IMI2 Project 101005122 - DRAGON

The RapiD and SecuRe AI enhAnced DiaGnosis, Precision Medicine and Patient EmpOwerment Centered Decision Support System for Coronavirus PaNdemics

WP6 – Multifactorial analysis

D6.2 Initial knowledge graph based upon existing knowledge

Lead contributor	1. Maastricht University
Other contributors	-
Deliverable submission date	26-03-2021
Deliverable type	Report
Dissemination level	PUBLIC

Abstract

Description of the deliverable; A knowledge graph produced by systematic review of existing knowledge and the landscape of new therapeutic development that can be used to support decision making and the explainability of the analysis.

Searching through the COVID-19 research literature to gain actionable clinical insight is a formidable task, even for experts. The usefulness of this corpus in terms of improving patient care is tied to the ability to see the big picture that emerges when the studies are seen in conjunction rather than in isolation. When the answer to a search query requires linking together multiple pieces of information across documents, simple keyword searches are insufficient. To answer such complex information needs, an innovative AI technology called a knowledge graph (KG) could prove to be effective.

Introduction

Due to the truly global nature of the COVID-19 pandemic, there has been an explosion of academic literature on this subject in 2020. Faced with a mountain of data, we often turn to machines for analysis. What we are really after is extracting knowledge from this data, and despite their prodigious computational power, machines are unable to understand (let alone answer) our complex questions. Even for the most basic questions, the search has to often be reduced to keywords, which reduces the sophistication of the query (loss of semantics). A human easily knows the difference between “Which drugs reduce the severity of COVID-19” and “Which drugs increase the severity of COVID-19”, but for a purely keyword-based search with no semantics, this difference cannot be conveyed to a machine. Now imagine a question like "Which are the top 3 drugs being trialed for treating COVID-19 in terms of total number of enrolled patients" or "For which clinical outcome is predictive modeling for COVID-19 most successful"? If a machine could answer such questions, it would significantly accelerate scientific progress by providing answers to complex questions that may today require many hours of reading, even by subject matter experts. KG is an AI innovation to bring us closer to this vision.

When the DRAGON proposal was first submitted to IMI, the partner list included Aladdin Healthcare Technologies, who were responsible for an early deliverable (M6) of “a knowledge graph produced by systematic review of existing knowledge and the landscape of new therapeutic development that can be used to support decision making and the explainability of the analysis.” During the consortium approval process, Aladdin was unable to remain within the consortium owing to financial constraints. While this meant that DRAGON was no longer privy to the exact vision of KG that Aladdin had planned, it spurred us to conduct this thorough review of the existing work related to KGs for COVID-19. This review allows us to identify the unmet clinical needs and refocus our efforts to produce graphs that actually add to the existing corpus, rather than merely duplicating efforts of other research groups.

Methods

Before delving into the methods used for this study, we would like to (1) emphasize that this is not a systematic review, but an exploratory literature review, and (2) explain the reason for making this choice. A defining feature of a systematic review is that it uses a repeatable analytical method to answer a well-defined research question. This translates to using databases like PubMed, MEDLINE, Web of Science, and clinical trial registries, and having pre-defined inclusion criteria that should ideally be formulated into a study protocol and published before the review starts. Systematic reviews are a great way of synthesizing various information sources in a mature discipline to guide evidence-based medicine. They are often meant to be an exhaustive summary of available evidence, where evidence is defined as peer-reviewed literature indexed in the databases mentioned above. A great example would be a systematic review of the clinical effectiveness of proton therapy. Since a systematic review is often meant to inform clinical practice, the inclusion criteria are much stricter than what is permissible in an exploratory review.

An exploratory review, by contrast, is not meant to follow a repeatable analytical method or be an exhaustive summary. It typically provides a broad overview of work that has been done in a certain research domain, and uses this to define the scope of future research. In other words, it is meant to define research objectives rather than change or inform clinical practice. Almost every original research paper begins with a short exploratory review of the current state-of-the-art. By its very nature, KG for COVID-19 is a nascent field of research. While peer-reviewed literature does exist in this field, there is work being done both in academia and in industry that has yet to be published in journals. Thus, using only indexing databases like PubMed is an inadequate way to capture the current extent of the research.

The aim of this review is to identify the different applications of KGs with respect to COVID-19, even if such research is not mature enough to have been published in peer-reviewed journals that are indexed by PubMed (which does not index all journals). For example, the first citation (an excellent review article) used in this paper cannot be found using PubMed, because the source (Harvard Data Science Review, published by MIT Press online) is not indexed on PubMed. Unlike PubMed, Google Scholar is not limited to clinical and biomedical journals, and includes conference proceedings, books, and reports, that are not included in Web of Science or PubMed. Google Scholar searches full text of articles but PubMed and Web of Science search only the citation,

abstract, and tagging information. The superiority of Google Scholar over PubMed with respect to the ability to retrieve relevant articles using a quick search has been studied before. The advantages of PubMed over Google Scholar, which mainly stem from PubMed using human curation, are irrelevant for this review, because the sources identified by Google Scholar are perused by us before inclusion in the results. It would be unacceptable to use Google Scholar for a systematic review because the process must be repeatable, and human judgement used for quality evaluation is subjective (and thus not repeatable). However, for an exploratory review, this does not pose a problem, and using Google Scholar allows access to a larger number of sources (sometimes referred to as 'grey literature'). An up-to-date comparison of these different search approaches from the perspective of a librarian can be found elsewhere.

The search term used for finding original sources for this review was "covid-19 knowledge graph", and the search was conducted using Google and Google Scholar. The reason for using Google in addition to Google Scholar was to identify companies or consortiums that are working in this field but have not published any literature, peer-reviewed or otherwise. The first five pages of results were considered in both these platforms. Domain expertise was used to reduce this to unique sources, which were then used to obtain the results. This reduction consisted of removing duplicates, verification of the relevance, and qualitative assessment of the rigor of publicly accessible work (whether in the form of articles or websites). The results constituted a broad overview of the field, separated the KGs for COVID-19 into clusters based on their intended use, and then briefly summarized the information pertaining to each original source.

Results

Table 1 summarizes the papers we found from our search and the main application for their KG (divided into use clusters). These use clusters (and their associated papers) are described in the rest of this section. In addition, our search also pointed us to the EU Datathon 2020 which organized two meetups of The Knowledge Graph Conference in April. The associated recordings and slides can be found in the following link. We also found the CovidGraph project (https://covidgraph.org), an interdisciplinary collaboration between academia and industry. In addition to literature data, they connected information from genes and proteins and their function, using open-source knowledge bases such as the Gene Ontology and the NCBI Gene Database. An important advantage of this project is that it uses Neo4j for modeling, storing, and exposing the KG, which considerably simplifies adoption by a large body of data scientists and app developers, as it is both powerful and intuitive. However, since there is no paper associated with this project yet, we cannot provide further detail in this review.

Table 1: Summary of papers resulting from our literature search.

Authors	Title	Application
Kejriwal	Knowledge Graphs and COVID-19: Opportunities, Challenges, and Implementation	KG overview
Steenwinckel et al	Facilitating the analysis of covid-19 literature through a knowledge graph	Literature review
Wise et al	COVID-19 knowledge graph: accelerating information retrieval and discovery for scientific literature	Literature review
Michel et al	Covid-on-the-Web: Knowledge graph and services to advance COVID-19 research	Literature review
Stebbing et al	COVID-19: combining antiviral and anti-inflammatory treatments	Drug repurposing
Wang et al	COVID-19 literature knowledge graph construction and drug repurposing report generation	Drug repurposing
Domingo-Fernandez et al	COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology	Drug repurposing
Zhou et al	Artificial intelligence in COVID-19 drug repurposing	Drug repurposing
Chen et al	Coronavirus knowledge graph: A case study	Multi-purpose
Reese et al	KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response	Multi-purpose

Knowledge graphs for literature review

We found three articles that used KGs to facilitate literature review. In the first paper by Steenwinckel et al, the Kaggle dataset of 63,000+ papers (also known as CORD-19, released to allow recent advances in NLP and other AI techniques to generate new insights to fight the pandemic) was used to create a KG. The authors started with a summary of initiatives by other research groups who are using the same dataset, identifying the CovidGraph project as the largest such initiative, which links the CORD-19 dataset to the NCBI Gene Database and other gene ontologies to enable scientific analysis. The authors then discussed the steps needed to construct their KG. In the CORD-19 dataset, information about each paper is provided in the form of a CSV file. For more than 51,000 of these papers, a JSON file is provided that contains detailed information about the authors, the content, and the other studies that were cited. The authors semantically enriched the data by mapping it to the Resource Description Framework (RDF) using the RDF Mapping Language (RML), which was convenient because the initial data was already structured (CSV and JSON). Before the conversion from JSON to RDF, the JSON. files were extended to include additional information from external resources, including DBpedia, BioPortal, CrossRef and ORCID.

To make the transformation from JSON to RDF, a mapping document was created that contained rules on how each element in the JSON can be mapped on a corresponding semantic output value. The mapping document was created with YARRRML, a human-readable text-based representation that can be used to represent RML rules. As this YARRRML document is only a human-readable text-based representation of RML rules, they converted this YARRRML document to an RML document by using the YARRRML Parser. While it is possible to write RML rules in this setup directly, by using YARRRML, they created the ability to let others extend the mapping documents with reduced human effort and without requiring extensive specific knowledge about semantic web formats. The RMLMapper takes both the extended JSON files and the RML document generated using the above YARRRML document as input and produces a set of N-Triples for each paper. All these N-Triple files were concatenated to form a single KG.

The authors discussed the applications of such a KG. One can perform network analysis by converting the KG into a regular directed graph. The conversion is needed as existing network analysis tools cannot deal with different labeled edges. The converted graph consists of nodes that represent the papers and edges between these nodes that represent citations from one paper to the other. Clustering analysis reveals information on how tightly some groups of publications are interconnected through citations. Node centrality analysis can identify publications that are influential with respect to COVID-19, rather than influential in general (for which looking at number of citations would suffice); several metrics can be used to estimate the centrality of a node. KGs cannot be directly used for machine learning. To tackle this issue, knowledge graph embeddings have been proposed, where components of a knowledge graph, including entities and relations, are embedded into continuous vector spaces. The most common technique to build such embeddings is RDF2Vec. Once converted into these vectors, it becomes easy to search for nearest neighbors, which allows one to easily find similar or related papers in a much more powerful way than a keyword search. These vectors can also be used for clustering papers, which is more powerful than the network clustering analysis previously described, which only uses citation links. The second paper, by Wise et al, also uses the CORD-19 dataset, is a demonstration of Amazon Web Services AI, and is conceptually more advanced than the first paper. Unlike the first paper, the second paper does not support the FAIR (Findable, Accessible, Interoperable, and Reusable) principle, and does not make any of its code public. However, their KG is used to power a search engine (https://www.cord19.aws/), which is available for public use. The authors provide a succinct definition of a KG: “Knowledge graphs (KGs) are structural representations of relations between real-world entities where relations are defined as triplets containing a head entity, a tail entity, and the relation type connecting them.” Their KG contains five types of entities: paper (with attributes of title, publication date, journal, and Digital Object Identifier (DOI) link), author (with attributes of first, middle, and last names), institution (with attributes of name, country, and city), concept, and topic. Figure 1 illustrates the directed property graph structure for a small subgraph of their KG.

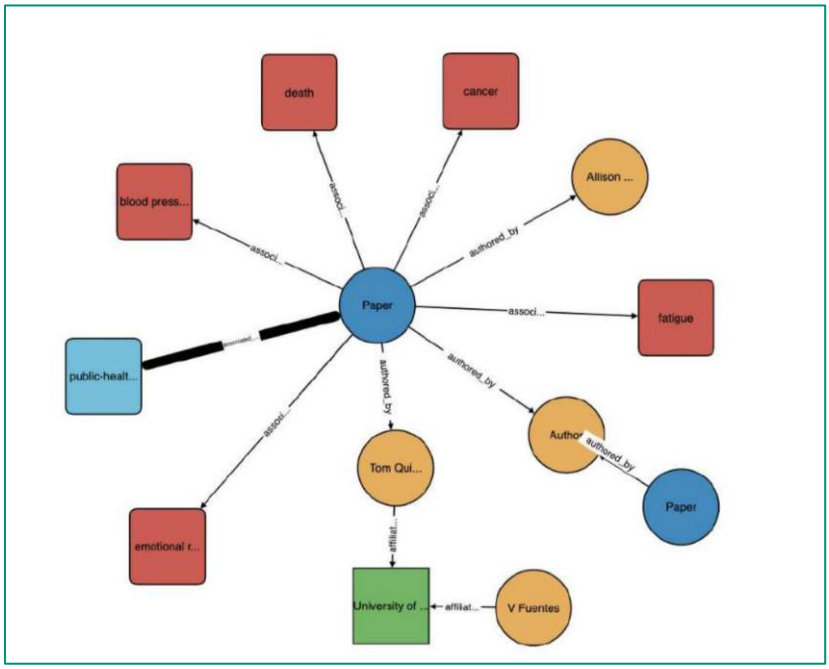

Figure 1: Visualization of KG. Paper entities (blue) connect to Concepts (red), topics (light blue), and authors (gold) through directed relations. Authors connect to institutions (green). Taken from Wise et al.

Concept entity: They used their proprietary NLP system called Comprehend Medical Detect Entities V2 for medical language entity recognition and relationship extraction. Given the example text "Abdominal ultrasound noted acute appendicitis, recommend appendectomy followed by several series of broad spectrum antibiotics", the system extracts Abdominal (Anatomy), ultrasound (Test Treatment Procedure), acute appendicitis (Medical Condition), appendectomy (Test Treatment Procedure), and antibiotics (Medication) as recognized entities along with entity types and model confidence scores. Entity names e.g. acute appendicitis, form concept entities while entity type and model confidence score are the entities' attributes. Topic entity: They defined 10 topics using expert knowledge: Vaccines/Immunology,

Genomics, Public Health Policies, Epidemiology, Clinical Treatment, Virology, Influenza, Healthcare Industry, Lab Trials (human) and Pulmonary infections. Since manually labeling a topic model is inefficient, they manually labeled only a subset of the papers and used this to train a multi-label classifier (an extension of Latent Dirichlet Allocation termed Z-LDA) using the title, abstract and body text from each paper. The resulting classifier achieved an average F1 score of 0.92 with on average 2.37 labels per document. To validate their topic model, they checked that generated topics of papers from Journal of Virology, e.g., virology, genomics, and lab-trials-human, were highly related to virology and the generated topics of papers from Journal of Vaccine, e.g., vaccines-immunology, were highly related to vaccinology.

To curate their KG, they applied data normalization techniques which eliminated duplicate entities and noisy linkages. Denoising included thresholding on the confidence scores, pruning concepts that occur in

less than 0.0001% of papers, and flagging concepts that appear in greater than 50% of papers for manual assessment. The KG was then used for two main tasks: information retrieval and article recommendations. For information retrieval, an example query "What papers discussing COVID-19 risk factors are most often cited by researchers within the CORD-19 dataset?" results in two steps: first, the articles which contain the risk factors as entities are retrieved, and then these articles are ranked based on citation counts withing the dataset. The authors combined article semantic information with KG topological information to quantify similarity between articles and construct a similarity-based recommendation system (given a paper, the engine retrieves a list of top-k most similar papers using cosine distance).

To capture semantic information, they used SciBERT that has shown strong transfer learning performance on a wide variety of NLP tasks. To capture KG topological information, they generated vector embeddings for each paper by using the algorithm TransE and Deep Graph Library Knowledge Embedding library (DGL-KE). Besides finding similar papers to a given paper, the recommendation engine can also be used to identify the most popular papers, where popularity captures the number of occurrences of an individual paper in the top-5 most similar items list for all papers in the dataset.

The third paper, by Michel et al, has grander ambitions than just literature review, as the Covid-on-the-Web Dataset created by this team can be put to other uses in the future (such as helping clinicians to get argumentative graphs to analyze clinical trials and make evidence-based decisions). We categorized Michel et al under literature review as this is what is explicitly demonstrated in their current work. The authors are strong proponents of open and reproducible science goals, and the FAIR principles. Like the previous papers mentioned in the results, they also used the CORD-19 dataset, and enriched it using DBpedia, BioPortal, and Wikidata to create the CORD-19 Named Entities Knowledge Graph. In addition, each abstract of the CORD-19 corpus was analyzed by Argumentative Clinical Trial Analysis (ACTA) and translated into RDF to yield the CORD-19 Argumentative Knowledge Graph. ACTA is a tool designed to analyze clinical trials for argumentative components and PICO (patients/population (P), intervention (I), control/comparison (C) and outcome (O)) elements. Finally, they provided several visualization and exploration tools based on the Corese Semantic Web platform (https://project.inria.fr/corese/) and MGExplorer visualization library (https://github.com/frmichel/morph-xr2rml/).

ACTA goes far beyond basic keyword-based search by retrieving the main claim(s) stated in the trial, as well as the evidence linked to this claim, and the PICO elements. In the context of clinical trials, a claim is a concluding statement made by the author about the outcome of the study. It generally describes the

relation of a new treatment (intervention arm) with respect to existing treatments (control arm). Accordingly, an observation or measurement is an evidence which supports or attacks another argument component. Observations comprise side effects and the measured outcome. Two types of relations can hold between argumentative components. The attack relation holds when one component is contradicting the proposition of the target component, or stating that the observed effects are not statistically significant. The support relation holds for all statements or observations justifying the proposition of the target component. The ACTA pipeline consists of four steps: (i) the detection of argumentative components, i.e. claims and evidence, (ii) the prediction of relations holding between these components, (iii) the extraction of PICO elements, and (iv) the production of the RDF representation of the arguments and PICO elements.

Knowledge graphs for drug repurposing

We found four articles related to using KGs for drug repurposing, which is a technique of using existing drugs to treat emerging and challenging diseases, thereby reducing development timelines and overall costs. The first article, by Stebbing et al, was published as a comment in Lancet Infectious Diseases near the beginning of the pandemic (April 1, 2020). The authors had earlier described how BenevolentAI's proprietary KG, queried by a suite of algorithms, enabled identification of baricitinib, a numb-associated kinase (NAK) inhibitor, to suppress clathrin-mediated endocytosis and thereby inhibit viral infection of cells. In this work, they re-examined the affinity and selectivity of all the approved drugs in their KG to identify those with both antiviral and anti-inflammatory properties, since the host inflammatory response becomes a major cause of lung damage and subsequent mortality for severe cases of COVID-19. This yielded three candidates: baricitinib, fedratinib, and ruxolitinib. Other AI-algorithm-predicted NAK inhibitors included a combination of the oncology drugs sunitinib and erlotinib, shown to reduce the infectivity of a wide range of viruses. However, sunitinib and erlotinib would be difficult for patients to tolerate at the doses required to inhibit NAK. Baricitinib emerged as the best choice, especially given its once-daily oral dosing and acceptable side-effect profile. In addition, the potential for combination therapy with baricitinib was high, including combining baricitinib with the direct-acting antivirals (lopinavir or ritonavir and remdesivir) currently being used in the COVID-19 outbreak to reduce viral infectivity, viral replication, and the aberrant host inflammatory response. This work demonstrates that a KG can facilitate rapid drug development. A trial of baricitinib plus remdesivir has already been conducted and was superior to remdesivir alone in reducing recovery time and accelerating improvement in clinical status.

The second article, by Wang et al, used KGs for drug repurposing report generation. For a given drug, such a report consists of 11 typical questions they identified: (1) Current indication: what is the drug class? What is it currently approved to treat? (2) Molecular structure (symbols desired, but a pointer to a reference is also useful), (3) Mechanism of action i.e., inhibits viral entry, replication, etc. (w/ a pointer to data), (4) Was the drug identified by manual or computation screen? (5) Who is studying the drug? (Source/lab name), (6) In vitro Data available (cell line used, assays run, viral strain used, cytopathic effects, toxicity, LD50, dosage response curve, etc.), (7) Animal Data Available (what animal model, LD50, dosage response curve, etc.), (8) Clinical trials on going (what phase, facility, target population, dosing, intervention etc.), (9) Funding source, (10) Has the drug shown evidence of systemic toxicity? (11) List of relevant sources to pull data from. The summary of their framework can be seen in Fig. 2.

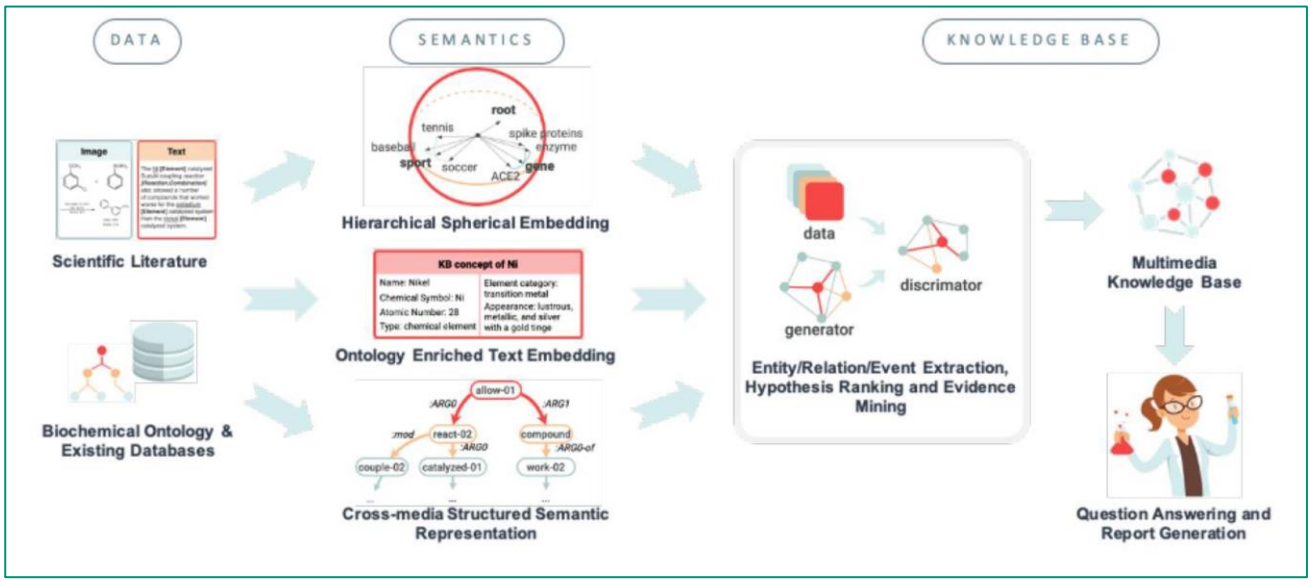

Figure 2: Framework used by Wang et al

They built a multimedia KG by combining (1) coarse-grained text knowledge extraction, (2) fine-grained text entity extraction, (3) image processing and cross-media entity grounding, and (4) knowledge graph semantic visualization. A KG constructed after just step (1) can be seen in Fig. 3. A demonstration of steps (2) and (3) can be seen in Figs. 4 and 5 respectively. Step (4) enhances the exploration and discovery of the information in the KG by allowing user interactivity that surpasses directed keyword searches or simple unigram word cloud or heatmap displays. Several clinicians and medical school students in their team reviewed the drug repurposing reports for three drugs that were used as a case study for the paper (Benazepril, Losartan, and Amodiaquine), and also the KGs connecting 41 drugs and COVID-19 related chemicals/genes. Preliminary results show that most of their output was informative and valid.

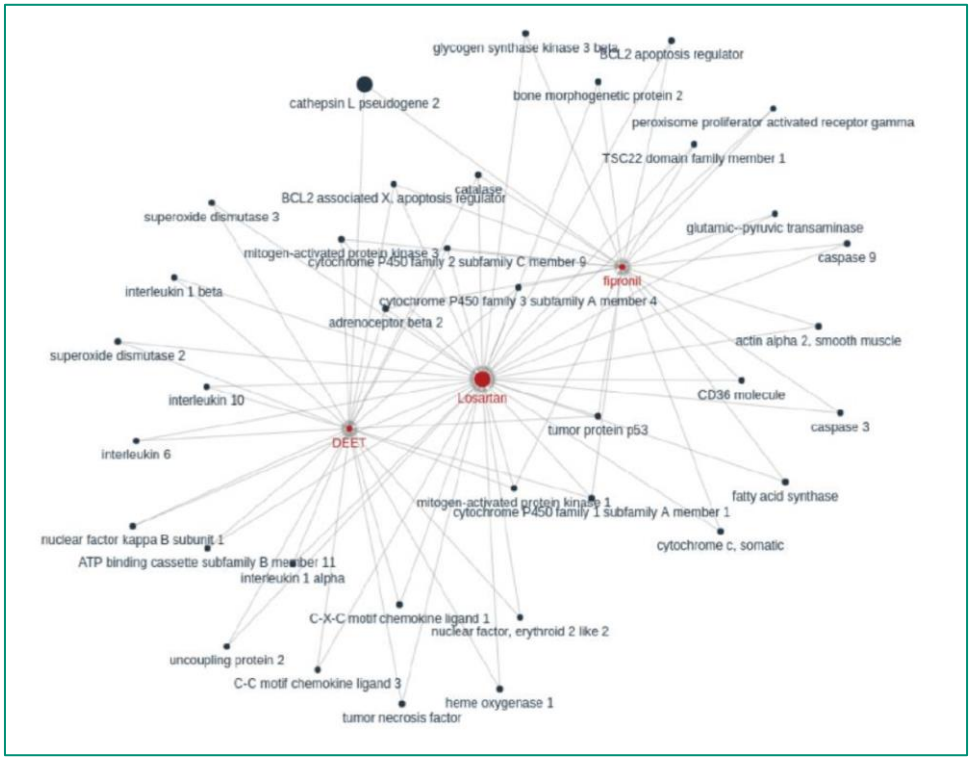

Figure 3: Constructed KG connecting Losartan (candidate drug in COVID-19) and cathepsin L pseudogene 2 (gene related to coronavirus). Taken from Wang et al

Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 **CORONAVIRUS** receptor: molecular mechanisms and potential therapeutic target. SARS-CoV-2 **CORONAVIRUS** has been sequenced [3]. A phylogenetic **EVOLUTION** analysis [3, 4] found a bat **WILDLIFE** origin for the SARS-CoV-2 **CORONAVIRUS**. There is a diversity of possible intermediate hosts for SARS-CoV-2 **CORONAVIRUS**, including pangolins **WILDLIFE**, but not mice **EUKARYOTE** and rats **EUKARYOTE** [5]. There are many similarities of SARS-CoV-2 **CORONAVIRUS** with the original SARS-CoV **CORONAVIRUS**. Using computer modeling, Xu *et al.* [6] found that the spike proteins **GENE_OR_GENOME** of SARS-CoV-2 **CORONAVIRUS** and SARS-CoV **CORONAVIRUS** have almost identical 3-D structures in the receptor binding domain that maintains Van der Waals forces **PHYSICAL_SCIENCE**. SARS-CoV spike proteins **GENE_OR_GENOME** has a strong binding affinity to human ACE2 **GENE_OR_GENOME**, based on biochemical interaction studies and crystal structure analysis [7]. SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** share identity in amino acid sequences and ……

Figure 4: Example of Fine-grained Entity Extraction. Taken from Wang et al

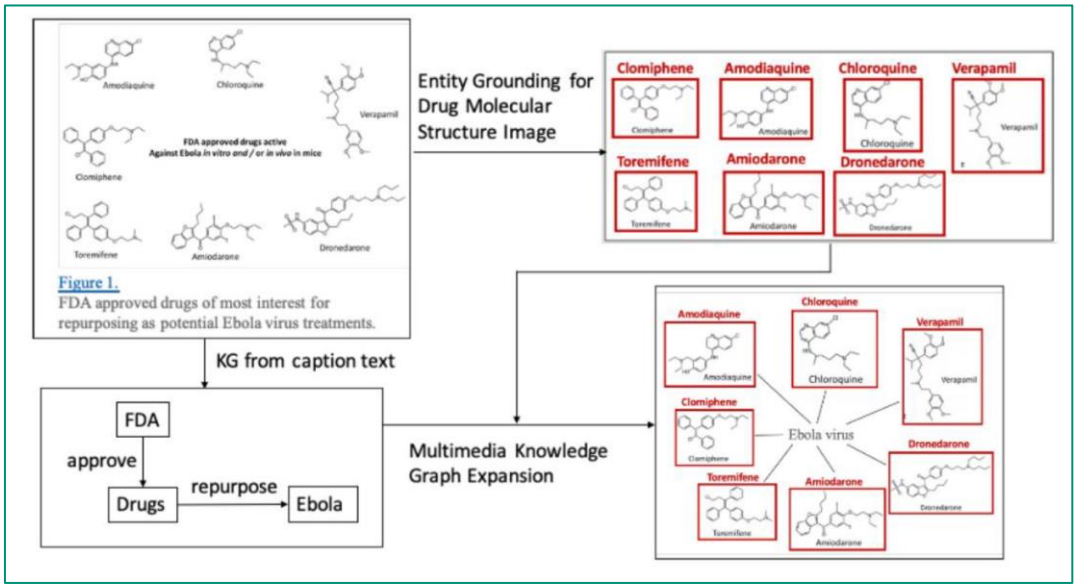

Figure 5: Expanding KG through Subfigure Segmentation and Cross-modal Entity Grounding. Taken from Wang et al

The third article, by Domingo-Fernandez et al, created a KG that is a cause-and-effect knowledge model of COVID-19 pathophysiology, which could then be applied for drug repurposing. The authors point out that although KGs were originally developed to describe interactions between entities, novel machine learning techniques can generate latent, low-dimensional representations of the KG which can then be utilized for downstream tasks such as clustering or classification. For the creation of the KG, scientific literature related to COVID-19 was retrieved from open access and freely available journals (PubMed, EuropePMC, and additional COVID-19 specific corpuses like LitCovid). This corpus was then filtered based on available information about potential drug targets for COVID- 19, biological pathways in which the virus interferes to replicate in its human host, and information on the various viral proteins along with their functions. Finally, the articles were prioritized based on the level of information that could be captured in the modeling language used to build the KG. Evidence text from the prioritized corpus was manually encoded in Biological Expression Language (BEL) as a triple including metadata about the nodes and their relationships as well as corresponding provenance and contextual information. BEL involves encoding mechanistic information such as protein-protein interactions, observed correlations between phenotypes and molecules, or effect of drugs on a given target. Therefore, only BEL encodable articles were selected. The authors explained in the Supplementary Material why they favored this manual curation over a text-mining approach, arguing that the manual approach provides better quality in terms of contextualization (i.e., finding the proper relation between two entities due to the complexity of scientific writing) and understandability of the KG. They mentioned the possibility of using a semi-automatic pipeline to combine the advantages of manual curation and text-mining.

Their KG summarizes mechanistic information on COVID-19 published in 160 original research articles. In its current state, the COVID-19 KG incorporates 4016 nodes, covering 10 entity types (e.g. proteins, genes, chemicals and biological processes) and 10,232 relationships (e.g. increases, decreases and association). Given the selected corpora, these cause-and-effect relations primarily denote host-pathogen interactions as well as comorbidities and symptoms associated with COVID-19. Furthermore, the KG contains molecular interactions related to host invasion (e.g. spike glycoprotein and its interaction with the host via receptor ACE2) and the effects of the downstream inflammatory, cell survival and apoptosis signaling pathways. The authors have identified over 300 candidate drugs currently being investigated in the context of COVID-19, including proposed repurposing candidates and drugs under clinical trial. The fourth article, by Zhou et al, is a review article for Lancet Digital Health. In the review, the authors introduced guidelines on how to use various forms of AI for accelerating drug repurposing, with COVID-19 as an example. With regard to KGs in particular, they mention that KGs can be reduced to low-dimensional feature vectors. Using the feature vectors of drugs and diseases, we can then measure their similarities and thus identify effective drugs for a given disease. One challenge for the graph embedding method is scalability.

The number of entities in a medical KG could be as many as several million. Several systems have been specifically designed for learning representations from large-scale graphs (e.g., GraphVite). The authors identified two works which evaded our search strategy: Gysi et al (which did not use the term knowledge graph in the paper and Zeng et al. Zeng et al's KG included 15 million edges across 39 types of relationships connecting drugs, diseases, proteins, genes, pathways, and expressions of genes and proteins from a large scientific corpus of 24 million PubMed publications. Using Amazon Web Services' computing resources and graph representation learning techniques (DGL-KE, mentioned earlier in this paper in the context of literature review), they identified 41 repurposable drug candidates (including dexamethasone, thalidomide, and melatonin) whose therapeutic associations with COVID-19 were validated by transcriptomic and proteomics data in SARS-CoV-2 infected human cells and data from ongoing clinical trials.

Knowledge graphs for clinical trials

The pre-eminent effort to synthesize the results of clinical trials related to the prevention and treatment of COVID-19 is the COVID-NMA initiative (https://covid-nma.com/). This project aims to provide a complete, high-quality, and up-to-date synthesis of evidence as soon as results are available as well as a living mapping of registered randomized controlled trials. The vast majority of work involved in curating the database is done by human volunteers. This synthesis will allow evidence-based decision-making and planning of future research. We would like to mention that this initiative cannot be classified as a KG in the AI sense, because the concept of triples (which is central to an AI KG) is not used. However, we still include information about this initiative in this review because the results of this approach are exactly in line with the goals of a KG. The living mapping of trials (i.e., trials registered on the WHO platform) is updated weekly, and contained 2358 RCTs at the end of 2020. The living synthesis of published trials (including both articles and preprints) is updated daily, and contained 157 RCTs with results at the end of 2020. The highly interactive data visualizations that have been developed as a result of this initiative constitute some of the most useful summaries of COVID-19 research. Some examples are shown in Figs. 6-8, but to fully appreciate the flexibility provided by the visualization tools, we encourage the reader to visit the website. Potentially this high-quality human curation can be replaced in the future by AI to ensure sustainability.

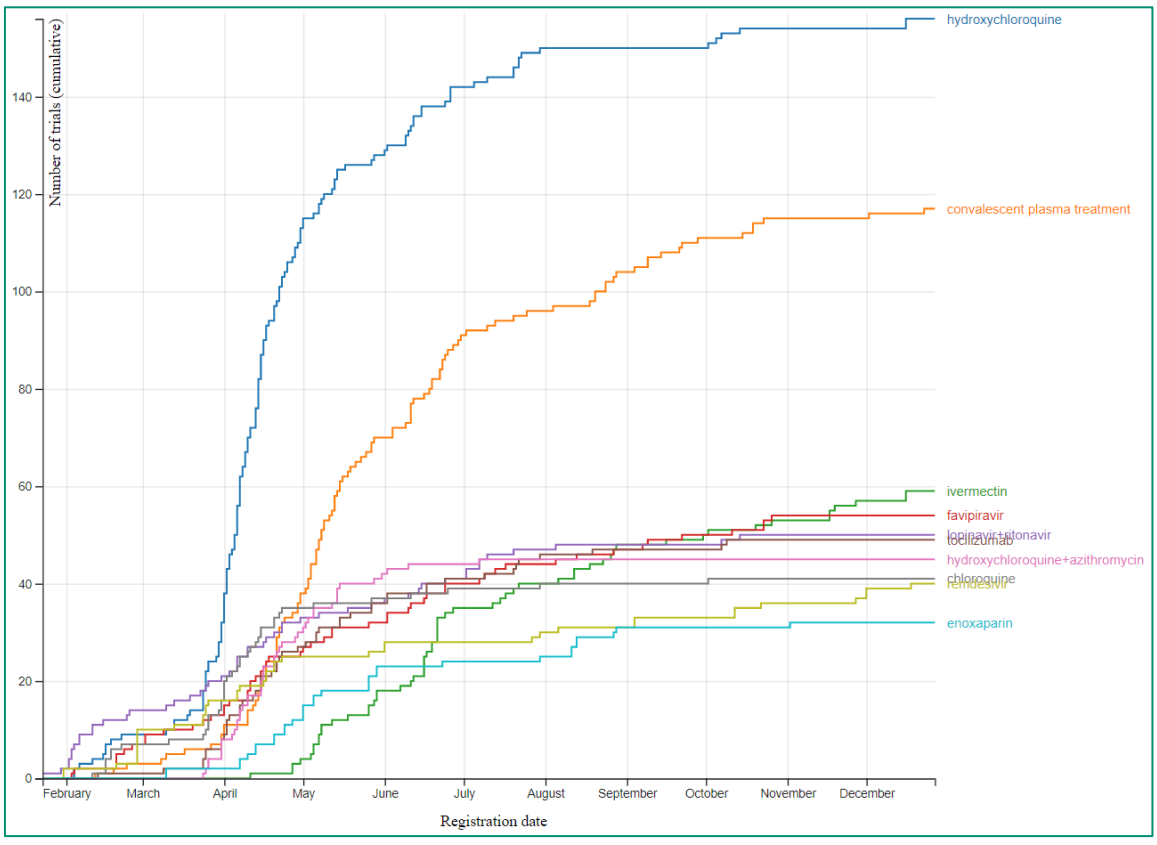

Figure 6: Trend of number of trials registered by treatment name.
Taken from https://covid-nma.com/.

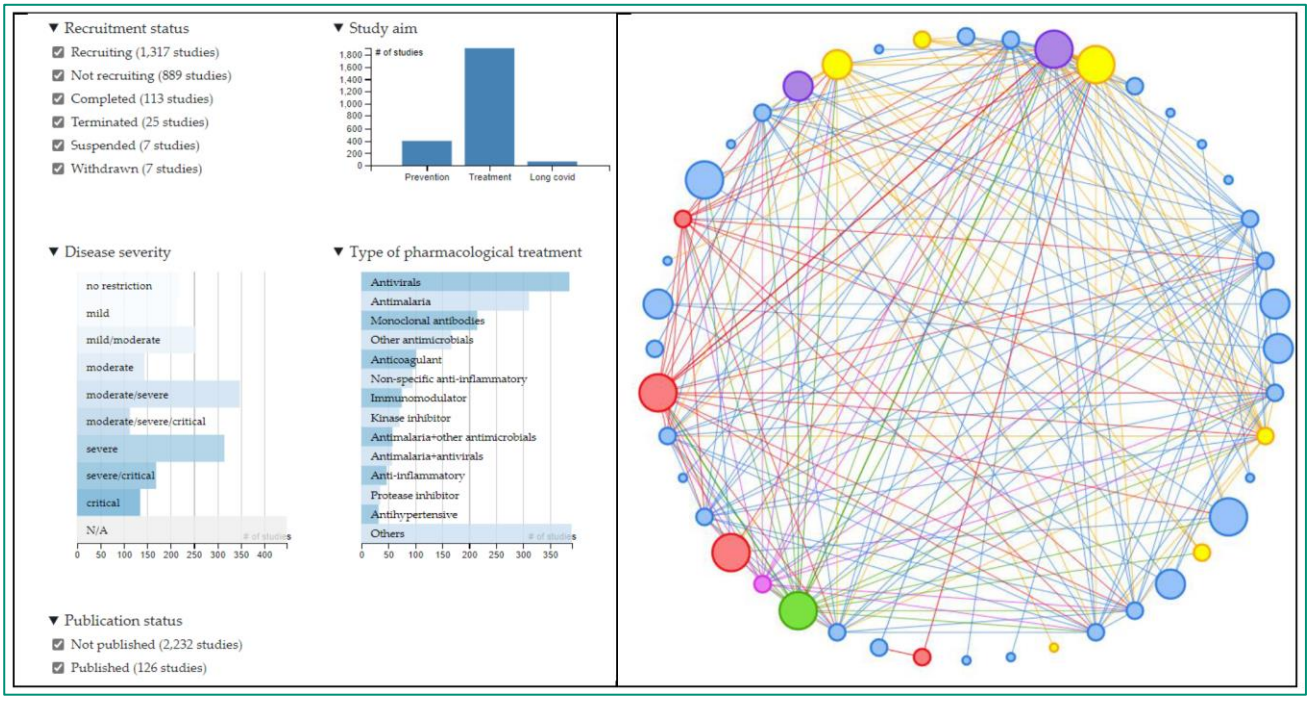

Figure 7: The diagram on the right describes the network of RCTs evaluating pharmacologic treatments for COVID-19 which fulfill criteria that the user selects. The nodes in the diagrams represent the different treatments evaluated in these RCTs and the lines represent the direct comparisons made in the studies. When two nodes are connected with a line, it means there is at least one study that compares the corresponding treatments. Whereas, when they are not connected, it means there is no study comparing them. The size of the nodes is proportional to the number of participants allocated to each intervention and the thickness of the lines is proportional to the number of studies that compare each pair of treatments. Taken from https://covid-nma.com/.

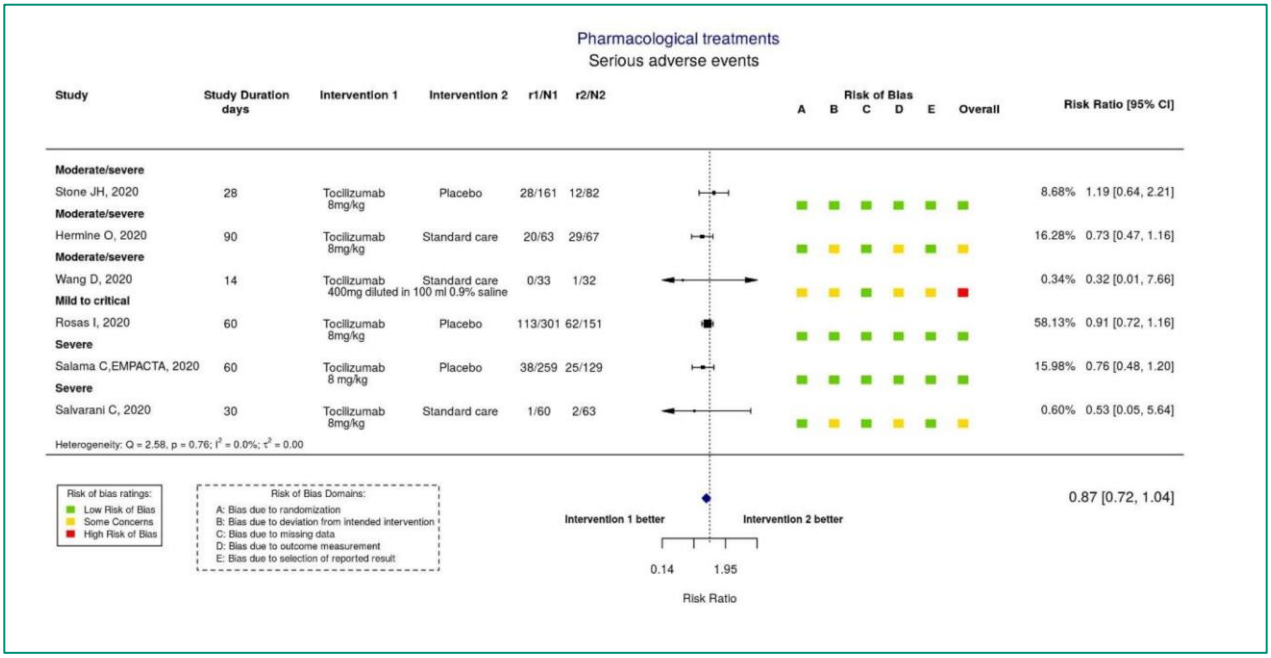

Figure 8: A Forest Plot comparing two interventions chosen by the user, in this case tocilizumab and placebo/standard of care. Taken from https://covid-nma.com/.

Multi-purpose knowledge graphs

We found two papers that use KGs for multiple tasks, including literature review and drug repurposing. The first, by Chen et al, does a case study on the application possibilities of KGs. The introduction of their paper provides an excellent summary of the emergence of KGs in the field of AI. They point out that in the past, KGs have been curated manually, but the move towards natural language understanding through semantic technologies has accelerated in the past decade, promoting Named Entity Recognition (NER) to a central NLP task. NER has been crucial for building and constructing KGs as the primary method of extracting entities and possibly relations from free text. Also, tasks such as link prediction, relation extraction, and graph completion on KGs are aided by NER. In the early 2000s, biomedical NER relied on feature engineering and graphical models such as Hidden Markov Models (HMM) and Conditional Random Fields (CRF), which had poor accuracy compared to the current state-of-the art which uses deep learning. Bidirectional Encoder Representations from Transformers (BERT) is the foundational work from Google that has made deep-learning-based NER possible. BioBERT is a biomedical language representation model based on BERT used by the authors to mine the CORD-19 dataset, as well as the PubMed database and PubMed KG.

To illustrate the utility of KGs, the authors performed several experiments, the most basic of which was compiling a list of most-published authors in the CORD-19 dataset. In an experiment using BioBERT, they found that BioBERT can easily recognize the common bio-entities with a high occurrence rate in the corpus, but fails to recognize rare biomedical terms. They used two metrics to find the strength of KG associations (i.e., weights) between source and target nodes: co-occurrence frequency and cosine similarity. Figure 9 shows KGs related to remdesivir based on co-occurrence frequency. While this is a promising approach, a major limitation of co-occurrence frequency is that it cannot reflect the relationship between the source node and the target node well. For example, if “A has nothing to do with B” is mentioned often in documents, its co-occurrence frequency will be high. Cosine similarity has the benefit of being a normalized metric unlike co-occurrence frequency, but it still has the same limitation.

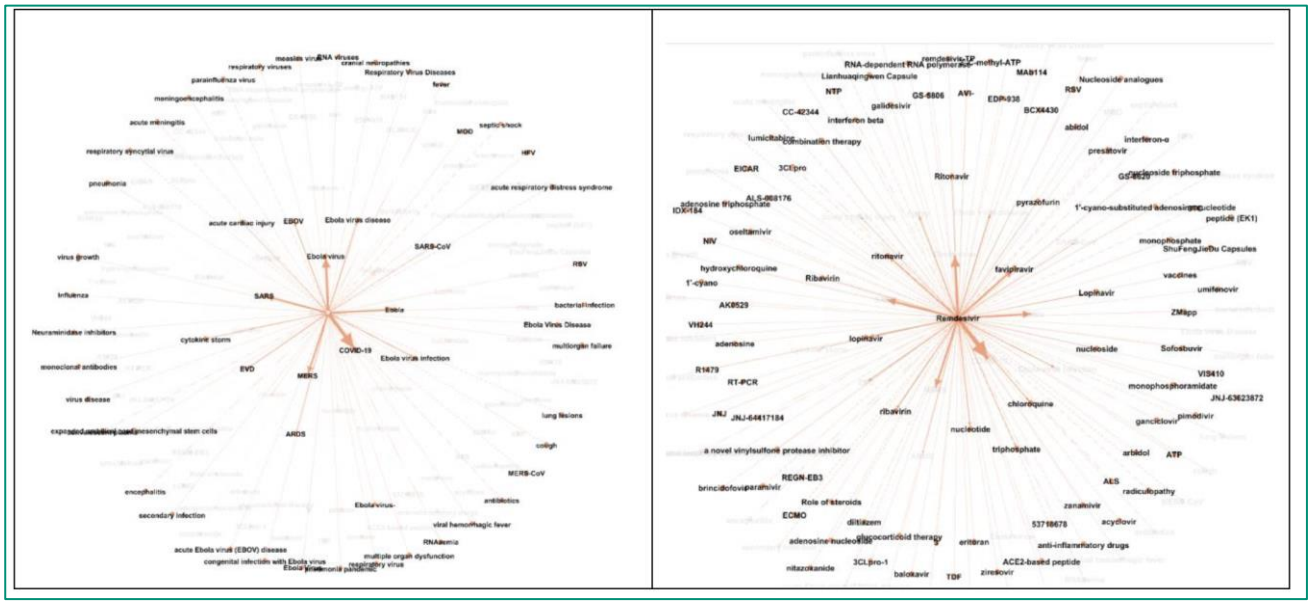

Figure 9: Remdesivir-related KGs: associated diseases (left) and associated drugs (right) based on cooccurrence frequency. Taken from Chen et al

The second paper, by Reese et al, is a framework for producing knowledge graphs that can be customized for downstream applications including machine learning tasks, hypothesis-based querying, and browsable user interface. For example, a drug repurposing application would make use of protein data linked with approved drugs, while a biomarker application could utilize data on gene expression linked with pathways. The authors explain that researchers are confronted with a number of technical challenges when trying to use existing data to discover actionable knowledge about COVID-19. The data needed to address a given question are typically siloed in different databases and employ different identifiers, data formats, and licenses. For example, to examine the function of proteins targeted by FDA-approved antiviral drugs, one must download and integrate drug, drug target, and FDA approval status data (from Drug Central, for example, in a bespoke TSV format) and functional annotations (from, for example, Gene Ontology in GPAD format). Furthermore, many datasets are updated periodically, which requires researchers to re-download and re-harmonize data. KGs are a way to represent and integrate heterogeneous data and their interrelationships using a hierarchical system such as an ontology. This kind of representation is amenable to complex queries (e.g. “which drugs target a host protein that interacts with a viral protein?”) and also to graph-based ML techniques.

Their workflow is divided into three steps: data download (fetch the input data), transform (convert the input data to KGX interchange format), and merge (combine all transformed sources). The ingested data are focused on sources relevant to drug repurposing for downstream querying and machine learning applications, prioritizing drug databases, protein interaction databases, protein function annotations, COVID-19 literature, and related ontologies. From the final merged graph, training and test data sets suitable for machine learning applications are created. Embiggen, their implementation of node2vec and related algorithms, is applied to this KG to generate embeddings, vectors in a low dimensional space which capture the relationships in the KG. Embiggen is trained iteratively to identify optimal node2vec hyperparameters (walk length, number of walks, p, q etc.) and to then train classifiers (e.g., logistic regression, random forest, support vector machines) that can be used for link prediction. The trained classifiers can then be applied to produce actionable knowledge: drug to disease links, drug to gene links, and drug to protein links. Besides machine learning, the authors have also used the KG for hypothesis-based querying. For example, they have queried the KG to identify host proteins that are known to interact with viral proteins, and these are further filtered according to whether these host proteins are targets of approved drugs. In the framework created by the authors, each data source is transformed and output as a separate graph, which is later combined with graphs for other data sources according to the needs of the user. Although the subgraphs from the various data sources (e.g., Drug Central) are produced locally by their framework, they could easily incorporate graphs generated by other researchers. The exchange of data via a ‘KG-Hub’ would eliminate the duplication of effort that occurs when researchers separately transform and prepare data, and might also facilitate the formation of a data sharing portal.

Conclusion

In this work, we have provided an exploratory review on knowledge graphs in the context of COVID-19. By providing links between disparate datasets that are stuck in silos, KGs enable the user to effectively search the overwhelming volume of COVID-19 research and gain actionable insight which would either be extremely tedious or impossible to achieve in the absence of such emerging uses of AI.

References

1. Kejriwal, M: Knowledge Graphs and COVID-19: Opportunities, Challenges, and Implementation. Harvard Data Science Review, https://doi.org/10.1162/99608f92.e45650b8 (2020).

2. Shariff, S. Z., Bejaimal, S. A., Sontrop, J. M., Iansavichus, A. V., Haynes, R. B., Weir, M. A., & Garg, A. X.: Retrieving clinical evidence: a comparison of PubMed and Google Scholar for quick clinical searches. Journal of medical Internet research, 15(8), e164 (2013).

3. Kendall, S. Which one is best: PubMed, Web of Science, or Google Scholar? https://libguides.lib.msu.edu/pubmedvsgooglescholar

4. Webber, J.: A programmatic introduction to neo4j. In Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity (pp. 217–218). Association for Computing Machinery (2012)

5. Steenwinckel B, Vandewiele G, Rausch I, Heyvaert P, Taelman R, Colpaert P, Simoens P, Dimou A, De Turck F, Ongenae F. Facilitating the analysis of covid-19 literature through a knowledge graph. In International Semantic Web Conference 2020 (pp. 344-357). Springer, Cham.

6. AI, A.I.F.: Covid-19 open research dataset challenge (cord-19). https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

7. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web, pp. 722–735. Springer (2007)

8. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research 37 (2009)

9. Lammey, R.: Crossref text and data mining services. Science Editing (2015)

10. Haak, L.L., Fenner, M., Paglione, L., Pentz, E., Ratner, H.: Orcid: a system to uniquely identify researchers. Learned Publishing 25(4), 259–264 (2012)

11. Heyvaert, P., De Meester, B., Dimou, A., Verborgh, R.: Declarative Rules for Linked Data Generation at Your Fingertips! In: European Semantic Web Conference. pp. 213–217. Springer (2018)

12. Dimou, A., De Meester, B., Heyvaert, P., Verborgh, R., Latré, S., Mannens, E.: RMLMapper: a tool for uniform Linked Data generation from heterogeneous data

13. Ristoski, P., Paulheim, H.: Rdf2vec: Rdf graph embeddings for data mining. In: International Semantic Web Conference. pp. 498–514. Springer (2016)

14. Wise C, Ioannidis VN, Calvo MR, Song X, Price G, Kulkarni N, Brand R, Bhatia P, Karypis G. COVID-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. arXiv preprint arXiv:2007.12731. (2020).

15. Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3606–3611 (2019).

16. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In Advances in neural information processing systems, pages 2787–2795 (2013).

17. Zheng, D., Song, X., Ma, C., Tan, Z., Ye, Z., Dong, J., Xiong, H., Zhang, Z., and Karypis, G.: Dgl-ke: Training knowledge graph embeddings at scale. arXiv preprint arXiv:2004.08532 (2020).

18. Michel F, Gandon F, Ah-Kane V, Bobasheva A, Cabrio E, Corby O, Gazzotti R, Giboin A, Marro S, Mayer T, Simon M. Covid-on-the-Web: Knowledge graph and services to advance COVID-19 research. In International Semantic Web Conference 2020 (pp. 294-310). Springer, Cham.

19. Mayer T, Cabrio E, Villata S. ACTA: a tool for argumentative clinical trial analysis. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence Demos. Pages 6551-6553. https://doi.org/10.24963/ijcai.2019/953 (2019).

20. Stebbing J, Phelan A, Griffin I, Tucker C, Oechsle O, Smith D, Richardson P. COVID-19: combining antiviral and anti-inflammatory treatments. The Lancet Infectious Diseases 20(4):400-2 (2020).

21. ACTT-2 Study Group: Baricitinib plus Remdesivir for Hospitalized Adults with Covid-19, NEJM, 2020.

22. Wang Q, Li M, Wang X, Parulian N, Han G, Ma J, Tu J, Lin Y, Zhang H, Liu W, Chauhan A. COVID-19 literature knowledge graph construction and drug repurposing report generation. arXiv preprint arXiv:2007.00576. (2020).

23. Domingo-Fernandez D, Baksi S, Schultz B, Gadiya Y, Karki R, Raschka T, Ebeling C, Hofmann-Apitius M. COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. BioRxiv (2020).

24. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F. Artificial intelligence in COVID-19 drug repurposing. The Lancet Digital Health (2020).

25. Zhu Z, Xu S, Qu M, Tang J. GraphVite: a high-performance CPU-GPU hybrid system for node embedding. arXiv preprint arxiv:1903.00757 (2019).

26. Gysi DM, Valle ÍD, Zitnik M, Ameli A, Gan X, Varol O, Sanchez H, Baron RM, Ghiassian D, Loscalzo J, Barabási AL. Network medicine framework for identifying drug repurposing opportunities for covid-19. arXiv preprint arXiv:2004.07229 (2020).

27. Zeng X, Song X, Ma T, Pan X, Zhou Y, Hou Y, Zhang Z, Li K, Karypis G, Cheng F. Repurpose open data to discover therapeutics for COVID-19 using deep learning. Journal of proteome research 19(11):4624-36 (2020).

28. Chen C, Ebeid IA, Bu Y, Ding Y. Coronavirus knowledge graph: A case study. arXiv preprint arXiv:2007.10287 (2020).

29. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

30. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234-40 (2020).

31. Reese JT, Unni D, Callahan TJ, Cappelletti L, Ravanmehr V, Carbon S, Shefchek KA, Good BM, Balhoff JP, Fontana T, Blau H. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. Patterns: 100155 (2020).