Bridging the Innovation Gap: Leveraging Patent Information for Scientists by Constructing a Patent-centric Knowledge Graph

Hidir Aras[1,*], Rima Dessi[1], Farag Saad[1] and Lei Zhang[1]

[1]FIZ Karlsruhe, Germany

Abstract

The purpose of patents is to allow companies and individuals to disclose inventions and legally protect intellectual property. Despite their recognition as a crucial source of scientific and technical knowledge in industrial contexts, the use of patent information by scientists remains low. Research surveys have identified several key barriers that hinder the effective use of patents by the scientific community. These include the complexity of the patent information landscape, characterized by domain-specific terminology and the extensive and varied nature of content. Additionally, challenges in accessing and comprehending patent knowledge further discourage widespread use among researchers. To address these issues, it is crucial to improve the integration of relevant information sources. Enhancing the interlinking of related objects and focusing on the targeted extraction of pertinent sections from patent documents through the use of semantic information derived from knowledge graphs (KG) could significantly ease the process. In this paper, we present our efforts to construct a Patent Knowledge Graph (PKG) by annotating, linking and integrating essential knowledge from patent text with scientific literature and domain-specific knowledge by leveraging semantics from the Linked Open Data (LOD) cloud and domain-specific ontologies for several scientific domains. Such advancements would enable scientists to more effectively access and leverage the essential information that is often concealed within patents.

Keywords

Patent Knowledge, LOD, Entity Linking, Named Entity Recognition, Domain-specific Knowledge

1. Introduction

Patents contain important scientific and technical information and serve as an important source for innovations in the field of intellectual property and related industries. Though their informational value is widely acknowledged, their actual use in non-industrial settings among scientists is rather low. The findings from several surveys [1, 2] indicate that the complexity of the information space, such as domain-specific vocabulary, the large amount of heterogeneous content with mixed domains and efficiency constraints for accessing and using patent information are the biggest obstacles. This fact also poses several challenges for the integration of such diverse knowledge for the application of AI systems. In order to overcome these challenges, interlinking patent information with scientific literature and domain-specific information sources by converting relevant textual content into a semantic representation can

2nd Workshop on Semantic Technologies for Scientific,Technical and Legal Data (SemTech4STLD'24)

*Corresponding author.

hidir.aras@fiz-karlsruhe.de (H. Aras); rima.dessi@fiz-karlsruhe.de (R. Dessi); farag.saad@fiz-karlsruhe.de (F. Saad); lei.zhang@fiz-karlsruhe.de (L. Zhang)

be regarded as important steps for enhancing the understandability and use of patent knowledge. In this paper, we describe a novel approach for forming a patent-centric KG based on explicit semantic information (entities) from large-scale KGs in the LOD cloud, e.g. Wikidata, and domain-specific knowledge bases. First, we extract the relevant conceptual knowledge for the regarded domains from Wikidata in order to enrich (annotate) and link patents, scientific literature and domain-specific sources. In order to link such explicit knowledge with relevant textual parts in patent text, we employ named entity recognition (NER) for the reliable detection of entity mentions. Our PKG is based on an extended ontological model for representing core information in patent documents and its fulltext. Beside semantic knowledge from the LOD cloud we exploit additional conceptual knowledge from domain-specific ontologies, e.g. for the plasma technology or material science and engineering domains. Herewith our approach allows to answer complex semantic queries/research questions of scientists, e.g. related to plasma medicine such as "*Which patents describe a plasma source that can be used for decontamination of room air and produces a plasma with virucidal efficacy?*"

2. Related Work

In the literature, patent documents firstly were encoded as LOD in the work of [1]. In the same year, the meta-data (incl. bibliographic data) of all EP patents since 1978 has been made available as "Linked open EP data" by the European Patent Office (EPO) comprising several millions of triples freely available with open license. The underlying semantic data model was described by [3]. The usage of a knowledge-based approach for patent retrieval from multiple sources was presented earlier by [4], where U.S. patents, court litigations, scientific publications and domain-specific information from the biomedical technology domain were covered. A first knowledge infrastructure that used Semantic Web standards to enable semantic interoperability for drug discovery was presented in [5]. In [6], the authors describe their approach of interlinking datasets from SciGraph and DBpedia using link discovery and named entity recognition (NER). Furtermore, semantic modeling of domain-specific knowledge for the plasma technology (PT) was researched in the QPTDat project[1], where a PT ontology serves as a basis for construction a KG for plasma research data based on the Plasma-MDS schema [7] in its core. In the work of [8] the PMD Core Ontology (PMDco) builds a foundation for domain-specific ontology development in the material science and engineering (MSE) domains. Herewith, data interoperability for the diverse material science domains following FAIR and LOD standards can be established. To the best of our knowledge, there is currently no existing research focused on providing in-depth semantic access to interlinked patent information. Our approach aims to integrate patent data with scientific literature and domain-specific knowledge bases using established Semantic Web standards, along with LOD and FAIR data principles. This integration will enhance the accessibility and utility of patent knowledge, connecting it more effectively with relevant academic and industry resources.

[1] https://zenodo.org/records/4350287

3. A Patent-centric Knowledge Graph Model

A semantic representation of patent knowledge involves creating a comprehensive and adaptable model that clearly defines the structure and components of a patent. Such a model aims to facilitate the enrichment of crucial textual content from patents by adding meaning and context, hence, making it easier to integrate with external scientific and domain-specific knowledge bases. By systematically enhancing and linking this content semantically, the model allows for a deeper and more effective connection between patent information and relevant external resources. Together with the core semantic model of a patent incl. meta-data and its (linked) entity annotations, we form a patent-centric KG, allowing us to answer research questions such as the aforementioned one, expressed as semantic queries in SPARQL.

3.1. The Patent4Science Ontology

The ontology behind the PKG consists of classes and properties that are tailored to a semantic representation of an enriched (and structured) patent document. The current version encompasses 19 classes, 21 object properties (semantic relations), and 47 datatype properties. Figure 1 illustrates the compact overview of the ontology.

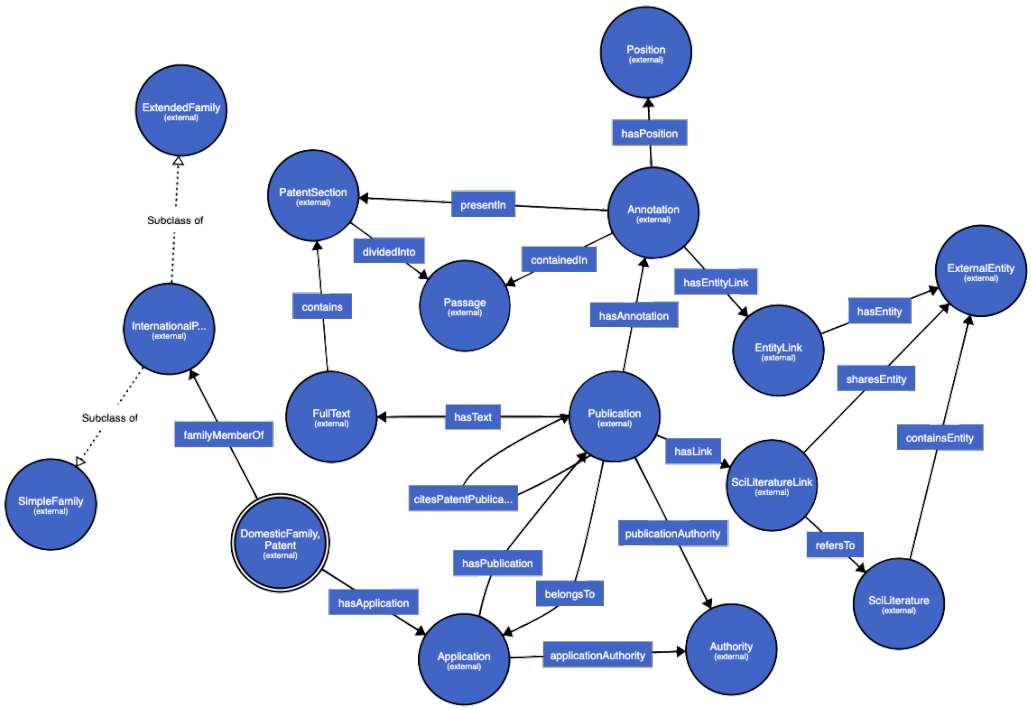

Figure 1: The general overview of the ontology.

In a first step, patent documents belonging to the PT, AM and BM domains are collected from the several professional patent databases. Then the PKG is formed by structuring and encoding

patent metadata (e.g., patent's application, number, classification code, publication date, etc.) in the Resource Description Framework (RDF) using the specified ontological model. Each patent is represented as an instance of `:Patent` class and the properties are also utilized to depict the patent's metadata. Typically, a patent belongs to a patent family which consists of multiple patent applications related to the same invention. Additionally, we model the corresponding patent publication in a dedicated class `:Publication` and its textual content, i.e., the textual content of each patent is associated with an instance of the class `:FullText` which consists of patent sections (i.e., abstract, claim, description).

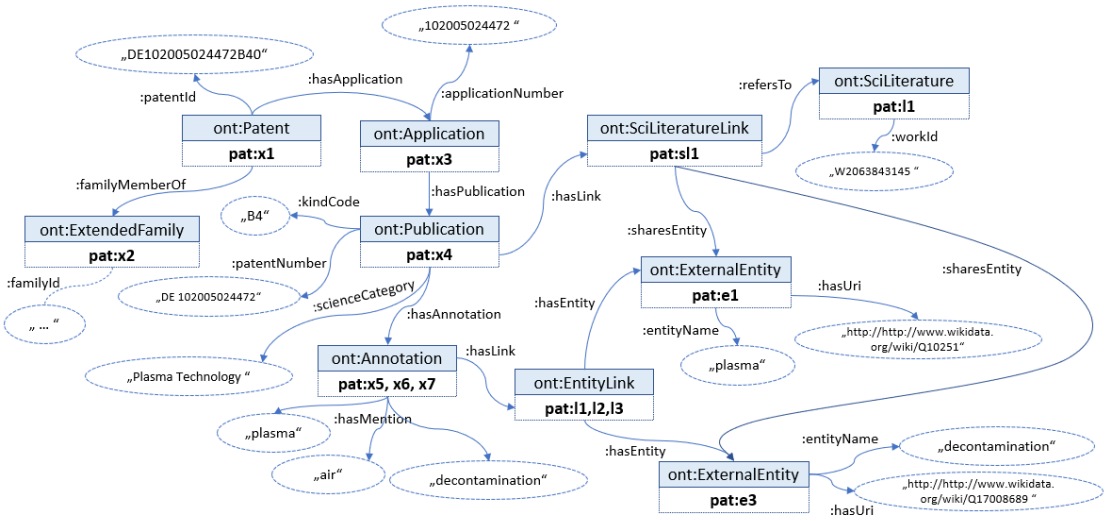

Figure 2: Linking patents, sci literature and domain-specific knowledge.

These are instances of the `:PatentSection` class. Each section (or even paragraph) is enriched (annotated) with the domain-specific entities sourced from Wikidata (see Section 4.3), enabling a connection to the LOD cloud. These annotations are then linked to the corresponding patent publication instances in the PKG with the help of the `:hasAnnotation` property. In this way, annotations of each section and their paragraphs are easily accessible which helps advanced analysis of the patent knowledge. To further augment the PKG, relevant scientific literature sourced from OpenAlex[2] is linked to each patent publication based on their relevancy score (see Section 4.4) with the `:hasLink` property. Overall, this graph captures the relations between various elements, including patent publications, their entity annotations, related scientific publications, and more. In the example fragment shown in Figure 2, the textual content of the patent "DE-102005024472-B4" was annotated with the Wikidata entities Q10251 (plasma), Q7391292 (air) and Q17008689 (decontamination) and linked to the scientific literature work W2063843145 from OpenAlex by calculating shared entities Q10251, Q17008689. In a similar way, further domain-specific knowledge bases for each of the aforementioned domains can be integrated and linked in the PKG employing the approaches described in the next chapter.

[2]https://openalex.org/

4. Linking Domain Knowledge and Scientific Literature

Large-scale KGs like Wikidata contain general and specific forms of explicit semantic information (conceptual entities and their relations) that can be used to enrich domain-specific information sources such as for plasma technology or material science. One way to extract relevant entities from Wikidata is to make use of Wikipedia categories for each domain to be covered. To link relevant explicit knowledge from such external large-scale KGs, mentions of entities in patent text must be identified employing rule-based or machine learning methods. Identifying such entities in patent text is not a trivial task due to many factors such as the complexity of the semantic structures of the entities, fuzzy entity boundaries, abundant use of synonyms, hyphens, digits, characters, and ambiguous abbreviations, etc. Therefore, we employ entity recognition and entity linking for the semantic annotation and disambiguation of patent knowledge. Hereby, links to relevant scientific literature can be established based on shared entities in patent text and scientific article. In a similar way, we link conceptual knowledge from a domain-specific KG e.g. for the plasma technology domain utilizing shared conceptual knowledge.

4.1. Domain-specific Named Entity Recognition and Classification

A central part of creating a PKG is to automatically identify named entities in the patent text. These named entities will be used as part of the PKG creation and linking. The aim of NER is to determine entity mentions, i.e., which terms in patent text might refer to an entity. In addition, we also classify these entity mentions into pre-defined entity types, such as 'plasma source' in the domain of plasma technology. There are some well-known existing tools for named entity recognition and classification, e.g. Stanford NER[3] and OpenNLP[4]. While these tools can deal with general text and entity types very well, they are not suitable for domain-specific patent text. Therefore, we have developed a NER model customized for patent text by applying a Bi-LSTM deep neural network [9], which has been originally designed for the biomedical domain and extended to other scientific domains in this work, e.g., plasma technology.

4.2. Extraction of Domain-specific Entities

To enrich the PKG based on external knowledge, we use Wikidata as our major external KG from the LOD cloud. The reason we choose Wikidata is due to the fact that patents generally cover a wide range of topics and Wikidata is a quite comprehensive KG that contains millions of entities in many domains. However, for a specific target domain, e.g., plasma technology, a large number of entities in Wikidata might not be needed. Therefore, we extract only entities that are related to the target domains by exploiting relevant Wikipedia categories, which are used to link articles under a common topic so that all categories form a hierarchy. Based on that, we firstly extracted Wikipedia articles from the given categories, e.g., `Plasma_physics`[5].

[3] https://nlp.stanford.edu/software/CRF-NER.shtml
[4] https://opennlp.apache.org/
[5] https://en.wikipedia.org/wiki/Category:Plasma_physics

4.3. Domain-specific Entity Linking

To enrich the PKG with domain-specific entity annotations, we apply entity linking (mention detection and entity disambiguation) to annotate patent texts with the Wikidata entities extracted for each target domain. For mention detection, we firstly rely on the named entities detected by the domain-specific NER described in 4.1. In addition, we use another POS tagging based method [10] to detect all noun phrases in patent texts that match against any labels of the extracted domain entities for the target domains. For all these mentions detected by both methods, their candidate entities are then extracted based on entity labels from Wikidata. After that, entity disambiguation aims to map ambiguous entity mentions onto entities from Wikidata, where we use an entity disambiguation method based on the PageRank algorithm. In addition, we compute a confidence score for each entity link. We firstly assign an entity-mention score for each candidate entity, which represents how often a mention is used to refer to the candidate entity, and based on that we re-weight each entity link using the PageRank algorithm and the final scores are normalized by the maximal score of all links in a patent text [11].

4.4. Scientific Literature Linking

In order to integrate and link scientific literature in the PKG, we utilize OpenAlex[6], a free and open catalog of the world's scholarly publications, which contains extensive metadata across scientific works, authors, publication venues, institutions, and concepts. It indexes over 240M works, which are scholarly documents like papers, journal articles, books and theses. These works are linked to more than 65,000 concepts, which are entities from Wikidata. Since the PKG focuses on specific target domains, we collect OpenAlex works using the previously extracted entities for each of the target domain. In other words, only the OpenAlex works that are linked with these extracted and annotated entities in patent text are considered. Hence, we create links between patents and scientific literature from OpenAlex based on their shared Wikidata entities. More specifically, for each patent in a target domain, we use the domain entities linked with this patent generated by the entity linking step to search for the works from OpenAlex that are also linked with the same domain entities, and once the number of common entities linked with the patent and a work exceeds a threshold (i.e., 5 chosen in this work based on a manual comparison of the results), we create a link between the patent and the work.

5. Summary and Outlook

In this paper, we have described our approach to building a patent-centric KG based on a semantic data model of patents that is enriched and linked by using explicit semantics from external KGs. Hereby, we used domain-specific semantics from the PT and MSE domains to annotate and link patent knowledge with scientific literature and domain-specific knowledge sources. At the current state, knowledge integration in the PKG is based on Wikidata entities shared between patents, scientific literature, and other domain-specific sources. In the future, we will use and integrate explicit semantics from dedicated knowledge bases for selected target domains. By exploiting semantics from their underlying domain-specific ontologies such as

[6]https://openalex.org/

the PTO[7] or PMDCo[8], we will enable a more consistent coverage of patent knowledge for the benefit of applications for patent retrieval and analysis.

6. Acknowledgments

This work was partly funded by the DFG project Patents4Science[9], Project id: 496963457.

References

[1] Z. A. Hassan, M., J. Lehmann, A linked open data representation of patents registered in the us from 2005–2017, Scientific Data (2018).

[2] S. Senger, Assessment of the significance of patent-derived information for the early identification of compound–target interaction hypotheses, J. of Cheminformatics (2017).

[3] M. Kracker, European patent information and the cpc taxonomy as linked open data, in: Proceedings of the Posters and Demos Track of the Semantics 2018 Conference, 2018.

[4] S. Taduri, G. T. Lau, K. H. Law, J. P. Kesan, A patent system ontology for facilitating retrieval of patent related information, Association for Computing Machinery, New York, NY, USA, 2012.

[5] A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble, B. Mons, Open phacts: semantic interoperability for drug discovery, Drug Discovery Today 17 (2012) 1188–1198.

[6] B. Yaman, M. Pasin, M. Freudenberg, Interlinking scigraph and dbpedia datasets using link discovery and named entity recognition techniques, in: M. Eskevich, G. de Melo, C. Fäth, J. P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek, M. Dojchinovski (Eds.), 2nd Conference on Language, Data and Knowledge,LDK 2019, May 20-23, 2019, Leipzig, Germany, volume 70 of *OASIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019, pp. 15:1–15:8.

[7] S. Franke, L. Paulet, J. Schäfer, D. O'Connell, M. M. Becker, Plasma-mds, a metadata schema for plasma science with examples from plasma technology, Scientific Data 7 (2020) 439.

[8] B. Bayerlein, M. Schilling, H. Birkholz, M. Jung, J. Waitelonis, L. Mädler, H. Sack, Pmd core ontology: Achieving semantic interoperability in materials science, Materials Design 237 (2024) 112603.

[9] F. Saad, H. Aras, R. Hackl-Sommer, Improving named entity recognition for biomedical and patent data using bi-lstm deep neural network models, in: E. Métais, F. Meziane, H. Horacek, P. Cimiano (Eds.), NLDB 2020 - 25th Int. Conference on Applications of Natural Language to Information Systems, Saarbrücken, Germany, June 24-26, 2020, Proceedings, volume 12089 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 25–36.

[10] L. Zhang, A. Rettinger, P. Philipp, Context-aware entity disambiguation in text using markov chains, in: 2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016, Omaha, NE, USA, October 13-16, 2016, IEEE Computer Society, 2016, pp. 49–56.

[11] L. Zhang, A. Rettinger, X-lisa: Cross-lingual semantic annotation, Proc. VLDB Endow. 7 (2014) 1693–1696.

[7] https://zenodo.org/records/3758137

[8] https://github.com/materialdigital/core-ontology

[9] https://www.pat4sci.org/