Skills and Expertise in Large Organizations

An Enterprise Knowledge Graph Approach

Blerina Spahiu[1,*], Anna Lisa Gentile[2], Chad DeLuca[2] and Andrea Maurino[1]

[1]Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126, Milan, Italy
[2]IBM Research - Almaden Lab, 650 Harry Rd, San Jose, CA 95120, United States

Abstract

Big organizations have a complex ecosystem of entities: products, people, skills, and intellectual properties. Formally capturing, maintaining, and serving this knowledge is a complex challenge. Enterprise Knowledge Graphs (EKG) are an effective method to represent enterprise information in ways that can be more easily interpreted by both humans and machines. In this study, we concentrate on the EKG's section related to individuals' skills and expertise. We present a method to determine the topics that employees are knowledgeable about, using the text from their scholarly publications and patents. We use publicly available datasets on US patents and scholarly publications and apply Information Extraction techniques to extract skills from the text and represent them in the EKG format. The resulting EKG proves valuable for querying and analyzing employees' skills, helping to identify experts in specific domains.

Keywords

knowledge graph construction, skills extraction, scholarly data

1. Introduction

Knowledge Graphs (KGs) are powerful tools for organizing and representing information in a structured and semantically rich manner [1]. In addition to the generic and open-world KGs such as DBpedia [2], Wikidata [3], most of the current KGs are domain-specific that focus on specific topics or areas of interest [4] such as economy [5], medicine [6], social science [7], etc. These KGs focusing on a specific topic have a narrower scope, but they can be more detailed and accurate in the coverage of the particular domain that they are developed to represent. Despite the numerous available KGs, there is no "one fits all" solution, and it is often necessary to build, enhance, refine, and enrich domain-specific knowledge graphs to serve specific use cases.

Innovative organizations need to stay current with rapidly evolving research areas [8]. Access to tools that offer up-to-date insights is invaluable for executives making strategic decisions and researchers looking to improve the state of the art [9]. In addition, for organizations with thousands of employees and diverse scientific disciplines, having a centralised resource to integrate and manage all needed data is crucial. Our idea is that semantically representing and enriching each asset, project, scientific publication, intellectual property - and eventually employee profile - can significantly enhance and facilitate all downstream tasks related to skills discovery, trend analysis, expertise matching, and so on.

The Semantic Web community has proposed various scholarly KGs like OAG [10], ORKG [11], OpenAlex [12], CS-KG [13], and AIDA [14]. Notable ontologies for organizational resources and employee knowledge include ORG Vocabulary[1], W3C Organization Ontology[2], and Schema.org[3]. Domain-specific ontologies, such as ESCO[4] and SARO[5], focus on occupations, skills, and recruitment.

*Corresponding author.
blerina.spahiu@unimib.it (B. Spahiu); annalisa.gentile@ibm.com (A. L. Gentile); delucac@us.ibm.com (C. DeLuca); andrea.maurino@unimib.it (A. Maurino)

[1]https://epimorphics.com/public/vocabulary/org.html
[2]https://www.w3.org/TR/vocab-org/
[3]https://schema.org/
[4]https://esco.ec.europa.eu/en/use-esco/download
[5]https://elisasibarani.github.io/SARO/

However, such KGs and ontologies have some limitations related to (i) *coverage* - most do not include information about patents; (ii) *representation* - none of them represents skills as a direct property of employees; (iii) *functionalities* - i.e. do not offer advanced query capabilities.

We create a Scholarly Enterprise Knowledge Graph (sEKG) containing an augmented representation of internal assets (publications, patents, projects, etc) as well as publicly available external assets from other research and innovation institutions. sEKG is effectively a collection of scholarly data (papers and patents), integrated in a machine-readable format and enriched with external knowledge resources, to enable efficient analysis and inference of implicit skills and organisational hierarchy. We bootstrap the knowledge population task with standard information about each asset, i.e. type of asset (Intellectual Property, Service, R&D product, Scientific paper, etc.), names, textual descriptions, owners, etc. Then we enrich each asset, extracting relevant concepts and linking them to external related ontologies and Linked Data concepts. We extended the W3C Organisational Ontology to represent employees' knowledge and department organisations. By leveraging the collected and enriched data, we can infer the skills of individual authors and use this information in different use cases, including: (i) identifying the right experts for specific projects, (ii) empowering skills growth among employees, (iii) identifying external competitors, and (iv) supporting employers in performing daily tasks. The primary contributions of this research encompass the following:

- Introduction of sEKG, the scholarly Enterprise Knowledge Graph that integrates scholarly and patent data, and enriches such data with potential skills for each employee.
- Exploration of diverse sEKG application scenarios.
- Public release of our extended ontology, derived from the W3C Organisational Ontology, tailored to meet the specific requirements of our sEKG.
- Presentation of the outcomes of a user study, illustrating the utility and potential of sEKG.

The remainder of this paper is structured as follows. After reviewing the related work in Section 2 we describe the sEKG creation in Section 3. Section 4 discusses potential application scenarios and describes a small pilot study. We discuss lessons learned and the potential evolution of this work in Section 5.

2. Related Work

2.1. Scholarly Knowledge Graphs

Several large knowledge graphs have been proposed in the scholarly field, and they cover vast information about entities such as publications, authors, and venues [10, 11, 12] or scholarly KGs that focus on a particular domain of research such as computer science [13], science [15], medicine [16], etc.

The first considerable effort to offer comprehensive semantic descriptions of conference events is represented by the metadata projects at ESWC 2006 and ISWC 2006 conferences [17], This projected generated the first version of the Semantic Web Conference Ontology[6] which has been later refactored [18] and is still used to collect conference data[7].

Open Academic Graph[8] (OAG) is a large knowledge graph unifying two billion-scale academic graphs: Microsoft Academic Graph (MAG) [10] and AMiner [19]. As of July 2020, the snapshot of such KG contains metadata for more than 239 million publications from all scientific disciplines, as well as over 1.38 billion references between publications making it one of the largest freely available scholarly knowledge graphs. OAG is built automatically by using machine learning algorithms and natural language processing techniques to extract and link information from academic papers. OAG does not provide APIs or tools for querying and analyzing data; it only provides links to download bulk data[9].

[6]Semantic Web Conference Ontology http://data.semanticweb.org/ns/swc/swc_2009-05-09.html
[7]http://www.scholarlydata.org/
[8]https://www.microsoft.com/en-us/research/project/open-academic-graph/
[9]https://www.aminer.cn/oag-2-1

Open Research Knowledge Graph[10] (ORKG) is an open and collaborative platform that has the aim to integrate research and academic knowledge in a structured way [11]. The ORKG contains not only entities regarding research papers, authors, institutions, research topics, and concepts, but also the relationships between them. Designed to be a community-driven platform that encourages the sharing of research knowledge and facilitates collaboration among researchers, ORKG provides a search functionality that allows users to explore the knowledge graphs and discover new research topics and connections. Moreover, it provides APIs that enable developers to build applications that use different aspects of the data it contains.

OpenAlex[11] is a heterogeneous directed graph, composed of five types of scholarly entities (authors, institutions, concepts, publishers, and sources), and the connections between them. It includes more than 248M works and it contains important identifiers including ORCIS, ROR, ISSN, etc. OpenAlex data can be used to build scholarly search engines, recommender services, or domain-specific knowledge graphs. It can help manage research by tracking citation impact, spotting emerging areas, etc [12].

Domain-specific knowledge graphs regarding science and computer science fields represent structured information about concepts, topics, entities, and relationships in the field of computer science such as CS-KG [13, 14]. The Academia/Industry DynAmics (AIDA) Knowledge Graph describes 21M publications and 8M patents according to the research topics drawn from the Computer Science Ontology[12].

Despite these continuous efforts, it has been argued that a great deal of information about academic conferences is still missing or spread across several sources in a largely chaotic and non-structured way [20]. Besides the problem of missing content, one of the other major challenges with scholarly data is to ensure data quality, which means dealing with data-entry errors, disparate citation formats, lack of (enforcement of) standards, imperfect citation-gathering software, ambiguous author names, and abbreviations of publication venue titles [21].

Although many generic or domain-specific scholarly KGs have been developed in the state-of-the-art, they have several drawbacks with regard to the aim of this paper: (i) most KGs do not include patent information, apart from AIDA, which itself does not offer rich query capabilities, including filtering by author name; (ii) none of the KGs include authors' skills. Moreover, to the best of our knowledge, none of the available literature showcases a real use-case scenario of such KGs.

2.2. Organizational and Skills Ontologies

Organizational ontologies represent entities, relationships, and properties that are relevant to organizations with the aim to facilitate the sharing and integration of organizational knowledge.

The ORG Vocabulary and W3C Organization Ontology are two well-known ontologies to describe the organisation of a company. The ORG ontology was developed as part of data.gov.uk initiative and is a small and generic ontology with the aim to publish information for the organizational structure of government institutions. It provides minimal basic terms to support representations of: (i) organizational structure, (ii) reporting structure, (iii) location information, and (iv) organizational history (merger, renaming, re-purposing). However, the ORG ontology is limited to some core base concepts and does not provide category structures for organization type, organization purpose, or roles. The W3C Organization Ontology instead is a vocabulary for describing organizational structures and relationships within and between organizations. Its design allows domain-specific extensions that support additional classification of organizations and roles, as well as extensions to support neighbouring information such as organizational activities. Differently from the ORG that was built with the aim to represent governmental institutions, the W3C Organizational Ontology can be used to represent a wide range of organizational structures, including companies, government agencies, non-profit organizations, and more. The ORG Vocabulary and the W3C Organization Ontology are closely related. In fact, the ORG Vocabulary was developed as an extension of the W3C Organization Ontology, with the goal of providing additional classes and properties that were specific to organizational structures and

[10] https://orkg.org/
[11] https://docs.openalex.org/
[12] https://cso.kmi.open.ac.uk/schema/cso

relationships. Despite the fact that both ontologies are used to describe organisational structures, ORG Vocabulary focuses more on the formal and legal aspects of organizations, while the W3C Organization Ontology is more general and can be applied to a broader range of organizational types.

Schema.org vocabulary is a set of schemas used to structure web content in a semantically meaningful way. It is used for marking up web pages in a way that search engines can easily understand the content and provide more relevant results to users. It includes a wide range of types, such as products, recipes, people, events, and more, with properties that describe their attributes and relationships.

On the other hand, there are several ontologies proposed to represent competences and skills of organisation' employees. The European Skills, Competences, Qualifications and Occupations[13](ESCO) ontology focuses on the EU labour market, describing skills and qualifications specific to the region. It covers three different domains – the three "pillars" of ESCO: i) occupations, ii) knowledge, skills and competences, and iii) qualifications[22]. The data model[14] is based on the Simple Knowledge Organization System (SKOS)[15] ontology which is used for representing knowledge organization systems, like thesauri, taxonomies and classification schemes. ESCO concepts are subclasses of SKOS concepts, with some additional metadata properties to structure the ESCO pillars. ESCO defines more than 10 000 concepts using 24 EU languages.

The Skills and Recruitment Ontology[16] (SARO) is a domain ontology representing occupations, skills and recruitment. Inspired by ESCO and Schema.org[17], SARO covers four dimensions: job posts, skills, qualifications, and users. It extends the ESCO `SkillandQualification` concept and introduces around 1000 concrete skill instances. However, in contrast with ESCO, SARO also describes the proficiency level for each skill. Such an ontology has been evaluated on the TOBIE system that comprises processing pipelines that extract the desired set of skills and job posting attributes and create a knowledge base that can be used for analysing the skill demand in the labor market domain [23].

While ESCO and SARO ontologies are rich resources for describing skills and competencies, they go beyond the scope of this paper and beyond the requirements for our use cases. Instead, the W3C Organisational Ontology and Schema.org provide a good starting point for our use cases, although we needed to design a minimal extension of the the W3C ontology to cover our application scenarios.

3. sEKG construction

sEKG is constructed in four steps: (i) dataset collection, (ii) skills extraction, (iii) ontology extension, and (iv) knowledge graph population. The overall pipeline is depicted in Figure 1.

3.1. Dataset collection

The two sources of data that we consider for the construction of sEKG are patent and publication data.

Patent data are downloaded from the US Patent and Trademark Office (USPTO)[18]. For the scope of this paper, we collected all granted patents in the US since 2013. Patent data are made available by the USPTO as bulk zip documents, containing XML descriptions of patents and the DTD to define their structure. To keep our system independent of the input data format, we transform each XML document into a JSON representation model, only extracting the attributes needed for the scope of this work. At the end of preprocessing, we have 3, 566, 517 US patents.

Scholarly data are metadata about the scientific publications by the company employees. Data include title, authors, publication venue, keywords, abstract, and publication year. We limit our approach to these attributes because they are typically available, even for non-open-access papers, allowing us to generalize our method.

[13]https://esco.ec.europa.eu/en/use-esco/download
[14]https://ec.europa.eu/esco/lod/static/model.html
[15]https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html
[16]https://elisasibarani.github.io/SARO/
[17]https://schema.org/
[18]https://www.uspto.gov/

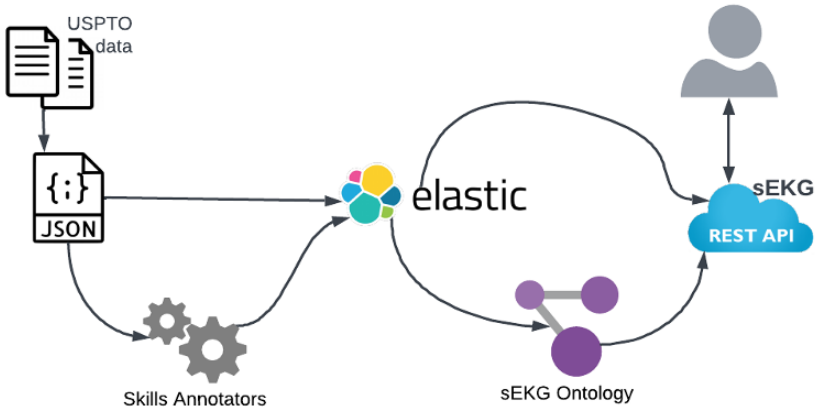

Figure 1: sEKG pipeline.

For each paper document, we enrich the initial metadata and produce a total of 21 attributes, adding additional information such as author aliases, external IDs for papers and authors, topical categories, etc. Each patent document has a total of 23 attributes, among which the official USPTO categorization (first, second, and third level category), abstract, author name, author affiliation, organization, claim, description, publication year, etc. We align these attributes to ensure that the corresponding data from both sources can be compared and analyzed efficiently.

3.2. Skills extraction

The ability to extract skills from scientific papers and patents can provide insights into trends and support decision-making in different domains. Extracting skills from such resources might be challenging due to the use of domain-specific jargon and the varying level of detail provided in the analyzed text. At the time of this manuscript, we implemented two extractors: one based on Latent Dirichlet Allocation (LDA), and one based on a Wikidata annotation API.

LDA, an unsupervised generative probabilistic method for topic modeling [24], is applied in this paper for skills extraction, inspired by [25]. The rationale behind such a decision is that LDA assumes that the set of words that have the main contribution in representing a topic are conceptually related and they all are talking about the same concept (skill, or competency). LDA extracts the most relevant words (concepts) from the text. The most common groups of words that co-occur can be interpreted as skills, competencies, or experience.

We pre-process the text of each patent/paper through standard tasks, including removal of HTML tags, punctuation, digits, stop words, case normalization, and tokenization. Bi-grams are collected, and part-of-speech tagging and lemmatization are applied. Bag-of-words representations are generated using both CountVectorizer and TfidfVectorizer, popular techniques for converting text documents into numerical formats suitable for machine learning algorithms like LDA. CountVectorizer creates a matrix representing documents as a bag-of-words model, while TfidfVectorizer considers word frequency in the corpus, assigning weights to each word based on its occurrence in the document and corpus. LDA models are trained on document representations from CountVectorizer and TfidfVectorizer. Finally, the top 5 skills are extracted for each document (abstract), and a list of all skills for each author is compiled and stored in JSON documents.

The second approach for extracting skills involves using state-of-the-art tools for automatically detecting named entities in free text and aligning them to a predefined knowledge base. Examples of such tools are Spotlight [26], X-Lisa [27], Babelfy [28], and Wikifier [29]. Specifically, we selected *Wikifier* [29], which produces linkages to Wikidata - and Wikipedia. This approach benefits from the extensive coverage of Wikidata/Wikipedia and can identify skills that may not have been explicitly mentioned in the text but are related to the concepts discussed. Similarly, to the use of LDA, extracted

skills for each patent/paper are stored in sEKG.

3.3. Ontology extension

The Semantic Web's potential hinges on ontology reuse, facilitating shared data understanding and minimizing redundancy. This paper employs a recommended ontology development approach [30], extending well-established ontologies like W3C Organisational Ontology and Schema.org. Utilizing a bottom-up approach, we identified general requirements and expanded concepts and predicates to cover specific use cases. The ontology, expressed in RDF (Figure 2), integrates Schema.org and W3C Organisational Ontology properties and introduces new ones (orange) for scholarly and patent data, maintaining compliance and enhancing data interoperability.

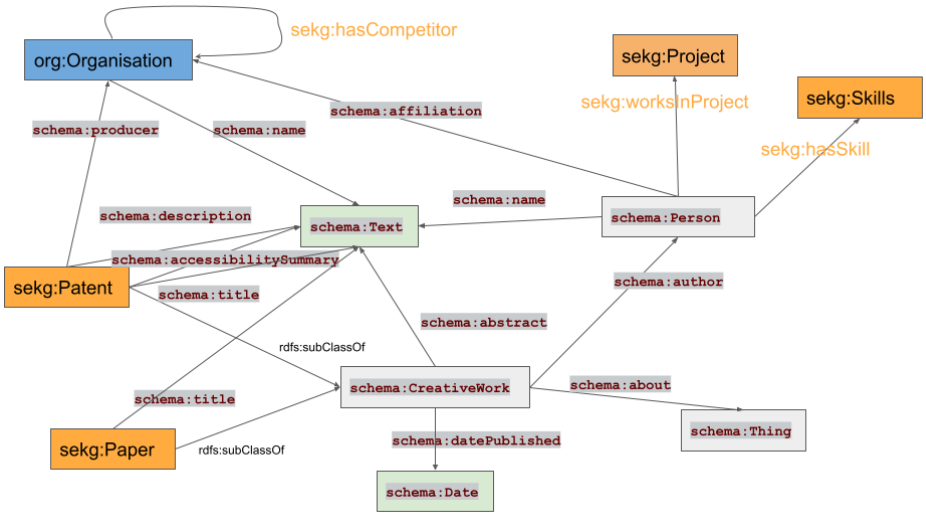

Figure 2: Main types used in sEKG ontology and their mutual relations.

The main types considered in the ontology are derived from Schema.org (schema namespace) to represent scientific papers and patents, while the main types and properties from W3C Org (org namespace) are derived to represent the organisational organogram of the company. Among the most frequent used types are:

- schema:CreativeWork is used to represent generic kind of creative work, including books, movies, photographs, software programs, etc;
- org:Organisation is used to represent a collection of people organized together into a community or other social, commercial, or political structure.
- sekg:Paper is a new type introduced in the ontology as a specialization of schema:CreativeWork to represent scholarly data. Schema.org has types to represent journal papers, conference papers, and also workshop papers. However, for the aim of this paper, we do not make any distinction on the type of scholarly data, thus, we introduced a new generic one named Paper.
- sekg:Patent is a new type introduced in the ontology as a specialization of schema:CreativeWork to represent patent data. Neither Schema.org, nor W3C Org ontologies provide this type to represent patent data.
- sekg:Project is a new type to represent potential projects that an employee is working on. Schema.org has a type to represent projects, called schema:Project, however, such type is a subclass of Organisation to describe an enterprise (potentially individual but typically collaborative), planned to achieve a particular aim. As the semantics of such a type, is different from the one referred to in this paper, we introduce such a concept as new.
- sekg:Skills is a new type to represent the skills of each employee. Neither W3C Org, nor Schema.Org defines a concept to represent skills or competency that a person might have. For this

Table 1
Examples of potential use cases that can be served by sEKG.

Use Case	Intuitive Query & Response
Find the expert	QUERY: Given a skill, for example retrieve all peoples name and department - API request: /sEKG/people?hasSkill=information%20extraction&fields=name&OrganisationalUnit
	RESPONSE: From the requirements, all scholar data about papers and patents for a given author retrieved and stored inside an array. For each paper and patent abstract and the title has been processed and relevant keywords are extracted. From sEKG the predicate `sekg:hasSkill` is used to represent the skills that the employee has.
Find competitors	QUERY: Given a keyword that regards a specific technology, entity or element, select organisations working on that specific technology. - API request: /sEKG/organisation?hasSkill=blockchain&name=IBM&fields=hasCompetitor
	RESPONSE: The predicate `sekg:hasCompetitor` is introduced to represent the competitor organisations for that particular keyword.
Employee progress	QUERY: Given an employee name working on a domain, retrieve data about his skills over a 2 year time - API request: /sEKG/people?author.name=Blerina%20Einstein&year[2021,2023]&fields=Skills
	RESPONSE: All skills and expertise that belong to an author are stored and annotated with the type `sekg:Skills` that will be used to analyse employees progress in a timespan by identifying new trends for specific categories.
Finding Related Work	QUERY: Given a keyword, retrieve related papers and patents in a structured way - API request: /sekg/documents?keyword=artificial%20intelligence&doc_type=paper,patent&fields=title,authors,datePublished,about
	RESPONSE: All the papers and patent data relevant for the searched keyword in the considered time period are returned by the API to the caller. Among the various pieces of information, the one that is used to retrieve such data are text annotated with the predicate `sekg:similarTopic`.

reason, we introduced it as a new type, with the future intention of adding macro-categorizations of skills.
- sekg:worksInProject is a new property used to represent and describe employees involved in a particular project. Not only paper and patent data are important when extracting competencies or skills of employees but also projects on which they are working.
- sekg:hasSkill is a new property used to represent and describe skills and competencies that an employee has.
- all the other properties to describe paper and patent attributes are considered from Schema.org.

Domain and range restrictions are introduced for properties where only one class/datatype was specified as the value of the domainIncludes and rangeIncludes properties. Users willing to extend the ontology can look at the recommended types specified in Schema.org in the annotation properties. All data, and textual data in particular, are represented using Unicode UTF-8 character encoding to support interoperability across languages at the alphabet level. The ontology is available for further extension or improvement[19].

3.4. Knowledge Graph Population

All collected and generated data about papers and patents is collated in an Elasticsearch index. Separately, we use the sEKG ontology to describe a subset of the company employees, for the pilot study. Data about such employees is retrieved from the underlying index and served internally via a REST API.

Table 1 reports examples of interactions (queries/responses) that the sEKG mediated API can support.

4. Application Scenarios and Pilot Study

We highlight the usage of sEKG for several application scenarios, including but not limited to *Expert Finding* and *Literature Review*.

Expert Finding is a key challenge in large companies focused on innovation and technological progress - finding colleagues who can collaborate to provide feedback or guidance on cutting-edge projects and research. This is especially true when receiving requests from external customers, coming with specific needs and requirements: sometimes it is sufficient to match their requests to existing products/services that the company provides [31], but other times the problem is novel and we want to initiate a research effort - and identify researchers that have the correct expertise. sEKG can be used to retrieve areas of expertise for each employee, but also given a particular "skill", it can identify employees that have exhibited that skill in any of their artifacts.

[19] https://anonymous.4open.science/r/sEKGOntology-37EA/sekg.ttl

Reviewing the state of the art is also a key activity for innovative enterprises. In the process of submitting a patentable model or algorithm, an employee wants to make sure they could review all the state-of-the-art. When reviewing available patents this can be especially challenging, both because of the specific language used and because available search tools - e.g. those provided by USPTO[20] - can be limited to mere keyword search.

We conducted a pilot study of the effectiveness of sEKG at inferring skills for a subset of employees of a big company. The goal of this evaluation is to determine whether the extracted skills were accurate by gathering feedback from the employees. We recruited 7 volunteers to whom we gave the set of skills in sEKG obtained with the two extractors defined in Section 3.2. Almost 67% of skills obtained by linking to Wikidata were considered as correct by the users. Instead, the number of correct skills extracted by applying LDA as a baseline is 44%. The errors fall into two categories. For the Wikidata skills, some of the extracted concepts - while being relevant for the user - are not necessarily skills or fields of expertise. For the skills extracted via LDA, we often have some very generic keywords. For example, "dataset" could be a skill if it refers to data analysis, but not if it is simply a reference to a data file mentioned in a text. We analyzed user feedback, which indicated overall satisfaction with the extracted skills from both approaches. Users particularly appreciated the accuracy of skills associated with Wikidata. This study revealed a limitation of the method: difficulty in accurately attributing skills when papers or patents involve multiple authors. For instance, collaboration on diverse projects could result in inaccurate skill assignments, particularly when extracting information solely from text. We plan to mitigate the issue either with a human-in-the-loop extractor, based on dictionary expansion techniques, similarly to [32] to allow each user to refine their skills in the sEKG, or by integrating CRediT[21] taxonomy. Additionally, inaccuracies in skill inference can also result from poor text quality. We could observe that there exist several cases where text is very short, and contains typos, incomplete sentences, or irrelevant content that could lead to the extraction of non-skills or irrelevant keywords.

5. Conclusions and Future Work

In this work, we presented scholarly enterprise knowledge graph (sEKG), a KG constructed of patent and scientific papers. For the construction of such KG, only English papers were considered and patents from the USPTO office. The data from both resources are in JSON comprising different attributes. In the first step, we align these attributes to ensure that the corresponding data from both resources could be compared and analyzed efficiently. In order to enrich and give semantics to such data, we used and extend the W3C Org ontology. The ontology was extended by including concepts and properties from Schema.org, and new concepts and properties were defined for portions of the data not covered by existing ontologies or when their use in the present study was different. We depict a few use cases and run a small pilot study to assess the feasibility and utility of sEKG. We mined skills from paper and patent text using two different approaches, one based on classical NLP techniques, one on semantic concept extraction and linking to Wikidata. A user evaluation involving seven volunteers indicated that both methods were effective in extracting skills, but users preferred the Wikidata annotation approach due to its higher accuracy.

Given the preliminary nature of this work, we anticipate several future directions, including exploiting the richness of semantic typing from Wikidata to filter skills as well as machine learning methods. Moreover, large language models (LLMs) will be employed to enhance multiple steps of the sEKG pipeline. For example, LLMs could automate and improve the precision of skill extraction from patents and publications by analyzing text at a deeper semantic level, identifying complex relationships, and understanding context beyond keyword matching. Additionally, LLMs could be used to refine the knowledge graph population process, providing more accurate and contextually aware mappings between entities, ultimately improving the overall quality and coverage of the sEKG Finally, we plan to use existing human-in-the-loop techniques to refine and enrich user profiles.

[20] https://ppubs.uspto.gov/pubwebapp/
[21] https://credit.niso.org/origins/

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. d. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al., Knowledge graphs, ACM Computing Surveys (CSUR) 54 (2021) 1–37.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings, Springer, 2007, pp. 722–735.

[3] T. Pellissier Tanon, D. Vrandečić, S. Schaffert, T. Steiner, L. Pintscher, From freebase to wikidata: The great migration, in: Proceedings of the 25th international conference on world wide web, 2016, pp. 1419–1428.

[4] B. Abu-Salih, Domain-specific knowledge graphs: A survey, Journal of Network and Computer Applications 185 (2021) 103076.

[5] D. Cheng, F. Yang, X. Wang, Y. Zhang, L. Zhang, Knowledge graph-based event embedding framework for financial quantitative investments, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 2221–2230.

[6] I. Y. Chen, M. Agrawal, S. Horng, D. Sontag, Robustly extracting medical knowledge from ehrs: a case study of learning a health knowledge graph, in: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020, World Scientific, 2019, pp. 19–30.

[7] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, K. Todorov, Claimskg: A knowledge graph of fact-checked claims, in: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18, Springer, 2019, pp. 309–324.

[8] N. Mihindukulasooriya, M. Sava, G. Rossiello, M. F. M. Chowdhury, I. Yachbes, A. Gidh, J. Duckwitz, K. Nisar, M. Santos, A. Gliozzo, Knowledge graph induction enabling recommending and trend analysis: a corporate research community use case, in: The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings, Springer, 2022, pp. 827–844.

[9] M. A. Hossain, Y. K. Dwivedi, N. P. Rana, State-of-the-art in open data research: Insights from existing literature and a research agenda, Journal of organizational computing and electronic commerce 26 (2016) 14–40.

[10] M. Färber, The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data, in: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18, Springer, 2019, pp. 113–129.

[11] M. Y. Jaradeh, A. Oelen, K. E. Farfar, M. Prinz, J. D'Souza, G. Kismihók, M. Stocker, S. Auer, Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture, 2019, pp. 243–246.

[12] J. Priem, H. Piwowar, R. Orr, Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, arXiv preprint arXiv:2205.01833 (2022).

[13] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Cs-kg: A large-scale knowledge graph of research entities and claims in computer science, in: The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings, Springer, 2022, pp. 678–696.

[14] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, The aida dashboard: a web application for assessing and comparing scientific conferences, IEEE Access 10 (2022) 39471–39486.

[15] S. Auer, V. Kovtun, M. Prinz, A. Kasprzik, M. Stocker, M. E. Vidal, Towards a knowledge graph for science, in: Proceedings of the 8th international conference on web intelligence, mining and semantics, 2018, pp. 1–6.

[16] J. Xu, S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J. F. Rousseau, X. Li, W. Xu, V. I. Torvik, et al., Building a pubmed knowledge graph, Scientific data 7 (2020) 205.

[17] K. Möller, T. Heath, S. Handschuh, J. Domingue, Recipes for semantic web dog food: The eswc

and iswc metadata projects, in: Proc. of ISWC'07/ASWC'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 802–815.

[18] A. G. Nuzzolese, A. L. Gentile, V. Presutti, A. Gangemi, Conference linked data: The scholarlydata project, in: P. Groth, E. Simperl, A. J. G. Gray, M. Sabou, M. Krötzsch, F. Lécué, F. Flöck, Y. Gil (Eds.), The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II, volume 9982 of *Lecture Notes in Computer Science*, 2016, pp. 150–158. URL: https://doi.org/10.1007/978-3-319-46547-0_16. doi:10.1007/978-3-319-46547-0_16.

[19] J. Tang, Aminer: Toward understanding big scholar data, in: Proceedings of the ninth ACM international conference on web search and data mining, 2016, pp. 467–467.

[20] V. Bryl, A. Birukou, K. Eckert, M. Kessler, What is in the proceedings? combining publisher's and researcher's perspectives, in: Proc. of SePublica 2014, Anissaras, Greece, May 25th, 2014, 2014.

[21] D. Lee, J. Kang, P. Mitra, C. L. Giles, B.-W. On, Are your citations clean?, Communications of the ACM 50 (2007) 33–38.

[22] J. De Smedt, M. le Vrang, A. Papantoniou, Esco: Towards a semantic web for the european labor market., in: Ldow@ www, 2015.

[23] E. Sibarani, S. Scerri, N. Mousavi, S. Auer, Ontology-based skills demand and trend analysis, 2016.

[24] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, Multimedia Tools and Applications 78 (2019) 15169–15211.

[25] S. Momtazi, F. Naumann, Topic modeling for expert finding using latent dirichlet allocation, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3 (2013) 346–353.

[26] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, Dbpedia spotlight: shedding light on the web of documents, in: Proceedings of the 7th international conference on semantic systems, ACM, 2011, pp. 1–8.

[27] L. Zhang, A. Rettinger, X-LiSA: cross-lingual semantic annotation, VLDB 7 (2014) 1693–1696.

[28] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: a unified approach, Transactions of the Association for Computational Linguistics 2 (2014) 231–244.

[29] J. Brank, G. Leban, M. Grobelnik, Annotating documents with relevant wikipedia concepts, Proceedings of SiKDD 472 (2017).

[30] N. F. Noy, D. L. McGuinness, et al., Ontology development 101: A guide to creating your first ontology, 2001.

[31] B. Shbita, A. L. Gentile, P. Li, C. DeLuca, G.-J. Ren, Understanding Customer Requirements: An Enterprise Knowledge Graph Approach, in: Proceedings of ESWC 2023, Lecture Notes in Computer Science, 2023, p. to appear.

[32] A. Alba, A. Coden, A. L. Gentile, D. Gruhl, P. Ristoski, S. Welch, Language agnostic dictionary extraction, in: N. Nikitina, D. Song, A. Fokoue, P. Haase (Eds.), Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017, volume 1963 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-1963/paper611.pdf.