A Survey on Covid-19 Knowledge Graphs and Their Data Sources

Hanieh Khorashadizadeh[1(✉)], Sanju Tiwari[2], and Sven Groppe[1]

[1] Institute of Information Systems (IFIS), Universität zu Lübeck, Lübeck, Germany
{khorashadizadeh,groppe}@ifis.uni-luebeck.de
[2] Universidad Autonoma de Tamaulipas, Reynosa, Mexico
sanju.tiwari@uat.edu.mx

Abstract. Since the onset of Covid-19 there has been a tremendous number of research to help deal better with the disease. This has produced a huge pile of data. The produced data might be contrary to each other and it is a tough task finding a specific data among the huge landscape of available data sources. Also answering a question that requires relating some research papers needs getting through piles of data and reading lots of literature. Knowledge graphs are very helpful tools to handle and structure all sort of data that is being produced every day. It also helps to find the uncovered relations between data. In this paper we reviewed the knowledge graphs that has dealt with Covid-19. Covid-19 has various aspects and each Knowledge graph normally deals with a specific domain of Covid-19 like drug repurposing, drug-drug interaction, etc. As they have different domains, their data sources also differs. In this paper an analysis over the knowledge graph data sources has been done. Mostly knowledge graphs have considered biomedical and medical aspects of Covid-19 and there has been rare knowledge graphs dealing with societal aspect and almost no work on economical and climate change aspects of the disease.

Keywords: Covid-19 · Knowledge graph · Data source · Data · Covid-19 aspects

1 Introduction

As Covid-19 impacted all over the world, there has been a rising number of research since the occurrence of covid-19 to be able to combat the disease. A huge data is being produced everyday regarding that. This data includes medical and biological research data, infection and death rate statistics, social media data, policy and protocol rules data and so on. There must be a way to get knowledge from the data and do further analysis. Further analysis encompasses the effectiveness of prescribed medication, relation between various policy and protocols and Covid-19 status, etc. Another issue is that some of this vast research endeavors since the occurrence of covid-19 might be redundant or contrary to each other, and some even wrong. So that would make it really tough for the

S. Nandan Mohanty et al. (Eds.): ICISML 2022, LNICST 470, pp. 142–152, 2023.
https://doi.org/10.1007/978-3-031-35078-8_13

scientists to get the right intuition for further analysis and development. There also seems to be a need to turn this knowledge into a structured format and integrate all data types. Knowledge graphs provide the information in a structured and organized format. Various scientific papers are being published every week, which leads to tremendous scientific data. There seems a need for a tool to search among this vast knowledge and obtain particular data or see how this data is related and visualize the relationship between the data [44].

If a scientist asks a complex covid-19 query that needs a chain of related documents to be answered, knowledge graphs are a helpful choice and simple keyword search is no use [6].

Even if a scientist is seeking a particular data, she needs to retrieve some data from multiple databases and integrate them all to get that particular data. Like if she would like to know about an FDA approved drug protein role, she needs to access and integrate various databases including drug, drug target and FDA drug status while making sure the sources are all updated [33]. So with the help of knowledge graphs all this heterogeneous data sources can be integrated. Another advantage of utilizing knowledge graphs is that we can further use them for machine learning analysis tasks [28] like link prediction, clustering and classification [13,37]. Other tasks include recommendation systems, searching and entity relations.

There also exists a correlation between people's point of view and attitudes towards covid-19 reflected in social media and the progress of the disease. With the use of knowledge graphs we can detect the people's community type and observe their attitudes contrast towards covid-19 [16]. So with the help of knowledge graphs we can better structure and shape data and do better analysis [41].

The paper is organized as follows. In section two there is a summary of Covid-19 knowledge graphs. In section three knowledge graph use-cases and applications has been discussed. Section four deals with knowledge graph data sources. Section five covers the challenges and issues with existing knowledge graphs and section six contains the conclusion.

2 Related Work

Knowledge4COVID-19 [34] is a knowledge graph that works on the interaction between the covid-19 treatments and coexisting diseases like high blood pressure or diabetes medication. Data sources are divided into two types, scientific open data and scientific publications. Scientific open data includes data extracted from open scientific databases like DrugBank. DrugBank is a database that holds data about drugs mechanisms, interactions and proteins. They have also utilised a data operator named FALCON [35] to detect drug entities, their side effects and the interaction consequences. So FALCON is in charge of NER and NEL. Scientific publications includes data extracted from scientific publications like PubMed, Cord-19, etc. Two NLP tools named MetaMAp and SemRep have been utilised to distinguish diseases and drugs either from abstract, title or from the whole content of the papers. With the help of some classification techniques,

unrecognized drug-drug interactions are predicted. Data is extracted with the help of Entity Recognition and Linking methods both from scientific open data and scientific publications.

KG-COVID-19 was made in three steps: download, transform and merge. In the download step, all data is fetched from various sources. In the transform step, the downloaded data will be transformed into graph based representation. Then in the merge step all components are merged. This means that two nodes with the same identifier will be merged and two edges will be consolidated if the source, target and edge type are identical [33].

Covid19KG utilised LitCovid and CORD-19 datasets and processed their texts to identify the entities and relationships. Nlp tools include iTextMine, PubTatur and SemRep. DrugBank, UniProtKB, Protein Ontology, STRING, iPTMnet and CoV-AbDab curated biomedical databases were also included [7].

COVID-KG is a multimedia knowledge graph that extracts entities with the help of a Coarse-grained Text Knowledge Extraction, Fine-grained Text Entity Extraction system called CORD-NER and Image Processing and Cross-media Entity Grounding. They have also made a visual information extraction system that will attach figures to the entities in knowledge graph utilising figure text, caption and the body context [43].

CKG-COVID-19 employs Amazon Comprehend Medical service to get the entity relationships from papers text. They have developed a document similarity engine leveraging semantic embeddings and graph embeddings, so the top k most similar articles will be found [44].

Covid-19KG, a knowledge graph on covid-19 pathophysiology, converts the data sources as a triple in Biological Expression Language (BEL) [14].

TweetsCov19 uses Los Angeles area Twitter data along with federal and state policy announcements and disease spread statistics. Tweets containing the covid related data is gathered with the help of a list of keywords. Then some pre-processing steps are done to clean the text. In order to extract entities and relations, natural language processing techniques including Sentiment Analysis, NMF and NCPD Topic Modeling, Time Series Analysis and Word Embedding are applied as the next step. The knowledge graph is then built with the topic modeling and word embedding models output, twitter dataset and temporal and event data like number of cases and number of new deaths [17].

Cone-KG is a knowledge graph concerning news articles and is one of the rare knowledge graphs on covid-19 social data analytics. Data sources are RDFized with Fandet Ontology [1].

OpenKG-COVID19 consists of ten connected subgraphs covering 10 various aspects of Covid-19 like encyclopedia, concept, medical, prevention, goods, research, epidemiology, health, and character to events. Knowledge is extracted from structured, semi-structured and plain text data sources [41].

We provide in Table 1 an overview of the existing knowledge graphs.

Table 1. Knowledge Graphs Overview.

Knowledge Graph Name	Number of RDF Triples/ Edges
Knowledge4COVID-19	80,570,440 RDF Triples
KG-COVID-19	21,433,063 Edges
covid-19-community	54,181,701 Edges
DRKG	5,874,261 RDF Triples
Covid19KG	10,232 relations
COVID-KG	3,332,151 relations
CKG-COVID-19	336,887 entities and 3,332,151 relations
Covid-19KG	more than 1.2 billion RDF triples
Twitter-COVID-19-Knowledge-Graph	5,215,237 Relations
Cone-KG	101,467,747 Triples
CASBiomedicalKG	18 million relationships
OpenKG-COVID19	2,687,329 facts(RDF Triples)

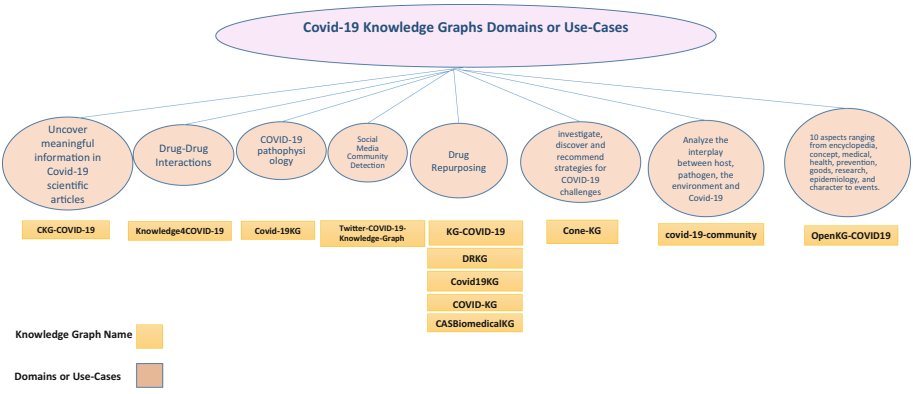

Fig. 1. Covid-19 Knowledge Graphs domains and use-cases

3 The KG Use-Cases and Applications

Each of the Covid-19 knowledge graphs address a specific domain of Covid-19. It might be interaction between Covid-19 medications and common comorbidities, Drug-Repurposing or other related medical issues. There has also been some research based on social media and news domains. Figure 1 depicts Covid-19 knowledge graphs use cases or domains and knowledge graphs names. As it shows there has already been five knowledge graphs focusing exclusively on drug repurposing. Drug repurposing is the process of using an existing drug for another therapeutic purposes. As Covid-19 was kind of a new disease and there has been no designated medication for the disease, it seems relevant that five out of twelve knowledge graphs so far have drug-repurposing application.

4 Knowledge Graph Data Sources

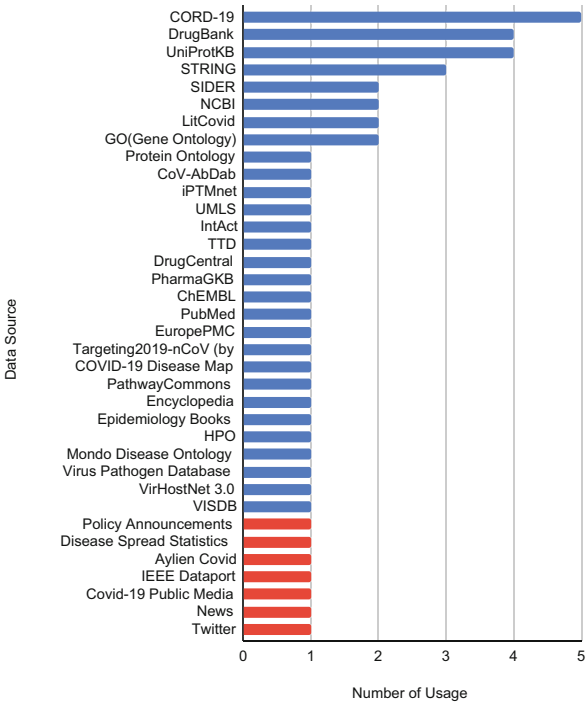

Fig. 2. Covid-19 Knowledge Graphs Data Sources

Table 2 lists Covid-19 knowledge graphs data sources, the data content explanation and the number of times each have been cited in Google Scholar. In Fig. 2, all data sources that have been used in covid-19 knowledge graphs are shown along with the number of times they have been used. Our analysis shows that mostly knowledge graphs have focused on biological and medical aspects of the disease and only a small number of them have dealt with social aspects of the disease. CORD-19 dataset gathered data from three sources: Medline and PubMed's PMC open access with covid queries, COVID-19 research articles from the WHO database adn arXIv, bioRxiv and medRxiv pre-print with the help of queries [10]. COVID-19KG is the only knowledge graph in biomedical aspect of COVID that queried PubMed on their own and haven't utilised CORD-19. This knowledge graph focuses on the disease pathophysiology. As the figure depicts, only 19,44 percent of the data sources are based on news and social media. As the disease has lots of aspects like climate change, societal, economical, etc., there seems a lack of work on these areas. Figure 3 depicts all data sources used in each Covid-19 knowledge graph. As it shows CORD-19 has been used in 5 knowledge graphs, DrugBank, UniProtKB, STRING, ... were also commonly used in several knowledge graphs.

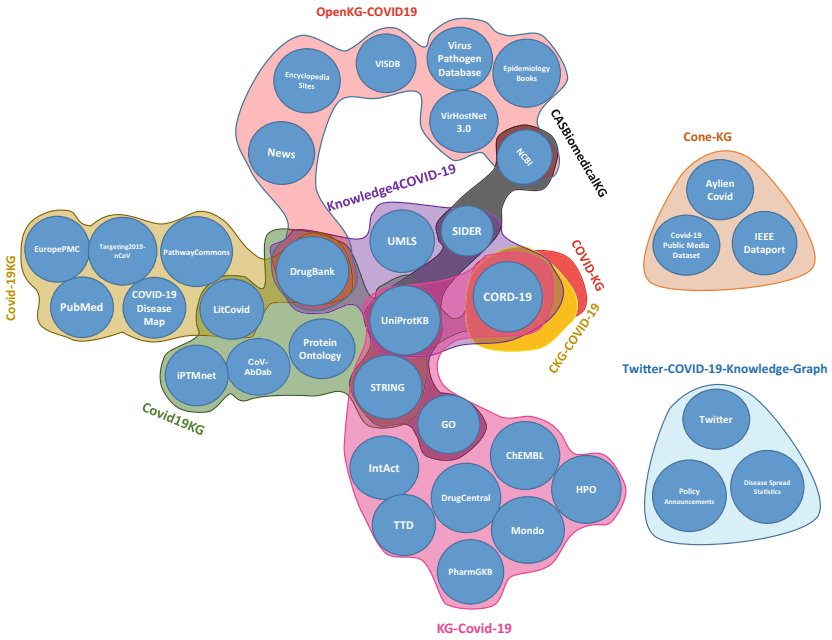

Fig. 3. An analysis over Covid-19 knowledge graph data sources

Table 2. Knowledge Graphs Data Sources. The number of citations were recorded on Nov 10, 2022, from Google Scholar.

Data Source	Data Type	Citations
CORD-19	A resource of over 1,000,000 scholarly articles about COVID-19	605 [42]
DrugBank	A dataset on drugs and drug targets	2693 [45]
UniProtKB	resource of protein sequence and functional information	677 [4]
STRING	Protein-Protein Interaction Networks	1659 [40]
SIDER	medicines and their recorded adverse drug reactions	913 [24]
NCBI	a series of databases relevant to biotechnology and biomedicine	1196 [15]
LitCovid	a curated literature hub for tracking up-to-date scientific information about the 2019 novel Coronavirus	173 [8]

(*continued*)

Table 2. (*continued*)

Data Source	Data Type	Citations
Protein Ontology	an ontological representation of protein-related entities, ranging from protein families to proteoforms to complexes	159 [26]
CoV-AbDab	the coronavirus antibody database	158 [32]
iPTMnet	an integrated resource for protein post-translational modification network discovery	85 [22]
UMLS	a repository of biomedical vocabularies	4149 [3]
IntAct	molecular interaction data	1099 [20]
TTD	Therapeutic Target Database	619 [9]
DrugCentral	online drug compendium	212 [38]
PharmaGKB	The Pharmacogenomics Knowledge Base	453 [21]
GO(Gene Ontology)	structured, computable knowledge regarding the functions of genes and gene products	4269 [12]
ChEMBL	a large-scale bioactivity database for drug discovery	3267 [18]
PubMed	more than 34 million citations and abstracts of biomedical literature	- [31]
EuropePMC	an open-access repository which contains millions of biomedical research works	110 [11]
Targeting2019-nCoV(by GHDDI)	public information sharing portal and data repository for the drug discovery community	- [19]
COVID-19 Disease Map	a knowledge repository of molecular mechanisms of COVID-19 as a broad community-driven effort	123 [29]
PathwayCommons	a web resource for biological pathway data	1175 [5]
Aylien Covid	Free Coronavirus News Dataset	– [2]
IEEE Dataport	CORONAVIRUS (COVID-19) TWEETS DATASET	– [25]
Covid-19 Public Media Dataset	Not available at the moment in Kaggle	–
Encyclopedia	encyclopedia sites to get COVID-19 related terms	–
HPO	The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease	322 [23]
Mondo Disease Ontology	The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species	9 [39]
Virus Pathogen Database	viPR Virus Pathogen Resource	140 [30]
VirHostNet 3.0	towards systems biology of virus/host interactions	164 [27]
VISDB	VISDB Viral Integration Site DataBase	35 [36]

5 Challenges and Issues with Existing Knowledge Graphs

Most covid-19 knowledge graphs so far dealt with a specific aspect of covid-19, like drug repurposing, drug-drug interaction or social media attitude towards

covid-19. Only one knowledge graph to date comprise a wider aspect of the disease [41]. It includes aspects such as research, prevention, epidemiology, etc.

Mostly knowledge graphs covered biomedical and medical dimensions of Covid-19. It seems relevant as the community made its best endeavours to promote vaccines or help cure the disease. Since Covid-19 vaccines have been released, there now seems a lack of work on the broad spectrum consequences of Covid-19, whether economical, climate change or mental health related. There also seems a need to process and answer questions related to quarantine and stay-at-home strategies. As knowledge graphs are AI tools that support question answering and uncover the hidden relations between data, decision makers would be able to understand about their decisions positive and negative consequences.

6 Summary and Conclusions

This paper has provided a review of the papers that have worked on Covid-19 knowledge graphs along with an analysis over the data sources they have utilised and areas each knowledge graph covers. There seems an inadequate endeavor on the societal, climate change and economical aspects of the disease. There also remains a need for a knowledge graph to structure the relation between Covid-19 protocols and policies and the number of infections and consequences afterwards. In future we are planning to construct a knowledge graph that covers a broader spectrum of the Covid-19 and also provide an assessment over the quality of existing Covid-19 knowledge graphs.

Acknowledgments. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 490998901.

Funded by

DFG Deutsche Forschungsgemeinschaft

German Research Foundation

References

1. Al-Obeidat, F., Adedugbe, O., Hani, A.B., Benkhelifa, E., Majdalawieh, M.: Conekg: a semantic knowledge graph with news content and social context for studying COVID-19 news articles on social media. In: 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 1–7 (2020). https://doi.org/10.1109/SNAMS52053.2020.9336541
2. Aylien: Aylien News API (2020). https://aylien.com/blog/free-coronavirus-news-dataset

3. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**(suppl_1), D267–D270 (2004)
4. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A.: Uniprotkb/swiss-prot. In: Edwards, D. (ed.) Plant Bioinformatics, pp. 89–112. Springer, New York (2007). https://doi.org/10.1007/978-1-59745-535-0_4
5. Cerami, E.G., et al.: Pathway commons, a web resource for biological pathway data. Nucleic Acids Res. **39**(suppl_1), D685–D690 (2010)
6. Chatterjee, A., Nardi, C., Oberije, C., Lambin, P.: Knowledge graphs for COVID-19: An exploratory review of the current landscape. J. Personal. Med. **11**(4) (2021). https://doi.org/10.3390/jpm11040300, https://www.mdpi.com/2075-4426/11/4/300
7. Chen, C., Ross, K.E., Gavali, S., Cowart, J.E., Wu, C.H.: COVID-19 Knowledge Graph from semantic integration of biomedical literature and databases. Bioinformatics **37**(23), 4597–4598 (2021). https://doi.org/10.1093/bioinformatics/btab694
8. Chen, Q., Allot, A., Lu, Z.: LITCOVID: an open database of COVID-19 literature. Nucleic Acids Res. **49**(D1), D1534–D1540 (2021)
9. Chen, X., Ji, Z.L., Chen, Y.Z.: TTD: therapeutic target database. Nucleic Acids Res. **30**(1), 412–415 (2002)
10. Colavizza, G., Costas, R., Traag, V.A., van Eck, N.J., van Leeuwen, T., Waltman, L.: A scientometric overview of cord-19. PLoS ONE **16**(1), e0244839 (2021)
11. Consortium, E.P.: Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Res. **43**(D1), D1042–D1048 (2015)
12. Consortium, G.O.: The gene ontology (go) database and informatics resource. Nucleic Acids Res. **32**(suppl_1), D258–D261 (2004)
13. Domingo-Fernández, D.: COVID-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. Bioinformatics **37**(9), 1332–1334 (2021)
14. Domingo-Fernández, D., et al.: COVID-19 knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. Bioinformatics **37**(9), 1332–1334 (2020). https://doi.org/10.1093/bioinformatics/btaa834, https://doi.org/10.1093/bioinformatics/btaa834
15. Federhen, S.: The NCBI taxonomy database. Nucleic Acids Res. **40**(D1), D136–D143 (2012)
16. Flocco, D., et al.: An analysis of COVID-19 knowledge graph construction and applications. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 2631–2640. IEEE (2021)
17. Flocco, D., et al.: An analysis of COVID-19 knowledge graph construction and applications. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 2631–2640 (2021). https://doi.org/10.1109/BigData52589.2021.9671479
18. Gaulton, A., et al.: Chembl: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. **40**(D1), D1100–D1107 (2012)
19. GHDDI: Targeting COVID-19: GHDDI Info Sharing Portal (2020). https://ghddi-ailab.github.io/Targeting2019-nCoV/
20. Hermjakob, H., et al.: Intact: an open source molecular interaction database. Nucleic Acids Res. **32**(suppl_1), D452–D455 (2004)
21. Hewett, M., et al.: Pharmgkb: the pharmacogenetics knowledge base. Nucleic Acids Res. **30**(1), 163–165 (2002)
22. Huang, H., et al.: IPTMNet: an integrated resource for protein post-translational modification network discovery. Nucleic Acids Res. **46**(D1), D542–D550 (2018)
23. Köhler, S., et al.: The human phenotype ontology in 2021. Nucleic Acids Res. **49**(D1), D1207–D1217 (2021)

24. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The sider database of drugs and side effects. Nucleic Acids Res. **44**(D1), D1075–D1079 (2016)
25. Lamsal, R.: Coronavirus (COVID-19) tweets dataset (2020). https://doi.org/10.21227/781w-ef42, https://dx.doi.org/10.21227/781w-ef42
26. Natale, D.A., et al.: The protein ontology: a structured representation of protein forms and complexes. Nucleic Acids Res. **39**(suppl_1), D539–D545 (2010)
27. Navratil, V., et al.: Virhostnet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. Nucleic Acids Res. **37**(suppl_1), D661–D668 (2009)
28. Olisah, C.C., et al.: Data-driven approach to COVID-19 infection forecast for Nigeria using negative binomial regression model. In: Data Science for COVID-19, pp. 583–596. Elsevier (2021)
29. Ostaszewski, M., et al.: COVID-19 disease map, building a computational repository of SARS-COV-2 virus-host interaction mechanisms. Sci. Data **7**(1), 1–4 (2020)
30. Pickett, B.E., et al.: Virus pathogen database and analysis resource (VIPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses **4**(11), 3209–3226 (2012)
31. PMC: PubMed Central® (PMC) (2020). https://www.ncbi.nlm.nih.gov/pmc/
32. Raybould, M.I., Kovaltsuk, A., Marks, C., Deane, C.M.: COV-ABDAB: the coronavirus antibody database. Bioinformatics **37**(5), 734–735 (2021)
33. Reese, J.T., et al.: KG-COVID-19: a framework to produce customized knowledge graphs for covid-19 response. Patterns **2**(1), 100155 (2021). https://doi.org/10.1016/j.patter.2020.100155, https://www.sciencedirect.com/science/article/pii/S2666389920302038
34. Sakor, A., et al.: Knowledge4covid-19: a semantic-based approach for constructing a COVID-19 related knowledge graph from various sources and analysing treatments' toxicities. arXiv preprint arXiv:2206.07375 (2022)
35. Sakor, A., Singh, K., Patel, A., Vidal, M.E.: Falcon 2.0: an entity and relation linking tool over Wikidata. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 3141–3148 (2020)
36. Tang, D., et al.: VISDB: a manually curated database of viral integration sites in the human genome. Nucleic Acids Res. **48**(D1), D633–D641 (2020)
37. Tiwari, S., Gaurav, D., Srivastava, A., Rai, C., Abhishek, K.: A preliminary study of knowledge graphs and their construction. In: Tavares, J.M.R.S., Chakrabarti, S., Bhattacharya, A., Ghatak, S. (eds.) Emerging Technologies in Data Mining and Information Security. LNNS, vol. 164, pp. 11–20. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-9774-9_2
38. Ursu, O., et al.: Drugcentral: online drug compendium. Nucleic Acids Res. gkw993 (2016)
39. Vasilevsky, N., et al.: Mondo disease ontology: harmonizing disease concepts across the world. In: CEUR-WS, vol. 2807 (2020)
40. Von Mering, C., et al.: String: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. **33**(suppl_1), D433–D437 (2005)
41. Wang, H., et al.: Construction of a linked data set of COVID-19 knowledge graphs: development and applications. JMIR Med. Inform. **10**(5), e37215 (2022)
42. Wang, L.L., et al.: Cord-19: the COVID-19 open research dataset. ArXiv (2020)
43. Wang, Q., et al.: COVID-19 literature knowledge graph construction and drug repurposing report generation. arXiv preprint arXiv:2007.00576 (2020)

44. Wise, C., et al.: Covid-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. arXiv preprint arXiv:2007.12731 (2020)
45. Wishart, D.S., et al.: Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. **36**(suppl_1), D901–D906 (2008)