Knowledge Graph for Covid-19: a data fusion approach

André Cerri [1], Diogo Nolasco[1], Jonice Oliveira[1]

[1]Programa de Pós graduação em Informática, PPGI
Universidade Federal do Rio de Janeiro (UFRJ)
Cidade Universitária – Rio de Janeiro – RJ – Brasil

andre.cerry@ufrj.br, diogo.sousa@ppgi.ufrj.br, jonice@ic.ufrj.br

Abstract. *The Covid-19 pandemic generated a rapid and heterogeneous growth of scientific, technological, and social data, making it difficult for researchers and decision-makers to follow the available evidence in a timely manner. To address this problem, this paper presents the Knowledge Graph for Covid-19 (KGC-19), a knowledge graph designed to integrate information from PubMed articles, WIPO, EPO and LATIPAT patents, and X (formerly Twitter) posts. The proposed pipeline uses ETL scripts to collect, filter, standardize, and load the data into Neo4j according to a previously defined ontology. Named Entity Recognition models, including Flair and BioNER, are then applied to identify additional entities in texts and enrich the graph. The resulting graph connects scientific, technological, and social entities, enabling semantic queries and multidimensional analysis in the Covid-19 domain. KGC-19 therefore offers a promising approach to organize large volumes of unstructured pandemic-related data and support the discovery of relationships between research production, technological development, and social perception.*

1. Introduction

Since the beginning of 2020, the Covid-19 pandemic has caused severe health, social, and economic impacts, claiming approximately 5.1 million lives by November 2021. In response, the scientific community produced an unprecedented volume of publications, patents, and public discussions related to the disease, including studies on social isolation, diagnostics, treatments, and vaccines. Although this effort expanded the available knowledge, it also created an information overload: the volume and heterogeneity of the data made it difficult for humans to analyze the evidence quickly enough to support decision-making.

Knowledge graphs are useful in this context because they organize heterogeneous data as entities and relationships, allowing semantic queries and inferences over integrated sources [Xiong et al. 2017]. They have been applied in domains such as medicine, drug discovery, and social analysis, helping researchers identify implicit relationships between compounds, publications, technologies, and public reactions [Li et al. 2020, Wise et al. 2020, Zhang et al. 2021, Al-Obeidat et al. 2020]. Thus, a knowledge graph can support a broader understanding of the pandemic by connecting scientific production, technological development, and social perception.

Although several Covid-19 knowledge graphs were proposed during the pandemic, many of them emphasize biomedical, pharmaceutical, or scientific evidence, such as publications, clinical concepts, drugs, and molecular interactions [Li et al. 2020, Wise et al. 2020, Zhang et al. 2021, Wang et al. 2021]. KGC-19 differs from these initiatives by explicitly combining three complementary perspectives in the same graph: scientific production, represented by biomedical publications; technological development, represented by patent records; and social perception, represented by X (formerly Twitter) posts. This integration allows questions that go beyond a single information source. For example, a researcher could investigate whether a topic related to vaccines, treatments, or diagnostic methods appears first in scientific articles, later in patent applications, and subsequently in social media discussions, or whether public debates about subjects such as chloroquine are aligned with the scientific and technological evidence available in the same period.

This paper proposes the Knowledge Graph for Covid-19 (KGC-19), a graph that integrates scientific articles, patents, and social network posts. Data were collected from PubMed, WIPO, EPO, LATIPAT, and the X API, processed through ETL scripts, enriched with Named Entity Recognition (NER), and inserted into Neo4j according to a domain ontology. The goal is to enable researchers to perform multidimensional analyses and semantic inferences about Covid-19 and its impacts. Therefore, the main contribution of this work is not only the construction of another Covid-19 knowledge graph, but the design of a data fusion approach that relates scientific, technological, and social dimensions in a unified semantic structure. The remainder of this paper is organized as follows: Section 2 presents the theoretical background; Section 3 describes the proposal and implementation; Section 4 presents the results; Section 5 discusses related work; and Section 6 concludes the paper.

2. Background

This section presents the main concepts used in the construction of KGC-19: Knowledge Graphs, Named Entity Recognition, and Data Fusion.

2.1. Knowledge Graph

A knowledge graph is a knowledge base structured as a graph, in which real-world entities, such as people, events, objects, documents, and organizations, are represented as nodes connected by relationships. These relationships are commonly expressed as triples composed of subject, predicate, and object [Qiao et al. 2016], as illustrated in Figure 1.

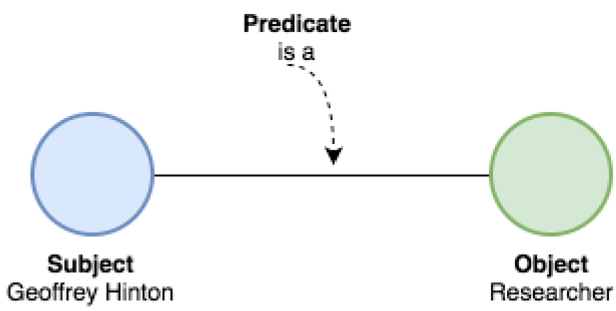

Figure 1. Structure of a triple

Because they represent explicit relationships among heterogeneous entities, knowledge graphs support information retrieval, knowledge discovery, and semantic inference. In this work, this structure is used to connect Covid-19-related scientific publications, patents, social media posts, authors, inventors, users, locations, dates, and entities extracted from text.

2.2. Named Entity Recognition

Named Entity Recognition (NER) is a Natural Language Processing task that identifies and classifies entities mentioned in text, considering the context in which each term appears [Mohit 2014]. NER is widely used in information extraction, question answering, and machine translation [Singh 2018]. Current approaches commonly rely on machine learning models trained on domain-specific corpora, including transformer-based models such as BERT [Devlin et al. 2019]. In KGC-19, NER is used to extract additional entities from article titles and abstracts, patent texts, and X posts.

2.3. Data Fusion

Data fusion is the process of integrating data from multiple sources and formats into a unified representation. Its goal is to combine complementary information and generate more consistent knowledge than would be possible by analyzing each source separately [Castanedo 2013]. In this work, data fusion supports the integration of biomedical publications, patent records, and social media posts into the same ontology-based graph.

3. Proposal

KGC-19 aims to support the analysis of Covid-19 from scientific, technological, and social perspectives. To do so, it integrates data from scientific articles, patents, and X posts into a single ontology-based graph, enabling semantic inferences and multidimensional analyses about the pandemic.

Our data collection focuses on the 2019–2021 period to capture the acute onset of the pandemic, allowing a rigorous analysis of the initial interactions between scientific production, technological development, and public perception.

In this section, all the processes carried out for the development of the KGC-19 are presented, such as the procedure for collecting data from the patents, scientific articles and social network posts databases, as well as the methods and tools used in the data processing and insertion into the triplestore.

3.1. Data Collection

The first step was the collection of heterogeneous data related to Covid-19. For patents, we used WIPO Patentscope, EPO Espacenet, and LATIPAT. These sources were selected because patent records are decentralized and each platform provides different coverage and export mechanisms. The searches used keywords such as *Coronavir**, *MERS-CoV**, *SARS-CoV**, *COVID*, *HCoV*, *SARS*, *MERS*, and *vaccine*, covering documents published between 2019 and June 2021. WIPO results were split by publication period because of download limits, EPO data were exported in several date intervals through advanced search, and LATIPAT was used to retrieve Brazilian patent records through Espacenet.

X data were collected through the X API using a web client application. The collection covered posts published from December 2019 to June 2021 and used Covid-19-related keywords, including *Covid-19*, *MERS-CoV**, *SARS-CoV**, *COVID*, *HCoV*, *SARS*, *MERS*, *vaccine*, *early treatment*, and *chloroquine*. The API returned JSON files containing tweet attributes such as user, identifier, text, retweet information, and location.

The data collected from the X platform was strictly limited to public posts. No direct personally identifiable information (PII) of the users was stored, ensuring compliance with privacy regulations such as the "Lei Geral de Proteção de Dados" (LGPD).

Scientific data were collected from PubMed, which provides access to biomedical citations and abstracts from sources such as MEDLINE, PubMed Central, and Bookshelf. The PubMed search used the same Covid-19-related terms and publication dates between June 2019 and June 2021.

Table 1 summarizes the data sources used in KGC-19 and their role in the proposed graph. This summary helps distinguish the scientific, technological, and social dimensions integrated, while keeping the detailed extraction procedures concise.

Table 1. Summary of data sources used in KGC-19

Source	Dimension	Coverage	Format	Role in KGC-19
PubMed	Scientific	Publications from June 2019 to June 2021	Structured metadata and abstracts	Provides biomedical evidence about Covid-19, including article metadata, authors, journals, dates, keywords, and textual fields for NER.
WIPO Patentscope	Technological	Patent records from 2019 to June 2021	Exported patent records	Expands the technological perspective with international patent documents related to coronavirus, vaccines, treatments, and diagnostic methods.
EPO Espacenet	Technological	Patent records from 2019 to June 2021	CSV files exported by date intervals	Complements the patent collection with European and international patent data, inventors, applicants, titles, abstracts, dates, and IPC information.
LATIPAT	Technological	Brazilian patent records from 2019 to June 2021	CSV files exported from Espacenet	Adds regional coverage by retrieving Brazilian patent requests related to Covid-19 and associated terms.
X API	Social	Posts from December 2019 to June 2021	JSON files	Represents public discussion and social perception through tweets, users, hashtags, mentions, locations, engagement metadata, and textual content.

3.2. Ontological Analysis and Data Fusion

After collection, the files were analyzed to identify relevant entities, attributes, and relationships. Empty or irrelevant fields were removed, and the remaining elements were modeled according to the domain and the Unified Foundational Ontology (UFO) [Guizzardi et al. 2008]. The data fusion process was then implemented through Python scripts that filtered the selected fields, standardized X JSON files into CSV format, and prepared the data for insertion into Neo4j. CSV was adopted because it facilitates partitioned loading and reduces memory issues during large insertions.

Table 2 summarizes the main ontology components. The implementation preserves the detailed attributes from each source, while the table groups them by their role in the graph to make the ontology easier to read within the paper.

Table 2. In this table we have our ontology components

Entity	**Attribute**	**Relationship**
Post	id, favorited, full_text, retweeted, truncated, creation date, reply_count, retweeted_status, favorite_count, conversation_id, is_quote_status, quoted_status_id, quote_status, display_text, in_reply_to_user_id, in_reply_to_status_id, in_reply_range, to_screen_name, possibly_sensitive_editable, possibly_sensitive	(Post)-[created_at]->(Date) (Post)-[posted_by]->(User) (Post)-[written_in]->(Language) (Post)-[posted_from]->(Place) (Post)-[posted_via]->(Device) (Post)-[has]->(Hashtag) (Post)-[has]->(Symbol) (Post)-[has]->(Url) (Post)-[has]->(User Mention)
Hashtag	index, text	(Hashtag)<-[has]-(Post)
Symbols	index, text	(Symbols)<-[has]-(Post)
URL	display_url, expanded_url, index, url, status, title, description	(Url)<-[has]-(Post)
User Mention	name, screen name	(User Mention)-[has_id]->(User)
User	id, name, username, description, pinned_tweet_id, protected, media_count, screen_name, listed_count, friends_count, statuses_count, default_profile, verified, followers_count, favourites_count, profile_link_color, contributors_enable, fast_followers_count, has_custom_timelines, has_extended_profile, default_profile_image, business_profile_state, normal_followers_count, advertiser_account_type, profile_background_color, profile_sidebar_fill_color, background_tile, profile_text_color, profile_sidebar_border_color, profile_use_background_image, advertiser_account_service_levels	(User)-[created_at]->(Date) (User)-[publishes]->(Post) (User)-[located_in]->(City) (User)-[use]->(Symbol) (User)-[create]->(Hashtag) (User)-[makes]->(User Mention) (User)-[has]->(Image) (User)-[has_profile_in]->(Url) (User)-[banner_has]->(Url)
Language	lang	
Country	name, code	(Country)<-[located_in]-(State)
State	name, state_code	(State)-[located_in]->(Country) (State)<-[located_in]-(City)
City	name, code	(City)-[located_in]->(State)
Date	date	
Device	name_device	

Entity	Attribute	Relationship
Patent	application id, title, publication date, device, abstract, publication number, application number, application date, country, IPC, priority data, FamilyNumber, national phase entries, EarliestPublication	(Patent)-[invented_by]->(Inventor) (Patent)-[published_in]->(Date) (Patent)-[created_in]->(Place) (Patent)-[has]->(Image) (Patent)-[required_by]->(Applicant)
Inventor	name	(Inventor)-[worked_with]->(Inventor)
Applicant	name	
Image	format_device, url_link	(Image)<-[has]-(User)
Article	id, source, doi, pmcid, pubModel, pubTypeList, keywordList, isOpenAccess, hasPDF, hasBook, hasSuppl, citedByCount, hasData, hasReferences, hasTextMinedTerms, hasDbCrossReferences, hasLabsLinks, authMan, epmcAuthMan, nihAuthMan, hasTMAccessionNumbers, license	(Article)-[published_by]->(Author) (Article)-[published_in]->(Date) (Article)-[written_in]->(Language) (Article)-[created_in]->(Date) (Article)-[revised_in]->(Date) (Article)-[publishedOnline_in]->(Date) (Article)-[indexed_in]->(Date) (Article)-[loaded_in]->(Date) (Article)-[published_by]->(Journal)
Foundation	name	(Foundation)-[has]->(Website)
Website	url	(Website)-[has]->(Url)
Author	fullname, firstname, lastname, initials, country	(Author)-[published]->(Article) (Author)-[worked_with]->(Author)
Journal	title, issue, volume, journalIssueId, dateOfPublication, printPublicationDate, ISOAbbreviation, medlineAbbreviation, NLMid, ISSN, ESSN	(Journal)-[publishes]-(Article)

3.3. Triplestore Configuration and Graph Creation

Neo4j was selected for its property graph structure and Cypher query language. The remote database was configured with the APOC library for data integration, alongside HTTP and Bolt connectors to allow external script access during loading and NER processing.

The graph was created by placing the standardized files in Neo4j's import directory and executing Cypher scripts partitioned by data source, such as WIPO, EPO, LATIPAT, PubMed, and X. These scripts loaded CSV records, created or updated entities with `MERGE`, and established relationships according to the ontology.

The following example illustrates the Cypher script used to create patent and inventor nodes and connect them according to the ontology. Similar scripts were created for the other sources, using fields and relationships of each dataset.

Listing 1. Example of Cypher insertion for patent and inventor data

```
LOAD CSV WITH HEADERS FROM 'file:///Espacenet.csv' AS row
WITH split(row['Inventors'], '\n') AS inventors, row
UNWIND inventors AS inventor
MERGE (i:Inventor {name: inventor})
MERGE (p:Patent {title: row['Title']})
```

```
SET p.abstract = row['Abstract'],
    p.publication_date = row['Publication date'],
    p.ipc = row['IPC']
MERGE (i)-[:Invented]->(p)
```

This modular organization facilitates maintenance and allows new sources to be added in future versions of KGC-19.

3.4. Named Entity Recognition

After the initial graph construction, NER was applied to enrich the graph with entities mentioned in titles, abstracts, patent descriptions, and X posts. Two complementary models were used: Flair [Akbik et al. 2019], selected for its general-domain coverage and variety of entity tags, and BioNER [Khan et al. 2020], selected for its biomedical focus and ability to identify entities such as diseases, drugs, and symptoms. A Python program connects to Neo4j, retrieves textual fields, applies the models, filters entities according to confidence thresholds, and inserts the accepted entities and relationships into the graph.

4. Results

This section presents the current results of KGC-19 and examples of semantic inferences supported by the graph. Figure 2 shows the ontology graph obtained in Neo4j, composed of 27 entities and 46 relationships. These entities connect social media data, scientific publications, patent records, dates, places, authors, inventors, users, and textual entities extracted through NER.

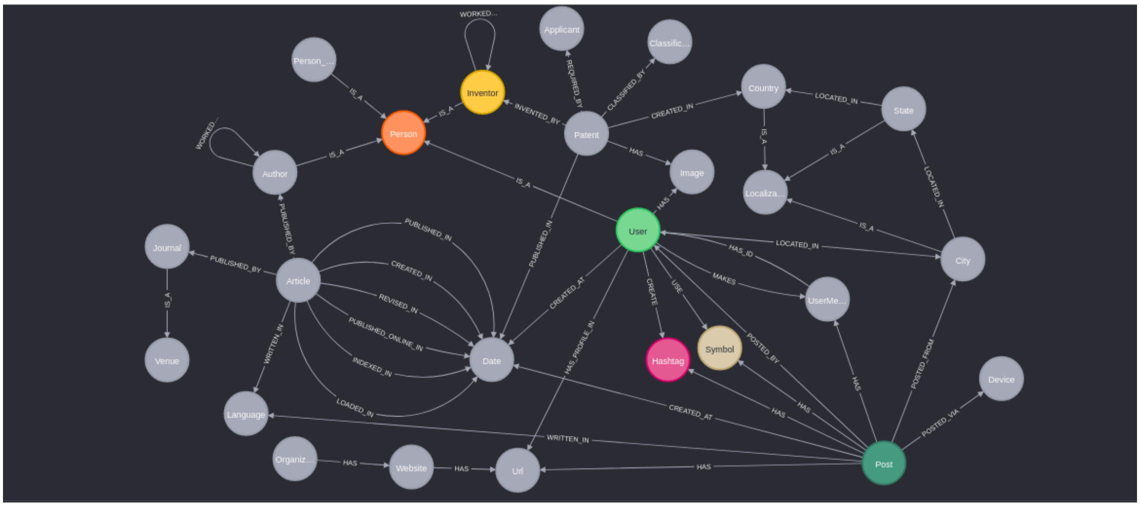

Figure 2. Ontology graph in Neo4j

Figure 3 illustrates a query that retrieves inventors associated with patents whose titles contain the term “COVID-19”. Although simple, this example shows how KGC-19 can connect technological records to people and organizations, enabling more complex analyses about research trajectories, institutional participation, and possible relationships between scientific production, technological development, and public dissemination.

Listing 2 presents a simplified version of the semantic query used to retrieve inventors associated with Covid-19 patent titles. The query shows how the graph structure allows the analyst to move from a technological artifact, such as a patent, to the people involved in its development. This same pattern can be extended to include applicants, countries, publication dates, IPC classes, or entities extracted from abstracts.

Listing 2. Query for inventors associated with Covid-19 patents

```
MATCH (i:Inventor)-[:Invented]->(p:Patent)
WHERE toLower(p.title) CONTAINS 'covid-19'
RETURN i.name, p.title, p.publication_date
```

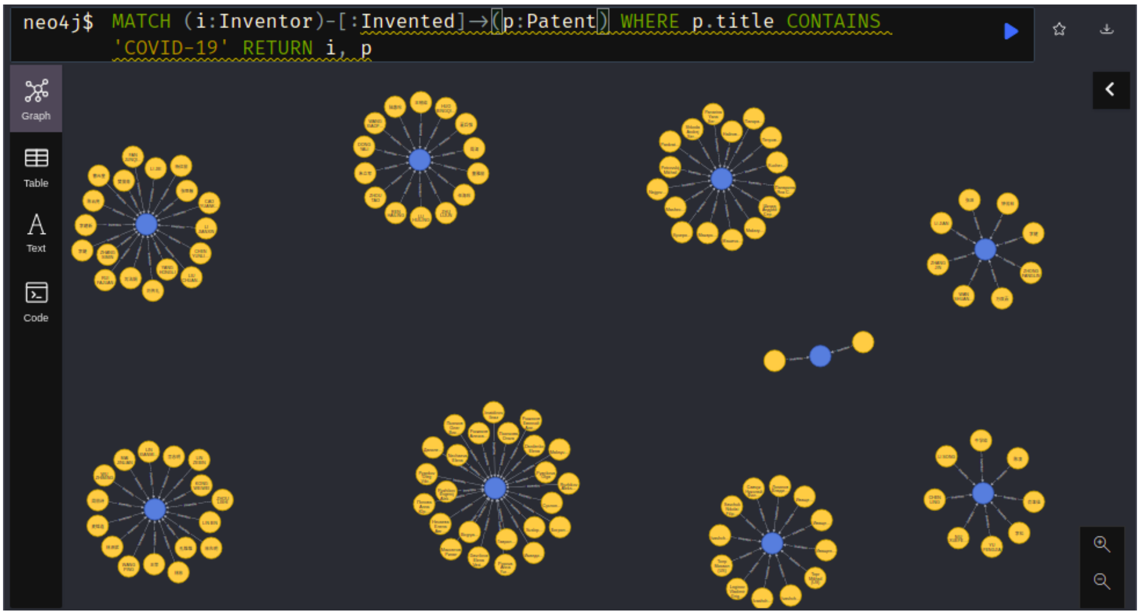

Figure 3. Consultation of all inventors holding patents related to covid-19

The integration of heterogeneous sources allows the graph to support questions that would be difficult to answer from isolated datasets. For example, KGC-19 may support analyses of how scientific publications, patents, and X posts evolve over time; how actors from academia, industry, and society are connected; and how biomedical or technological topics are reflected in public debate.

5. Related work

Knowledge graphs have been used in recent years for knowledge extraction and information retrieval in several domains [Wang et al. 2018, Manica et al. 2019]. Their popularization, especially through applications such as web search, encouraged the development of techniques for entity extraction, semantic retrieval, and ontology-based data integration [Wang et al. 2017].

During the Covid-19 pandemic, several knowledge graphs were proposed to organize biomedical, pharmaceutical, and scientific information [Li et al. 2020, Li et al. 2021, Wise et al. 2020, Wang et al. 2021, Zhang et al. 2021]. These works benefited from resources such as CORD-19 [Wang et al. 2020] and pre-trained language models, including BERT, BioNER, and BioBERT [Devlin et al. 2019, Khan et al. 2020, Lee et al. 2020]. KGC-19 differs from many of these approaches by combining biomedical, technological, and social data in the same graph, enabling analyses that relate scientific production, patent activity, and public discussion about Covid-19.

5.1. Limitations

Although KGC-19 integrates heterogeneous sources in a single graph, some limitations should be considered. First, the quality of the graph depends on the completeness and consistency of the original databases, which may contain missing fields, duplicated records, or differences in metadata standards. Second, the use of keywords may not retrieve all relevant documents and may also include records that only mention Covid-19-related terms tangentially. Third, the social dimension is represented through X data, which captures an important but partial view of public perception. Finally, NER models may introduce false positives or miss domain-specific entities. These limitations motivate the future inclusion of name disambiguation, topic extraction, and additional validation procedures.

6. Conclusions

The Covid-19 pandemic generated information at a speed and volume that made manual analysis increasingly difficult. In this paper, we presented KGC-19, a knowledge graph designed to integrate scientific articles, patent records, and social media posts related to Covid-19. The proposed pipeline collected data from PubMed, WIPO, EPO, LATIPAT, and X, processed them through ETL scripts, modeled them according to an ontology, inserted them into Neo4j, and enriched the graph through NER models.

The resulting graph supports semantic inferences about relationships between science, technology, and society during the pandemic. It may help answer questions such as whether political discourse influenced social media discussions about hydroxychloroquine, whether criticism of social distancing was stronger in specific locations, whether anti-vaccine discourse was associated with contexts of lower scientific development, and whether countries with more Covid-19 publications showed different patterns of denialism online. These analyses can contribute to better science communication, improved dissemination strategies, and responses to misinformation.

KGC-19 is still under development. Future work includes applying name disambiguation techniques to improve retrieval involving authors and inventors, extracting topics from textual data to characterize the subjects discussed, and expanding the graph with new sources and temporal analyses.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Al-Obeidat, F., Adedugbe, O., Hani, A. B., Benkhelifa, E., and Majdalawieh, M. (2020). Cone-kg: A semantic knowledge graph with news content and social context for studying covid-19 news articles on social media. In *Proceedings of the 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–7. IEEE.

Castanedo, F. (2013). A review of data fusion techniques. *The scientific world journal*, 2013(1):704504.

Devlin, J., Chang, M.-W., Lee, K. w., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Guizzardi, G., de Almeida Falbo, R., and Guizzardi, R. S. (2008). Grounding software domain ontologies in the unified foundational ontology (ufo): The case of the ode software process ontology. In *Proceedings of the XI Ibero-American Conference on Software Engineering (CIbSE)*, pages 127–140.

Khan, M. R., Ziyadi, M., and AbdelHady, M. (2020). Mt-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers. *arXiv preprint arXiv:2001.08904*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., Sun, Z., Tang, B., Chang, T. H., Wang, S., and Liu, Y. (2020). Real-world data medical knowledge graph: construction and applications. *Artificial Intelligence in Medicine*, 103:101817.

Li, X., Zhang, K., Li, G., and Zhu, B. (2021). A chinese knowledge graph for cardiovascular disease. In Liang, Q., Wang, W., Liu, X., Na, Z., Li, X., and Zhang, B., editors, *Communications, Signal Processing, and Systems*, pages 1816–1826, Singapore. Springer Singapore.

Manica, M., Auer, C., Weber, V., Zipoli, F., Dolfi, M., Staar, P., Laino, T., Bekas, C., Fujita, A., Toda, H., Hirose, S., and Orii, Y. (2019). An information extraction and knowledge graph platform for accelerating biochemical discoveries. *arXiv preprint arXiv:1907.05167*.

Mohit, B. (2014). *Named Entity Recognition*, pages 221–245. Springer Berlin Heidelberg, Berlin, Heidelberg.

Qiao, L., Yang, L., Hong, D., Yao, L., and Zhiguang, Q. (2016). Knowledge graph construction techniques. *Journal of computer research and development*, 53(3):582–600.

Singh, S. (2018). Natural language processing for information extraction. *arXiv preprint arXiv:1807.02383*.

Wang, C., Ma, X., Chen, J., and Chen, J. (2018). Information extraction and knowledge graph construction from geoscience literature. *Computers and Geosciences*, 112:112–120.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A., Wang, K., Wang, N. X. R., Wilhelm, C., Xie, B., Raymond, D., Weld, D. S., Etzioni, O., and Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. *arXiv preprint arXiv:2004.10706*.

Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, H., Liu, W., Chauhan, A., Guan, Y., Li, B., Li, R., Song, X., Fung, Y. R., Ji, H., Han, J., Chang, S.-F., Pustejovsky, J., Rah, J., Liem, D., Elsayed, A., Palmer, M., Voss, C., Schneider, C., and Onyshkevych, B. (2021). Covid-19 literature knowledge graph construction and drug repurposing report generation. *arXiv preprint arXiv:2007.00576*.

Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Wise, C., Calvo, M. R. S., Bhatia, P., Ioannidis, V., Karypus, G., Price, G., Song, X., Brand, R., and Kulkani, N. (2020). Covid-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. In *Proceedings of knowledgeable NLP: the first workshop on integrating structured knowledge and neural networks for NLP*, pages 1–10.

Xiong, C., Power, R., and Callan, J. (2017). Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279.

Zhang, R., Hristovski, D., Schutte, D., e Marcelo Fiszman, A. K., and Kilicoglu, H. (2021). Drug repurposing for covid-19 via knowledge graph completion. *Journal of Biomedical Informatics*, 115:103696.