaches for concept extraction achieve F1 scores ranging from 56.6% to 96.9% (see Table 2), for coreference resolution F1 scores range from 46.0% to 61.4% [17,25,70], and for binary relation extraction from 28.0% to 83.6% (see Table 3). The task of n-ary relation extraction with an F1 score from 28.7% to 56.4% [56,58] is especially challenging, since such relationships usually span beyond sentences or even sections and thus, machine learning models require an understanding of the whole document. The inter-coder agreement for the task of concept extraction ranges from 0.6 to 0.96 (Table 2), for relation extraction from 0.6 to 0.9 (see also Table 3), while for coreference resolution the value of 0.68 was reported in two different studies [17,70]. The results suggest that these tasks are not only difficult for machines but also for humans in most cases.

### 2.4 Quality of knowledge graphs

KGs may contain billions of machine-readable facts about the world or a certain domain. However, do the KGs have also an appropriate quality? Data quality (DQ) is defined as *fitness for use by a data consumer* [107]. Thus, to evaluate data quality, it is important to know the needs of the data consumer since, in the end, the consumer judges whether or not a product is fit for use. Wang et al. [107] propose a data quality evaluation framework for information systems consisting of 15 dimensions grouped into four categories, i.e.:

1. **Intrinsic DQ**: accuracy, objectivity, believability, and reputation.
2. **Contextual DQ**: value-added, relevancy, timeliness, completeness, and an appropriate amount of data.
3. **Representational DQ**: interpretability, ease of understanding, representational consistency, and concise representation.
4. **Accessibility DQ**: accessibility and access security.

Bizer [10] and Zaveri [111] propose further dimensions for the Linked Data context like consistency, verifiability, offensiveness, licensing and interlinking. Pipino et al. [86] subdivide completeness into schema completeness, i.e. the extent to which classes and relations are missing in the ontology to support a certain use, column completeness (also known as Partial Closed World Assumption [46]), i.e. the extent to which facts are not missing, and population completeness, i.e. the extent to which instances for a certain class are missing. Führer et al. [38] comprehensively evaluate and compare the data quality of popular KGs (e.g. DBpedia, Freebase, Wikidata, YAGO) using such dimensions.

To evaluate the correctness of instance data (also known as precision), the facts in the KG have to be compared against a ground truth. For that, humans annotate a set of facts as true or false. YAGO found to be 95% correct [101]. The automatically populated AI-KG has a precision of 79% [32]. The KG automatically populated by the Never-Ending Language Learner (NELL) has a precision of 74% [21].

To evaluate the completeness of instance data (also known as coverage and recall), small collections of ground-truth capturing all knowledge for a certain ontology is necessary, that are usually difficult to obtain [108]. However, some studies estimate the completeness of several KGs. Galarrage et al. [45] suggest a rule mining approach to predict missing facts. In Freebase [12] 71% of people have an unknown place of birth, and 75% have an unknown nationality [35]. Suchanek et al. [100] report that 69%-99% of instances in popular KGs (e.g. YAGO, DBPedia) do not have at least one property that other instances of the same class have. The AI-KG has a recall of 81.2% [32].

### 2.5 Systematic literature reviews

Literature reviews are one of the main tasks of researchers, since a clear identification of a contribution to the present scholarly knowledge is a crucial step in scientific work [50]. This requires a comprehensive elaboration of the present
3 Requirements analysis

As the discussion of related work reveals, existing knowledge graphs for research information focus on specific use cases (e.g. search engines, help to reproduce research results) and mainly manage metadata and research artefacts about articles. We envision a KG in which research articles are linked through a deep semantic representation of their content to enable further use cases. In the following, we formulate the problem statement and describe our research method. This motivates our use case analysis in Section 3.1 from which we derive requirements for an ORKG.

Problem statement: Scholarly knowledge is very heterogeneous and diverse. Therefore, an ontology that conceptuallyises scholarly knowledge comprehensively does not exist. Besides, due to the complexity of the task, the population of comprehensive ontologies requires domain and ontology experts. Current automatic approaches can only populate rather simple ontologies and achieve moderate accuracy (see Section 2.3 and Appendix A). On the one hand, we desire an ontology that can comprehensively capture scholarly knowledge, and instance data with high correctness and completeness. On the other hand, we are faced with a “knowledge acquisition bottleneck”.

Research method: To illuminate the problem statement, we perform a requirements analysis. We follow the design science research (DSR) methodology [52,18]. The requirements analysis is a central phase in DSR, as it is the basis for design decisions and selection of methods to construct effective solutions systematically [13]. The objective of DSR in general is the innovative, rigorous and relevant design of information systems for solving important business problems, or the improvement of existing solutions [14,50].

To elicit requirements, we studied guidelines for (a) systematic literature reviews (see Section 2.5), (b) data quality requirements for information systems (see Section 2.4), and (c) interviewed members of the ORKG and Visual Analytics team at TIB, who are software engineers and researchers in the field of computer science and environmental sciences. Based on the requirements, we elaborate possible approaches to construct an ORKG, which were identified through a literature review (see Section 2.3). To verify our assumptions on the presented requirements and approaches, ORKG and Visual Analytics team members reviewed them in an iterative refinement process.

3.1 Overview of the use cases

We define functional requirements with use cases which are a popular technique in software engineering [13]. A use case describes the interaction between a user and the system from the user’s perspective to achieve a certain goal. Furthermore, a use case introduces a motivating scenario to guide the design of a supporting ontology and the use case analysis helps to figure out which kind of information is necessary [29].

https://projects.tib.eu/orkg/project/team/
https://www.tib.eu/en/research-development/
visual-analytics/staff
There are many use cases (e.g. literature reviews, plagiarism detection, peer reviewer suggestion) and several stakeholders (e.g. researchers, librarians, peer reviewers, practitioners) that may benefit from an ORKG. Nguyen et al. discuss some research-related tasks of scientists for information foraging at a broader level. In this study, we focus on use cases that support researchers (a) conducting literature reviews (see also Section 2.5), (b) obtaining a deep understanding of a research article and (c) reproducing research results. A full discussion of all possible use cases of graph-based knowledge management systems in the research environment is far beyond the scope of this article. With the chosen focus, we hope to cover the most frequent, literature-oriented tasks of scientists.

Figure 2 depicts the main identified use cases, which are described briefly in the following. Please note that we focus on how semantic content can improve these use cases and not further metadata.

**Get research field overview:** Survey articles provide an overview of a particular research field, e.g. a certain research problem or a family of approaches. The results in such surveys are sometimes summarised in structured and comparative tables (an approach usually followed in domains such as computer science, but not as systematically practised in other fields). However, once survey articles are published they are no longer updated. Moreover, they usually represent only the perspective of the authors, i.e. very few researchers of the field. To support researchers to obtain an up-to-date overview of a research field, the system should maintain such surveys in a structured way, and allow for dynamics and evolution. A researcher interested in such an overview should be able to search or to browse the desired research field in a user interface for ORKG access. Then, the system should retrieve related articles and available overviews, e.g. in a table or a leaderboard chart.

An ORKG should support the answering of queries related to such questions, which can be fine-grained or broad search intents. Preferably, the system should support natural language queries as approached by semantic search and question answering engines. The system has to return a set of relevant articles.

**Assess relevance:** Given a set of relevant articles the researcher has to assess whether the articles match the criteria of interest. Usually researchers skim through the title and abstract. Often, also the introduction and conclusions have to be considered, which is cumbersome and time-consuming. If only the most important paragraphs in the article are presented to the researcher in a structured way, this process can be boosted. Such information snippets might include, for instance, text passages that describe the problem tackled in the research work, the main contributions, the employed methods or materials, or the yielded results.

**Find related work:** Finding relevant research articles is a daily core activity of researchers. The primary goal of this use case is to find research articles which are relevant to a certain research question. A broad research question is often broken down into smaller, more specific sub-questions which are then converted to search queries. For instance, in this paper, we explored the following sub-questions: (a) Which ontologies do exist to represent scholarly knowledge? (b) Which scientific knowledge graphs do exist and which information do they contain? (c) Which datasets do exist for scientific information extraction? (d) What are current state-of-the-art methods for scientific information extraction? (e) Which approaches do exist to construct a knowledge graph?

**Find related work:**

| Use Case | Example Sources |
|----------|-----------------|
| Find related work | e.g. DataCite, Dataset Search, GitHub, WikiData, TIB AV-portal |
| Assess relevance | e.g. Wikipedia, beaker.org, TIB AV-portal |
| Get research field overview | e.g. ORKG backend |
| Obtain deep understanding | e.g. ORKG backend |
| Extract relevant information | e.g. ORKG backend |
| Obtain recommended articles | e.g. ORKG backend |
| Reproduce results | e.g. ORKG backend |
| Reproduce results | e.g. ORKG backend |

Fig. 2: UML use case diagram for the main use cases between the actor researcher, an Open Research Knowledge Graph (ORKG), and external systems.
Extract relevant information: To tackle a particular research question, the researcher has to extract relevant information from research articles. In a systematic literature review, the information to be extracted can be defined through a data extraction form (see Section 2.5). Such extracted information is usually compiled in written text or comparison tables in a related work section or survey articles. For instance, for the question “Which datasets do exist for scientific sentence classification?” a researcher who focuses on a new annotation study could be interested in (a) domains covered by the dataset and (b) the inter-coder agreement (see Table 2 as an example). Another researcher might follow the same question but focusing on machine learning, and thus could be more interested in (c) evaluation results and (d) feature types used.

The system should support the researcher with tailored information extraction from a set of research articles: (1) the researcher defines a data extraction form as proposed in systematic literature reviews (e.g. the fields (a)-(d)), and (2) the system presents the extracted information as suggestions for the corresponding data extraction form and articles in a comparative table. Figure 3 illustrates a data extraction form with corresponding fields in form of questions, and a possible approach to visualise the extracted text passages from the articles for the respective fields in a tabular form.

Get recommended articles: When the researcher focuses on a particular article, further related articles could be recommended by the system utilising an ORKG, for instance, articles that address the same research problem or apply similar methods.

Obtain deep understanding: The system should help the researcher to obtain a deep understanding of a research article (e.g. equations, algorithms, diagrams, datasets). For this purpose, the system should connect the article with artefacts such as conference videos, presentations, source code, datasets, etc., and visualise the artefacts appropriately. Also text passages can be linked, e.g. explanations of methods in Wikipedia, source code snippets of an algorithm implementation, or equations described in the article.

Reproduce results: The system should offer researchers links to all necessary artefacts to help to reproduce research results, e.g. datasets, source code, virtual research environments, materials describing the study, etc. Furthermore, the system should maintain semantic descriptions of domain-specific and standardised evaluation protocols and guidelines such as in machine learning reproducibility checklists and bioassays in the medical domain.

3.2 Knowledge graph requirements

As outlined in Section 2.4 data quality requirements should be considered within the context of a particular use case (“fitness for use”). In this section, we first describe dimensions we used to define non-functional requirements for an ORKG. Then, we discuss these requirements within the context of our identified use cases.
3.2.1 Dimensions for KG requirements

In the following, we describe the dimensions that we use to define the requirements for ontology design and instance data. We selected these dimensions since we assume that they are most relevant and also challenging to construct an ORKG with appropriate data to support the various use cases.

For ontology design, i.e. how comprehensively should an ontology conceptualise scholarly knowledge to support a certain use case, we use the following dimensions:

A) Domain specialisation of the ontology: How domain-specific should the concepts and relation types be in the ontology? An ontology with high domain specialisation targets a specific (sub-)domain and uses domain-specific terms. An ontology with low domain specialisation targets a broad range of domains and uses rather domain-independent terms. For instance, various ontologies (e.g. Scholarly Ontology [83]) propose domain independent concepts (e.g. Process, Method, Material). In contrast, Klampanos et al. [61] present a very domain-specific ontology for artificial neural networks.

B) Granularity of the ontology: Which granularity of the ontology is required to conceptualise scholarly knowledge? An ontology with high granularity conceptuallyis scholarly knowledge with a lot of classes that have very detailed and a lot of fine-grained properties and relations. An ontology with a low granularity has only a few classes and relation types. For instance, the annotation schemes for scientific corpora (see Section 2.3) have a rather low granularity, as they do not have more than 10 classes and 10 relation types. In contrast, various ontologies (e.g [49,83]) with more than 20 to 35 classes and over 20 to 70 relations and properties are fine-grained and have a relatively high granularity.

Although there is usually a correlation between domain specialisation and granularity of the ontology (e.g. an ontology with high domain-specialisation has also a high granularity), there exist also rather domain-independent ontologies with a high granularity, e.g. Scholarly Ontology [83], and ontologies with high domain-specialisation and low granularity, e.g. the PICO criterion in Evidence Based Medicine [59,91] which stands for Population (P), Intervention (I), Comparison (C), and Outcome (O). Thus, we use both dimensions independently. Furthermore, a high domain specialisation requirement for a use case implies that each sub-domain requires a separate ontology for the specific use case. These domain-specific ontologies can be organised in a taxonomy.

For the instance data, we use the following dimensions:

C) Completeness of the instance data: Given an ontology, to which extent do all possible instances (i.e. instances for classes and facts for relation types) in all research articles have to be represented in the KG? Low completeness: it is tolerable for the use case when a considerable amount of instance data is missing for the respective ontology. High completeness: it is mandatory for the use case that for the respective ontology, a considerable amount of instances are present in the instance data. For instance, given an ontology with a class “Task” and a relation type “subTaskOf” to describe a taxonomy of tasks, the instance data for that ontology would be complete if all tasks mentioned in all research articles are present (population completeness) and “subTaskOf” facts between the tasks are not missing (column completeness).

D) Correctness of the instance data: Given an ontology, which correctness is necessary for the corresponding instances? Low correctness: it is tolerable for the use case, that some instances (e.g. 30%) are not correct. High correctness: it is mandatory for the use case, that instance data must not be wrong i.e. all present instances in the KG must conform to the ontology and reflect the content of the research articles properly. For instance, an article is correctly assigned to the task addressed in the article, the F1 score in the evaluation results are correctly extracted, etc.

It should be noted that completeness and correctness of instance data can be evaluated only for a given ontology. For instance, let A be an ontology having the class “Deep Learning Model” without properties, and let B be an ontology that also has a class “Deep Learning Model” and additionally further relation types describing the properties of the deep learning model (e.g. drop-out, loss functions, etc.). In this example, the instance data of ontology A would be considered to have high completeness, if it covers most of the important deep learning models. However, for ontology B, the completeness of the same instance data would be rather low since the properties of the deep learning models are missing. The same holds for correctness: if ontology B has, for instance, a sub-type “Convolutional Neural Network”, then the instance data would have a rather low correctness for ontology B if all “Deep Learning Model” instances are typed only with the generic class “Deep Learning Model”.

3.2.2 Discussion of the KG requirements

Next, we discuss the seven main use cases with regard to the required level of ontology domain specialisation and granularity, as well as completeness and correctness of instance data. Table 1 summarises the requirements for the use cases along the four dimensions at ordinal scale. The use cases are grouped together, when they have (1) similar justifications for the requirements, and (2) a high overlap in ontology concepts and instances.
### Table 1: Requirements and approaches for the main use cases.

The upper part describes the minimum requirements for the ontology (domain specialisation and granularity) and the instance data (completeness and correctness). The bottom part lists possible approaches for manual, automatic and semi-automatic curation of the KG for the respective use cases. “X” indicates that the approach is suitable for the use case while “(x)” denotes that the approach is only appropriate with human supervision. The left part (delimited by the vertical triple line) groups use cases suitable for manual, and the right side for automatic approaches. Vertical double lines group use cases with similar requirements.

|                        | Extract relevant info | Research field overview | Deep understanding | Reproduce results | Find related work | Recommend articles | Assess relevance |
|------------------------|-----------------------|-------------------------|--------------------|-------------------|-------------------|--------------------|------------------|
| **Ontology**           |                       |                         |                    |                   |                   |                    |                  |
| Domain specialisation  | high                  | high                    | med                | med               | low               | low                | med              |
| Granularity            | high                  | high                    | med                | med               | low               | low                |                  |
| **Instance data**      |                       |                         |                    |                   |                   |                    |                  |
| Completeness           | low                   | med                     | low                | med               | high              | high               | med              |
| Correctness            | med                   | high                    | high               | high              | low               | low                | med              |
| **Manual curation**    |                       |                         |                    |                   |                   |                    |                  |
| Maintain terminologies  | -                     | X                       | -                  | -                 | X                 | X                  | -                |
| Define templates       | X                     | X                       | -                  | -                 | -                 | -                  | -                |
| Fill in templates      | X                     | X                       | X                  | X                 | -                 | -                  | -                |
| Maintain overviews     | X                     | X                       | -                  | -                 | -                 | -                  | -                |
| **Automatic curation** |                       |                         |                    |                   |                   |                    |                  |
| Entity/linking         | (x)                   | (x)                     | (x)                | (x)               | X                 | X                  | X                |
| Sentence classification | (x)                   | -                       | (x)                | (x)               | X                 | -                  | X                |
| Template-based extraction | (x)                   | (x)                     | (x)                | (x)               | -                 | -                  | -                |
| Cross-modal linking    | -                     | -                       | (x)                | (x)               | -                 | -                  | -                |

**Extract relevant information & get research field overview:** The information to be extracted from relevant research articles for a data extraction form within a literature review is very heterogeneous and depends highly on the intent of the researcher and the research questions. Thus, the ontology has to be domain-specific and fine-grained to offer all possible kinds of desirable information. However, missing information for certain questions in the KG may be tolerable for a researcher. Furthermore, it is tolerable for a researcher if some of the extracted suggestions are wrong since the researcher can correct them.

Research field overviews are usually the result of a literature review. The data in such an overview has also to be very domain-specific and fine-grained. Also, this information must have high correctness, e.g. an F1 score of an evaluation result must not be wrong. Furthermore, an overview of a particular research field should have appropriate completeness and must not miss any relevant research papers. However, it is acceptable when overviews for some research fields are missing.

**Obtain deep understanding & reproduce results:** The information required for these use cases has to achieve a high level of correctness (e.g. accurate links to dataset, source code, videos, articles, research infrastructures). An ontology for the representation of default artefacts can be rather domain-independent (e.g. Scholix [20]). However, semantic representation of evaluation protocols require domain-dependent ontologies (e.g. EXPO [97]). Missing information is tolerable for these use cases.

**Find related work & get recommended articles:** When searching for related work, it is essential not to miss relevant articles. Previous studies revealed that more than half of search queries in academic search engines refer to scientific entities [109]. However, the coverage of scientific entities in general-purpose KGs (e.g. WikiData) is rather low, since the introduction of new concepts in research literature occurs at a faster pace than KG curation [1]. Despite the low completeness, Xiong et al. [109] could improve the ranking of search results in academic search engines by exploiting general-purpose KGs. Hence, the instance data for the “find related work” use case should have high completeness with fine-grained scientific entities. However, semantic search engines leverage latent representations of KGs and text (e.g. graph and word embeddings) [6]. Since a non-perfect ranking of the search results is tolerable for a researcher, lower correctness of the instance data could be acceptable. Furthermore, due to latent feature representations, the ontology can be kept rather simple and domain-independent. For instance, the STM corpus [15] introduces four domain-independent concepts.

Graph- and content-based research paper recommendation systems [8] have similar requirements since they also leverage latent feature representations and require fine-grained scientific entities. Also, non-perfect recommendations are tolerable for a researcher.

**Assess relevance:** To help the researcher to assess the relevance of an article according to her needs, the system should highlight the most essential zones in the article to get a quick overview. The completeness and correctness of the presented...
information must not be too low, as otherwise the user acceptance may suffer. However, it can be suboptimal, since it is acceptable for a researcher when some of the highlighted information is not essential or when some important information is missing. The ontology to represent essential information should be rather domain-specific (i.e. using terms that the researchers understands) and quite simple (cf. ontologies for scientific sentence classification in Section 2.3.2).

4 ORKG construction strategies

In this section, we discuss the implications for the design and construction of an ORKG and outline possible approaches, which are mapped to the use cases in Table 1. Based on the discussion in the previous section, we can subdivide the use cases into two groups: (1) requiring high correctness and high domain specialisation with rather low requirements on the completeness (left side in Table 1), and (2) requiring high completeness with rather low requirements on the correctness and domain specialisation (right side in Table 1). The first group requires manual approaches while the second group could be accomplished with fully automatic approaches. To ensure trustworthiness, data records should contain provenance information, i.e. who or what system curated the data.

Manually curated data can also support use cases with automatic approaches, and vice versa. Furthermore, automatic approaches can complement manual approaches by providing suggestions in user interfaces. Such synergy between humans and algorithms may lead to a “data flywheel” (also known as data network effects, see Figure 3): users produce data which enable to build a smarter product with better algorithms so that more users use the product and thus produce more data, and so on.

4.1 Manual approaches

Ontology design: The first group of use cases requires rather domain-specific and fine-grained ontologies. We suggest to develop novel or reuse ontologies that fit the respective use case and the specific domain (e.g. EXPO [97] for experiments). Moreover, appropriate and simple user interfaces are necessary for efficient and easy population.

However, such ontologies can evolve with the help of the community, as demonstrated by WikiData and Wikipedia with “infoboxes” (see Section 2.3). Therefore, the system should enable the maintenance of templates, which are predefined and very specific forms consisting of fields with certain types (see Figure 5). For instance, to automatically generate leaderboards for machine learning tasks a template would have the fields Task, Model, Dataset and Score, which can then be filled in by a curator for articles providing such kind of results in a user interface generated from the template. Such an approach is based on meta-modelling [13], as the meta-model for templates enables the definition of concrete templates, which are then instantiated for articles.

Knowledge graph population: Several user interfaces are required to enable manual population: (1) populate semantic content for a research article by (1a) choosing relevant templates or ontologies and (1b) fill in the values; (2) terminology management (e.g. domain-specific research fields); (3) maintain research field overviews by (3a) assigning relevant research articles to the research field, (3b) define corresponding templates, and (3c) fill in the templates for the relevant research articles.

Furthermore, the system should also offer Application Programming Interfaces (APIs) to enable population by third-party applications, e.g.: – Submission portals such as https://www.easychair.org/ during submission of an article. – Authoring tools such as https://www.overleaf.com/ during writing. – Virtual research environments [99] to store evaluation results and links to datasets and source code during experimenting and data analysis.

To encourage crowd-sourced content, we see the following options:

– Top-down enforcement via submission portals and publishers.
– Incentive models: Researchers want their articles to be cited; semantic content helps other researchers to find, explore and understand an article. This is also related to the concept of enlightened self-interest, i.e. act to further interests of others to serve the own self-interest.
– Provide public acknowledgements for curators.
– Bring together experts (e.g. librarians, researchers from different institutions) who curate and organise content for specific research problems or disciplines.

4.2 (Semi-)automatic approaches

Ontology design: The second group of use cases require a high completeness while a relatively low correctness and domain specialisation are acceptable. For these use cases, rather simple or domain-independent ontologies should be developed or reused. Although approaches for automatic ontology learning exist (see Section 2.3), the quality of their results is not sufficient to generate a meaningful ORKG with complex conceptual models and relations. Therefore, meaningful ontologies should be designed by human experts.

Knowledge graph population: Various approaches can be used to (semi-)automatically populate an ORKG. Methods for entity and relation extraction (see Section 2.3) can help to populate fine-grained KGs with high completeness and entity linking approaches can link mentions in text with entities in KGs. For cross-modal linking, Singh et al. [96] suggest an approach to detect URLs to datasets in research articles automatically, while the Scientific Software Explorer [51] connects text passages in research articles with code fragments. To extract relevant information at sentence level, approaches for sentence classification in scientific text can be applied (see Section 2.3). To support the curator fill in templates semi-automatically, template-based extraction can (1) suggest relevant templates for a research article and (2) pre-fill fields of templates with appropriate values. For pre-filling, approaches such as n-ary relation extraction [43,53,56,58] or end-to-end question answering [90,33] could be applied.

Furthermore, the system should enable to plug external information extractors, developed for certain scientific domains to extract specific types of information. For instance, as depicted in Figure 5 an external template information extractor has to implement an interface with three methods. This enables the system (1) to filter relevant template extractors for an article and (2) extract field values from an article.

5 Conclusions

In this paper, we have presented a requirements analysis for an Open Research Knowledge Graph (ORKG). An ORKG should represent the content of research articles in a semantic way to enhance or enable a wide range of use cases. We identified literature-related core tasks of a researcher that can be supported by an ORKG and formulated them as use cases. For each use case, we discussed specificities and requirements for the underlying ontology and the instance data. In particular, we identified two groups of use cases: (1) the first group requires instance data with high correctness and rather fine-grained, domain-specific ontologies, but with moderate completeness; (2) the second group requires a high completeness, but the ontologies can be kept rather simple and domain-independent, and a moderate correctness of the instance data is sufficient. Based on the requirements, we have described possible manual and semi-automatic approaches (necessary for the first group), and automatic approaches (appropriate for the second group) for KG construction. In particular, we propose a framework with lightweight ontologies that can evolve by community curation. Furthermore, we have described the interdependence with external systems, user interfaces, and APIs for third-party applications to populate an ORKG.

The results of our work aim to give a holistic view of the requirements for an ORKG and guide further research. The suggested approaches have to be refined, implemented and evaluated in an iterative and incremental process (see www.orkg.org for the current progress). Additionally, our analysis can serve as a foundation for a discussion on ORKG requirements with other researchers and practitioners.

Conflict of interest

The authors declare that they have no conflict of interest.
A Comparative Overviews for Information Extraction Datasets from Scientific Text

Table 2 and Table 3 show comparative overviews for some datasets from research papers of various disciplines for the tasks sentence classification, relation extraction, and concept extraction, respectively.

References

1. Ammar, W., Groeneveld, D., Bhagavatula, C., Bethard, S., Ford, M., Downey, D., Dunkelberger, J., Elghorahy, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., Ooi, H., Peters, M.E., Power, J., Skjongsbjerg, S., Wang, L.L., Wilhelm, C., Yuan, Z., van Zuylen, M., Etzioni, O.: Construction of the literature graph in semantic scholar. In: S. Bangalore, J. Chu- Carroll, Y. Li (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1–6, 2018, Volume 3 (Industry Papers), pp. 84–91. Association for Computational Linguistics (2018). DOI 10.18653/v1/n18-3011. URL https://doi.org/10.1085/v1/n18-3011

2. Aryani, A., Wang, J.: Research graph: Building a distributed graph of scholarly works using research data switchboard. In: Open Repositories CONFERENCE (2017). DOI 10.4225/03/58e696655da3a. URL https://figshare.com/articles/Research_Graph_Building_a_Distributed_Graph_of_Scholarly_Works_using_Research_Data_Switchboard/4742413

3. Auer, S., Mann, S.: Towards an open research knowledge graph. The Serials Librarian 76(1–4), 35–41 (2019). DOI 10.1080/0361526X.2019.1540272. URL https://doi.org/10.1080/0361526X.2019.1540272

4. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: The Serials Librarian 76(1–4), 35–41 (2019). DOI 10.1080/0361526X.2019.1540272. URL https://doi.org/10.1080/0361526X.2019.1540272

5. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: SemEval 2017 task 10: Sciencedoc - extracting keyphrases and relations from scientific publications. In: S. Bethard, M. Grüniger (eds.) Formal Ontology in Information Semantics, Proceedings of the Fifth International Conference, OII 2019, Saarbrücken, Germany, October 31st - November 3rd, 2019, Proceedings, vol. 15, pp. 89–93. IOS Press (2020). DOI 10.3233/978-1-80020-140-3-208

6. Balog, K.: Entity-oriented search. Springer (2018). DOI 10.1007/978-3-319-39335-3. URL https://eos-book.org

7. Bechhofer, S., Buchan, I.E., Roure, D.D., Missier, P., Ainsworth, J.D., Bhagat, J., Couch, P.A., Crickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D.T., Owen, S., Newman, D.R., Suli, S., Goble, C.A.: Why linked data is not enough for scientists. Future Gener. Comput. Syst. 29(2), 599–611 (2013). DOI 10.1016/j.future.2011.08.004. URL https://doi.org/10.1016/j.future.2011.08.004

8. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. Int. J. Digit. Libr. 17(4), 305–338 (2016). DOI 10.1007/s10799-015-0156-0. URL https://doi.org/10.1007/s10799-015-0156-0

9. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: K. Inui, J. Jiang, V. Ng, X. Wan (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, pp. 3613–3618. Association for Computational Linguistics (2019). DOI 10.18653/v1/D19-1373. URL https://doi.org/10.18653/v1/D19-1373

10. Brack, A.: Quality-Driven Information Filtering - In the Context of Web-Based Information Systems. VDM Verlag, Saarbrücken, DEU (2007)

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

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

13. Boos, G., Rumbaugh, J., Jacobson, I.: Unified Modeling Language User Guide, The (2nd Edition) (Addison-Wesley Object Technology Series). Addison-Wesley Professional (2005)

14. Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. J. Assoc. Inf. Sci. Technol. 66(11), 2215–2222 (2015). DOI 10.1002/asi.23329. URL https://doi.org/10.1002/asi.23329

15. Brack, A., D’Souza, J., Poppe, A., Auer, S., Ewerth, R.: Domain-independent extraction of scientific concepts from research articles. In: J.M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M.J. Silva, F. Martins (eds.) Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I. Lecture Notes in Computer Science, vol. 12035, pp. 251–266. Springer (2020). DOI 10.1007/978-3-030-45439-5_17. URL https://doi.org/10.1007/978-3-030-45439-5_17

16. Brack, A., Poppe, A., Stocker, M., Auer, S., Ewerth, R.: Requirements analysis for an open research knowledge graph. In: M.M. Hall, T. Mercun, T. Risse, F. Dachateau (eds.) Digital Libraries for Open Knowledge - 24th International Conference on Theory and Practice of Digital Libraries, TDPL 2020, Lyon, France, August 25-27, 2020, Proceedings, Lecture Notes in Computer Science, vol. 12246, pp. 3–18. Springer (2020). DOI 10.1007/978-3-030-54956-5_1. URL https://doi.org/10.1007/978-3-030-54956-5_1

17. Brack, A., Müller, D., Poppe, A., Ewerth, R.: Coreference resolution in research papers from multiple domains. In: Proceedings of ECIR 2021 (accepted for publication) (2021)

18. Braun, R., Benedikt, M., Wendler, H., Esswein, W.: Proposal for requirements driven design science research. In: B. Donnellan, M. Helfert, J. Kenneally, D.E. VanderMeer, M.A. Rothenberger, R. Winter (eds.) New Horizons in Design Science: Broadening the Research Agenda - 10th International Conference, DESRIST 2015, Dublin, Ireland, May 20-22, 2015, Proceedings, Lecture Notes in Computer Science, vol. 9073, pp. 135–151. Springer (2015). DOI 10.1007/978-3-19-18714-3_9. URL https://doi.org/10.1007/978-3-19-18714-3_9

19. Brodaric, B., Reitsma, F., Qiang, Y.: Slicing with DOLCE: toward an e-science knowledge infrastructure. In: C. Eschenbach, M. Grüning (eds.) Formal Ontology in Information Systems, Proceedings of the Fifth International Conference, FOIS 2008, Saarbrücken, Germany, October 31st - November 3rd, 2008, Frontiers in Artificial Intelligence and Applications, vol. 183, pp. 208–219. IOS Press (2008). DOI 10.3233/978-1-58603-923-3-208. URL https://doi.org/10.3233/978-1-58603-923-3-208
Table 2: Characteristics of datasets and performance measures for sentence classification in research papers.

| Dataset       | Domains                                | # Papers | Coverage | Sentence Classes                                                                 | Inter-coder Agreement | Performance |
|---------------|----------------------------------------|----------|----------|----------------------------------------------------------------------------------|------------------------|-------------|
| PubMed-20k    | Biomedicine                            | 20,000   | abstracts| Background, Objective, Methods, Results, Conclusion                               | n/a                    | 92.9% F1    |
| NICTA-PIBOSO  | Biomedicine                            | 1,000    | abstracts| Background, Intervention, Study, Population, Outcome, Other                      | 62.0% k                | 84.7% F1    |
| CSABSTRUCT   | Computer Science                       | 2,189    | abstracts| Background, Objective, Method, Result, Other                                      | 75.0% k                | 83.1% F1    |
| CS-Abstracts  | Computer Science                       | 654      | abstracts| Background, Objective, Methods, Result, Conclusions                               | n/a                    | 74.6% F1    |
| Emerald 100k  | Management Information Science          | 103,457  | abstracts| Purpose, Design/methodology/approach, Findings, Originality/value, Social implications, Practical implications, Research limitations/implications | n/a                    | n/a         |
| MAZEA         | Physics                                | 1,335    | abstracts| Background, Gap, Purpose, Method, Result, Conclusion                               | 59.4% k                | 66.0% accuracy |
| Safder et al. | Computer Science                       | 92       | full text| Algorithmic Efficiency, Dataset Description, Algorithmic Time Complexity, Other    | n/a                    | 78.5% accuracy |
| Dr. Inventor  | Computer Graphics                      | 40       | full text| Background, Challenge, Approach, Outcome, Future Work                              | 66.7% k                | 72.5% accuracy |
| ART/CoreSC    | Chemistry, Computational Linguistic    | 225      | full text| Background, Motivation, Goal, Hypothesis, Object, Model, Method, Experiment, Result, Observation, Conclusion | 57.0% k                | 51.6% F1    |
Table 3: Characteristics of datasets and performance measures for binary and n-ary relation extraction in research papers.

| Dataset     | Domain          | # Papers | Coverage  | Cardinality | Relation Types                                                                 | Scope                      | Inter-coder Agreement | Mean Deviation | Mean | Median | # Papers | # Relations | # Relations Performed | # Relations Performed | Precision | Recall | F1   |
|-------------|-----------------|---------|-----------|-------------|--------------------------------------------------------------------------------|---------------------------|------------------------|-------------------|-----|--------|---------|------------|-----------------------|-----------------------|----------------------|---------|-------|------|
| SemEval17   | Computer Science| 500     | abstract  | binary      | synonym-of, hyponym-of                                                             | intra-sentence            | 60.0% k                | 34.3              | 4.5 | 4.5    | 56.4%   | 56.4%      | 56.4%                  | n/a                    |         |       |      |
| SemEval18   | Comp. Linguistics| 500    | abstract  | binary      | usage, result, model, part-whole, topic, comparison                               | intra-sentence            | 90.8% F1               | 73.1              | 67.8%| 67.8% | 49.3%   | 49.3%      | 49.3%                  | n/a                    |         |       |      |
| ChemProt    | Biomedicine     | 2482    | abstract  | binary      | UPREGULATOR, ACTIVATOR, DOWNREGULATOR, INHIBITOR, AGONIST, ANTICONSTITUTIVE, INHIBITOR, ACTIVATOR, DOWNREGULATOR, UPREGULATOR | n/a                       | n/a                   | 500               | 83.6%| 83.6% | 83.6%   | 83.6%      | 83.6%                  | n/a                    |         |       |      |
| SciERC      | Art. Intelligence| 500    | abstract  | binary      | hyponym-of, compare, part-of, conjunction, evaluate-for, feature-of, used-for    | cross-sentence            | 67.8% F1               | 4.716             | 67.8%| 67.8% | 39.3%   | 39.3%      | 39.3%                  | n/a                    |         |       |      |
| PWC         | Art. Intelligence| 731     | full text | n-ary       | (Task, Dataset, Metric, Score)                                                   | document-level            | n/a                   | 2,295             | 28.7%| 28.7% | 28.7%   | 28.7%      | 28.7%                  | n/a                    |         |       |      |
| CKB         | Biomedicine     | 343     | full text | n-ary       | (Drug, Gene, Mutation)                                                            | document-level            | n/a                   | 2,025             | 52.8%| 52.8% | 52.8%   | 52.8%      | 52.8%                  | n/a                    |         |       |      |
| SOFC-Exp    | Material Sciences| 45      | full text | n-ary       | (AnodeMaterial, CathodeMaterial, Device, ElectrolyteMaterial, FuelUsed, InterlayerMaterial, OpenCircuitVoltage, PowerDensity, Resistance, WorkingTemperature) | document-level            | n/a                   | n/a               | 56.4%| 56.4% | 56.4%   | 56.4%      | 56.4%                  | n/a                    |         |       |      |

* For SOFC-Exp corpus, performance values were obtained with bound full concept mentions.
Table 4: Characteristics of datasets and performance measures for scientific concept extraction in research papers. * For SOFC-Exp corpus, performance values were obtained with ground truth sentences describing experiments.

| Dataset     | Domains                  | # Papers | # Concepts | Coverage | Concept Types                   | Inter-coder Agreement | Performance |
|-------------|--------------------------|----------|------------|----------|---------------------------------|------------------------|-------------|
| SemEval17   | Computer Science, Material Sciences, Physics | 500      | 9,946      | abstract | Process, Task, Material         | 60.0% κ               | 56.9% F1    |
| STM         | Agriculture, Astronomy, Biology, Chemistry, Computer Science, Earth Science, Engineering, Materials Science, Science, Medicine | 110      | 6,127      | abstract | Process, Method, Material, Data | 76.0% κ               | 65.5% F1    |
| SciERC      | Art. Intelligence        | 500      | 8,089      | abstract | Task, Method, Metric, Material, Other, Generic | 76.9% κ               | 75.2% F1    |
| ACL2        | Comp. Linguistics        | 300      | 6,818      | abstract | Method, Tool, Language Resource (LR), LR product, Tool, Model, Measures/Measurements, Other | 63.0% F1               | 69.9% F1    |
| B5CDR       | Biomedicine              | 1500     | 28,785     | abstract | Chemical, Disease               | 91.8% F1              | 88.9% F1    |
| NCBI-disease| Biomedicine              | 793      | 6,892      | abstract | Disease                         | 88.0% F1              | 96.9% F1    |
| SOFC-Exp    | Material Sciences        | 45       | 4,004      | full text | Material, Device, Value         | 95.8% F1              | 81.5% F1    |

* For SOFC-Exp corpus, performance values were obtained with ground truth sentences describing experiments.
Requirements Analysis for an Open Research Knowledge Graph

30, 2019, Proceedings, Part II, Lecture Notes in Computer Science, vol. 11779, pp. 113–129. Springer (2019). DOI 10.1007/978-3-030-30796-7_8. URL https://doi.org/10.1007/978-3-030-30796-7_8

38. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of dbpedia, freebase, opencyc, wikidata, and YAGO. Semantic Web 9(1), 77–129 (2018). DOI 10.3233/SW-170275. URL https://doi.org/10.3233/SW-170275

39. Fathalla, S., Vahdani, S., Auer, S., Lange, C.: Towards a knowledge graph representing research findings by semantifying survey articles. In: J. Kamps, G. Tsakonas, Y. Manolopoulos, L.S. Iliaidis, I. Karydis (eds.) Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings, Lecture Notes in Computer Science, vol. 10450, pp. 315–327. Springer (2017). DOI 10.1007/978-3-319-67008-9_25. URL https://doi.org/10.1007/978-3-319-67008-9_25

40. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication, MIT Press, Cambridge, MA (1998)

41. Fink, A.: Conducting Research Literature Reviews: From the Internet to Paper. SAGE Publications (2014)

42. Fisas, B., Saggion, H., Ronzano, F.: On the discoursive structure of computer graphics research papers. In: A. Meyers, I. Rehbein, H. Zinsmeister (eds.) Proceedings of The 9th Linguistic Annotation Workshop, L@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA, pp. 42–51. The Association for Computer Linguistics (2015). DOI 10.3115/v1/w15-1605. URL https://https://doi.org/10.3115/v1/w15-1605

43. Friederich, A., Adel, H., Tomazic, F., Hingerl, J., Fortuna, A., Benteau, R., Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Q. 28(1), 75–105 (2004). URL http://misq.org/design-science-in-information-systems-research.html

44. Fries, B., Saggion, H., Ronzano, F.: On the discoursive structure of computer graphics research papers. In: A. Meyers, I. Rehbein, H. Zinsmeister (eds.) Proceedings of The 9th Linguistic Annotation Workshop, L@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA, pp. 42–51. The Association for Computer Linguistics (2015). DOI 10.3115/v1/w15-1605. URL https://https://doi.org/10.3115/v1/w15-1605

45. Friederich, A., Adel, H., Tomazic, F., Hingerl, J., Fortuna, A., Benteau, R., Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Q. 28(1), 75–105 (2004). URL http://misq.org/design-science-in-information-systems-research.html

46. Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.M.: AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In: D. Jurafsky, J. Chai, N. Schluter, J.R. Tetreault (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 1255–1268. Association for Computational Linguistics (2020). DOI 10.18653/v1/2020.acl-main.116. URL https://https://doi.org/10.18653/v1/2020.acl-main.116

47. Gaonkar, S., Cortez, P., Moro, S.: A deep learning classifier for sentence classification in biomedical and computer science abstracts. Neural Comput. Appl. 32(11), 6793–6807 (2020). DOI 10.1007/s00521-019-04334-2. URL https://doi.org/10.1007/s00521-019-04334-2

48. Gómez, T., Handschuh, S., Möller, K., Decker, S.: SALT - semantically annotated latex for scientific publications. In: E. Franconi, M. Kifer, W. May (eds.) The Semantic Web: Research and Applications, 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3-7, 2007, Proceedings, Lecture Notes in Computer Science, vol. 4519, pp. 518–532. Springer (2007). DOI 10.1007/978-3-540-72667-8_37. URL https://https://doi.org/10.1007/978-3-540-72667-8_37

49. Hars, A.: Structure of scientific knowledge, pp. 83–185. Springer Berlin Heidelberg, Berlin, Heidelberg (2003). DOI 10.1007/978-3-540-24737-1_3. URL https://https://doi.org/10.1007/978-3-540-24737-1_3

50. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Q. 28(1), 75–105 (2004). URL http://misq.org/design-science-in-information-systems-research.html

51. Hof, H., Jochim, C., Gleize, M., Bonin, C., Ganguly, D.: Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In: A. Korhonen, D.R. Traum, L. Marquez (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 5203–5213. Association for Computational Linguistics (2019). DOI 10.18653/v1/p19-1513. URL https://https://doi.org/10.18653/v1/p19-1513

52. Horvath, I.: Comparison of three methodological approaches of design research. In: S. (ed.) Proceedings of the 16th International Conference on Engineering Design, ICED’07, pp. 1–11. Ecole Central Paris (2007). Null : Conference date: 28-08-2007 Through 30-08-2007

53. Hou, Y., Jochim, C., Gleize, M., Bonin, C., Ganguly, D.: Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In: A. Korhonen, D.R. Traum, L. Marquez (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 5203–5213. Association for Computational Linguistics (2019). DOI 10.18653/v1/p19-1513. URL https://https://doi.org/10.18653/v1/p19-1513

54. Jain, S., van Zuylen, M., Hajiﬁshiri, H., Beltagy, I.: Sciex: A challenge dataset for document-level information extraction. In: D. Jurafsky, J. Chai, N. Schuter, J.R. Tetreault (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 1255–1268. Association for Computational Linguistics (2020). DOI 10.18653/v1/2020.acl-main.116. URL https://https://doi.org/10.18653/v1/2020.acl-main.116

55. Jaradhi, M.Y., Oelen, A., Prinz, M., Stocker, M., Auer, S.: Open research knowledge graph: A system walkthrough. In: A. Doucet, A. Isaac, K. Golub, T. Aalberg, A. Jatowt (eds.) Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings, Lecture Notes in Computer Science, vol. 11799, pp. 348–351. Springer (2019). DOI 10.1007/978-3-030-30760-8_31. URL https://https://doi.org/10.1007/978-3-030-30760-8_31

56. Jia, R., Wong, C., Poon, H.: Document-level n-ary relation extraction with multiscale representation learning. In: J. Burstein, C. Doran, T. Solorio (eds.) Proceedings of the 2019 Conference
77. Oelen, A., Jaradeh, M.Y., Stocker, M., Auer, S.: Generate Knowledge Graphs with Scholarly Literature Surveys. In: R. Huang, D. Wu, G. Marchionini, D. He, S.J. Cunningham, P. Hansen (eds.) JCDL ’20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020, pp. 97–106. ACM (2020). DOI 10.1145/3383583.3398520. URL https://doi.org/10.1145/3383583.3398520

80. Park, S., Caragea, C.: Scientific keyphrase identification and classification by pre-trained language models intermediate task transfer learning. In: D. Scott, N. Bel, C. Zong (eds.) Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, pp. 5409–5419. International Committee on Computational Linguistics (2020). DOI 10.18653/v1/2020.coling-main.472. URL https://doi.org/10.18653/v1/2020.coling-main.472

82. Peroni, S., Shotton, D.M.: Fabio and cito: Ontologies for describing bibliographic resources and citations. J. Web Semant. 17, 33–43 (2012). DOI 10.1016/j.websem.2012.08.001. URL https://doi.org/10.1016/j.websem.2012.08.001

83. Pertsa, V., Constantopoulos, P.: Scholarly ontology: modeling scholarly practices. Int. J. Digit. Libr. 18(3), 173–190 (2017). DOI 10.1007/s10799-016-0169-3. URL https://doi.org/10.1007/s10799-016-0169-3

85. Petasis, G., Karkalevits, V., Palioras, G., Kritthara, A., Zavitsanos, E.: Ontology population and enrichment: State of the art. In: D. Demner-Fushman, K.B. Cohen, S. Ananiadou, I. Tsuji (eds.) Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019, pp. 58–65. Association for Computational Linguistics (2019). DOI 10.18653/v1/w19-5006. URL https://doi.org/10.18653/v1/w19-5006

88. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: J. Su, X. Carreras, K. Duh (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. pp. 2383–2392. The Association for Computational Linguistics (2016). DOI 10.18653/v1/d16-1264. URL https://doi.org/10.18653/v1/d16-1264

90. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: J. Su, X. Carreras, K. Duh (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. pp. 2383–2392. The Association for Computational Linguistics (2016). DOI 10.18653/v1/d16-1264. URL https://doi.org/10.18653/v1/d16-1264

91. Richardson, S., Wilson, M., Nishikawa, J., Hayward, R.: The well-built clinical question: a key to evidence-based decisions. ACP journal club 123(3), A12–13 (1995)

94. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Birukou, A., Osborne, F., Motta, E.: The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. Data Intell. 2(3), 115–130 (2018). DOI 10.1007/s12306-018-00055-2. URL https://doi.org/10.1007/s12306-018-00055-2

95. Say, A., Fathalla, S., Vahdati, S., Lehmann, J., Auer, S.: Semantic representation of physics research data. In: D. Aveiro, J.L.G. Dietz, J. Filipe (eds.) Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2020, Volume 2: KEOD, Budapest, Hungary, November 2-4, 2020, pp. 64–75. SCITEPRESS (2020). DOI 10.5220/00110100640075. URL https://doi.org/10.5220/00110100640075

96. Singh, M., Barua, B., Palod, P., Garg, M., Satapathy, S., Bushi, S., Ayush, K., Rohith, K.S., Gamidi, T., Goyal, P., Mukherjee, A.: OCR++: A robust framework for information extraction from scholarly articles. In: N. Calzolari, Y. Matsumoto, R. Prasad (eds.) COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, pp. 3390–3400. ACL (2016). URL https://www.aclweb.org/anthology/C16-1320/

97. Soldatova, L.N., King, R.D.: An ontology of scientific experiments. Journal of The Royal Society Interface 3(11), 795–803 (2006). DOI 10.1098/rsif.2006.0134. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2006.0134
