Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Knowledge4COVID-19: A semantic-based approach for constructing a COVID-19 related knowledge graph from various sources and analyzing treatments' toxicities

Ahmad Sakor [a,b], Samaneh Jozashoori [a,b], Emetis Niazmand [a,b], Ariam Rivas [a,b], Konstantinos Bougiatiotis [c,d], Fotis Aisopos [c,*], Enrique Iglesias [a,b], Philipp D. Rohde [a,b], Trupti Padiya [a,b], Anastasia Krithara [c], Georgios Paliouras [c], Maria-Esther Vidal [a,b,**]

[a] TIB Leibniz Information Centre for Science and Technology, Welfengarten 1 B, Hannover, Germany
[b] L3S Research Center, University of Hannover, Appelstraße 9a, Hannover, Germany
[c] Institute of Informatics & Telecommunications, NCSR Demokritos, Patr. Grigoriou & Neapoleos Str, Ag. Paraskevi, Athens, Greece
[d] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Panepistimiou 30, Athens, Greece

ARTICLE INFO

Article history:
Received 8 December 2021
Received in revised form 8 September 2022
Accepted 5 October 2022
Available online 13 October 2022

MSC:
00-01
99-00

Keywords:
Knowledge graphs
COVID-19
Drug–drug interactions

ABSTRACT

In this paper, we present Knowledge4COVID-19, a framework that aims to showcase the power of integrating disparate sources of knowledge to discover adverse drug effects caused by drug–drug interactions among COVID-19 treatments and pre-existing condition drugs. Initially, we focus on constructing the Knowledge4COVID-19 knowledge graph (KG) from the declarative definition of mapping rules using the RDF Mapping Language. Since valuable information about drug treatments, drug–drug interactions, and side effects is present in textual descriptions in scientific databases (e.g., DrugBank) or in scientific literature (e.g., the CORD-19, the Covid-19 Open Research Dataset), the Knowledge4COVID-19 framework implements Natural Language Processing. The Knowledge4COVID-19 framework extracts relevant entities and predicates that enable the fine-grained description of COVID-19 treatments and the potential adverse events that may occur when these treatments are combined with treatments of common comorbidities, e.g., hypertension, diabetes, or asthma. Moreover, on top of the KG, several techniques for the discovery and prediction of interactions and potential adverse effects of drugs have been developed with the aim of suggesting more accurate treatments for treating the virus. We provide services to traverse the KG and visualize the effects that a group of drugs may have on a treatment outcome. Knowledge4COVID-19 was part of the Pan-European *hackathon#EUvsVirus* in April 2020 and is publicly available as a resource through a GitHubrepository and a DOI.

1. Introduction

In early December 2019, an outbreak of a novel virus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) occurred in China, causing a rapid spread of the coronavirus disease 2019 (COVID-19). SARS-CoV-2 can be transmitted during the asymptomatic phase of infection and poses a global health emergency because of the intricacy of tracing mild or presymptomatic phases. The disease spectrum of SARS-CoV-2 infection varies in severity from asymptomatic to mild respiratory tract infection and severe or fatal pneumonia. The virus infection landscape poses serious challenges that have to be addressed by the research community to come up with the tools that efficiently combat the pandemic. Specifically, the aggregation of heterogeneous data (e.g., publications and open scientific databases) into a common knowledge base will enable the development of data-driven tools. Moreover, data governance, interoperability and data quality issues, and efficient query processing and data exploration are relevant challenges demanded to be solved efficiently. More importantly, it is crucial to explore adverse effects of the treatments commonly prescribed for pre-existing conditions and the potential treatments for COVID-19.

Our Resource: We address the problem of data integration and propose a resource named Knowledge4COVID-19, which transforms COVID-19 and SARS-CoV-2 related data into a KG. The

[*] Corresponding author.
[**] Corresponding author.Leibniz Information Centre for Science and Technology, Welfengarten 1 B, Hannover, Germany
E-mail addresses: fotis.aisopos@iit.demokritos.gr (F. Aisopos), Maria.Vidal@tib.eu (M.-E. Vidal).

https://doi.org/10.1016/j.websem.2022.100760

Knowledge4COVID-19 resource is composed of a data ecosystem (DE) and the Knowledge4COVID-19 KG, both allow for a unified view of the data sources in terms of the unified schema. The different components of the Knowledge4COVID-19 DE enable entity extraction and linking, data curation, and the resolution of the heterogeneity conflicts across the data sources. Moreover, they facilitate the integration of the heterogeneous data into a uniform view. Mapping rules expressed in the RDF mapping language (RML) describe these correspondences [1]. In addition, knowledge extraction methods make use of knowledge encoded in diverse sources for extracting drug–drug interactions. These data sources include controlled vocabularies (e.g., Unified Medical Language System-UMLS[1]), scientific publications (e.g., CORD-19[2]) and scientific open databases (e.g., DrugBank[3]). Machine learning methods are also employed to predict interactions between drugs. The Knowledge4COVID-19 framework is publicly available as a resource in GitHub[4] and Zenodo.[5] Additionally, diverse services are offered to access and explore the KG (e.g., an API[6] and a public SPARQL endpoint[7]). The detailed instructions to access are provided at the project repository.[8] The Knowledge4COVID-19 KG can be created locally following the guidelines.[9]

In summary, the scientific contributions of this work are as follows:

- A novel infrastructure to transform heterogeneous data sources into a knowledge graph based on a unified schema. The implementation of this infrastructure provides a software pipeline that includes Named Entity Recognition and Named Entity Linking methods, as well as novel mapping rules for aggregating various data retrieved under a unified KG. The resulting KG can be traversed following referenceable resources or queried using SPARQL endpoints or a federated query engine.
- A publicly available KG resource related to COVID-19 integrating information from Scientific Open Data and Publications. This is a product of the aforementioned infrastructure, and allows for the exploration of various sources and data.
- A deductive system to discover drug–drug interactions in a COVID-19 treatment. This system is built on top of fine-grained representation of Pharmacokinetics drug–drug interactions extracted from scientific open data sources (e.g., DrugBank).
- A machine learning based drug–drug interaction prediction method, identifying non-documented interactions for treatments related to a specific disease. This produces the predicted COVID-19 related drug–drug interactions that are included in the Knowledge4COVID-19 Data Ecosystem.
- An analysis of the effectiveness and toxicity of COVID-19 treatments, providing drug–drug interactions deduced from the Knowledge4COVID-19 KG and adverse effects of these interactions.

This paper is structured in eight additional sections. Section 2 reports the worldwide statistics that summarize the infection situation and presents an overview of the preliminaries. Section 3 defines Knowledge4COVID-19 as a data ecosystem and Section 4 presents the process of knowledge graph creation from the declarative definition using RML mapping rules. Section 5 describes the Web APIs that enable the traversal of the Knowledge4COVID-19 KG, and the results of the empirical evaluations are reported in Section 6. The state of the art is summarized in Sections 7 and 8 describes Knowledge4COVID-19 as a resource. Finally, Section 9 wraps up and outlines future work.

2. Context and preliminaries

Fig. 1 depicts world statistics available at Worldometer;[10] numbers of infections by June 2020, April 2021, and November 2021 are summarized. In all three snapshots, at least 98% of the active infections reported develop mild symptoms, and at most 2% can be in serious or critical conditions. As can also be observed in the latter, nearly 97% of the patients who have suffered from COVID-19 have been either categorized as those with a mild condition or have already recovered.

Worldometers[11] also reports weekly new cases concerning the last week of November 2021. The perspective is different when analyzing these reports. The number of infections has increased to 16% versus 12% of weekly recovered. Also, the number of new deaths increases by 8%. The high mortality rate and new cases of infections indicate the unexpected spreading of the virus, still lack knowledge on the infection behavior of populations. Despite the intensity of statistical analyses and related research efforts dedicated to studying the outcome of these infections in certain countries, COVID-19 progression is still unpredictable for most patients, while being many times abrupt for the ones with a severe or critical condition.

According to World Health Organization (WHO) statistics[12] a broad spectrum of demographic, clinical, and molecular conditions appear to affect the evolution of the disease. Although age and sex seem not to be associated with the infection rate, once infected, the mortality rate in men is much higher than in women. Moreover, significant percentages of deaths represent patients above certain ages, as could be expected. Lifestyle variables such as smoking habits also play an essential role. Although regular smokers occur to be significantly underrepresented among those requiring hospital treatment for the illness, smoking emerges to be associated with rapid progression and increased mortality rates. Another factor that seriously affects the fatality rate for COVID-19 seems to be comorbidities, such as cardiovascular diseases, cancer, hypertension, etc. In particular, 80% of deaths are related to patients with at least one comorbidity, while COVID-19 patients suffering a serious disease (e.g., cancer) seem to develop more rapid progression and appear an increased mortality rate, in contrast to those with no pre-existing chronic medical conditions. Furthermore, the WHO guidelines[13] urge clinicians for careful consideration of adverse effects of medications that may be used in the context of COVID-19 and encourage medications that carry the least risk possible of drug–drug interactions with other medicines that a patient with specific comorbidities may be receiving. Researchers should address this need by detecting the risk of documented or even unknown interactions related to specific comorbidities and medications, though looking into big data and identifying relevant patterns.

[1] https://www.nlm.nih.gov/research/umls/index.html.

[2] https://www.semanticscholar.org/cord19.

[3] https://go.DrugBank.com/.

[4] https://github.com/SDM-TIB/Knowledge4COVID-19.

[5] https://zenodo.org/record/4702125#.YH4ACu8zaV4.

[6] https://github.com/SDM-TIB/Knowledge4COVID-19/tree/main/Exploration-API.

[7] https://labs.tib.eu/sdm/covid19-kg/sparql.

[8] https://github.com/SDM-TIB/Knowledge4COVID-19/wiki.

[9] https://github.com/SDM-TIB/Knowledge4COVID-19/wiki/Running-Knowledge4COVID-19-KG-locally.

[10] https://www.worldometers.info/coronavirus/.

[11] Data from November 28th, 2021.

[12] https://globalhealth5050.org/covid19/age-and-sex-data/.

[13] https://www.who.int/publications-detail.

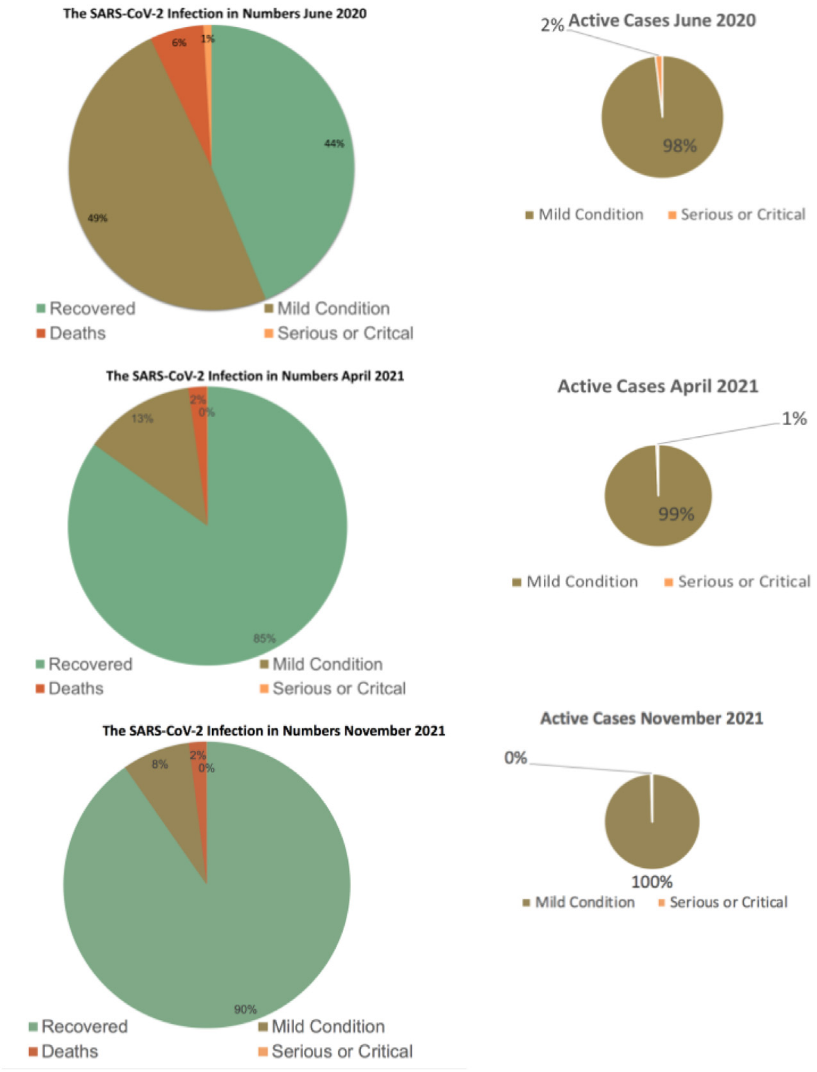


Fig. 1. SARS-CoV-2 Infections. Comparison of the severity of infections in June 2020, April 2021, and November 2021. Although the percentages of recovered cases increases significantly, the percentage of new deaths still remains above 2%.

2.1. Basic concepts

Data Ecosystems Data ecosystems (DEs) are data-driven infrastructures that allow different stakeholders to exchange data [2]. DEs are furnished with various computational methods to solve interoperability and integrate data while preserving data privacy, security, and sovereignty. DEs can be centralized, and one single node maintains all the data sources shared by the providers. The node also hosts all the services implemented on top of the DE data sources. Contrary, whenever data cannot be moved to a single node and data privacy regulations hinder the materialized and complete data integration of the DE data sources, DEs will be decentralized, i.e., they will be composed of several nodes. Each DE node will be able to perform services and share data management and analytical results. Semantic data models or ontologies provide the meaning of the data sources in a DE. Moreover, mapping rules relating to how data sources are defined in terms of the semantic data models are included.

Knowledge Graphs (KG): Knowledge graphs [3] are data structures that represent factual knowledge as entities and their relationships using a graph data model. Metadata is part of the KG, as well as taxonomies of entities, relationships, and classes. A KG contributes to the development of a common understanding of the meaning of entities in a domain and provides a formal specification of the properties of these entities. A KG $\mathcal{G}$ can be defined as a data integration system $DIS_{\mathcal{G}} = \langle O, S, M \rangle$ where O corresponds to the unified schema, S is a set of data sources, and M corresponds to mapping assertions defining concepts in O as conjunctive queries over sources in S. The instances of $\mathcal{G}$ are the result of the execution of the M rules over the data sources in S.

RDF Mapping Language - RML: The RDF Mapping Language (RML) [1] extends the W3C-standard mapping language R2RML to manage heterogeneous data sources represented in various formats, e.g., CSV, XML, JSON, and relational tables. These rules, named as *RML triples maps*, define the instances of RDF classes and their properties in terms of a logical source. Attributes from the logical data source of a triples map describe the resources of the corresponding class. RML is an RDF triple-oriented mapping language, where a triples map comprises mapping assertions [4] that define the instances of a class (a.k.a. subjectMap), and the property and object (a.k.a. predicateObjectMap) of the RDF triples where these instances participate as a subject. RML triples maps are expressed in RDF. This allows the exploration and tracing of the definition of the process of KG creation.

3. The Knowledge4COVID-19 data ecosystem

The Knowledge4COVID-19 framework is a data ecosystem [5]. A data ecosystem is defined as a 4-tuple *DE*=⟨*Data Sources, Data*

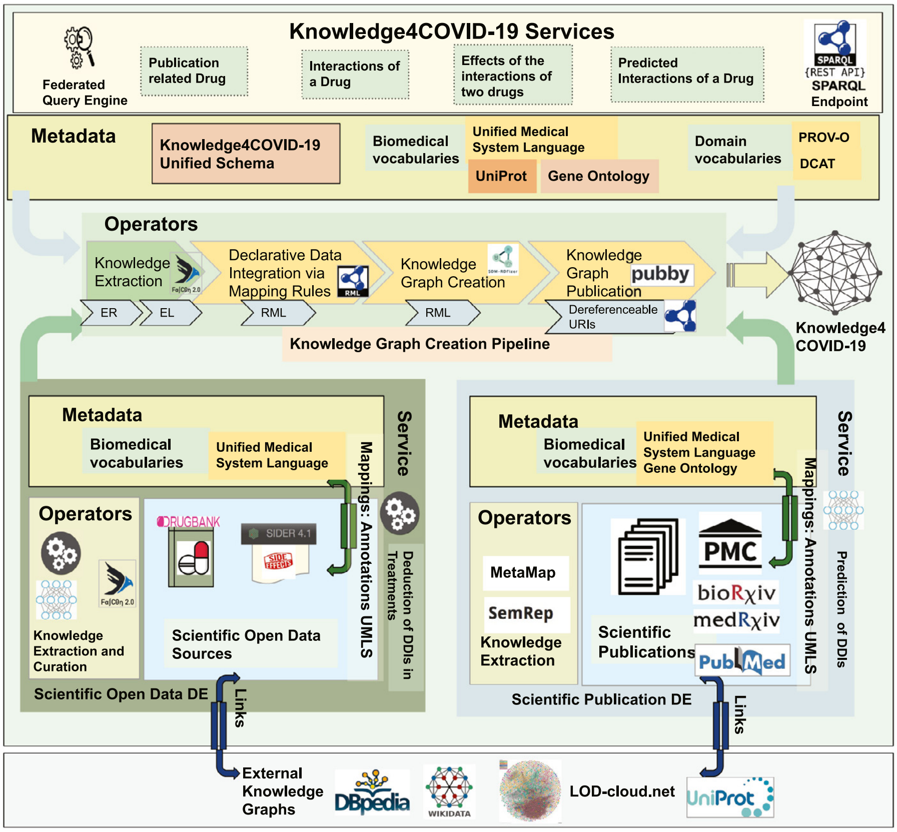


Fig. 2. Knowledge4COVID-19 as a Data Ecosystem (DE). A Nested Data Ecosystem comprising the Scientific Open Data and Publication DEs. Each DE processes comprises data sources, metadata, and operators to annotate their respective data sources.

Operators, Metadata, Mappings, Services)[6]. Data sources represent the collections from where data and knowledge are retrieved. Data operators correspond to functions used for data management (e.g., entity recognition and linking). The metadata component facilitates the specification of the meaning of the data collected from the data sources and the annotation with controlled vocabularies; it also comprises a unified schema that provides an integrated view of the data sources. The mappings align the data sources with the unified schema and describe their meaning. Lastly, services exploit the knowledge encoded in the metadata and data operators to satisfy user requirements. Services include federated query processing, interactions between the drugs of a treatment, predicted drug interactions, or mapping generation.

Fig. 2 depicts the components of the Knowledge4COVID-19 DE; it comprises two data ecosystems, one for scientific publications and another for scientific open data. The scientific publication DE includes COVID-19 related literature from PubMed,[14] bioRxiv,[15] medRxiv,[16] and PubMed Central (PMC).[17] Scientific open data DE collects data about COVID-19 drugs, their side effects and the adverse events generated by their interactions. The scientific open data DE integrates data extracted from textual descriptions from DrugBank[3] and SIDER.[18] This section describes these two DEs in detail, while Section 4 describes the pipeline that creates the Knowledge4COVID-19 knowledge graph from the data and knowledge extracted from these two DEs.

3.1. *The scientific open data DE*

This data ecosystem makes available data and knowledge about drugs extracted from open data sources.

3.1.1. *Data sources*

The Scientific Open Data DE integrates medical concepts extracted from open scientific databases. Albeit structured, these datasets may comprise textual attributes that encode relevant entities and relations. For example, the drug–drug interaction between Metformin and Hydroxychloroquine is described like “The therapeutic efficacy of Metformin can be increased when used in combination with Hydroxychloroquine.”.[19] Additionally, the indication of Hydroxychloroquine is presented like “Hydroxychloroquine is indicated for the prophylaxis of malaria where chloroquine resistance is not reported, treatment of uncomplicated malaria (caused by P. falciparum, P. malariae, P. ovale, or P. vivax), chronic discoid lupus erythematosus, systemic lupus erythematosus, acute rheumatoid arthritis, and chronic rheumatoid arthritis.“[20] These descriptions encode relevant facts that can be read and understood by humans. However, further analysis is required to make them understandable by machines. This DE makes used of data operators for named entity recognition to identify entities that correspond to drug related concepts. Table 1 describes the data collected from DrugBank [7], SIDER [8], and UMLS [9].

DrugBank is a Web-accessible database containing information about drugs and their administration routes, mechanisms,

[14] https://pubmed.ncbi.nlm.nih.gov/.
[15] https://www.biorxiv.org/.
[16] https://www.medrxiv.org/.
[17] https://www.ncbi.nlm.nih.gov/pmc/.
[18] http://sideeffects.embl.de/.
[19] https://go.DrugBank.com/drugs/DB00331.
[20] https://go.drugbank.com/drugs/DB01611.

Table 1
Data sources for the Scientific Open Data DE.

Data source	Data type	#Instances
DrugBank	Pharmacokinetic DDIs	769,352
2022-01-04	Pharmacodynamics DDIs	503,700
	Drug indications	2421
	Drug toxicities	1533
SIDER 2021	Drug side effects	58,945
UMLS Nov 2021	Medical concepts	4,864,162
CRD	Pair of drugs that target a CYP protein [10]	345,116
NCRD	Pair of drugs that target a No CYP protein [10]	5513

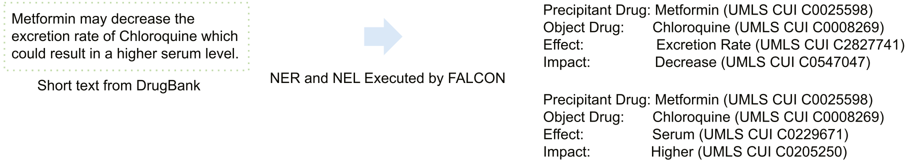

Metformin may decrease the excretion rate of Chloroquine which could result in a higher serum level.
Short text from DrugBank
NER and NEL Executed by FALCON
Precipitant Drug: Metformin (UMLS CUI C0025598)
Object Drug: Chloroquine (UMLS CUI C0008269)
Effect: Excretion Rate (UMLS CUI C2827741)
Impact: Decrease (UMLS CUI C0547047)
Precipitant Drug: Metformin (UMLS CUI C0025598)
Object Drug: Chloroquine (UMLS CUI C0008269)
Effect: Serum (UMLS CUI C0229671)
Impact: Higher (UMLS CUI C0205250)

Fig. 3. FALCON Recognizes Relevant Entities and Predicates. As a result, a Fine-Grained Representation of Drug–Drug Interactions is part of the Knowledge4COVID-19 KG.

proteins, and interactions. Drug–drug interactions can be Pharmacodynamics and Pharmacokinetics. A pharmacodynamic drug–drug interaction between drugs A and B indicates that both drugs influence in their effects directly, e.g., "The risk or severity of QTc prolongation can be increased when Hydroxychloroquine is combined with Acetophenazine". On the other hand, if drug A has a pharmacokinetic drug–drug interaction with drug B, A alters the disposition (absorption, distribution, elimination) of B, and ends up in the increase or the decrease of B plasma drug concentrations. For example, Abatacept has a pharmacokinetic drug–drug interaction with Hydroxychloroquine, because "The metabolism of Hydroxychloroquine can be increased when combined with Abatacept.". The Scientific Open Data DE has collected 769,352 and 503,700 Pharmacokinetic and Pharmacodynamics DDIs, respectively. Moreover, 2421 drug indications and 1532 toxicities have been collected and processed from DrugBank. SIDER is also a Web-accessible database which makes available mechanisms of actions of drugs and their possible adverse effects; 58,945 side effects are collected. UMLS is a controlled vocabulary that comprises terminology, classification, and semantic types and groups of biomedical concepts; 4,536,579 terms are collected together with their definitions, and semantic types and groups. Lastly, following the method proposed by Sridhar et al. [10], two data sources with pairs of drugs that shared at least one protein are computed. CRD are drugs from DrugBank that target at least one protein of the family CYP, while the NCRD drugs also target at least one protein, but it is not of the family CYP.

3.1.2. Data operators

The data operators enable the recognition of entities corresponding to drugs, their side effects, and the adverse events caused by their interactions. FALCON [11] recognizes the words corresponding to the drugs that interact and the effect and impact of these interactions. Additionally, the extracted words are linked to terms in UMLS. As illustrated in Fig. 3, "Metformin" and "Chloroquine" correspond to the extracted entities from the short text collected from DrugBank. At the same time, "excretion rate" and "decrease" represent, respectively, the effect and impact of the interaction of "Metformin" and "Chloroquine". The UMLS identifiers C0025598 and C0020336 are linked to "Metformin" and "Hydroxychloroquine", while C2827741 and C0547047 are related to "excretion rate" and "decrease", respectively. FALCON also connects "Metformin" and "Chloroquine" to their corresponding resources in DBpedia and Wikidata.

FALCON [12] is also used to extract the Drug–Drug Interactions (DDIs) reported in DrugBank as short texts. We customize FALCON for analyzing the DDI text. Since the DDI text is related to the medical domain, UMLS is utilized as the background knowledge for FALCON. In this case, in addition to recognizing words that correspond to two drugs that interact, FALCON identifies the effect and impact of an interaction. FALCON resorts to the catalog of rules for extracting the previously mentioned types of entities; additionally, a background knowledge base is utilized to determine the semantic type of the extracted entities. Since most of the descriptions of the interactions share similar patterns, i.e., the structure of the sentences is very repetitive, only few extra rules are required to be added to the catalog of rules. The rules were created by replacing each drug mention with a variable (DrugX, DrugY). Out of 1,273,052 drug–drug interactions collected from DrugBank, 320 patterns were recognized; Table 2 shows a sample of the extracted patterns.

As a result of the knowledge extraction process executed by FALCON, the Scientific Open Data DE makes available fine-grained representation of DDIs. This representation enables the deduction of new drug–drug interactions implemented as a service of this DE. Moreover, these descriptions are also used to validate the prediction tasks implemented in the Scientific Publications DE.

3.1.3. Data services

The Scientific Open Data DE implements a deductive system that enables to deduce drug–drug interactions among a multi-drug treatment whose interactions may reduce the effectiveness of the treatment or increase the number of toxicities. The deductive system is defined in terms of Datalog rules; it exploits the fine-grained representation of the DDIs interactions generated by FALCON. The execution of this deductive system is grounded on the results of deductive databases [13] to compute the minimal model that includes the instances of the deduced drug–drug interactions in a treatment. The minimal model corresponds to the fixed-point of the assignments of the values of variables in the deductive system rules. Since rules free of negations compose the deductive system, the minimal model is computed in polynomial

Table 2
Overview of extracted DDI patterns. Drugs mentions are in bold. Effect is in Italic. Impact is underlined.

DDI patterns
DrugY may increase the *anticoagulant activities* of **DrugX**. The *risk or severity of bleeding and hemorrhage* can be increased when **DrugX** is combined with **DrugY**. The *risk or severity of gastrointestinal bleeding* can be increased when **DrugX** is combined with **DrugY**. The *risk or severity of bleeding* can be increased when **DrugY** is combined with **DrugX**. The *metabolism* of **DrugY** can be decreased when combined with **DrugX**. **DrugX** may decrease the *vasoconstricting activities* of **DrugY**. **DrugX** may decrease the *excretion rate* of DrugY which could result in a higher serum level. **DrugY** may increase the *constipating activities* of **DrugX**.
The *risk or severity of gastrointestinal bleeding and gastrointestinal ulceration* can be increased when **DrugX** is combined with **DrugY**.

Table 3
Summary of datalog predicates. Extensional predicates are ddi(A,E,I,B), member(A,T), treatment(T), rule1(E,I), and rule2(E,I). Intensional predicates are ddi(A,E,I,B,T), toxicity(A,increase,B,T), and effectiveness(A,decrease,B,T).

Predicate	Explanation
ddi(A,E,I,B)	Pharmacokinetic drug–drug interaction between *A* and *B*. Precipitant drug *A* generates effect *E* (e.g., absorption, excretion, metabolism, serum concentration) with impact *I* (e.g., increase or decrease) in object drug *B*.
ddi(A,E,I,B,T)	Pharmacokinetic drug–drug interaction between *A* and *B* in treatment *T*. Precipitant drug *A* generates effect *E* (e.g., absorption, excretion, metabolism, serum concentration) with impact *I* (e.g., increase or decrease) in object drug *B*.
rule1(E,I)	Combinations of effect *E* with impact *I* that alter the toxicity of an object drug.
rule2(E,I)	Combinations of effect *E* with impact *I* that alter the effectiveness of an object drug.
treatment(T)	*T* is a medical treatment
member(A,T)	*A* is a drug in the medical treatment *T*
toxicity(A,increase,B,T)	The precipitant drug *A* increases the toxicity of object drug *B* in treatment *T*
effectiveness(A,decrease,B,T)	The precipitant drug *A* reduces the effectiveness of object drug *B* in treatment *T*

time in the size of the number of treatments and drug–drug interactions generated by FALCON. The approach proposed by Rivas and Vidal [14] is followed to implement this data service. The extensional database corresponds to statements about interactions between drugs extracted by FALCON. On the other hand, the intensional database comprises a set of Horn clauses that define the conditions to be met by the drugs whose interactions may reduce the effectiveness of a treatment or increase the number of toxicities. This intensional database relies on the fact that pharmacokinetic drug–drug interactions cause that the concentration of one of the interacting drugs (a.k.a. object) is altered when it is combined with the other drug (a.k.a. precipitant). Thus, the rate of absorption, distribution, metabolism, or excretion of the object drug is affected. Whenever the object drug absorption is decreased (resp. increased) the bioavailability of the drug is also affected. Furthermore, any alteration in the metabolism or excretion of the object drug has consequences on the therapeutic efficacy and toxicity of the drug. The following Datalog rules state the effect of pharmacokinetic DDIs. Considering the predicates in Table 3, the intensional database defines the toxicity effects of drug–drug interactions in a treatment:

$$ddi(A, E, I, B), treatment(T), member(A, T), member(B, T) \rightarrow ddi(A, E, I, B, T).$$

$$ddi(A, E, I, B, T), rule1(E, I) \rightarrow toxicity(A, increase, B, T).$$

$$toxicity(A, increase, B, T), toxicity(B, increase, C, T) \rightarrow toxicity(A, increase, C, T).$$

$$toxicity(A, increase, B, T), ddi(B, E, I, C, T) \rightarrow ddi(A, E, I, C, T).$$

The conditions to reduce effectiveness are defined as follows:

$$ddi(A, E, I, B, T), rule2(E, I) \rightarrow effectiveness(A, decrease, B, T).$$

$$effectiveness(A, decrease, B, T), effectiveness(B, decrease, C, T) \rightarrow effectiveness(A, decrease, C, T).$$

The extensional database includes the following ground predicates:

$$rule1(serum, increase).$$
$$rule1(metabolism, decrease).$$
$$rule1(absorption, increase).$$
$$rule1(excretion, decrease).$$
$$rule2(serum, decrease).$$
$$rule2(metabolism, increase).$$
$$rule2(absorption, decrease).$$
$$rule2(excretion, increase).$$

Additionally, a graph traversal method is implemented to compute the drugs that affect the most the effectiveness or toxicity of a treatment drug. The implemented method creates a directed graph from drug–drug interactions with the extensional facts and deduced of the Datalog rules. The direction of an edge from node A to B denotes that A is the precipitant and B is the object of the interaction. Drugs that affect the most the effectiveness or toxicity of a treatment drug are defined in terms of the middle-vertices in the wedges [14], or paths with two directed edges [15], in the directed graph that represents the drug–drug interactions among the drugs of a treatment. The middle-vertex of a wedge is both the object drug of one interaction, and the precipitant drug of the other interaction. Thus, drugs that correspond to middle-vertices of N wedges in a treatment T, correspond to drugs that cause $2 * N$ different drug–drug interactions in that treatment.

Fig. 4 depicts an exemplar treatment composed of five drugs. DDIs are represented by the predicates summarized in Table 3. By evaluating the Datalog program, a new DDI is deduced and represented in red. This evaluation deduces that doxycycline decreases the metabolism of montelukast in the treatment T1 and that doxycycline increases toxicity of montelukast. For the sake of simplicity, a single deduced DDI is depicted, even if the Datalog program deduces five new DDIs. Computing the absolute frequency of a drug being the wedges middle-vertex identifies the

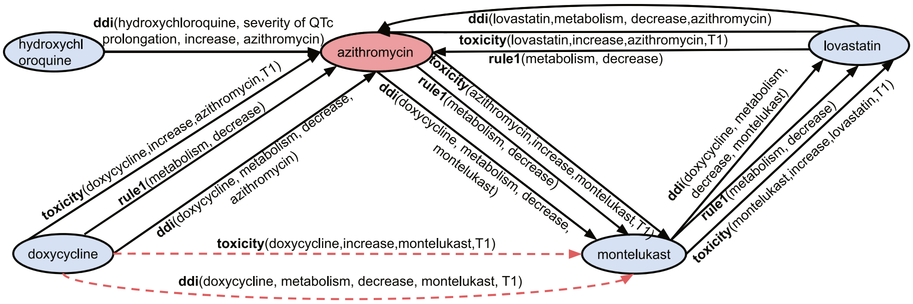

hydroxychl
oroquine
azithromycin
lovastatin
doxycycline
montelukast
ddi(hydroxychloroquine, severity of QTc prolongation, increase, azithromycin)
ddi(lovastatin,metabolism, decrease,azithromycin)
toxicity(lovastatin,increase,azithromycin,T1)
rule1(metabolism, decrease)
toxicity(doxycycline,increase,azithromycin,T1)
rule1(metabolism, decrease)
ddi(doxycycline, metabolism, decrease, azithromycin)
toxicity(azithromycin,increase,montelukast,T1)
rule1(metabolism, decrease)
ddi(doxycycline, metabolism, decrease, montelukast)
ddi(doxycycline, metabolism, decrease, montelukast)
rule1(metabolism, decrease)
toxicity(montelukast,increase,lovastatin,T1)
toxicity(doxycycline,increase,montelukast,T1)
ddi(doxycycline, metabolism, decrease, montelukast, T1)

Fig. 4. COVID-19 treatment. Example of deducing DDIs and computing wedges. The red arrows represent the DDIs deduced, and the red node represents the drug with the higher absolute frequency of being the wedges middle-vertex.

Table 4
The full list of data sources and ontologies used for the Scientific Publications DE.

Sources	#publications	Ontologies	#annotations
PubMed	106,150	MeSH	1,356,578
PMC	26,105	Gene Ontology	125,629
CORD-19	460,772	Disease Ontology	5129

drugs that affect the most the effectiveness or toxicity of drug treatment. In this case, azithromycin is the drug with the higher absolute frequency of being the middle-vertex of the wedges in the graph (i.e., absolute frequency of three); it is followed by montelukast with a value of two and lovastatin with a value of one. Note that after removing azithromycin from the treatment, there is only one DDI between montelukast and lovastatin, i.e., 83.3% of the DDIs are eliminated.

In Section 6.3, we compare the drug–drug interactions deduced by the previously described deduction system and existing tools that discover drug–drug interactions in a treatment. The results of this evaluation suggest that middle-vertices, with high frequency in the directed graph of a therapy, correspond to the drugs that produce more toxicities. Therefore, identifying frequent middle-vertices in the directed graph that models a treatment provides a computational method for discovering toxic medications in treatment.

3.2. The scientific publications DE

The Data Ecosystem of Scientific Publications comprises the components extracting relevant medical concepts from scientific publications.

3.2.1. Data sources

The Scientific Publications DE collects data from the following data sources: CORD-19 [16], PubMed,[21] and PubMed Central (PMC),[22] enriched with information from certain ontologies. CORD-19 is a collection of scientific papers about COVID-19 and related coronavirus; the version by 2021-03-01 includes 460,772 publications. PubMed is Web-accessible engine to primary access scientific publications from the MEDLINE database. The Scientific Publications DE has and harvested articles from PubMed and PubMed Central (PMC) until April 2022, including the MeSH topic 'covid-19'. Table 4 describes the full list of data sources and ontologies used by the Scientific Publications DE.

3.2.2. Data operators

The natural language processing (NLP) tools MetaMap,[23] and SemRep[24] are utilized to recognize drugs and diseases from the titles and abstracts of the integrated articles[25] and also from the full texts of articles that are available in PMC. The Unified Medical Language System (UMLS) is used to describe the extracted medical entities using a controlled vocabulary of medical terms. Moreover, the Medical Subject Headings (MeSH) thesaurus, along with some Open Biological and Biomedical Ontology (OBO) Foundry ontologies, are also harvested in order to retrieve topic annotations and hypernymic relations of drugs and diseases.

In total, 542,672 publications are annotated with semantics relations from UMLS[26] (e.g., ASSOCIATED_WITH, TREATS, CAUSES), adverse events (e.g., Dyspnea increase, Confusion increase), disorders (e.g., Colorectal cancer, Bladder cancer), phenotypes (e.g., Allergic Reaction, Hemorrhage), and drugs (e.g., Becaplermin, Naloxone). Furthermore, metadata of the processed publications (e.g., title, authors, publication date, journal name, and citation number) describes the main attributes of the scientific publications.

3.2.3. Data services

Despite the wide adoption of MetaMap and SemRep tools, their effectiveness is far from perfect [17]. Thus, triples resulting from applying those NLP tools on publications tend to be the most noisy part of the knowledge graph. To overcome this quality challenge, we need to apply some kind of error detection mechanism for the Scientific Publications Graph refinement [18,19]. In our case, we have experimented with various approaches, such as graph embeddings [20], path ranking solutions (PaTyBRED) [21] and a hybrid approach called PRGE (Path Ranking Guided Embeddings) [22]. PRGE method uses the PaTyBRED path ranking technique, in order to produce confidence scores for all the triples of a graph. It then uses those scores in order to guide the TransE embedding method focusing on the probably correct triples, during the graph embedding creation. This is realized by incorporating triple confidence scores in the embedding Loss function, guiding thus the training procedure to put less emphasis on noisy triples. The selected approach results in a final confidence score for each triple of the graph in the range of [0-1]. Deciding a cut-off confidence threshold below which all triples will be considered as erroneous provides a trade-off between quality and the amount of data that will be produced. In our

[21] https://pubmed.ncbi.nlm.nih.gov/.

[22] https://www.ncbi.nlm.nih.gov/pmc/.

[23] https://metamap.nlm.nih.gov/.

[24] https://semrep.nlm.nih.gov/.

[25] https://github.com/SDM-TIB/Knowledge4COVID-19/wiki/CORD-19-Publication-Processing.

[26] https://www.nlm.nih.gov/research/umls/META3_current_relations.html.

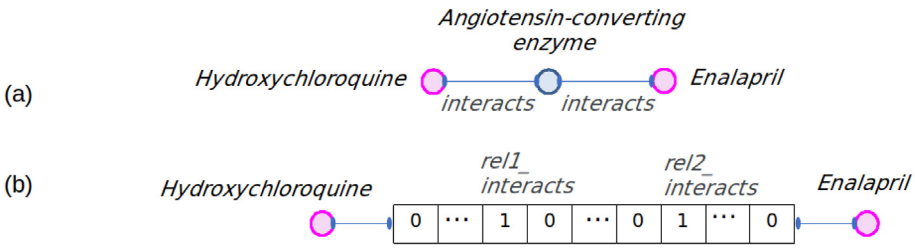


Fig. 5. An example of (a) a semantic path between drugs Hydroxychloroquine and Enalapril, (b) transformation into a feature vector.

case, we selected a median threshold of 0.5, in order to keep the majority of the graph triples. Applying this method to the Scientific Publications DE Graph resulted into a 40% of the total triples identified as possibly erroneous.

Scientific Publications DE analysis: As a next step, we apply a predictive analysis on the Scientific Publications DE, in order to identify previously unknown adverse effects of drug combinations, in the form of drug–drug interaction relations. For this purpose, a machine learning method that exploits patterns unveiled from contextual information of the Scientific Publications DE to predict potential drug–drug interactions is implemented. This method is based on the analysis of the Scientific Publications Graph [23] that results from the natural language processing and semantic indexing of biomedical publications and open resources, as described above. The Scientific Publications Graph constitutes an integral part of the Knowledge4COVID-19 KG, representing the structured information extracted from relevant publications in the form of triples. Drugs included in DrugBank are also considered a part of this graph, relating these with specific targets, diseases, and other biomedical entities identified in literature text, through a set of semantic relations from the UMLS Semantic Network[26].

Prediction of new DDIs: The problem of predicting new drug–drug interactions is addressed as a binary classification problem for interacting/non-interacting drug pairs in the Scientific Publications Graph. The result of this classification provides a set of drug pairs with none previously known interaction, marked as False Positives, that our classifier identifies as interacting with a certain confidence score. These predictions can provide an indication of potential interactions to pharmaceutical experts that have not been previously documented. To this end, the aforementioned machine learning technique focuses on the analysis of the undirected semantic paths connecting different pairs of drugs in the Scientific Publications Graph. This method is called Drug–Drug Interaction prediction on a Biomedical Literature Knowledge Graph (DDI-BLKG) [24]. Each one of these paths includes a sequence of semantic relations of length n that are aggregated into feature vectors representing the frequency of each relation in a specific position $(1, n)$. As an example, if Hydroxychloroquine and the Diabetes-related Enalapril both interact with the target Angiotensin-converting enzyme, this provides the undirected path and the respective feature vector (Fig. 5).

Let D be the number of relevant drugs examined, where relevance is determined by the existence of such drugs in COVID-19 related publications. Aggregating all possible paths between pairs of drug nodes, we generate a big dataset of $(D-1)!$ feature rows that denote relations' frequency in specific positions, as illustrated above. Each feature row is of size $(n\ x\ r)$, where r denotes the number of different relation types. In our case, maximum path length is set to 3 $(n = 3)$, as this has provided the best trade-off between data size and accuracy. Also, 35 unique relation types are used from the UMLS Semantic Network $(r = 35)$. Therefore, $(3\ x\ 35)$ features are calculated for every pair and are used to train a Random Forest classifier that is able to effectively discriminate between two classes: interacting and non-interacting pairs, based on the respective label extracted from a gold dataset.

In order to generate the final set of predictions, the Random Forest classifier is trained using all COVID-19 related pairs (where at least one of the two drugs is mentioned in Drugbank as COVID-19 experimental treatments[27]), denoted as positives in DrugBank. Testing the classifier for all possible remaining COVID-19 related drug pairs, which are not known to be interacting, produces 8925 unknown drug–drug interaction predictions in total, with a certain confidence score within a range of [0,1]. The critical threshold of this score is considered to be 0.5, meaning that drug pairs with a score < 0.5 are less possible to be interacting, while pairs with a score > 0.5 represent the most possible interactions.

4. The Knowledge4COVID-19 knowledge graph

This section the Knowledge4COVID-19 DE in terms of the pipeline for the creation of Knowledge4COVID-19 knowledge graph (KG), the linking to existing KGs (e.g., DBpedia and Wikidata), and the techniques of federated query processing implemented on top of Knowledge4COVID-19 KG. The Knowledge4COVID-19 DE relies on annotations from UMLS, DBpedia, and Wikidata to solve entity alignment. The execution of 221 RML mapping assertions – manually defined by two knowledge engineers and curated by two more – transforms the structured representation of the data sources, annotations, and alignments into the Knowledge4COVID-19 KG.

Fig. 6 depicts the steps of the KG creation process. Steps 1 and 2 are done at the level of Scientific Open Data and Publications DEs, while steps 3 and 4 are conducted at the level of Knowledge4COVID-19 DE (Fig. 2) to create the Knowledge4-COVID-19 KG. First, data is ingested and described in terms of metadata (step 1), e.g., title and abstract of the publications, and drug–drug interactions. Knowledge extraction methods recognize biomedical entities from textual data and link them to UMLS, and to resources in DBpedia, Wikidata, Uniprot, and DrugBank. A total of 12,223,409 UMLS annotations have been extracted by FALCON. These annotations are used for solving entity alignment and semantic data integration of biomedical entities in the Knowledge4COVID-19 KG (e.g., drugs, phenotypes, side effects, and adverse events). Moreover, there are 3,739,445 links to DBpedia, 3,476,435 links to Wikidata, 5248 links to the Uniprot RDF KG, and 3427 links to DrugBank.

The shared data sources are mapped to the Knowledge4-COVID-19 unified schema. SDM-RDFizer [25] transforms these shared data into an RDF graph by executing the RML mapping rules. SDM-RDFizer implements optimized data structures that are

[27] https://go.drugbank.com/covid-19#drugs.

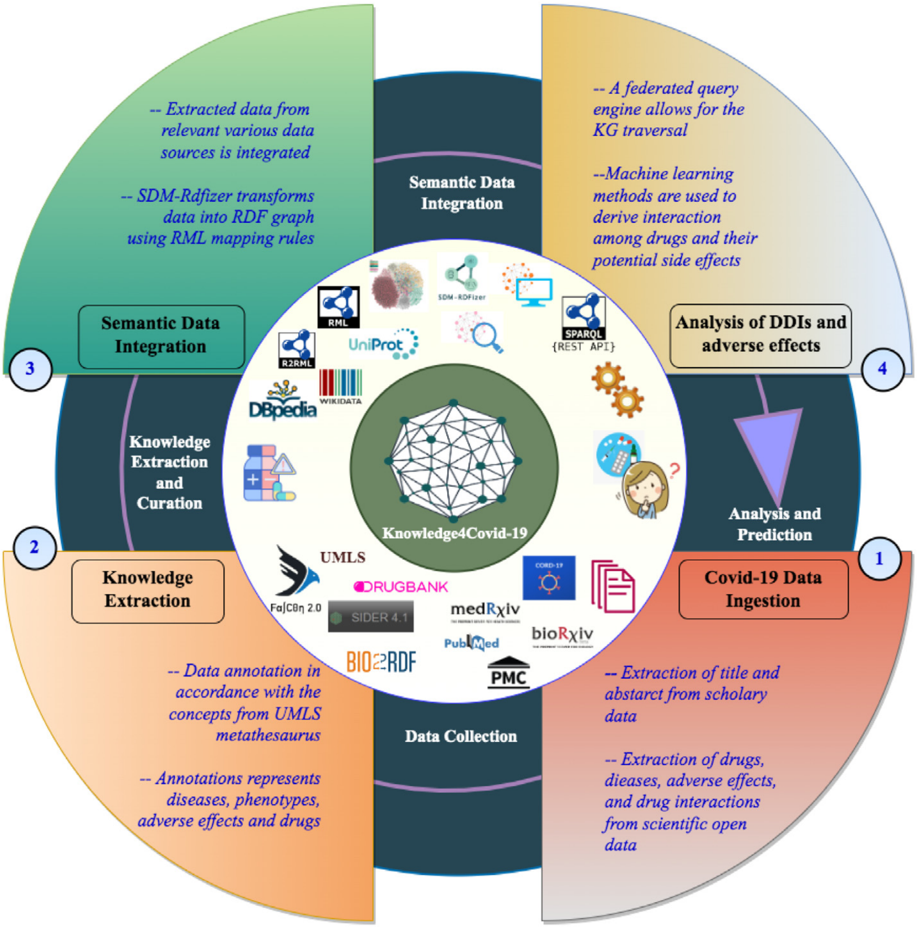

-- Extracted data from relevant various data sources is integrated
-- SDM-Rdfizer transforms data into RDF graph using RML mapping rules
Semantic Data Integration
Semantic Data Integration
3
-- A federated query engine allows for the KG traversal
--Machine learning methods are used to derive interaction among drugs and their potential side effects
Analysis of DDIs and adverse effects
4
Knowledge Extraction and Curation
Analysis and Prediction
RML
R2RML
UniProt
SDM-RDFizer
SPARQL
{REST API}
WIKIDATA
DBpedia
Knowledge4Covid-19
2
Knowledge Extraction
UMLS
FalCon 2.0
DRUGBANK
SIDER 4.1
BIO2RDF
medRxiv
PubMed
bioRxiv
PMC
CORD-19
1
Covid-19 Data Ingestion
-- Data annotation in accordance with the concepts from UMLS metathesaurus
-- Annotations represents diseases, phenotypes, adverse effects and drugs
Data Collection
-- Extraction of title and abstarct from scholary data
-- Extraction of drugs, dieases, adverse effects, and drug interactions from scientific open data

Fig. 6. The Knowledge4COVID-19 KG Pipeline. Steps followed during the transformation of heterogeneous data into the Knowledge4COVID-19 KG. UMLS annotations provide the basis for entity alignment and data integration.

exploited during the execution of RML mapping rules to speed up the KG creation process [25]. The Knowledge4COVID-19 KG is published following the Linked Data principles. A linked data interface using Pubby[28] is provided; thus, all the URIs can be dereferenced. Additionally, a SPARQL endpoint allows for querying processing on top of the Knowledge4COVID-19 KG, while the federated query engine, DeTrusty [26], evaluates SPARQL queries over the federation of the Knowledge4COVID-19 KG, DBpedia, Wikidata, and UniProt RDF. Additionally, various API REST services are offered to traverse the Knowledge4COVID-19 KG, and analyze drug–drug interactions and side effects (step 4).

4.1. The Knowledge4COVID-19 unified schema

The Knowledge4COVID-19 unified schema comprises concepts that provide abstract representations of the entities present in the data sources. Each generic concept of a type or category is defined as a Class in OWL. These concepts represent annotations from controlled vocabularies, drugs, COVID-19 treatments and drugs, disorders, phenotypes, adverse events, enzymes, targets, side effects, scientific publications, and interactions between drugs, drugs and side effects, and drugs and their targets. The current version of the unified schema is composed of 67 classes, 37 object properties, 49 data type properties, and eight annotation properties. Fig. 7 shows examples of classes and properties of the Knowledge4COVID-19 unified schema. The inner circle in Fig. 7 displays 17 classes of the unified schema; each class is shown in a different color. The outer circle, however, illustrates examples of the properties categorized by the classes. Each group of properties shown in the same color as one class represents all the properties which domains are the same class; in average, a class has in 3.7 properties. Following the Global as View (GAV) modeling approach [27], we define the classes in the unified schema such that they involve all the concepts represented in data sources and recognized by a domain expert. Similarly, the properties are defined considering the domain specific relations between the concepts residing in different data sources.

In defining the unified schema concepts, we exploit two available unified schemas corresponding to two different biomedical knowledge graphs: iASiS.[29] and BigMedilytics[30] Additionally, the Knowledge4COVID-19 unified schema concepts (i.e., classes and properties) are related via the `owl:equivalentClass` and `owl:equivalentProperty` predicates to concepts in DBpedia, Wikidata, Uniprot, the Open Biological and Biomedical Ontology, the Semanticscience Integrated Ontology, and Dublin Core. In total, 17 concepts are mapped to at least one concept in these ontologies.

The unified schema is publicly available as a VoCol repository supported by TIB.[31] VoCol [28] provides a loose coupling of components for validation, querying, analytics, visualization,

[28] https://github.com/cygri/pubby.

[29] http://ontology.tib.eu/iasis/.

[30] http://ontology.tib.eu/bigmedilytics/.

[31] http://ontology.tib.eu/K4COVID-19/.

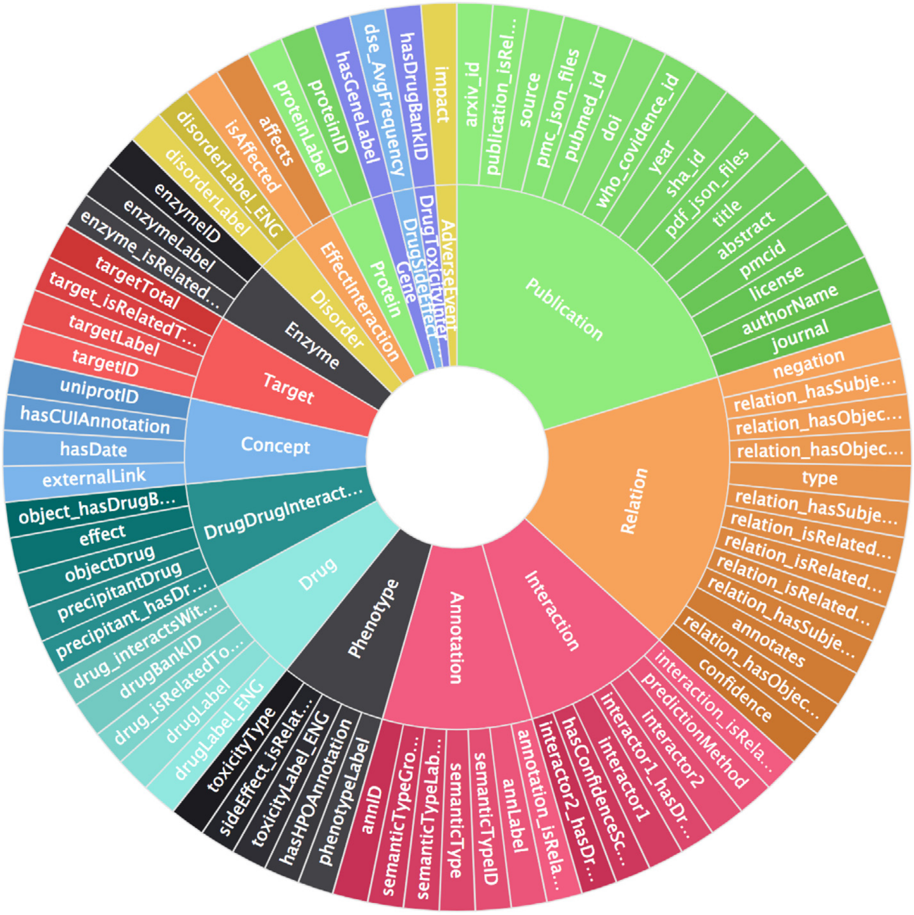
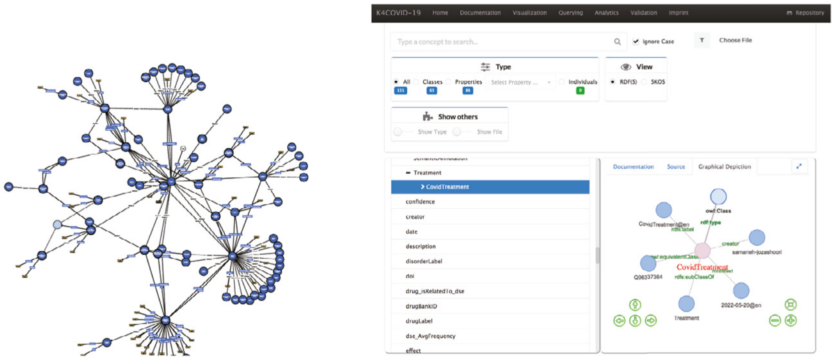


Fig. 7. The Unified Schema. Classes and properties.



(a) Classes and properties

(b) Metadata of the class covid-19:CovidTreatment

Fig. 8. The Knowledge4COVID-19 unified schema. VoCol Visualization of the classes, and data and object properties.

and documentation on top of a standard Git repository. VoCol also provides an interface for specifying queries against the unified schema and ontology management features that enable the visualization and exploration of the ontology. Finally, the documentation describing the metadata of each class and property can be consulted, as well as basis analysis describing the number of classes and properties that comprise the unified schema. Fig. 8(a) depicts the Knowledge4COVID-19 unified schema visualized by VoCol. The metadata describing each of the depicted concepts can be accessed at VoCol.[32] Fig. 8(b) presents the description of the

[32] http://ontology.tib.eu/K4COVID-19/documentation.

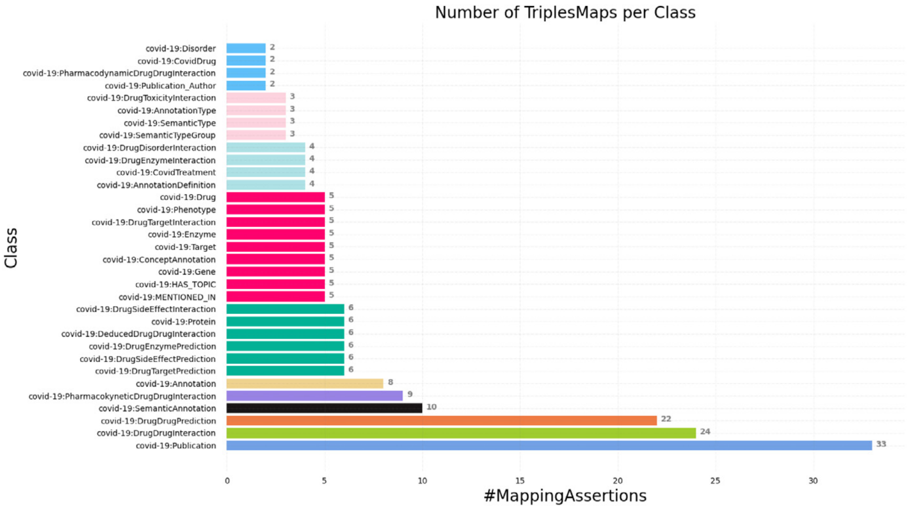


Fig. 9. Mapping Assertions in RML Triples Maps. Number of Mapping Assertions (i.e., subject and object maps) per classes.

```
PREFIX rr:   <http://www.w3.org/ns/r2rml#>
PREFIX rml: <http://semweb.mmlab.be/ns/rml#>
PREFIX COVID-19:  <http://research.tib.eu/covid-19/vocab/>
SELECT DISTINCT  ?mappingRule ?logicalSource ?predicate ?sourceAttribute
WHERE {
?mappingRule  rml:logicalSource  ?ls.
?ls           rml:source         ?logicalSource.
?mappingRule  rr:subjectMap      ?subject.
?subject      rr:class           COVID-19:Publication.
OPTIONAL {    ?mappingRule  rr:predicateObjectMap  ?pObjectMap  .
              ?pObjectMap   rr:predicate           ?predicate  .
              ?pObjectMap   rr:objectMap           ?objectMap  .
              ?objectMap    ?mode                  ?sourceAttribute}}
```

Fig. 10. SPARQL Query to retrieve the RML rules that define `COVID-19:Publication`.

class `covid-19:CovidTreatment`, which groups of COVID-19 drugs and the drugs of common comorbidities.

4.2. Mapping the data sources into the unified schema

Classes and properties in the unified schema are defined in terms of the attributes in the data sources by means of RML triples maps. The Knowledge4COVID-19 KG is defined using 57 RML triples maps that comprise 223 mapping assertions (i.e., subject or object Map). Fig. 9 presents the number of mapping assertions of the RML triples maps that define each of the unified schema classes and their properties. For example, the class `COVID-19:DrugDrugInteractionPrediction` is defined using 22 mapping assertions, and `COVID-19:Publication` is the class with the greater number of properties and is defined by 33 mapping assertions. A SPARQL endpoint with the unified schema and the triples maps is publicly available.[33]

Fig. 10 presents a SPARQL query that collects the information about the mapping rules that define the class `COVID-19:Publication`. The results of this query evaluation include the data source from where the data is collected, and per predicate of the class, the attribute(s) of the corresponding data source used to populate the predicate.

4.3. The Knowledge4COVID-19 KG in numbers

The current version of the Knowledge4COVID-19 KG comprises 80,570,440 RDF triples. Fig. 11 depicts the number of resources per class in the Knowledge4COVID-19 KG. As observed, `covid-19:Annotation` comprises 4,536,579 resources, 542,672 resources in `covid-19:Publication`, 503,700 for `covid-19:PharmacokyneticDrugDrugInteraction`. The Knowledge4COVID-19 KG includes 87 COVID-19 drugs; 68 drugs are from DrugBank.[34] and the rest have been extracted from the Mayo Clinical website[35] Additionally, the Knowledge4COVID-19 KG integrates 216 COVID-19 treatments that comprise COVID-19 drugs and drugs for the most common comorbidities that impact on the survival of COVID-19 patients [29]: hypertension, depressive syndrome anxiety, obesity, cardiopathy, diabetes mellitus, hepatitis disease, chronic obstructive pulmonary disease, renal disease, asthma, dyslipidemia hypercholesterolemia, neurodegenerative disorder, gastrointestinal disease, vascular disease, benign prostatic hyperplasia, and obstructive sleep apnea. There are 923 deduced DDIs (a.k.a. DeducedDDIs). In average, each

[33] https://labs.tib.eu/sdm/covid19-kg-mappings/sparql.

[34] https://go.drugbank.com/covid-19.

[35] https://www.mayoclinic.org/diseases-conditions/coronavirus/expert-answers/coronavirus-drugs/faq-20485627.

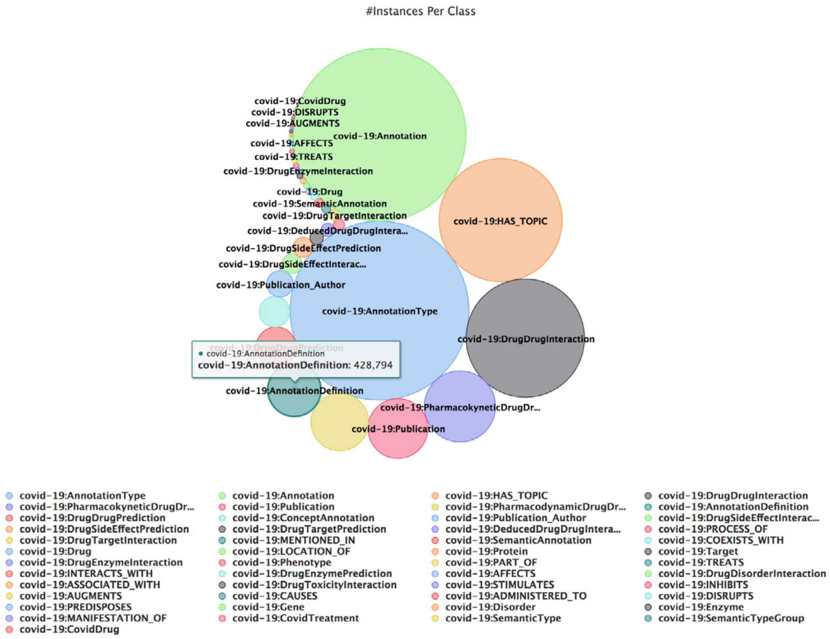


Fig. 11. Knowledge4COVID-19 KG. Number of Resources per classes; 4,864,162 annotations encode the meaning of 542,672 scientific publications and open data.

COVID-19 treatment has 10.63 drugs, 1.58 COVID-19 drugs, and 9.11 comorbidity drugs. Additionally, COVID-19 treatments have in average two comorbidities and 121.43 DeducedDDIs; the same DDI can produce different effects, and they are counted as different DDIs. Moreover, the Knowledge4COVID-19 KG integrates 345,116 CRD and 5513 NCRD pairs of drugs, and 124,537 instances of predicted DDIs (i.e., instances of the class `covid-19:DrugDrugPrediction`). Specifically, 8925 of the predicted DDIs are generated by the DDI-BLKG method, 5907 have a score equal or greater than 0.5 (a.k.a. DDI-BLKG-0.5). The rest of the DDIs are discovered by state-of-the-art methods; they are included in the KG to provide a baseline for future benchmarking. These DDIs are predicted from the DDIs extracted from DrugBank, and are as follows: (**i**) TransE [30] 28,752 DDIs generated by TransE. (**ii**) RESCAL [31] 28,752 DDIs generated by RESCAL. (**iii**) HolE [32] 28,752 DDIs generated by HolE. (**iv**) DistMult [33] 28,752 DDIs generated by DistMult.

5. Exploring the Knowledge4COVID-19 KG

This section describes the services implemented to facilitate the traversal and data retrieval on top of the Knowledge4COVID-19 KG.

5.1. Relevant adverse effects detected on Knowledge4COVID-19

This service aims at providing the support for analyzing relevant adverse effects that may be produced as a result of interactions among drugs to treat COVID-19 and conditions. As a proof of concept, we illustrate the results of the analysis of the most common comorbidities, i.e., hypertension, asthma, and diabetes. These comorbidities are linked to the ACE-2 receptor expression and may facilitate the entry of the virus into the host cells as a consequence of releasing the proprotein convertase. More importantly, this effect may fire a "vicious infectious circle" which may result in the increase of morbidity and mortality [34]. Nevertheless, a more detailed analysis of the impact of the combination of drugs can be executed on the public available Jupyter Notebook.[36] Exemplar drug–drug interactions represented in the Knowledge4COVID-19 KG can also be visualized.[37]

Figs. 12, 13, and 14 depict adverse effects that can be triggered in COVID-19 patients who receive treatments for hypertension, asthma, or diabetes. Each plot reports a labeled directed graph, nodes represent drugs and an edge between two drugs, represent an interaction. The label of an edge, denoted by the line color and the figure legend, indicate the type of side effect.

Fig. 12 presents 14 types of drug–drug interactions that may occur among the COVID-19 drugs Hydroxychloriquine, Zinc, and Chloroquine, and asthma drugs. The pharmacokinetic drug–drug interactions between a pair of drugs A and B indicate that A impacts B's absorption, metabolism, excretion when both drugs are administrated together. As a result, A may reduce the effectiveness or increase toxicities. The rest of the interactions are pharmacodynamic, i.e., their pharmacological outcome may be affected. Six out of the 14 reported drug–drug interactions are pharmacokinetic. Chloroquine may reduce the metabolism of Zafirlukast, Mometasone, and Fluticasone; it can also decrease the excretion rate of Levosalbutamol. Hydroxychloriquine also impacts the metabolism of Theophylline. Furthermore, the serum concentration of Chloroquine may be increased with asthma drugs by Methylprednisolone, Prednisone, and Budesonide. Thus, the effectiveness of the treatment was negatively affected. Four drugs may increase the severity of the side effects of Hydroxychloriquine. At the pharmacodynamic level, it can be observed that Montelukast and Chloroquine may increase the risk of myopathy, and Salmeterol and Hydroxychloriquine may increase the risk of QT prolongation. Since the risk of cardiac events during QT syndrome is high, these results suggest that the combinations of the treatments need to be administrated with great precaution. Similarly, Fig. 13 reveals a more significant number of interactions among the drugs Hydroxychloriquine, Zinc, and Chloroquine and the drugs typically prescribed to Type 2 diabetes patients. All the drugs affect the efficacy of Hydroxychloriquine and the

[36] https://colab.research.google.com/drive/146-oQTxDpZQoOifKY6iafaEwuupH7q3t?usp=sharing.

[37] https://youtu.be/7YsTYJzRfR0.

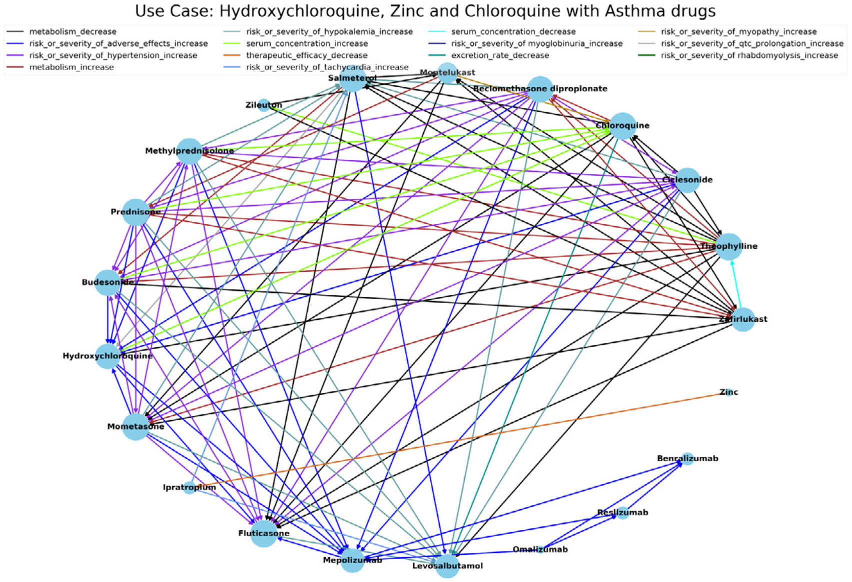

Use Case: Hydroxychloroquine, Zinc and Chloroquine with Asthma drugs
metabolism_decrease
risk_or_severity_of_adverse_effects_increase
risk_or_severity_of_hypertension_increase
metabolism_increase
risk_or_severity_of_hypokalemia_increase
serum_concentration_increase
therapeutic_efficacy_decrease
risk_or_severity_of_tachycardia_increase
serum_concentration_decrease
risk_or_severity_of myoglobinuria_increase
excretion_rate_decrease
risk_or_severity_of_myopathy_increase
risk_or_severity_of_qtc_prolongation_increase
risk_or_severity_of rhabdomyolysis_increase
Salmeterol
Montelukast
Beclomethasone dipropionate
Zileuton
Chloroquine
Methylprednisolone
Ciclesonide
Prednisone
Theophylline
Budesonide
Zafirlukast
Hydroxychloroquine
Zinc
Mometasone
Benralizumab
Ipratropium
Reslizumab
Fluticasone
Omalizumab
Mepolizumab
Levosalbutamol

Fig. 12. The adverse effects generated as the result of the interactions among COVID-19 drugs (Hydroxychloroquine, Zinc, and Chloroquine) with treatments for Asthma. Relations retrieved from the Knowledge4COVID-19 KG.

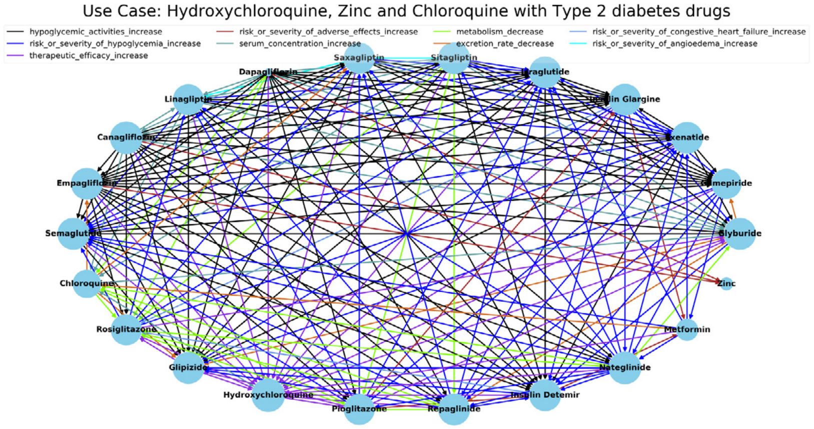

Use Case: Hydroxychloroquine, Zinc and Chloroquine with Type 2 diabetes drugs
hypoglycemic_activities_increase
risk_or_severity_of_hypoglycemia_increase
therapeutic_efficacy_increase
risk_or_severity_of_adverse_effects_increase
serum_concentration_increase
metabolism_decrease
excretion_rate_decrease
risk_or_severity_of_congestive_heart_failure_increase
risk_or_severity_of_angioedema_increase
Saxagliptin
Sitagliptin
Dapagliflozin
Liraglutide
Linagliptin
Insulin Glargine
Canagliflozin
Exenatide
Empagliflozin
Glimepiride
Semaglutide
Glyburide
Chloroquine
Zinc
Rosiglitazone
Metformin
Glipizide
Nateglinide
Hydroxychloroquine
Insulin Detemir
Pioglitazone
Repaglinide

Fig. 13. The adverse effects generated as the result of the interactions among COVID-19 drugs (Hydroxychloroquine, Zinc, and Chloroquine) with treatments for Type 2 Diabetes. Relations retrieved from the Knowledge4COVID-19 KG.

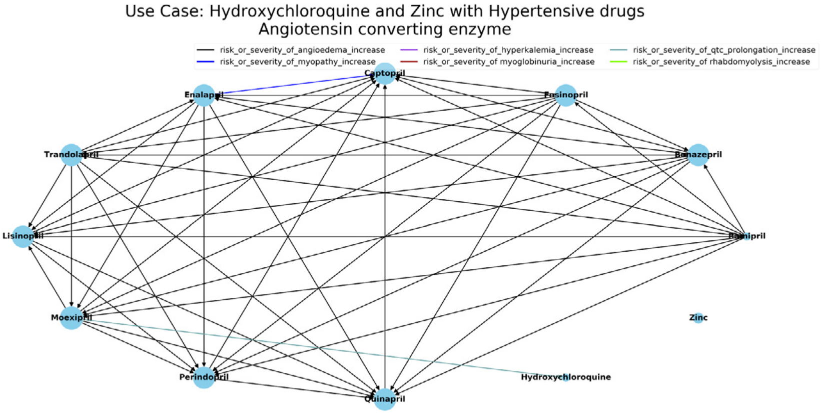

Use Case: Hydroxychloroquine and Zinc with Hypertensive drugs
Angiotensin converting enzyme
risk_or_severity_of_angioedema_increase
risk_or_severity_of_myopathy_increase
risk_or_severity_of_hyperkalemia_increase
risk_or_severity_of myoglobinuria_increase
risk_or_severity_of_qtc_prolongation_increase
risk_or_severity_of rhabdomyolysis_increase
Captopril
Enalapril
Fosinopril
Trandolapril
Benazepril
Lisinopril
Ramipril
Moexipril
Zinc
Perindopril
Hydroxychloroquine
Quinapril

Fig. 14. The adverse effects generated as the result of the interactions among COVID-19 drugs (Hydroxychloroquine, Zinc, and Chloroquine) with treatments for Hypertension. Relations retrieved from the Knowledge4COVID-19 KG.

combination of Rosiglitazone in treatments with Insulin Determir or Insulin Glargine. Additionally, the therapeutic efficacy of Rosiglitazone can be increased when used in combination with Hydroxychloroquine, and Chloroquine may reduce the effectiveness of Metformin. They should be administrated with precaution because their therapeutic efficacy may be reduced. Drug interactions of hypertension treatments based on drugs Angiotensin converting enzyme, with the drugs Hydroxychloriquine and Zinc are reported in Fig. 14. As reported, the combination of these drugs may cause pharmacodynamic interactions that can critically affect the function of nerve and muscle cells, including those in the heart. The above results suggest that COVID-19 patients receiving treatments for pre-existing conditions need to be carefully treated.

5.2. Web APIs to traverse the Knowledge4COVID-19 KG

The Knowledge4COVID-19 KG can be explored by executing SPARQL queries against the public SPARQL endpoint[7]. Additionally, specific Web Application Programming Interfaces (APIs)[6] allow for the execution of specific requests. They include (**i**) the Publications related to drugs; (**ii**) the Drug–Drug Interactions between two or several drugs; (**iii**) the Predicted Drug–Drug Interactions between two or several drugs. The source code and the description of how to use the API is available on our GitHub repository[6]. The Web APIs were executed 20 times, and the average execution time is reported.

Publications related to drugs retrieves the scientific publications annotated with UMLS Concept Unique Identifiers (aka CUIs) of the input drugs.

Input: CUI ids for one or several drugs.

Output: All the properties of the publications annotated with input drugs.

Pre-conditions: Publications are correctly annotated with CUIs.

Post-conditions: Returned publications have mentions of the input drugs with respect to the CUI annotations in the abstract or title.

Average response time: 50 ms.

Example SPARQL Query: Appendix A.1.

Drug–Drug Interactions (DDI) retrieves the DDI of the input drugs.

Input: Drug CUIs and a variable "`target`" to indicate the output mode.

Output: Drug–Drug Interactions related to the input drugs with all the properties defined in the KG. Interactions of the related drugs are returned as an output. Each interaction includes the effector drug, the affected drug, the effect, and the impact of the effect. If the variable `target` = DDI, then return the DDI of each input drug individually. If `target` = drug–drug interactions, then return the DDI of all the possible pairs of the input drugs.

Pre-conditions: Drugs have interactions in the KG; these interactions are extracted from DrugBank or the literature.

Post-conditions: Returned interactions are related to the drugs in the input.

Average response time: 62 ms.

Example SPARQL Query: Appendices A.2.1 and A.2.2.

Predicted Drug–Drug Interactions retrieves predicted DDI of input drugs.

Input: Drug CUI and a variable "`target`" to indicate the output mode.

Output: Predicted Drug–Drug Interactions related to the input drugs with all the properties defined in the KG. Predicted interactions of the related drugs are returned as an output. Each interaction consists of the effector drug, the affected drug, a confidence score of the interaction, and the provenance. If `target` = DDIP, then the API returns the predicted DDI of each drug individually. If `target` = DDIPS, then the API returns the predicted DDI of all the possible pairs of the input drugs.

Pre-conditions: Drugs have predicted interactions in the KG.

Post-conditions: Returned predicted interactions have a confidence score greater than zero wrt the CUI of the drugs in the input.

Average response time: 58 ms.

Example SPARQL Query: Appendices A.3.1 and A.3.2.

5.3. Federated query processing on top of the Knowledge4COVID-19 KG

DeTrusty [26] is a federated query engine for RDF sources. Hence, it allows querying the Knowledge4COVID-19 KG in conjunction with external sources like DBpedia, Wikidata, and Uniprot.[38] This in turn is only possible because entities in the Knowledge4COVID-19 KG are linked to those datasets. Fig. 15 shows an example of a federated query; providing information about treatments that involve drugs for COVID-19 and Asthma. DeTrusty contacts both KGs to retrieve the complete answer to the query. The Knowledge4COVID-19 KG delivers data about the treatments fulfilling the conditions; including `owl:sameAs` links for both drugs. DeTrusty also contacts DBpedia to get additional information about the COVID-19 drugs, e.g., the route of administration. DeTrusty decomposes the SPARQL query into star-shaped sub-queries around the subjects [35], i.e., each triple of a subquery has the same variable or constant in the subject position. For source selection in the presence of a SPARQL query without the `SERVICE` clause, DeTrusty uses a semantic source description with information about the classes and their predicates, like MULDER [36].

6. Evaluation of Knowledge4COVID-19

In this section, we report on the evaluation of the quality of the integrated data and the patterns discovered by exploiting the knowledge encoded in the Knowledge4COVID-19 KG. We aim to answer the following research questions: (**Q1**) What is the accuracy of the named entity recognition (NER) and named entity linking (NEL) performed over data from DrugBank to extract drug–drug interactions and the effects of these interactions? (**Q2**) What is the accuracy of the prediction methods that enhance the knowledge about drug–drug interactions? (**Q3**) What is the quality of the knowledge discovery methods implemented on top the Knowledge4COVID-19 KG to uncover drug–drug interactions among the multi-drug COVID-19 treatments?

6.1. Effectiveness of NER and NEL methods

Data about drug–drug interactions is collected from DrugBank release 2022-01-04 with 1,273,052 entries composed of pairs of drugs and the textual description of the effects of each interaction. In order to evaluate the performance of FALCON in this use case, 1198 DDI descriptions were manually annotated by twelve annotators; annotations correspond to CUIs from UMLS and constitute the gold standard of the evaluation. For example, for the DDI description: "The serum concentration of Lepirudin can be decreased when it is combined with Tipranavir"; Lepirudin and Tipranavir correspond to the extracted entities from the

[38] https://labs.tib.eu/sdm/k4covid-query-engine/sparql.

```
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX k4covid: <http://research.tib.eu/covid-19/vocab/>
PREFIX k4covide: <http://research.tib.eu/covid-19/entity/>

SELECT DISTINCT * WHERE {
  SERVICE <https://labs.tib.eu/sdm/covid19kg/sparql> {
    ?treatment k4covid:hasCovidDrug ?covidDrug.
    FILTER( ?comorbidity=k4covide:Asthma )
    ?treatment k4covid:hasComorbidity ?comorbidity.
    ?treatment k4covid:hasComorbidityDrug ?comorbidityDrug.
    ?comorbidityDrug k4covid:hasCUIAnnotation ?CUIComorbidityDrug.
    ?CUIComorbidityDrug owl:sameAs ?sameAsComorbidityDrug .
    ?covidDrug k4covid:hasCUIAnnotation ?CUICovidDrug.
    ?CUICovidDrug owl:sameAs ?sameAsCovidDrug .
  }
  SERVICE <https://dbpedia.org/sparql> {
    ?sameAsCovidDrug dbp:excretion ?excretation.
    ?sameAsCovidDrug dbp:metabolism ?metabolism.
    ?sameAsCovidDrug dbp:routesOfAdministration ?routes.
  }
}
```

Fig. 15. Example of a Federated Query.

above record, while decrease and serum concentration represent, respectively, the effect and impact of the interaction of Tipranavir with Lepirudin. One of the annotators was a senior researcher, two were experts in the biomedical domain, and the rest were Computer Science Ph.D. students. Disagreements among the annotators were solved by majority voting. A 2-fold cross-validation was followed. The evaluation indicates a precision of 98%. The 2% where FALCON failed to extract and link the terms correctly, are interactions which contain more than one interaction in the same sentence, where FALCON was only considering one interaction (Table 2 last pattern). All the drug–drug interactions, that followed this pattern, were corrected manually before integrating them into the Knowledge4COVID-19 KG.

6.2. Effectiveness of the predictive tasks for DDI identification

As explained, DrugBank presents an adequate source for retrieving potential adverse effects of treatments received by COVID-19 patients in combination with other medications, explicitly providing the interactions of each drug with other drugs in a structured way. However, many drug interactions observed in everyday practice are not currently recorded in medical databases like DrugBank that are continuously evolving and extended with new drugs and relevant information.[39] Thus, a Knowledge Graph completion challenge arises, for adding new drug–drug interaction links in the Knowledge4COVID-19 KG.

The effectiveness of the Random Forest classifier presented previously in Section 3.2, for predicting new interactions is assessed following a 10-fold cross-validation (cv) procedure, and DrugBank v5.0.3 is our gold dataset for drug–drug interactions. Existing techniques for knowledge graph embeddings available in the TorchKGE[40] library (i.e., TransE, RESCAL, HolE, and DistMult) are used as baselines. Each model is trained for a maximum of 100 epochs while early stopping was used, utilizing 10% of the data for validation. For each model, 100-sized embeddings were used, since an increase of the embedding size did not provide better results. The performance of the predictive models is measured using the area under the receiver-operating characteristic (ROC-AUC), as well as the macro-average of Precision, Recall and F1-score for the positive class. Fig. 16 suggests that our approach (DDI-BLKG) outperforms all mainstream embedding-based methods tested. DDI-BLKG can exploit knowledge encoded in the fine-grained representation of the publications in the Knowledge Graph. As a result, the DDI-BLKG prediction accuracy is enhanced compared to the baseline methods.

Moreover, Fig. 17 reports on the overlap between the DDIs deduced on the drugs of the COVID-19 treatments (a.k.a. DeducedDDIs), DDI-BLKG, DDI-BLKG-0.5 (DDI-BLKG with a prediction score equal or greater than 0.5), CRD, and NCRD. It is essential to highlight that CRD and NCRD are computed from the whole DrugBank dataset of drugs, while DDI-BLKG and DeducedDDIs are limited to COVID-19 drugs. The percentages of overlap of DeducedDDIs, DDI-BLKG, and DDI-BLKG-0.5 with CRD are 24.70%, 17.51%, and 22.60%. Thus, both methods (i.e., the deductive system and DDI-BLKG) can identify DDIs between drugs mediated by the CYP enzyme family, i.e., CRD pairs of drugs. CYP enzymes play an important role in catalyzing the metabolism of pharmaceuticals and their inhibition or induction causes clinically significant

[39] https://go.DrugBank.com/release_notes.

[40] https://torchkge.readthedocs.io.

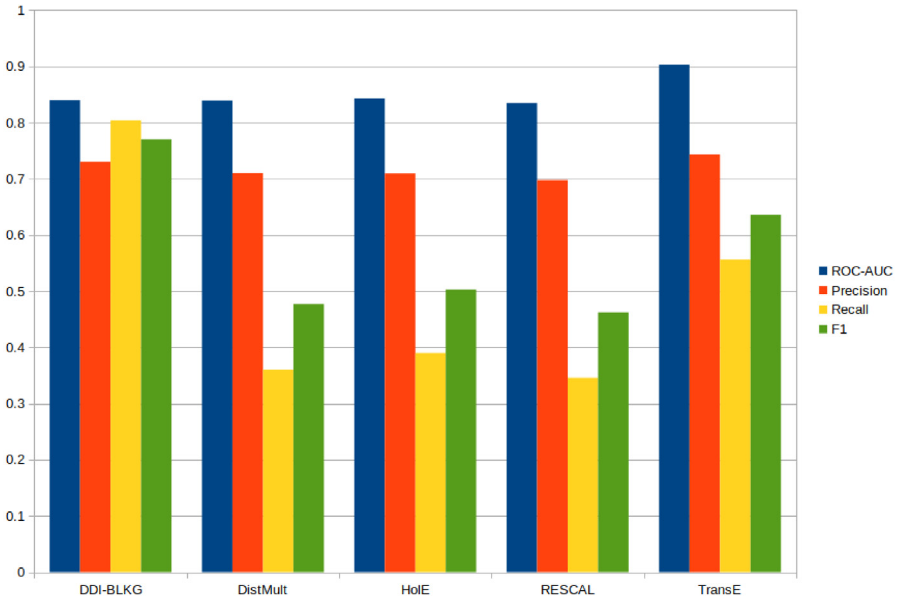


Fig. 16. Results of the 10-fold cross validation, comparing the current DDI prediction approach (DDI-BLKG) with various graph embedding methods. Only the TransE approach outperforms our approach in ROC-AUC and Precision, while in terms of the F1 score, DDI-BLKG outperforms all embeddings by far.

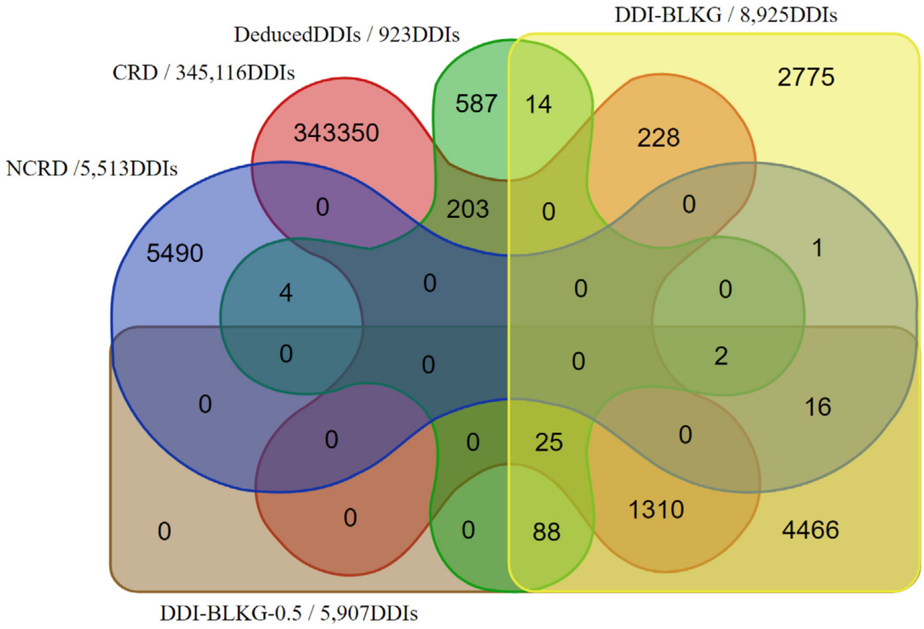


Fig. 17. Venn Diagram. It depicts the overlap among five sets of DDIs. 345,116 CRD pairs of drugs targeting at least one protein of the family CYP. 5513 NCRD are pairs of drugs targeting a No CYP protein. 8925 DDI-BLKG are DDIs predicted by the DDI-BLKG method, while 5907 DDI-BLKG-05 represents the subset of DDIs in DDI-BLKG with score equal or greater than 0.5. 923 DeducedDDIs generated by the deductive system.

pharmacokinetic drug–drug interactions [37]. Thus, these results suggest that even though these methods do not exploit any information about the drug's target enzymes, they can identify a good proportion of DDIs that are part of the CRD group.

6.3. Impact on the effectiveness and toxicity of COVID-19 treatments

The Knowledge4COVID-19 KG is a unique source of knowledge to identify patterns in the integrated networks of interactions, biomedical entities, and publications, e.g. adverse events generated by combining COVID-19 drugs and drugs prescribed for pre-existing conditions. Note that existing tools (e.g., COVID-19 Drug Interactions for University of Liverpool[41]) only identify pairwise interactions. In this section, we evaluate the drug–drug interactions that can be deduced over the Knowledge4COVID-19

[41] https://www.covid19-druginteractions.org/.

Table 5
Five COVID-19 treatments. Frequency distribution of wedges with knowledge capture. Treatments are evaluated in four interaction checker tools: COVID-19, WebMD, Medscape, and DrugBank (May 2nd, 2022). For each tool, it is shown the DDI-Reduction percentage that indicates how many DDIs are avoided in a treatment when the middle-vertex drug is removed. The DDI-reduction percentage is a higher-is-better metric. Middle-vertex drugs reduce DDIs, suggesting, thus, wedges and their middle vertices are part of DDIs that affect treatment effectiveness and toxicities. Best values in **bold**.

T	Knowledge capture		*D*	DDI-R rcentage			
	Middle-Vertex	*F*		COVID-19	WebMD	Medscape	Drugbank
T1	**Azithromycin**	**9**	45.45	**100.0**	**100.0**	**100.0**	42.9
	Montelukast	4					
	Lovastatin	4					
	Hydroxychloroquine	0					
	Doxycycline	0					
T2	**Ciprofloxacin**	**12**	52.17	33.3	**75.0**	**75.0**	44.4
	Metoprolol	**12**		33.3	25.0	25.0	33.3
	Hydroxychloroquine	9					
	Azithromycin	9					
	Linagliptin	7					
T3	**Hydroxychloroquine**	**5**	33.33	**100.0**	25.0	25.0	60.0
	Glyburide	**5**		0.0	50.0	50.0	60.0
	Simvastatin	3					
	Azithromycin	3					
	Ramipril	0					
T4	**Propranolol**	**8**	15.38	**100.0**	50.0	50.0	60.0
	Hydroxychloroquine	5					
	Azithromycin	5					
	Theophylline	4					
	Ramipril	1					
T5	**Timolol**	**11**	38.89	**50.0**	**50.0**	**50.0**	44.4
	Cyclophosphamide	**11**		0.0	0.0	0.0	44.4
	Azithromycin	7					
	Hydroxychloroquine	7					
	Bupropion	6					

KG and the effects of these interactions. We consider five COVID-19 treatments and the effects of including in these treatments drugs for comorbidities. The treatment for COVID-19 used in these five cases is recommended by the official guidelines.[42] The concomitant drugs used in the first treatment $T1$ are for the comorbidities asthma, high cholesterol, and pneumonia and for the second treatment $T2$ are diabetes, hypertension, and pneumonia. The comorbidities in the third treatment $T3$ are diabetes, high cholesterol, hypertension. The comorbidities in the fourth treatment $T4$ are asthma and hypertension, and for the fifth treatment, $T5$ are renal diseases, obesity, and hypertension.

Table 5 shows the percentage of DDIs deduced (D) and wedge absolute frequency (F) for each middle-vertex by the method [14] in existing treatments.

The middle-vertex of a wedge is highly important because the middle-vertex is both the object drug for one interaction and the precipitant drug for another interaction. Thus, drugs that correspond to the middle-vertex of wedges, represent drugs whose presence in the treatment may negatively impact effectiveness and toxicity. We can observe in Table 5 that over 15% of new DDIs are deduced in all the treatments. Table 5 shows the DDI-Reduction percentage for the drugs with higher wedge absolute frequency (F) for each treatment. The DDI-Reduction percentage was evaluated in four interaction checker tools on May 2nd, 2022, Liverpool COVID-19 Interactions,[43] WebMD,[44] Medscape,[45] and Drugbank.[46] The validation was done on the versions of Liverpool COVID-19 Interactions and Drugbank which correspond to 2022-04-13 and 2022-01-04, respectively. DDI-Reduction percentage is measured, and it indicates how many DDIs are avoided in a treatment when the middle-vertex drug is withdrawn. The evaluation suggests that withdrawing the middle-vertex with higher absolute frequency reduces most interactions. Thus, wedges and their middle-vertex represent DDIs that affect treatment effectiveness and toxicities. When more than one drug contains the higher wedge absolute frequency (F) in treatment, the clinicians have to decide which drug is withdrawn. The first COVID-19 treatment reported contains concomitant drugs for the comorbidities of asthma, high cholesterol, and pneumonia. The method proposed by [14] indicates Azithromycin as the drug with the highest absolute frequency of being the wedges middle-vertex. Therefore, it represents the DDIs that affect treatment effectiveness and toxicities, and withdrawing it reduces most interactions.

7. Related work

Data Ecosystems and Spaces The International Data Space (IDS) [38] exemplifies DEs where various W3C standards, technologies, and governance models allow for the description of the data sources to secure and standardize data exchange and integration. Data ecosystems provide reference architectures that comprise components to enable the description of the data sources to be exchanged and mappings between data sources with integrated views or unified schemas. Specifically, the networks of knowledge-driven data ecosystems (by Geisler and Vidal, et al. [6]) enable the nested definition of data ecosystems in terms of other data ecosystems whose connections induce a network. Metadata of each data ecosystem is described using controlled vocabularies and domain ontologies. Additionally, services are part of data ecosystems and can exploit metadata

[42] https://www.covid19treatmentguidelines.nih.gov/therapies/antiviral-therapy/chloroquine-or-hydroxychloroquine-and-or-azithromycin/.

[43] https://www.covid19-druginteractions.org/checker.

[44] https://www.webmd.com/interaction-checker/default.htm.

[45] https://reference.medscape.com/drug-interactionchecker.

[46] https://go.drugbank.com/drug-interaction-checker.

to enhance interoperability, traceability, data quality assessment, and integrity constraint validation. The Knowledge4COVID-19 data ecosystem implements the reference architecture proposed by Geisler and Vidal, et al. [6]; it comprises the data ecosystem for Scientific Open Data and Scientific Publications. Mappings between data sources and the unified schema are defined using RML, and the execution of these mappings results in the materialized Knowledge4COVID-19 KG. Services for knowledge extraction and prediction are implemented at each data ecosystem. A deductive system, developed on top of the Knowledge4COVID-19 KG, facilitates discovering new drug–drug interactions and their effects on treatment toxicities and effectiveness.

Knowledge Graphs have gained momentum as data structures able to model the convergence of data and knowledge as factual statements [39]. Despite being coined by the research community for several decades, KGs are playing an increasingly relevant role in scientific and industrial areas [40]. The research community has actively contributed to the problem of automatic knowledge graph creation. As a result, declarative specification of a KG creation process [25,41,42], techniques for semantic data integration [43,44], and virtualization of KGs [45,46] enable to merge data silos and provide an integrated view of data and metadata. Existing KG construction methods vary from crowdsourced (e.g., Wikidata [47]), extraction from existing knowledge bases (e.g., DBpedia [48] and YAGO [49]), and automatic generation (e.g., KnowledgeVault [50] and AI-KG [51]). Moreover, KG refinement includes methods for predicting relations, completing type assertions, and finding erroneous relations, external links, and values [52]. The creation process of Knowledge4COVID-19 KG is declarative using RML mapping rules, facilitating, thus, extensibility, modularity, and reusability of the KG creation process.

Knowledge Graphs and COVID-19: Several authors have proposed using knowledge graphs to make available expressive sources of data and knowledge about COVID-19. Specifically, several knowledge bases have been developed to provide an integrated view of COVID-19-related data. Exemplar approaches include COVID-19 Knowledge Graph,[47] Drugs4Covid [53]. Similarly, Knowledge4COVID-19 integrates CORD-19 scientific publications, but in addition, it models a fine-grained representation of drug–drug interactions and their adverse effects in the treatments of comorbidities. Additionally, Queralt-Rosinach et al. [54] present a knowledge graph that integrates clinical data collected in the context of the BEAT-COVID project.[48] These approaches put in perspective the protagonist role of knowledge graphs in understanding COVID-19. Similarly, Knowledge4COVID-19 aims to provide a resource that clinicians and patients can explore to understand the effects of interactions in a COVID-19. Thus, given the impact that pre-existing conditions seem to have on the outcome of a SARS-CoV-2 infection, Knowledge4COVID-19 represents a resource that can be linked to existing COVID-19 knowledge graphs to empower their analytical capacity.

Chatterjee et al. [55] present an exploratory review of recent works constructing knowledge graphs from different sources. For instance, Wang et al. [56] have produced a literature knowledge graph construction and drug repurposing approach, also working on the fine-grained text entity extraction, while more recently authors in [57] also construct a knowledge graph from scientific literature, focusing on cause-and-effect relations. Knowledge4COVID-19 follows the best practices of FAIR [58] and Linked Data principles,[49] and makes available a KG that integrates COVID-19 related data from various sources. UMLS is used to annotate biomedical entities; links to KGs (e.g., DBpedia and Wikidata) enhance interoperability.

Reese et al. [59] describe a knowledge graph for COVID-19 where biomedical concepts and publications are represented at symbolic and subsymbolic levels. Complementary, Knowledge4COVID-19 provides a fine-grained representation of biomedical concepts and publications. Well-known tools like MetaMap and SemRep are used to extract relevant biomedical entities and relations from scientific publications. At the same time, drug indications, side effects, and adverse events of drug–drug interactions are recognized by FALCON2.0. The extracted entities are linked to equivalent resources in existing KGs (i.e., DBpedia, Wikidata, DrugBank, and Uniprot) and annotated using UMLS terms and relations; networks of drug–drug, drug–target, and drug–side effect interactions predicted using diverse methods are also merged. This makes the Knowledge4COVID-19 KG a complementary source of knowledge that can be connected to existing COVID-19 KGs (e.g., the one implemented by Reese et al.) using the linking techniques implemented by FALCON2.0.

8. Knowledge4COVID-19 as a resource

8.1. Discussion of the Knowledge4COVID-19 framework

This section describes our resources and discusses our contributions:

The Knowledge4COVID-19 DE integrates data sources from the Scientific Open Data and Publications DEs. The pipeline for KG creation and management is available as a Docker container. It includes the Knowledge4COVID-19 unified schema, the RML triple maps, and the data sources processed by the NLP tools implemented by Scientific Open Data and Publications DEs. In addition, to create the Knowledge4COVID-19 KG, the pipeline uploads the KG to a Virtuoso endpoint and makes each resource available in the Knowledge4COVID-19 KG, following the Link Data principles. Moreover, the required configurations of the federated query engine DeTrusty are generated. DeTrusty is also available via its HTTP API like a regular SPARQL endpoint.

The Knowledge4COVID-19 KG comprises COVID-19 related data about drugs, DDIs (predicted and known), scientific publications, drugs' side effects, and interactions with targets. The KG can be explored through three APIs, a SPARQL endpoint, and a federated query engine.

DDI Prediction Methods employ machine learning techniques to identify previously unknown potential COVID-19 related drug–drug interactions with a certain confidence score. Predicted DDIs are not documented in open drug databases, such as Drugbank, and clinicians can use them as an indication of possible toxicities, during the treatment of a patient suffering from COVID-19.

Benchmarks of DDIs include known, deduced, and predicted DDIs. The known DDIs are extracted from DrugBank, while CRD, NCRD, and DeducedDDIs are deduced. Finally, a set of DDIs predicted by state-of-the-art machine learning methods is also part of the Knowledge4COVID-19 KG. These DDIs can be used to reproduce our reported results or for future comparisons.

8.2. The Knowledge4COVID-19 resource characteristics

Novelty: Knowledge4COVID-19 introduces a novel infrastructure to transform heterogeneous data sources into a KG. The mappings among the data sources and the unified schema are defined as

[47] https://covidgraph.org/.

[48] https://www.izi.fraunhofer.de/en/about-us/united-against-corona/beat-covid.html.

[49] https://www.w3.org/TR/ld-bp/.

RML mapping assertions. Moreover, the methods implemented in SDM-RDFizer allow for the efficient execution of the KG creation process. The Knowledge4COVID-19 KG occupies 13 GB and is created from 2.8 GB of raw data. Knowledge4COVID-19 KG executes 57 RML triples maps (comprising 223 mapping assertions) over the raw data in 82 min 55 s. Additionally, novel prediction methods are utilized to predict interactions between drugs. We hope that these results encourage the community to create declarative pipelines for KG creation that are able to scale up to the avalanche of data expected in the next years.

Availability: Knowledge4COVID-19 is released publicly by the Scientific Data Management (SDM) group at TIB, Hannover. TIB is one of the largest libraries for science and technology in the world. Following its policy of engaging open access to scientific artifacts, it will support Knowledge4COVID- 19 as a source of knowledge for SARS-CoV-2 and other viruses. The Knowledge4-COVID-19 DE is open source, written in Python 3, and uses RML; it is available under the Apache License 2.0 license in an open GitHub repository[4]. It will be regularly updated with new data sources, triples maps, and APIs for exploration. More importantly, respecting open science good practices, Knowledge4COVID-19 is registered at Zenodo. Thus, users can use and cite a specific version, ensuring reproducibility and traceability of any experimental evaluation.

Utility: A docker image of Knowledge4COVID-19 is available at;[50] it enables accessing the KG locally. The GitHub repository of the resource provides a detailed explanation of how to run the Docker container. From 24 to 26 April 2020, Knowledge4COVID-19 participated in the Pan-European hackathon #EUvsVirus[51] organized with the aim of connecting experts, investors, and civilian organizations to devise together innovative solutions to the coronavirus outbreak.[52]

Predicted Impact: Open pharmaceutical databases such as Drugbank or drugs.com are periodically updated, manually adding drug–drug interactions, since new unknown DDIs are frequently reported by clinicians and health institutions. Our methods can potentially deduce or predict such interactions for new or experimental drugs by analyzing contextual information in biomedical publications before being observed in practice and documented. This could support treatment decision-making, avoiding unnecessary side effects of drug combinations. Moreover, given the number of scientific publications and open data about drugs, disorders, and adverse events integrated into the Knowledge4COVID-19 KG, we are optimistic that it will be the starting point of future developments and benchmarking in the Semantic Web community. Lastly, the pipeline for KG management is domain agnostic, but there are still many opportunities to make it fully transparent. We hope this paper encourages the community to develop traceable and interpretable methods for transparent KG management.

Adoption and Reusability: We are reusing the same DE and core concepts of the unified schema and mapping rules in projects like EU H2020 projects like iASiS,[53] BigMedilytics - lung cancer pilot,[54] and P4-LUCAT.[55] As in Knowlege4COVID-19, the generated KGs include fine-grained representations of publications, and biomedical entities (e.g., drugs, side effects, targets, and interactions); the mapping rules that defined these core concepts have been reused with minor changes. This opens the spectrum of possibilities of reusability and adoption, and puts in perspective the relevance of DEs where KG creation is defined declaratively through mapping rules.

9. Conclusions and future work

This paper addresses the problem of providing an integrated view of heterogeneous sources of COVID-19 data. Following the reference architecture of networks of knowledge-driven data ecosystem (by Geisler and Vidal et al. [6]), we presented the Knowledge4COVID-19 framework as a data ecosystem (DE) where mappings among data sources and a unified schema are described in terms of RML. The Knowledge4COVID-19 DE uses the SDM-RDFizer to execute the RML mappings and create the Knowledge4COVID-19 KG. Tasks of Natural Language Processing enable recognizing relevant entities and predicates in the text describing drug–drug interactions and side effects. Additionally, a deductive system and KG predictive models allow the discovery and prediction of patterns to explain the impact of drug–drug interactions on treatment effectiveness and toxicity. As a result, the Knowledge4COVID-19 KG comprises factual statements about drugs, adverse events, and drug–drug interactions harvested from COVID-19 data sources and scientific publications.

A repertoire of Web APIs over the Knowledge4COVID-19 KG is made available. They enable exploring entities through their connections and discovering associations to enhance understanding of a SARS-CoV-2 infection and its progression. Thus, Knowledge4COVID-19 broadens the portfolio of semantic web technologies and provides the basis for developing interpretable analytical methods. In the future, we plan to connect the Knowledge4COVID-19 KG to other KGs that maintain COVID-19 related data. Additionally, we would like to extend the KG clinical data about COVID-19 patients and empower Knowledge4COVID-19 DE with the capacity of detecting patterns that can explain the correlation between survival, drug interactions, and adverse events.

CRediT authorship contribution statement

Ahmad Sakor: Worked on the Scientific Open Data Ecosystem, as well as the Knowledge4COVID-19 knowledge graph creation pipeline and exploration, Resources, Investigation, Data Curation, Writing – review & editing. **Samaneh Jozashoori:** Worked on the Scientific Open Data Ecosystem, as well as the Knowledge4COVID-19 knowledge graph creation pipeline and exploration, Resources, Investigation, Data curation, Writing – review & editing. **Emetis Niazmand:** Worked on the Scientific Open Data Ecosystem, as well as the Knowledge4COVID-19 knowledge graph creation pipeline and exploration, Computational analysis, Validation, Writing – review & editing. **Ariam Rivas:** Worked on the Scientific Open Data Ecosystem, as well as the Knowledge4COVID-19 knowledge graph creation pipeline and exploration, Computational analysis, Validation, Writing – review & editing. **Konstantinos Bougiatiotis:** Worked on the Scientific Publications Data Ecosystem creation and evaluation, Resources, Investigation, Data curation, Computational analysis, Validation, Writing – review & editing. **Fotis Aisopos:** Worked on the Scientific Publications Data Ecosystem creation and evaluation, Resources, Investigation, Data curation, Computational analysis, Validation, Writing – review & editing. **Enrique Iglesias:** Worked on the Scientific Open Data Ecosystem, as well as the Knowledge4COVID-19 knowledge graph creation pipeline and exploration, Computational analysis, Validation, Writing – review & editing. **Philipp D. Rohde:** Worked on the Scientific Open Data Ecosystem, as well as the Knowledge4COVID-19

[50] https://github.com/SDM-TIB/Knowledge4COVID-19/wiki/Running-Knowledge4COVID-19-KG-locally.

[51] https://www.euvsvirus.org/.

[52] https://devpost.com/software/COVID-19-kg.

[53] http://project-iasis.eu/.

[54] https://www.bigmedilytics.eu/.

[55] https://p4-lucat.eu/.

knowledge graph creation pipeline and exploration, Computational analysis, Validation, Writing – review & editing. **Trupti Padiya:** Worked on the Scientific Open Data Ecosystem, as well as the Knowledge4COVID-19 knowledge graph creation pipeline and exploration, Computational analysis, Validation, Writing – review & editing. **Anastasia Krithara:** Worked on the Scientific Publications Data Ecosystem creation and evaluation, Writing – review & editing. **Georgios Paliouras:** Worked on the Scientific Publications Data Ecosystem creation and evaluation, Supervision, Conceptualization, Methodology, Writing – review & editing. **Maria-Esther Vidal:** Worked on the Scientific Open Data Ecosystem, as well as the Knowledge4COVID-19 knowledge graph creation pipeline and exploration, Supervision, Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

An open API has been shared, providing access to the data of the Knowledge4COVID19 knowledge graph. Query examples have been provided in the project github repository.

Acknowledgments

This work has been supported by the EU H2020 RIA funded projects iASiS with grant agreement No. 727658, P4-LUCAT with GA No. 53000015, BigMedilytics with GA No. 780495, CLARIFY with GA No. 875160, and Federal Ministry for Economic Affairs and Energy of Germany in the project CoyPu (No 01MK21007[A-L]). Maria-Esther Vidal is also partially supported by Leibniz Association in the program "Leibniz Best Minds: Programme for Women Professors", project TrustKG-Transforming Data in Trustable Insights with grant P99/2020. Knowledge4COVID-19 was originally a project of the Pan-European hackathon #EUvsVirus in April 2020.

Appendix. KG exploration API queries

All the following queries are available on our GitHub repository.[56]

A.1. Publications related to drugs

```
PREFIX k4covid: <http://research.tib.eu/covid—19/vocab/>
PREFIX k4covide: <http://research.tib.eu/covid—19/entity/>
SELECT DISTINCT ?pub ?year ?journal ?title ?url ?drug ?drugLabel where {
        ?drug a k4covid:Drug.
        ?drug k4covid:hasCUIAnnotation ?drugCUI.
        Filter(?drugCUI in (k4covide:C0031623,
        k4covide:C0751995,
        k4covide:C0030106))
        ?drugCUI k4covid:annLabel ?drugLabel.
        ?ann a k4covid:ConceptAnnotation.
        ?ann k4covid:hasSemanticAnnotation ?semAnn.
        ?semAnn k4covid:hasCUIAnnotation ?drugCUI.
        ?ann k4covid:annotates ?pub.
        ?pub <http://purl.org/dc/terms/title> ?title.
        ?pub k4covid:year ?year.
        ?pub k4covid:journal ?journal.
        ?pub k4covid:externalLink ?url.
    }
```

A.2. Drug-drug interactions (DDI)

A.2.1. Get interactions of a drug

```
PREFIX k4covid: <http://research.tib.eu/covid—19/vocab/>
PREFIX k4covide: <http://research.tib.eu/covid—19/entity/>
SELECT DISTINCT ?effectorDrugLabel ?affectdDrugLabel ?effect AS ?effectLabel
?impactLabel   WHERE {
    ?interaction k4covid:precipitantDrug ?effectorDrugCUI.
    ?interaction k4covid:objectDrug ?affectdDrugCUI.
    ?effectorDrugCUI k4covid:annLabel ?effectorDrugLabel.
    ?affectdDrugCUI k4covid:annLabel ?affectdDrugLabel.
    ?interaction k4covid:effect ?effectCUI.
    ?effectCUI k4covid:annLabel  ?effect.
    ?interaction k4covid:impact ?impactLabel.
FILTER(?affectdDrugCUI in (k4covide:C0000970))}
```

A.2.2. Get all the interaction among the provided drugs

```
PREFIX k4covid: <http://research.tib.eu/covid—19/vocab/>
PREFIX k4covide: <http://research.tib.eu/covid—19/entity/>
SELECT * {
{SELECT DISTINCT ?effectorDrugLabel ?affectdDrugLabel ?effect AS ?effectLabel ?impactLabel
WHERE {
    ?interaction k4covid:precipitantDrug k4covide:C0000970.
    ?interaction k4covid:objectDrug k4covide:C0028978.
    k4covide:C0000970 k4covid:annLabel ?effectorDrugLabel.
    k4covide:C0028978 k4covid:annLabel ?affectdDrugLabel.
    ?interaction k4covid:effect ?effectCUI.
    ?effectCUI k4covid:annLabel ?effect.
    ?interaction k4covid:impact ?impactLabel.
}} UNION
{SELECT DISTINCT ?effectorDrugLabel ?affectdDrugLabel ?effect AS ?effectLabel ?impactLabel
WHERE {
    ?interaction k4covid:precipitantDrug k4covide:C0028978.
    ?interaction k4covid:objectDrug k4covide:C0000970.
    k4covide:C0028978 k4covid:annLabel ?effectorDrugLabel.
    k4covide:C0000970 k4covid:annLabel ?affectdDrugLabel.
    ?interaction k4covid:effect ?effectCUI.
    ?effectCUI k4covid:annLabel ?effect.
    ?interaction k4covid:impact ?impactLabel.
}}}
```

A.3. Predicted drug-drug interactions

A.3.1. Get the predicted interactions of a drug

```
PREFIX k4covid: <http://research.tib.eu/covid—19/vocab/>
PREFIX k4covide: <http://research.tib.eu/covid—19/entity/>
SELECT DISTINCT ?effectorDrugLabel ?affectdDrugLabel ?confidence ?provenance WHERE {
    ?interaction a k4covid:DrugDrugPrediction.
    ?interaction k4covid:hasInteractingDrug ?effectorDrug.
    ?interaction k4covid:hasInteractingDrug ?affectedDrug.
    FILTER (?effectorDrug != ?affectedDrug)
    ?affectedDrug k4covid:hasCUIAnnotation ?affectdDrugCUI.
    ?effectorDrug k4covid:hasCUIAnnotation ?effectorDrugCUI.
    ?effectorDrugCUI k4covid:annLabel ?effectorDrugLabel.
    ?affectdDrugCUI k4covid:annLabel ?affectdDrugLabel.
    ?interaction k4covid:confidence ?confidence.
    ?interaction k4covid:predictionMethod ?provenance.
FILTER(?affectdDrugCUI in (k4covide:C0000970))}
```

A.3.2. Get all the interaction among the provided drugs

```
PREFIX k4covid: <http://research.tib.eu/covid—19/vocab/>
PREFIX k4covide: <http://research.tib.eu/covid—19/entity/>
SELECT * {
{SELECT DISTINCT ?effectorDrugLabel ?affectdDrugLabel ?confidence ?provenance  WHERE {
    ?interaction k4covid:hasInteractingDrug ?effectorDrug.
    ?interaction k4covid:hasInteractingDrug ?affectedDrug.
    FILTER (?effectorDrug != ?affectedDrug)
    ?effectorDrug k4covid:hasCUIAnnotation k4covide:C0995188.
    ?affectedDrug k4covid:hasCUIAnnotation k4covide:C0765273.
    k4covide:C0000970 k4covid:annlabel ?effectorDrugLabel.
    k4covide:C0009214 k4covid:annlabel ?affectdDrugLabel.
    ?interaction k4covid:confidence ?confidence.
    ?interaction k4covid:predictionMethod ?provenance.
}} UNION
{SELECT DISTINCT ?effectorDrugLabel ?affectdDrugLabel ?confidence ?provenance  WHERE {
    ?interaction k4covid:hasInteractingDrug ?effectorDrug.
    ?interaction k4covid:hasInteractingDrug ?affectedDrug.
    FILTER (?effectorDrug != ?affectedDrug)
    ?effectorDrug k4covid:hasCUIAnnotation k4covide:C0765273.
    ?affectedDrug k4covid:hasCUIAnnotation k4covide:C0995188.
    k4covide:C0009214 k4covid:annLabel ?effectorDrugLabel.
    k4covide:C0000970 k4covid:annLabel ?affectdDrugLabel.
    ?interaction k4covid:confidence ?confidence.
    ?interaction k4covid:predictionMethod ?provenance.
}}}
```

[56] https://github.com/SDM-TIB/Knowledge4COVID-19/blob/main/Exploration-API/SPARQL/README.md.

References

[1] A. Dimou, M.V. Sande, P. Colpaert, R. Verborgh, E. Mannens, R.V. de Walle, RML: A generic language for integrated RDF mappings of heterogeneous data, in: Workshop on Linked Data on the Web Co-Located with WWW, 2014.

[2] C. Capiello, A. Gal, M. Jarke, J. Rehof, Data ecosystems: Sovereign data exchange among organizations (Dagstuhl seminar 19391), Dagstuhl Rep. (2020) http://dx.doi.org/10.4230/DagRep.9.9.66.

[3] 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 Comput. Surv. 54 (4) (2021) http://dx.doi.org/10.1145/3447772.

[4] M. Namici, G. De Giacomo, Comparing query answering in OBDA tools over W3C-compliant specifications, in: Description Logics, 2018.

[5] M.I.S. Oliveira, G.d.F.B. Lima, B.F. Lóscio, Investigations into data ecosystems: a systematic mapping study, Knowl. Inf. Syst. 61 (2) (2019) 589–630.

[6] S. Geisler, M. Vidal, C. Cappiello, B.F. Lóscio, A. Gal, M. Jarke, M. Lenzerini, P. Missier, B. Otto, E. Paja, B. Pernici, J. Rehof, Knowledge-driven data ecosystems towards data transparency, Spec. Issue Data Transpar. ACM J. Data Inf. Qual. (2022).

[7] D. Wishart, Y. Feunang, A. Guo, E. Lo, A. Marcu, J. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucl. Acids Res. (2018) http://dx.doi.org/10.1093/nar/gkx1037.

[8] M. Kuhn, I. Letunic, L. Jensen, P. Bork, The SIDER database of drugs and side effects, Nucl. Acids Res. (2015) http://dx.doi.org/10.1093/nar/gkx1037.

[9] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucl. Acids Res. (2004) http://dx.doi.org/10.1093/nar/gkh061.

[10] D. Sridhar, S. Fakhraei, L. Getoor, A probabilistic approach for collective similarity-based drug-drug interaction prediction, Bioinform. 32 (20) (2016) 3175–3182.

[11] A. Sakor, K. Singh, A. Patel, M.-E. Vidal, Falcon 2.0: An entity and relation linking tool over wikidata, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 3141–3148.

[12] A. Sakor, I.O. Mulang, K. Singh, S. Shekarpour, M.E. Vidal, J. Lehmann, S. Auer, Old is gold: linguistic driven approach for entity and relation linking of short text, in: Proceedings of the 2019 NAACL HLT (Long Papers), 2019, pp. 2336–2346.

[13] S. Ceri, G. Gottlob, L. Tanca, What you always wanted to know about datalog (and never dared to ask), IEEE Trans. Knowl. Data Eng. 1 (1) (1989) 146–166.

[14] A. Rivas, M. Vidal, Capturing knowledge about drug-drug interactions to enhance treatment effectiveness, in: K-CAP '21: Knowledge Capture Conference, Virtual Event, USA, December 2-3, 2021, 2021, pp. 33–40.

[15] Y. Yang, Y. Fang, M.E. Orlowska, W. Zhang, X. Lin, Efficient bi-triangle counting for large bipartite networks, Proc. VLDB Endow. 14 (6) (2021) 984–996, http://dx.doi.org/10.14778/3447689.3447702.

[16] L.L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. Wade, K. Wang, N.X.R. Wang, C. Wilhelm, B. Xie, D. Raymond, D.S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: The Covid-19 open research dataset, 2020, http://dx.doi.org/10.48550/ARXIV.2004.10706, ArXiv Preprint.

[17] H. Kilicoglu, G. Rosemblat, M. Fiszman, D. Shin, Broad-coverage biomedical relation extraction with SemRep, BMC Bioinformatics 21 (2020) 1–28.

[18] J. Yuan, Z. Jin, H. Guo, H. Jin, X. Zhang, T. Smith, J. Luo, Constructing biomedical domain-specific knowledge graph with minimum supervision, Knowl. Inf. Syst. 62 (1) (2020) 317–336.

[19] D. Zhang, D. He, N. Zou, X. Zhou, F. Pei, Automatic relationship verification in online medical knowledge base: a large scale study in SemMedDB, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2018, pp. 1673–1680.

[20] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, Adv. Neural Inf. Process. Syst. 26 (2013).

[21] A. Melo, H. Paulheim, Detection of relation assertion errors in knowledge graphs, in: Proceedings of the Knowledge Capture Conference, 2017, pp. 1–8.

[22] K. Bougiatiotis, R. Fasoulis, F. Aisopos, A. Nentidis, G. Paliouras, Guiding graph embeddings using path-ranking methods for error detection in noisy knowledge graphs, 2021, arXiv preprint arXiv:2002.08762.

[23] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, iASiS open data graph: Automated semantic integration of disease-specific knowledge, in: International Symposium on Computer-Based Medical Systems, CBMS, 2020.

[24] K. Bougiatiotis, F. Aisopos, A. Nentidis, A. Krithara, G. Paliouras, Drug-drug interaction prediction on a biomedical literature knowledge graph, in: International Conference on Artificial Intelligence in Medicine, 2020.

[25] E. Iglesias, S. Jozashoori, D. Chaves-Fraga, D. Collarana, M.-E. Vidal, SDM-RDFizer: An RML interpreter for the efficient creation of rdf knowledge graphs, in: ACM International Conference on Information & Knowledge Management, 2020.

[26] P.D. Rohde, DeTrusty v0.4.3, 2022, http://dx.doi.org/10.5281/zenodo.6570166.

[27] M. Lenzerini, Data integration: A theoretical perspective, in: ACM Symposium on Principles of Database Systems, 2002.

[28] L. Halilaj, N. Petersen, I. Grangel-González, C. Lange, S. Auer, G. Coskun, S. Lohmann, VoCol: An integrated environment to support version-controlled vocabulary development, in: Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW, 2016.

[29] I. Djaharuddin, S. Munawwarah, A. Nurulita, M. Ilyas, N.A. Tabri, N. Lihawa, Comorbidities and mortality in COVID-19 patients, Gac. Sanit. 35 (2021) http://dx.doi.org/10.1016/j.gaceta.2021.10.085.

[30] A. Bordes, N. Usunier, A. García-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: C.J.C. Burges, L. Bottou, Z. Ghahramani, K.Q. Weinberger (Eds.), Conference on Neural Information Processing Systems, 2013, pp. 2787–2795.

[31] M. Nickel, V. Tresp, H. Kriegel, A three-way model for collective learning on multi-relational data, in: L. Getoor, T. Scheffer (Eds.), Proceedings of the 28th International Conference on Machine Learning, 2011.

[32] M. Nickel, L. Rosasco, T.A. Poggio, Holographic embeddings of knowledge graphs, in: D. Schuurmans, M.P. Wellman (Eds.), Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Press, 2016, pp. 1955–1961.

[33] B. Yang, W.-t. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, 2014, http://dx.doi.org/10.48550/ARXIV.1412.6575, URL https://arxiv.org/abs/1412.6575.

[34] H. Ejaz, A. Alsrhani, A. Zafar, H. Javed, K. Junaid, A.E. Abdalla, K.O. Abosalif, Z. Ahmed, S. Younas, COVID-19 and comorbidities: Deleterious impact on infected patients, J. Infect. Public Health (2020).

[35] M. Vidal, E. Ruckhaus, T. Lampo, A. Martínez, J. Sierra, A. Polleres, Efficiently joining group patterns in SPARQL queries, in: The Semantic Web: Research and Applications. ESWC 2010, Springer, Berlin, Heidelberg, 2010, pp. 228–242, http://dx.doi.org/10.1007/978-3-642-13486-9_16.

[36] K.M. Endris, M. Galkin, I. Lytra, M.N. Mami, M.-E. Vidal, S. Auer, Querying interlinked data by bridging RDF molecule templates, in: Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXIX, Vol. 11310, Springer, Berlin, Heidelberg, 2018, pp. 1–42, http://dx.doi.org/10.1007/978-3-662-58415-6_1.

[37] J. Hakkola, J. Hukkanen, M. Turpeinen, et al., Inhibition and induction of CYP enzymes in humans: an update, Arch. Toxicol. 94 (2020) 3671–3722, http://dx.doi.org/10.1007/s00204-020-02936-7.

[38] S.R. Bader, J. Pullmann, C. Mader, S. Tramp, C. Quix, A.W. Müller, H. Akyürek, M. Böckmann, B.T. Imbusch, J. Lipp, S. Geisler, C. Lange, The international data spaces information model - An ontology for sovereign exchange of digital content, in: The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II, 2020, pp. 176–192.

[39] C. Gutiérrez, J.F. Sequeda, Knowledge graphs, Commun. ACM 64 (3) (2021) 96–104.

[40] N.F. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, J. Taylor, Industry-scale knowledge graphs: lessons and challenges, Commun. ACM 62 (8) (2019) 36–43.

[41] S. Jozashoori, D. Chaves-Fraga, E. Iglesias, M. Vidal, Ó. Corcho, FunMap: Efficient execution of functional mappings for knowledge graph creation, in: The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, 2020, pp. 276–293.

[42] D. Chaves-Fraga, E. Ruckhaus, F. Priyatna, M. Vidal, Ó. Corcho, Enhancing virtual ontology based access over tabular data with Morph-CSV, Semant. Web 12 (6) (2021) 869–902.

[43] P. Cudré-Mauroux, Leveraging knowledge graphs for big data integration: the XI pipeline, Semant. Web 11 (1) (2020) 13–17.

[44] M. Vidal, K. Endris, S. Jazashoori, A. Sakor, A. Rivas, Transforming heterogeneous data into knowledge for personalized treatments - A use case, Datenbank-Spektrum 19 (2) (2019) 95–106.

[45] G. Xiao, L. Ding, B. Cogrel, D. Calvanese, Virtual knowledge graphs: An overview of systems and use cases, Data Intell. 1 (3) (2019) 201–223.

[46] K.M. Endris, P.D. Rohde, M. Vidal, S. Auer, Ontario: Federated query processing against a semantic data lake, in: Database and Expert Systems Applications - 30th International Conference, DEXA 2019, Linz, Austria, August 26-29, 2019, Proceedings, Part I, 2019, pp. 379–395.

[47] D. Vrandecic, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85.

[48] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: A nucleus for a web of open data, in: Proceedings of ISWC + ASWC, 2007, pp. 722–735.

[49] J. Hoffart, F. Suchanek, K. Berberich, G. Weikum, YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, Artificial Intelligence 194 (2013) 28–61.

[50] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, W. Zhang, Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion, KDD '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 601–610, http://dx.doi.org/10.1145/2623330.2623623.

[51] D. Dessì, F. Osborne, D.R. Recupero, D. Buscaldi, E. Motta, H. Sack, AI-KG: An automatically generated knowledge graph of artificial intelligence, in: International Semantic Web Conference, 2020, pp. 127–143.

[52] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, Semant. Web 8 (3) (2017) 489–508.

[53] C. Badenes-Olmedo, et al., Drugs4Covid: Drug-driven knowledge exploitation based on scientific publications, 2020, CoRR abs/2012.01953.

[54] N. Queralt-Rosinach, R. Kaliyaperumal, C. Bernabé, et al., Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic, J. Biomed. Semant. 13 (12) (2022) http://dx.doi.org/10.1186/s13326-022-00263-7.

[55] A. Chatterjee, C. Nardi, C. Oberije, P. Lambin, Knowledge graphs for COVID-19: An exploratory review of the current landscape, J. Pers. Med. 11 (4) (2021) 300.

[56] Q. Wang, M. Li, X. Wang, N. Parulian, G. Han, J. Ma, J. Tu, Y. Lin, H. Zhang, W. Liu, et al., COVID-19 literature knowledge graph construction and drug repurposing report generation, 2020, arXiv preprint arXiv:2007.00576.

[57] D. Domingo-Fernández, S. Baksi, B. Schultz, Y. Gadiya, R. Karki, T. Raschka, C. Ebeling, M. Hofmann-Apitius, A.T. Kodamullil, COVID-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology, Bioinformatics 37 (9) (2021) 1332–1334.

[58] M. Wilkinson, et al., The FAIR guiding principles for scientific data management and stewardship, Sci. Data 3 (2016) http://dx.doi.org/10.1038/sdata.2016.18.

[59] J.T. Reese, D.R. Unni, T.J. Callahan, L. Cappelletti, V. Ravanmehr, S. Carbon, K.A. Shefchek, B.M. Good, J.P. Balhoff, T. Fontana, H. Blau, N. Matentzoglu, N.L. Harris, M.C. Munoz-Torres, M.A. Haendel, P.N. Robinson, M.P. Joachimiak, C.J. Mungall, KG-COVID-19: A framework to produce customized knowledge graphs for COVID-19 response, Patterns 2 (1) (2021) 100155, http://dx.doi.org/10.1016/j.patter.2020.100155.