DATA NOTE

Organizing knowledge to enable faster data interpretation in COVID-19 research [version 1; peer review: 2 approved with reservations]

Joseph Hearnshaw, Marco Brandizi, Ajit Singh, Chris Rawlings, Keywan Hassani-Pak

Rothamsted Research, Harpenden, AL5 2JQ, UK

Abstract
Enormous volumes of COVID-19 research data have been published and this continues to increase daily. This creates challenges for researchers to interpret, prioritize and summarize their own findings in the context of published literature, clinical trials, and a multitude of databases. Overcoming the data interpretation bottleneck is vital to help researchers to be more efficient in their quest to identify COVID-19 risk factors, potential treatments, drug side-effects, and much more. As a proof of concept, we have organized and integrated a range of COVID-19 and human biomedical data and literature into a knowledge graph (KG). Here we present the datasets we have integrated so far and the content of the KG which consists of 674,969 biological concepts and over 1.6 million relationships between them. The COVID-19 KG is available via KnetMiner, an interactive online platform for gene discovery and knowledge mining, or via RDF and Neo4j graph formats which can be searched programmatically through SPARQL and Cypher endpoints. KnetMiner is a road mapped ELIXIR UK service. We hope this integrated resource will enable faster data interpretation and discovery of linkages between genes, drugs, diseases and many more types of information relating to COVID-19.

Keywords
Coronavirus, COVID-19, SARS-CoV-2, knowledge graph, knowledge base, target discovery, knowledge mining, bioinformatics

This article is included in the ELIXIR gateway.
Corresponding author: Keywan Hassani-Pak (keywan.hassani-pak@rothamsted.ac.uk)

Author roles: Hearnshaw J: Data Curation, Formal Analysis, Software, Writing – Original Draft Preparation; Brandizi M: Methodology, Software, Writing – Review & Editing; Singh A: Software, Writing – Review & Editing; Rawlings C: Conceptualization, Writing – Review & Editing; Hassani-Pak K: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the UKRI Biotechnology and Biological Sciences Research Council (BBSRC) through the Designing Future Wheat ISP: BB/P016855/1 and the BBR grant: BB/S020020/1. CR, KHP, AS are additionally supported by strategic funding to Rothamsted Research from BBSRC. 
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Hearnshaw J et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Hearnshaw J, Brandizi M, Singh A et al. Organizing knowledge to enable faster data interpretation in COVID-19 research [version 1; peer review: 2 approved with reservations] F1000Research 2021, 10(ELIXIR):703
https://doi.org/10.12688/f1000research.54071.1

First published: 30 Jul 2021, 10(ELIXIR):703 https://doi.org/10.12688/f1000research.54071.1
**Introduction**

The global COVID-19 pandemic, caused by the SARS-CoV-2 virus, has infected millions worldwide, causing many more infections and deaths than the previous severe acute respiratory syndrome (SARS) outbreak that occurred in 2002 and 2003 (WHO, 2020). The race to find plausible treatment strategies and management plans for the coronavirus pandemic, concurrent with investigating the biological underpinnings of how SARS-CoV-2 operates has driven a very rapid rise in the number of SARS-CoV-2-related publication, pre-prints and biological data. The White House, alongside other groups, made a global call to action to find a way to rapidly sift through this mountain of COVID-19 literature. They released the COVID-19 Open Research Dataset (CORD-19), which has over 150,000 COVID-19-related full text articles (Kaggle, 2020). Additionally, the COVID-19 data portal contains over 360,000 publications and many other datasets related to the disease. The rate of growth and sheer volume of data related to the pandemic and the virus has become so large that it has led to something of a new phenomenon for many researchers: ‘too much information’. As such, there lies a challenge to organize the growing mountain of biomedical knowledge in order to overcome the data interpretation bottleneck (Good et al., 2014).

Our goal was to repurpose the KnetMiner software, which we originally developed for crop scientists to help identify the most important genes involved in complex plant traits (Hassani-Pak et al., 2021), to provide medical researchers with quick and intuitive access to all documented linkages between genes, potential therapeutic compounds, and the virus. KnetMiner contains fast algorithms for scanning millions of relationships from across a range of datasets and literature, scoring all evidence and displaying the links in easy to understand knowledge graphs. KnetMiner for COVID-19 would enable users to search for genes and keywords related to COVID-19 and explore the surrounding connected gene networks and pathways, for example, for negative downstream effects of drugs, or genetic associations with known diseases, or human-pathogen interactions.

KnetMiner requires as input a knowledge graph (KG) which is a flexible semantic data model to represent heterogeneous, interconnected data where graph nodes represent biological concepts and graph edges the relationships between them. Such biological concepts can include, but are not limited to, genes, drugs, diseases, and publications. There have been a few parallel efforts to construct KGs from COVID-19 data (Reese et al., 2020). These mostly differ in data coverage and modelling approaches and would require further work to be made usable by KnetMiner. Hence, we decided to use the toolkit provided by KnetMiner to build a unified knowledge graph from public datasets and the scientific literature.

Here, we describe how we organized the biomedical knowledge in a format that is compatible with KnetMiner and present a use case.

**Methods**

**Data sources**

Firstly, we identified the datasets that we considered as important for the first iteration of the COVID-19 KG, based on discussions held with participants of the COVID-19 Biohackathon (5–11 April 2020). We downloaded over 20 datasets including COVID-19 related publications and pre-prints, human genetics studies, virus-host protein interactions, drug-target interactions, pathway information, ontology annotations and mouse knock-out mutant data with links to diseases and phenotypes (see Table 1). The data download and processing steps were automated using custom Python scripts (see Extended data). Various data formats are supported by the down-stream integration pipeline, including PubMed XML, UniProt XML and OWL, and most importantly tabular format, which required the development of an XML-based configuration to map the tabular content to a labelled property graph representation.

**KnetMiner**

The KnetBuilder (https://github.com/Rothamsted/knetbuilder) software was used to integrate the various data sources into a KG with unified semantics and augment it with additional text mining-based relationships (Hassani-Pak et al., 2010). The data loading and integration steps are specified using an XML-based configuration. KnetBuilder uses an in-memory graph implementation (Köhler et al., 2006) and provides exports to RDF and tooling to convert RDF to Neo4j (Brandizi et al., 2018b).

KnetMiner is an open-source software allowing to index and serve the contents of a knowledge graph in order to accelerate gene discovery. It requires as input a dataset folder with configuration files and a compatible knowledge graph in OXL format. We used the version 4.0 of KnetMiner and deployed it with our COVID-19 KG (see Underlying data). The COVID-19 KG contains all human genes; however, KnetMiner was configured to index a subset of 9,141 COVID-19-related genes, derived from SciBite gene annotations of CORD-19 version 54 (Giles et al., 2020). In the presented use case, KnetMiner was searched with a list of differentially expressed genes (Table 3) and the following keywords:

"Defense Response To Virus" Remdesivir "Killing Of Cells Of Other Organism" "Antimicrobial Humoral Immune Response Mediated By Antimicrobial Peptide" "Response To Virus" "Cellular Response To Lipopolysaccharide" "Positive Regulation Of Transcription By RNA Polymerase II" "White blood cell count" "Blood protein levels"

**COVID-19 Knowledge Graph**

The combined COVID-19 KG has 674,969 biological concepts and over 1.6 million relationships between them, and is
**Table 1.** The key data sources used to construct the COVID-19 knowledge graph, their composite biological concept type, respective semantics, and URL links for the data.

| Datasource                                | Data Types                  | Concept Classes                                                                 | URLs                                                                                     |
|-------------------------------------------|-----------------------------|---------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| Mouse Genome Informatics data (MGI)       | Gene Disease (OMIM)         | Gene → Disease                                                                  | http://www.informatics.jax.org/downloads/reports/MGI_DO.rpt                              |
| InterPro                                  | Protein Domain              | Protein → Protein Domain                                                         | https://ftp.ebi.ac.uk/pub/databases/interpro/entry.list                                   |
| OMIM                                      | Disease                     | Drug → Disease                                                                  | https://www.omim.org/                                                                     |
| NHGRI-EBI GWAS Catalog                    | SNP (GWAS) Trait            | Gene → SNP → Trait                                                              | https://www.ebi.ac.uk/gwas/api/search/downloads/studies_alternative                        |
| Gene Ontology (GO)                        | Biological Processes        | Protein → BioProc                                                                | http://purl.obolibrary.org/obo/go.owl                                                     |
| IntAct                                    | Protein                     | Protein → Protein                                                                | www.ebi.ac.uk/intact/export?format=mitab_27&query=annot%3A%22dataset%3Acoronavirus%22&negative=false&spoke=false&ontology=false&sort=intact-miscore&asc=false |
| Ensembl (GRCh38.99)                        | Gene                        | Protein                                                                         | https://www.ensembl.org/info/data/ftp/index.html                                          |
| UniProt                                   | Protein Publication         | Gene → Protein → Publication                                                     | https://www.uniprot.org/uniprot?include=false&format=xml&compress=yes&force=true &query=proteome:UP000005640 |
|                                            | Gene Ontologies             | Protein → Gene Ontologies                                                        |                                                                                           |
| DrugBank                                  | Drug                        | Protein → Drug                                                                   | https://www.drugbank.ca/releases/latest                                                   |
| HumanCyc                                  | Pathway                     | Protein → Enzyme Complex → Reaction → Pathway                                   | http://brg-files.ai.sri.com/public/dist/humancyc.tar.gz                                  |
|                                            | Enzyme Complex              | → Reaction → Pathway                                                             |                                                                                           |
|                                            | Molecules                   | → Compound                                                                       |                                                                                           |
|                                            | Compound                    |                                                                                |                                                                                           |
| SciBite                                   | Gene Protein (SARS-CoV-2)   | Gene → Publication                                                               | https://github.com/SciBiteLabs/CORD19                                                    |
|                                            | Phenotype                   | Protein → Publication                                                             |                                                                                           |
|                                            |                             | Phenotype → Publication                                                           |                                                                                           |
|                                            |                             | Drug → Publication                                                                |                                                                                           |
available in OXL, RDF and Neo4j formats. It contains data on 27,599 human genes that are present in the KG as “type gene” concepts that are linked to a range of other concepts through different relation types and evidences. The content of the KG with a breakdown by concept type is shown in Table 2.

Programmatic access is available via public SPARQL and Cypher endpoints (Brandizi et al., 2018a) (see Underlying data). We have developed sample Jupyter notebooks to illustrate how the KG can be queried (Brandizi, 2020). This maximizes the adhesion to FAIR data principles, ensuring the data is reusable for a variety of applications. Availability, including pointers to data repositories on the Open Science Framework (OSF), is listed in the README file (see data availability section). Using Neo4j and Cypher, we can query the data and walk the knowledge graph in multiple directions. For example, we can look at the most common pathways related to proteins which are targeted by literature cited drugs using the following Cypher query:

```
MATCH (bp:BioProc)<-[[:participates_in]]-(prot:Protein)<-[[:has_target]]-(drug:Drug)<-[[:occ_in]]->(pub:Publication)
RETURN bp.prefName, COUNT(DISTINCT pub) AS pubNo
ORDER BY pubNo DESC
LIMIT 10
```

**COVID-19 KnetMiner**

COVID-19 KnetMiner (https://knetminer.com/COVID-19/) provides gene-centric and interactive access to the integrated knowledge base created by KnetBuilder. KnetMiner can be used to search the KG with keywords, gene lists, genomic regions of interest, and/or any combinations of these user inputs. We have created several example queries for different use cases and search strategies. These are available on the COVID-19 KnetMiner website for easy use.

In June 2020, a gene expression study was published which investigated the transcriptional response of human lung epithelial cells to SARS-CoV-2 infection (Enes & Pir, 2020). The data and the analysis between infected vs mock cells are available at the EBI Gene Expression Atlas (Table 3; https://www.ebi.ac.uk/gxa/experiments/E-GEOD-147507).

We chose this dataset because of its relevance and tractable size to showcase the use of KnetMiner for the interpretation of differentially expressed genes (DEGs). The approach can be equally applied to any data driven analysis and derived gene lists. All DEGs from the above study were used in KnetMiner to investigate linkages to drugs, diseases, biological processes, and the CORD-19 literature. As such the DEG IDs were entered into the Gene List search interface of KnetMiner and analyzed in several iterations (Figure 1): (i) without any keywords to get an unbiased view of the linked and enriched knowledge, (ii) with the keyword SARS-CoV-2 to review recent CORD-19 articles in relation to these genes, and (iii) with specific drugs, diseases and GO terms to visualize and share the organized knowledge with other scientists (Figure 2).

KnetMiner user stories usually follow a common pattern: search, prioritize, explore, and share knowledge. All views shown in Figure 1 can generate knowledge networks for selected genes and/or evidence terms. As an example, an interactive network was generated for the top 20 scoring DEGs and keywords related to white blood cell count (GWAS study), Remdesivir (Drug), Response to Virus (GO).

**Summary**

KnetMiner provides open-source tools to build, search, visualize and share knowledge graphs, and its species agnostic architecture has provided a cost-efficient toolkit for building the first COVID-19 KnetMiner resource. By organizing COVID-19 related data in one place, integrated through a clear semantic data model, the hope is that this will enable data interpretation using more reproducible and objective approaches and support.
Table 3. Differentially expressed genes from EBI Gene Expression Atlas ([https://www.ebi.ac.uk/gxa/experiments/E-GEOD-147507](https://www.ebi.ac.uk/gxa/experiments/E-GEOD-147507)).

| Gene ID               | Gene Name | 'Severe acute respiratory syndrome coronavirus 2 (USA-WA1/2020)' vs 'mock' in 'NHBE' at '24 hour'.foldChange | 'Severe acute respiratory syndrome coronavirus 2 (USA-WA1/2020)' vs 'mock' in 'NHBE' at '24 hour'.pValue |
|-----------------------|-----------|------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| ENSG00000008517       | IL32      | 1.1                                                                                                  | 4.58E-16                                                                                              |
| ENSG000000023445      | BIRC3     | 1.3                                                                                                  | 9.71E-17                                                                                              |
| ENSG000000089127      | OAS1      | 1.3                                                                                                  | 2.97E-14                                                                                              |
| ENSG000000090339      | ICAM1     | 1.5                                                                                                  | 1.75E-27                                                                                              |
| ENSG00000100985       | MMP9      | 1.6                                                                                                  | 7.97E-17                                                                                              |
| ENSG00000104368       | PLAT      | 1.2                                                                                                  | 3.66E-14                                                                                              |
| ENSG00000108342       | CSF3      | 1.4                                                                                                  | 4.23E-13                                                                                              |
| ENSG00000111331       | OAS3      | 1.1                                                                                                  | 1.32E-15                                                                                              |
| ENSG00000111335       | OAS2      | 1.1                                                                                                  | 3.70E-17                                                                                              |
| ENSG00000112096       | SOD2      | 1.3                                                                                                  | 2.43E-25                                                                                              |
| ENSG00000113070       | HBEGF     | 1.2                                                                                                  | 6.30E-21                                                                                              |
| ENSG00000115009       | CCL20     | 2.4                                                                                                  | 9.70E-63                                                                                              |
| ENSG00000118503       | TNFAIP3   | 1.5                                                                                                  | 9.51E-45                                                                                              |
| ENSG00000122641       | INHBA     | 1.5                                                                                                  | 8.55E-30                                                                                              |
| ENSG00000125730       | C3        | 1.4                                                                                                  | 7.66E-33                                                                                              |
| ENSG00000126709       | IFI6      | 1.3                                                                                                  | 1.80E-11                                                                                              |
| ENSG00000128342       | LIF       | 1.2                                                                                                  | 3.26E-21                                                                                              |
| ENSG00000134070       | IRAK2     | 1.2                                                                                                  | 9.16E-13                                                                                              |
| ENSG00000134339       | SAA2      | 2.2                                                                                                  | 8.78E-63                                                                                              |
| ENSG00000136244       | IL6       | 1.5                                                                                                  | 2.52E-16                                                                                              |
| ENSG00000136688       | IL36G     | 2.1                                                                                                  | 8.43E-40                                                                                              |
| ENSG00000140379       | BCL2A1    | 1.3                                                                                                  | 3.70E-10                                                                                              |
| ENSG00000140519       | RHCG      | 1.2                                                                                                  | 1.92E-24                                                                                              |
| ENSG00000143546       | S100A8    | 1.7                                                                                                  | 4.20E-33                                                                                              |
| ENSG00000145901       | TNIP1     | 1.2                                                                                                  | 3.36E-33                                                                                              |
| ENSG00000157601       | MX1       | 1.7                                                                                                  | 5.44E-22                                                                                              |
| ENSG00000162366       | PDZK1P1   | 1.4                                                                                                  | 1.17E-13                                                                                              |
| ENSG00000163216       | SPRR2D    | 2                                                                                                     | 7.40E-29                                                                                              |
| ENSG00000163218       | PGLYRP4   | 1.2                                                                                                  | 5.73E-10                                                                                              |
| ENSG00000163734       | CXCL3     | 1.1                                                                                                  | 4.65E-08                                                                                              |
| ENSG00000163735       | CXCL5     | 1.9                                                                                                  | 1.53E-23                                                                                              |
| ENSG00000163739       | CXCL1     | 1.3                                                                                                  | 1.70E-30                                                                                              |
# Expression Atlas
Transcriptional response of human lung epithelial cells to SARS-CoV-2 infection

# Query: Genes matching: ‘default query’, specifically up/down differentially expressed in 1/2 selected comparisons given the adjusted p-value cutoff 0.05 and log2-fold change cutoff 1 in experiment E-GEOD-147507

# Timestamp: Fri, 30-Apr-2021 12:55:29

| Gene ID          | Gene Name | 'Severe acute respiratory syndrome coronavirus 2 (USA-WA1/2020)' vs 'mock' in 'NHBE' at '24 hour'.foldChange | 'Severe acute respiratory syndrome coronavirus 2 (USA-WA1/2020)' vs 'mock' in 'NHBE' at '24 hour'.pValue |
|------------------|-----------|---------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| ENSG00000163874  | ZC3H12A   | 1.3                                                                                               | 7.73E-17                                                                                           |
| ENSG00000165949  | IFI27     | 1.9                                                                                               | 1.73E-30                                                                                           |
| ENSG00000166920  | C15orf48  | 1.1                                                                                               | 1.09E-19                                                                                           |
| ENSG00000169429  | CXCL8     | 2.1                                                                                               | 3.90E-89                                                                                           |
| ENSG00000173432  | SAa1      | 1.9                                                                                               | 1.72E-41                                                                                           |
| ENSG00000173918  | C1QTNF1   | 1.1                                                                                               | 1.31E-07                                                                                           |
| ENSG00000185215  | TNFAIP2   | 1.2                                                                                               | 1.64E-17                                                                                           |
| ENSG00000185479  | KRT6B     | 1.5                                                                                               | 3.38E-40                                                                                           |
| ENSG00000188375  | H3-5      | 2.3                                                                                               | 6.63E-35                                                                                           |
| ENSG000002241794 | SPRR2A    | 1.3                                                                                               | 5.16E-14                                                                                           |
| ENSG00000243649  | CFB       | 1.1                                                                                               | 3.99E-08                                                                                           |
| ENSG00000268104  | SLC6A14   | 1.1                                                                                               | 1.52E-20                                                                                           |
| ENSG00000276980  | AC008760.2| 1.1                                                                                               | 6.67E-09                                                                                           |

Figure 1. COVID-19 KnetMiner web application. A) Search result for gene list and search terms. B) Gene View: Ranked list of genes and evidence summaries. C) Map View: Interactive genomic map with related GWAS studies and top scoring genes. D) Evidence View: Enriched linked knowledge.
the international search for useful drugs, stop researchers repeating work done elsewhere, avoid harmful interventions, and ultimately, help pave the way to effective treatments. In an effort to maximize our data usefulness and grow the knowledge graph faster, we plan to offer mappings and conversions between our data and the KGX/Biolink ecosystem (Biolink, 2020). For instance, we are investigating methods to ingest the covid-19-kg produced by the Berkely Lab (Reese et al., 2020) and integrate it into the KnetMiner platform. This is work that will be conducted within the COVID-19 international research team (COV-IRT).

**Data availability**

**Underlying data**

COVID-19 KG: https://github.com/Rothamsted/covid19-kg/

RDF Endpoint: http://knetminer-data.cyverseuk.org/lodestar/sparql

Figure 2. Knowledge graph of 20 DEGs. Information from various sources are visualized in a single view. Users can add or hide information using the legend. In total, the network contains 1216 nodes and 1808 edges, of which 101 and 143 are shown respectively. For example, 2 out of 72 Trait concepts are visible and other linked Traits (GWAS studies) can simply be added by double-clicking the Trait symbol in the legend.
Neo4j Endpoint: http://knetminer-covid19.cyverseuk.org:7476/browser/

Extended data
Documentation and custom scripts are available from https://github.com/Rothamsted/covid19-kg (see README)
Archived scripts as at time of publication: https://doi.org/10.5281/zenodo.509469
License: AGPL-3.0

Software availability
KnetMiner web app: https://knetminer.com/COVID-19
ELIXIR bio.tools: https://bio.tools/covid-19_knetminer
Software available from: https://hub.docker.com/u/knetminer
Source code available from: https://github.com/Rothamsted/knetminer
Archived source code at time of publication (usually Zenodo): https://doi.org/10.5281/zenodo.3891097

License: AGPL-3.0

Acknowledgements
We are grateful to the organizers of COVID-19 Biohackathon and the participants of the Knowledge Graph topic, especially Justin Reese, Deepak Unni and the SciBite topic for their support with our queries. We would also like to thank the COVID-19 International Research Team (https://www.cov-irt.org/) for allowing us to present KnetMiner and for providing an excellent networking channel. We additionally thank CyVerse UK and acknowledge UKRI BBSRC support (BB/R000662/1) for providing us with resources to host our Neo4j and RDF COVID-19 knowledge graphs.

References
Biolink: Biolink Model. 2020. Reference Source
Brandizi M: The Power of Standardised and FAIR Knowledge Graphs. 2020. Reference Source
Brandizi M, Singh A, Rawlings C, et al.: Towards FAIRer Biological Knowledge Networks Using a Hybrid Linked Data and Graph Database Approach. J Integr Bioinform. 2018a; 15(3): 20180023. PubMed Abstract | Publisher Full Text | Free Full Text
Brandizi M, Singh A, Rawlings C, et al.: Getting the Best of Linked Data and Property Graphs: rdf2neo and the KnetMiner Use Case. 2018b. Publisher Full Text
Enes A, Pir P: Transcriptional response of signalling pathways to SARS-CoV-2 infection in normal human bronchial epithelial cells. bioRxiv. 2020; 2020.06.20.163006. Publisher Full Text
Giles O, Huntley R, Karlsson A, et al.: Reference ontology and database annotation of the COVID-19 Open Research Dataset (CORD-19). bioRxiv. 2020; 2020.10.04.325266. PubMed Full Text
Good BM, Ainscough BJ, McMichael JF, et al.: Organizing knowledge to enable personalization of medicine in cancer. Genome Biol. 2014; 15(8): 438. PubMed Abstract | Publisher Full Text | Free Full Text
Hassani-Pak K, Legare R, Canivet C, et al.: Enhancing Data Integration with Text Analysis to Find Proteins Implicated in Plant Stress Response. J Integr Bioinform. 2010; 7(3). PubMed Abstract | Publisher Full Text
Hassani-Pak K, Singh A, Brandizi M, et al.: KnetMiner: A Comprehensive Approach for Supporting Evidence-based Gene Discovery and Complex Trait Analysis across Species. Plant Biotechnol J. 2021. PubMed Abstract | Publisher Full Text
Kaggle: CORD-19. 2020. Reference Source
Köhler J, Baumbach J, Taubert J, et al.: Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics. 2006; 22(11): 1383–90. PubMed Abstract | Publisher Full Text
Reese J, Unni D, Callahan TJ, et al.: KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. bioRxiv. 2020; 2020.08.17.254839. PubMed Abstract | Publisher Full Text | Free Full Text
WHO: COVID-19 Situation Reports. 2020. Reference Source
Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 01 September 2021

https://doi.org/10.5256/f1000research.57521.r90852

© 2021 Hall K. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Keith Hall
Google Research, New York, NY, USA

- This article presents an automatically constructed and normalized Knowledge Graph of Covid-19 related literature and structured data sources. An existing tool KnetMiner is used to perform the ingestion and normalization of the data.

- It is not clear from the paper what the actual normalization algorithms are. When combining entities and relations from different sources, what is the algorithm which determines that two entities are actually the same entity? The same question applies to relations. While there are pointers to the available tool, without knowing the details, it's hard to determine the quality is of the created resource.

- An evaluation of the quality of the data returned from this resource would have been very helpful. This would fill in the gaps where we do not know the actual algorithm/models being used.

- I suspect the tool may be very useful, but without some notion of the quality of the data, it is hard to determine its utility. Without knowing the precision of the data, it would be difficult to rely on this resource for active research.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Natural Language Processing, Machine Learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 27 August 2021

https://doi.org/10.5256/f1000research.57521.r91031

© 2021 Reese J. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Justin Reese id
Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

The author describe a knowledge graph that integrates public datasets about COVID-19, as well as existing tooling to query and extract knowledge from the KG, i.e. Knetminer.

Transform code is publicly available on GitHub, and source data are well explained.

Database dumps are available on request, but I think these should be made available in some public place.

Knetminer code is also available on GitHub, and is usable under an MIT license that is permissive and favorable for academic use.

Authors do an excellent job of conforming to existing data standards (OXL, SPARQL/Neo4J endpoints, code to convert to Neo4J, etc) and demonstrating how to use their tools. Again though, I don't see why these RDF or other dumps should not be public at some stable URL.

The DEG use case the authors use to demonstrate their KG and tooling is compelling and interesting. The display components of this tool are particularly strong. For example, network output for results is a very expressive way of displaying results. Could this not be linked to some public instance of Knetminer, so readers can see this result for themselves, and interact with the network?

The authors should probably also compare their tools/KG with existing tools. They compare with Reese et al fairly well, but there are several other COVID-19 KG efforts that probably should be included in the comparison.
Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Knowledge graphs, machine learning, data science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com