COVIDGraph: Connecting biomedical COVID-19 resources and computational biology models

Lea Gütebier
University Medicine Greifswald
Greifswald, Germany
lea.guetebier@stud.uni-greifswald.de

Ron Henkel
University Medicine Greifswald
Greifswald, Germany
ron.henkel@uni-greifswald.de

Alexander Jarasch
German Center for Diabetes Research
Munich, Germany
jarasch@dzd-ev.de

Tim Bleimehl
German Center for Diabetes Research
Munich, Germany
tim.bleimehl@helmholtz-muenchen.de

Sebastian Müller
yWorks
Tübingen, Germany
sebastian.mueller@yworks.com

Jamie Munro
Munro Consulting
London, UK
jamie@munro.consulting

Martin Preusse, and the HealthEcco Team
Kaiser & Preusse
Freiburg, Germany
martin@kaiser-preusse.com

Dagmar Walthemath
University Medicine Greifswald
Greifswald, Germany
dagmar.waltemath@uni-greifswald.de

ABSTRACT

The COVID-19 pandemic has changed life across the globe. In January 2020, little was known about SARS-COV-2, but the vastly increasing number of infections and the uncontrolled spreading demanded fast medical action. Within a year, over 4 million publications relating to COVID-19 appeared in the scientific literature. Additionally, patents have been registered, ontologies have been extended, simulation studies for prediction of disease spread and underlying bioinformatics mechanisms have been built, and health studies have been designed. To support the exploration of COVID-19 data, the CovidGraph project was initiated as a non-profit, collaborative and open project driven by researchers, software developers, data scientists and medical professionals. In this article we outline the history, goals and scope of CovidGraph. Using the example of computational biology models, we show how additional resources can be integrated with the knowledge graph to extend the scope of the CovidGraph, for example, to systems biology data.

Reference Format:
Lea Gütebier, Ron Henkel, Alexander Jarasch, Tim Bleimehl, Sebastian Müller, Jamie Munro, Martin Preusse, and the HealthEcco Team, and Dagmar Walthemath. COVIDGraph: Connecting biomedical COVID-19 resources and computational biology models. In the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores (SEA Data 2021).

PVLDB Artifact Availability:
The source code, data, and/or other artefacts have been made available at https://github.com/covidgraph/documentation.

Published in the Proceedings of the 2nd Workshop on Search, Exploration, and Analysis in Heterogeneous Datastores, co-located with VLDB 2021 (August 16-20, 2021, Copenhagen, Denmark) on CEUR-WS.org.

1 INTRODUCTION

CovidGraph is a research and communication platform that encompasses publications, case statistics, genes and functions, molecular data and more. It is developed and maintained by HealthECCO, a non-profit collaboration of researchers, software developers, data scientists and medical professionals (https://healthecco.org/). Our aim is to help researchers quickly and efficiently find their way through COVID-19 datasets using tools that implement artificial intelligence methods, advanced visualisation techniques, and intuitive user interfaces. Through CovidGraph users can explore papers, patents, treatments and medications covering the family of corona viruses. In addition to literature data we connect information from biological entities - namely genes, proteins and their function - spanning a network of unparalleled size and knowledge. The latest addition to the CovidGraph are systems biology models (Fig. 1).

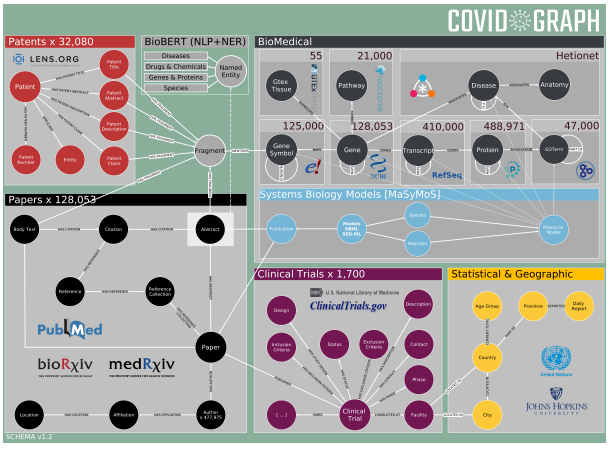


Figure 1: Overview: CovidGraph data sources with the integrated system biology nodes (cyan box).

Over the last years, NoSQL approaches such as Key-Value Stores, BigTable, document databases, triple stores, or graph databases [1], together with semantic web applications, became more popular within the life sciences. Graph databases offer a storage concept based on nodes, (directed) edges, properties and labels. Nodes can be labelled and are connected by edges, and both can contain properties. They also allow easy horizontal scaling and fast graph traversal. Finally, graph databases are schema optional – a feature that is much appreciated when storing heterogeneous, highly connected, cross-domain data items from different sources. The HealthECCO project integrates such heterogeneous resources and compiles a knowledge-base targeted at COVID-19 data (https://healthecco.org/covidgraph/), and potentially other diseases in future versions. The underlying graph database is Neo4j [18].

2 DATA RESOURCES

Previous versions of the CovidGraph already integrated data from five categories (Fig. 2 (A)): Patents, Papers, BioMedical (ontologies and controlled vocabularies), Clinical Trials and Statistical & Geographic. Categories are cross-linked by relationships. For example, items from the "Papers" category are linked to items from the "Patents" category. One paper source is the COVID-19 Open Research Dataset (CORD-19) – a collection of research papers relating to COVID-19 (and corona viruses) [24]. It is the main data source for information about papers in the CovidGraph and contains publications from PubMed, medRxiv and bioRxiv. Papers and related information are stored and linked in multiple nodes in the CovidGraph. Each paper node has author nodes connected to affiliation nodes that, in turn, are linked to location nodes. Papers can be linked to COVID-19 patents. The Lens (https://about.lens.org/covid-19/) provides datasets of patent documents and literature concerning human corona viruses and COVID-19. The CovidGraph furthermore contains information about clinical COVID-19 studies from the ClinicalTrials.gov registry. Studies are represented as clinical trials nodes which are linked to multiple other nodes representing more detailed information about each study. Also included in the CovidGraph are case statistics and case data from Johns Hopkins University [7] and population estimates from the United Nations World Population Prospects (https://population.un.org/wpp/). Nodes include city, country, province, daily report and age group. Biomedical data encodes information about genes, proteins, pathways and different diseases associated with COVID-19. The data comprises information from various biological and biomedical resources and is connected to Gene Ontology terms. The Gene Ontology is a resource for computational representation of the function of genes and gene products [4]. Information about genes from the NCBI Gene Database [2] is stored in Gene nodes which are connected to other nodes describing the underlying biology. Therefore, the connected nodes include Gene Symbols according to the Ensembl Genome Browser, a genome database [10]. The gene symbols are mapped to synonyms. Since genes are expressed in various tissues the gene nodes are linked to Gtex Tissue nodes containing gene expression data from the GTEx Portal [14]. For genes that are part of a pathway there exists a relation between the corresponding gene node and pathway node. The data included in the COVIDGraph describes which genes are members of a pathway according to the Reactome pathway knowledgebase, a database for molecular information about biological pathways [11]. As components of the transcription and translation process in humans genes code for transcripts which in turn code for proteins. In the CovidGraph these processes are described by relationships between gene nodes, transcript nodes and protein nodes. The data for the transcript nodes is taken from the NCBI Reference Sequence Database [17]; the Universal Protein Resource (UniProt) provides a resource of protein sequences and annotation data [5]. Proteins associated with annotation data from the Gene Ontology are linked to GO term nodes. The last node type connected with gene nodes are disease nodes. They are in turn associated with anatomy nodes. The corresponding data is provided by Hetionet, an integrative network of biomedical data including connections between diseases and anatomies [9].

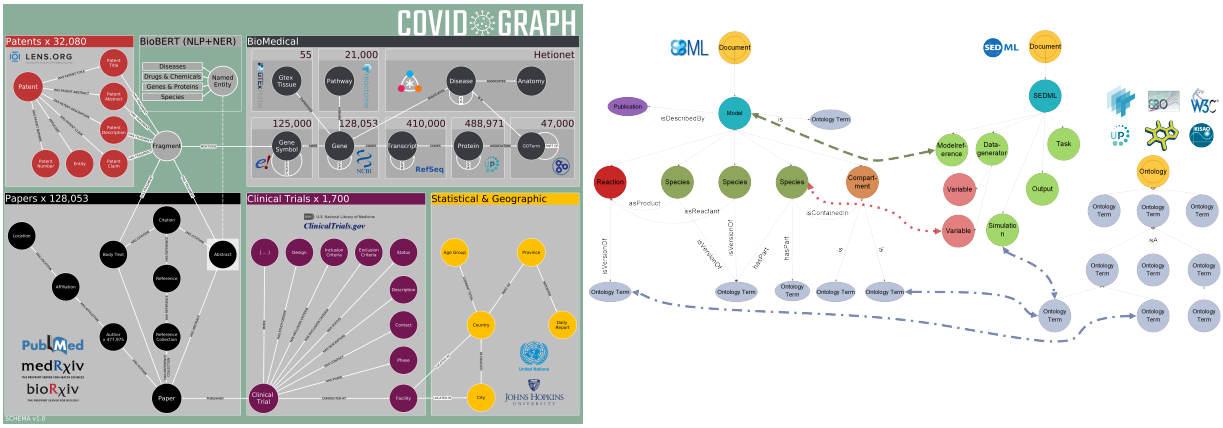
Knowledge is primarily centred around the domain of coronaviruses but is steadily extended to other connected diseases as part of the HealthECCO project. The latest addition to CovidGraph is a resource of computational biology models. We will introduce the systems biology node in detail in Section 4.

3 COVIDGRAPH FRAMEWORK

The CovidGraph infrastructure is built as a labelled property graph based on the Neo4j Enterprise edition v4.2. Textual information, such as publications, clinical studies or ontology term descriptions, is enriched and recognised by a pipeline based on natural language processing and named entity recognition (BioBERT [13]). The graph, as of now, contains 36 million nodes and 59 million relationships but is still growing as the modular software framework encourages to add and integrate new data sources. Server-wise, CovidGraph relies on Docker Container. To integrate a new data source, it needs to be wrapped in a container and it needs to provide information such as connection data and mapping information (https://github.com/covidgraph/data_template). An ETL-process (https://git.connect.dzd-ev.de/dzdtools/motherlode) subsequently extracts the data from the new source, transforms the data in accordance with the provided mapping information, and loads the data into the main CovidGraph.

4 INTEGRATION OF SIMULATION STUDIES

Via the aforementioned ETL-process, we connected the CovidGraph and the Management System for Models and Simulations (MaSyMoS, [8]). MaSyMoS is a Neo4j graph database for storing and retrieving data items describing biomedical simulation studies. The data is extracted from repositories for computational biology models (BioModels [15] and Physiome Model Repository2 [25]) and integrated in a single graph (Fig. 2 (B)). We consider a computational biology model a mathematical model written in a formal machine-readable language, such that it can be systematically parsed and employed by simulation and analysis software without further human translation [12]. A biomedical simulation study is considered any calculation performed on a model and describing evolution of the biological system represented, for instance, over spatial and/or temporal dimensions [23]. MaSyMoS links simulation studies, their results and corresponding models. Curated simulation studies are furthermore annotated with meta-data, primarily



(A) (B)

Figure 2: (A) Original CovidGraph data model with data from i) Patents, ii) an index for biomedical terms (BioBERT [13]), iii) BioMedical Ontologies [2, 4, 5, 9–11, 17, 21, 22], iv) COVID-19 related papers [3, 24], v) Clinical Trials [26]), vi) and a Statistical & Geographic information [7, 16]. (B) A simplified MaSyMoS [8] meta graph containing i) simulation models formerly encoded in SBML and CellML (not shown) [20], ii) simulation descriptions formerly encoded in SEDML [20], iii) bioontologies encoded in OWL, iv) and links to publications in PubMed.

reference publications and ontological terms from bio-ontologies [4–6, 11]. MaSyMoS provides access to over 1000 manually curated simulation studies originally published in BioModels. This set contains highly curated studies targeting COVID-19 disease and spreading (https://www.ebi.ac.uk/biomodels/covid-19). The resulting knowledge graph offers domain-specific retrieval and similarity measures, and it enables efficient access and reuse. As all model have been shown to reproduce the published results, they are a valuable resource for biomedical investigations.

The integration of MaSyMoS data with CovidGraph was twofolded: First we matched papers (publications) from both domains. Then we connected biomedical ontology terms from both resources thereby linking disease knowledge and biomedical simulation studies. The Paper data set (cmp. Fig. 2 (A)) in CovidGraph is represented by different nodes (e.g., the abstract, authors, paper ID). In MaSyMoS a paper is represented by a single publication node containing the same aforementioned set of information about a publication. Consequently, we mapped the corresponding IDs (PubMedID and DOI) from CovidGraph paper ID nodes and MaSyMoS publication nodes, thus connecting relevant publications from both data sets. This mapping resulted in 19 connections. This result is in our expected range, as the underlying publication corpus covers different areas of interest (e.g. cell cycle, MAPK and apoptosis for simulation models & clinical trials, respiratory studies and diseases for CovidGraph). The BioMedical data set in the CovidGraph represents different ontologies with relevance for COVID-19 research. These ontologies have possible connections and overlap with ontological terms used to annotate simulation studies in MaSyMoS (cmp. Figure 2 (B)). Our analyses showed that most overlap can be observed in gene information, chemical entities, proteins and diseases. Consequently, we mapped ontological terms in MaSyMoS and CovidGraph for Gene Ontology (1810 connections), ChEBI (1211 connections), UniProt (911 connections) and Disease Ontology (72 connections) by their IDs (cmp. Figure 1). For Gene Ontology, ChEBI and Disease Ontology more than 94% of the terms stored in MaSyMoS were connected to terms in the CovidGraph. The UniProt coverage reached 41%.

Example: COVID-19 spread in Wuhan city. The simulation study by Roda at al. [19] investigates the COVID-19 spread in Wuhan city in the beginning of 2020. Figure 3 shows a Neo4j excerpt of the model in MaSyMoS and the association to disease information in the CovidGraph. The association is build by a matching reference publication and a matching ontology entry from the Disease Ontology. More specifically, the model is linked (in the middle, dark green) to several resources (pink). For example, one annotation refers to an ontology term from the Disease Ontology and is associated to the corresponding entry in the CovidGraph (on the right, brown). Another example is the reference publication which links to the corresponding publication in the CovidGraph (on the right, blue). We consider this example a first step towards bridging the gap between medical research and systems biology.

5 TAKEAWAYS & FUTURE WORK

The CovidGraph project integrates COVID-related data from heterogeneous data sources, mainly from the medial and health domains, into a single knowledge graph. We demonstrate that even for fairly distinct scientific domains such as computational biology modeling and clinical research, it is possible to link knowledge graphs and thereby quickly provide new data sources. The presented version of CovidGraph provides a tool set and a single-access point to previously disconnected data sources. Biomedical and clinician scientists can explore a rich set of data items, which are not connected in any other resource. CovidGraph is only one example for rapid integration of knowledge. The HealthECCO infrastructure offers solutions for integration and exploration of other diseases, building on the same integration workflow showcased in this paper.

Figure 3: Simulation study by Roda at al. [19] represented in MaSyMoS (model in light blue) with links to CovidGraph.

The CovidGraph-Team hopes to motivate other data providers to link up with our resource, but we also like to discuss the applicability of our graph database infrastructure on existing data silos.

ACKNOWLEDGMENTS

The work presented here is the result of the HealthEcco Team (https://healthecco.org/team/). The COVID-19 collection in BioModels was built with the help of an EOSC COVID-19 Fast Track funding.

REFERENCES

[1] Renzo Angles and Claudio Gutierrez. 2008. Survey of graph database models. *ACM Computing Surveys (CSUR)* 40, 1 (2008), 1.

[2] Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, et al. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic acids research* 43, D1 (2015), D36–D42. https://doi.org/10.1093/nar/gku1055

[3] Kathi Canese and Sarah Weis. 2013. PubMed: the bibliographic database. *The NCBI Handbook* 2 (2013), 1.

[4] The Gene Ontology Consortium. 2021. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research* 49, D1 (2021), D325–D334. https://doi.org/10.1093/nar/gkaa1113

[5] UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* 47, D1 (2019), D506–D515. https://doi.org/10.1093/nar/gky1049

[6] Paula de Matos, Adriano Dekker, Marcus Ennis, Janna Hastings, Kenneth Haug, Steve Turner, and Christoph Steinbeck. 2010. ChEBI: a chemistry ontology and database. *Journal of cheminformatics* 2, 1 (2010), 1–1.

[7] Ensheng Dong, Hongru Du, and Lauren Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 20, 5 (2020), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

[8] Ron Henkel, Olaf Wolkenhauer, and Dagmar Waltemath. 2015. Combining computational models, semantic annotations and simulation experiments in a graph database. *Database* 2015 (2015), bau130.

[9] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* 6 (Sept. 2017), e26726. https://doi.org/10.7554/elife.26726

[10] Tim Hubbard, Daniel Barker, Ewan Birney, Graham Cameron, Yuan Chen, L Clark, Tony Cox, J Cuff, Val Curwen, Thomas Down, et al. 2002. The Ensembl genome database project. *Nucleic acids research* 30, 1 (2002), 38–41. https://doi.org/10.1093/nar/30.1.38

[11] Bijay Jassal, Lisa Matthews, Guilherme Viteri, Chuqiao Gong, Pascual Lorente, Antonio Fabregat, Konstantinos Sidiropoulos, Justin Cook, Marc Gillespie, Robin Haw, et al. 2020. The reactome pathway knowledgebase. *Nucleic acids research* 48, D1 (2020), D498–D503. https://doi.org/10.1093/nar/gkz1031

[12] Nicolas Le Novère, Andrew Finney, Michael Hucka, Upinder S Bhalla, Fabien Campagne, Julio Collado-Vides, Edmund J Crampin, Matt Halstead, Edda Klipp, Pedro Mendes, et al. 2005. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature biotechnology* 23, 12 (2005), 1509–1515.

[13] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[14] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. 2013. The genotype-tissue expression (GTEx) project. *Nature genetics* 45, 6 (2013), 580–585. https://doi.org/10.1038/ng.2653

[15] Rahuman S Malik-Sheriff, Mihai Glont, Tung VN Nguyen, Krishna Tiwari, Matthew G Roberts, Ashley Xavier, Manh T Vu, Jinghao Men, Matthieu Maire, Sarubini Kananathan, et al. 2020. BioModels—15 years of sharing computational models in life science. *Nucleic acids research* 48, D1 (2020), D407–D415.

[16] United Nations. 2019. World population prospects 2019: highlights.

[17] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* 35, suppl_1 (2007), D61–D65. https://doi.org/10.1093/nar/gki025

[18] Ian Robinson, Jim Webber, and Emil Eifrem. 2013. *Graph Databases*. O'Reilly Media, CA, USA.

[19] Weston C Roda, Marie B Varughese, Donglin Han, and Michael Y Li. 2020. Why is it difficult to accurately predict the COVID-19 epidemic? *Infectious Disease Modelling* 5 (2020), 271–281.

[20] Falk Schreiber, Björn Sommer, Tobias Czauderna, Martin Golebiewski, Thomas E Gorochowski, Michael Hucka, Sarah M Keating, Matthias König, Chris Myers, David Nickerson, et al. 2020. Specifications of standards in systems and synthetic biology: status and developments in 2020. *Journal of integrative bioinformatics* 17, 2-3 (2020).

[21] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. 2012. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research* 40, D1 (2012), D940–D946.

[22] The GTEx Portal. 2020. GTEx Portal Documentation. https://gtexportal.org/home/documentationPage. Online, accessed 12 October 2020.

[23] Dagmar Waltemath, Richard Adams, Daniel A Beard, Frank T Bergmann, Upinder S Bhalla, Randall Britten, Vijayalakshmi Chelliah, Michael T Cooling, Jonathan Cooper, Edmund J Crampin, et al. 2011. Minimum information about a simulation experiment (MIASE). *PLoS computational biology* 7, 4 (2011), e1001122.

[24] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020. Cord-19: The covid-19 open research dataset. *ArXiv* arXiv2004. (2020), 10706v2.

[25] Tommy Yu, Catherine M Lloyd, David P Nickerson, Michael T Cooling, Andrew K Miller, Alan Garny, Jonna R Terkildsen, James Lawson, Randall D Britten, Peter J Hunter, et al. 2011. The physiome model repository 2. *Bioinformatics* 27, 5 (2011), 743–744.

[26] Deborah A Zarin, Tony Tse, Rebecca J Williams, Robert M Califf, and Nicholas C Ide. 2011. The ClinicalTrials.gov results database - update and key issues. *New England Journal of Medicine* 364, 9 (2011), 852–860.