

OPEN
COMMENT

CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis

Yongqun He¹✉, Hong Yu^{2,3}, Edison Ong¹, Yang Wang^{1,2,3}, Yingtong Liu¹, Anthony Huffman¹, Hsin-hui Huang^{1,4}, John Beverley⁵, Junguk Hur⁶, Xiaolin Yang⁷, Luonan Chen^{8,9}, Gilbert S. Omenn¹, Brian Athey¹ & Barry Smith¹⁰

The Coronavirus Infectious Disease Ontology (CIDO) is a community-based ontology that supports coronavirus disease knowledge and data standardization, integration, sharing, and analysis.

Ontologies, as the term is used in informatics, are structured vocabularies comprised of human- and computer-interpretable terms and relations that represent entities and relationships. Within informatics fields, ontologies play an important role in knowledge and data standardization, representation, integration, sharing and analysis. They have also become a foundation of artificial intelligence (AI) research. In what follows, we outline the Coronavirus Infectious Disease Ontology (CIDO), which covers multiple areas in the domain of coronavirus diseases, including etiology, transmission, epidemiology, pathogenesis, diagnosis, prevention, and treatment. We emphasize CIDO development relevant to COVID-19.

Human coronaviruses have given rise to a series of major crises in global public health. Severe acute respiratory syndrome (SARS) emerged in China in November 2002, lasted for eight months and resulted in 8,098 confirmed human cases in 29 countries with 774 deaths (case-fatality rate: 9.6%)¹. Approximately ten years later in June 2012, the Middle East Respiratory Syndrome (MERS), another highly pathogenic coronavirus disease, was identified in Saudi Arabia. The MERS outbreak has caused 2,260 cases in 27 countries and 803 deaths (35.5%)². More recently, the World Health Organization (WHO) declared the Coronavirus Disease 2019 (COVID-19) outbreak as a pandemic on March 11, 2020, when there were 118,326 confirmed cases and 4,292 deaths. As of May 13, there have been over 4.4 million confirmed cases and over 295,000 deaths globally. Unfortunately, we still do not have available effective drugs and vaccines against these highly pathological coronaviruses.

Extensive studies have been conducted on coronaviruses, the results of many of which exist in publicly available data repositories such as GEO³. Publications concerning COVID-19 have exploded in recent months, and new clinical trials have been and are being conducted to develop drugs and vaccines against COVID-19, 1,430 of which have been registered in ClinicalTrials.gov as of May 13, 2020. As of May 13, 2020, a PubMed search of “SARS”, “MERS”, and “SARS-CoV-2 OR COVID-19” resulted in 12,993, 4,493 and 11,813 publications, respectively. A coordinated study of all such results would likely help with understanding and developing treatments for COVID-19. This coordinated study requires the integration of the large and exponentially growing data and research concerning COVID-19 to better understand its etiology, transmission, and pathogenesis mechanism. Moreover, we must be able to translate that understanding into rapid development of patient stratification

¹University of Michigan Medical School, Ann Arbor, MI, 48109, USA. ²People's Hospital of Guizhou Province, Guiyang, Guizhou, 550002, China. ³Guizhou University Medical College, Guiyang, Guizhou, 550025, China. ⁴National Yang-Ming University, Taipei, 112-21, Taiwan. ⁵Northwestern University, Evanston, IL, 60208, USA. ⁶University of North Dakota School of Medicine and Health Sciences, Grand Forks, ND, 58203, USA. ⁷Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences (CAMS) & School of Basic Medicine, Peking Union Medical College (PUMC), Beijing, China. ⁸Key Laboratory of Systems Biology, Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, 200031, China. ⁹Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, 650223, China. ¹⁰University at Buffalo, Buffalo, NY, 14260, USA. ✉e-mail: yongqunh@med.umich.edu

methods leveraging precision medicine, therapeutic drugs, and preventive vaccines. However, there are two bottlenecks to achieving these tasks:

First, the characteristic five V's of our Big Data⁴ era lead to disintegrated and non-interoperable data and knowledge. The amount of data (volume), speed at which it is produced (velocity), range of its sources (variety), quality and accuracy (veracity), and assessment of utility (value), result in large, complex, multidimensional, and diverse datasets. Disintegrated and non-interoperable data cannot be interpreted by computers and this inhibits computer-assisted reasoning, which is the essence of artificial intelligence. Consequently, our knowledge – data and information that embodies awareness and understanding – of domains represented by various datasets is seriously hindered. This is a familiar problem for biomedical research in general, which relies heavily on data acquisition, and for coronavirus research in particular, given the global challenge we currently face. The second bottleneck is the lack of bioinformatics tools that can efficiently and robustly integrate and analyze heterogeneous data and knowledge. This is likely a major stumbling block that is slowing the discovery of effective measures against coronaviruses even despite extensive effort across the globe.

A critical key to data/information/knowledge disintegration and big data analysis is ontologies. Ontologies are widely used in biomedical data and metadata standardization, and robustly support data integration, sharing, reproducibility, and computer-assisted data analysis. Ontologies are also regarded as the foundation of knowledge representation and reasoning (KR², KR&R), a major field of artificial intelligence. An important biomedical example is the Gene Ontology (GO)⁵, which was originally developed in the late 1990s by researchers studying the genomes of three model organisms: fruit fly, mouse, and yeast (*Saccharomyces cerevisiae*), but later extended to provide terms and relations used to annotate genes from humans, plants, animals, and microbes. GO includes three branches, for *cellular components*, *molecular functions*, and *biological processes*, forming a controlled vocabulary that can be used to represent attributes of gene products in a species-neutral way. Many GO-based tools and algorithms have been developed⁶. Since publication in 2000, the original GO paper⁵ has been cited some 25,000 times. GO makes possible consistent and reproducible annotations and analyses of genes and genomes from and across different organisms.

The success of GO inspired the development of hundreds of biomedical ontologies over the past two decades⁷. Included among those ontologies are many relevant to COVID-19: the Disease Ontology (DOID)⁸ classifies 18,000 human diseases now including COVID-19; the Human Phenotype Ontology (HPO)⁹ defines 26,000 human phenotypes; the Chemical Entities of Biological Interest ontology (ChEBI)¹⁰ includes 135,000 chemical entities; the Ontology for General Medical Sciences (OGMS) includes 200 terms for general medical classification; and the Ontology for Biomedical Investigations (OBI)¹¹ includes nearly 4,000 terms related to all aspects of biomedical investigations. All these ontologies can be applied to the study of coronavirus diseases. Of most relevance to us here is the Infectious Disease Ontology (IDO)¹², which defines 550 terms relating to infectious diseases in general and provides a basis for more specific IDO ontologies, for flu, malaria, brucellosis, and other diseases.

Ontology openness and interoperability are critical for data sharing and integration. The FAIR Guiding Principles propose that all research data should be Findable, Accessible, Interoperable and Reusable (FAIR) for both machine and human users¹³. “Interoperability” is the basis of the four FAIRness principles, which see ontology interoperability as the foundation of data/information/knowledge interoperability. With hundreds of ontologies developed, many ontologies overlap each other but, unfortunately, are not interoperable. Many ontologies, through lack of interoperability with other, more widely used ontologies, form silos and thereby fail to support integrative research. To foster interoperability, the Open Biomedical and Biological Ontologies (OBO) Foundry was initiated in 2007 by ontology developers who agreed to adopt a set of principles – including the commitment to collaboration and openness, use of definitions in both human- and computer-readable formats – specifying best practices in ontology development¹⁴. The OBO ontology library includes approximately 200 ontologies (including GO).

To meet the challenge of COVID-19, we recently initiated the development of CIDO, a community-driven open-source biomedical ontology in the area of coronavirus infectious disease (<https://github.com/CIDO-ontology/cido>). CIDO provides standardized human- and computer-interpretable annotation and representation of various coronavirus infectious diseases, including their etiology, transmission, epidemiology, pathogenesis, host-coronavirus interactions, diagnosis, prevention, and treatment. CIDO will be used as a state-of-the-art knowledge base for standard and logical representation of heterogeneous coronavirus knowledge. Having been accepted as an OBO library ontology, CIDO follows the OBO Foundry principles¹⁴, and uses an OBO-compatible extensible ontology development strategy¹⁵. To support data interoperability, CIDO reuses relevant coronavirus terms from existing reliable reference ontologies themselves aligning with OBO Foundry principles, and aligns these terms under the Basic Formal Ontology (BFO)¹⁶, an ISO/IEC standard 21838-2 (<https://www.iso.org/standard/74572.html>) top-level ontology. BFO is a realism-based ontology that covers all domains by providing highly general ontology classes such as material entity, process, role, site, and so forth. By using BFO as its upper-level architecture, CIDO is automatically interoperable and integrated with >300 other ontologies that also align with BFO. Currently, CIDO contains over 4,000 terms, imported from some 20 further ontologies such as ChEBI, Human Phenotype Ontology, Disease Ontology⁸, and the NCBI taxonomy ontology (NCBITaxon). Additionally, new CIDO-specific terms have been developed to meet the special needs arising in the research of COVID-19 and other coronavirus diseases.

Developing robust ontologies adequate for representing complex domains requires more than the simple construction of taxonomies. Taxonomies reflect important hierarchical relationships among class and subclass terms, represented using *is_a* relations. For example, instances of coronavirus are instances of viruses, which is to say coronavirus *is_a* virus. Extending beyond taxonomies, ontologies provide additional relations among entities within and across domains. Figure 1 illustrates a general design of how we can logically link ontology terms that may come from different branches of CIDO. The relations along the arrows in Fig. 1 are computer-understandable

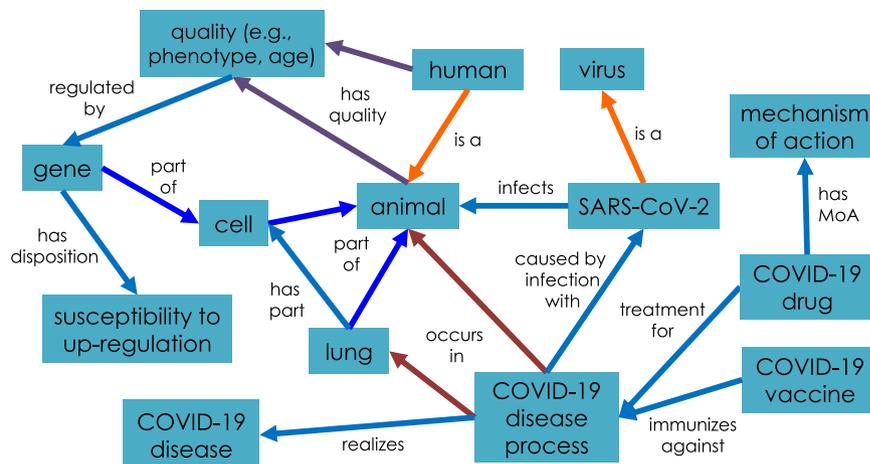


Fig. 1 The design pattern of CIDO for logically representing and linking different components related to a coronavirus disease, e.g., COVID-19. The terms presented in the figure are generated in CIDO or imported by CIDO from other ontologies. To reduce complexity, the ontology sources of the terms are not labeled.

links between ontology classes. For example, we can define a logical axiom using the ‘caused by’ relation to link the COVID-19 disease process and SARS-CoV-2 virus:

COVID-19 disease process: ‘caused by infection with’ some SARS-CoV-2

Such an axiom defines the causal relation between the COVID-19 disease process and the virus strain SARS-CoV-2. More specifically, the development of the COVID-19 disease process (or realization of the COVID-19 disease disposition) in a patient is causally induced by an infection of the SARS-CoV-2 virus in the patient, involving viral invasion of and replication in host cells. In CIDO, diseases such as COVID-19 and pathogens such as SARS-CoV-2 have distinct hierarchies, with relations linking terms in these hierarchies. The inclusion of the relations (e.g., ‘caused by infection with’) in addition to *is_a* greatly expands expressiveness, reasoning capabilities, and expected inferences.

Figure 1 illustrates many other key relations. Particularly, COVID-19 occurs in the lung, and some genes in the cells of the lung would have the disposition of being susceptible up- or down-regulated in the cells of SARS-CoV-2-infected lung. Such genes may function as gene markers and play important roles in pathogenesis. In addition, the infected patient will display different phenotypes after manifesting the disease, and such phenotypes may be associated with other patient attributes (e.g., biological sex, age) and the patient’s gene profile. CIDO thus provides semantically interoperable representations of host-coronavirus interaction mechanisms. Although Fig. 1 provides only a high-level overview of some CIDO resources, more details, such as specific signature genes in some cells of the lung that are susceptible to be up- or down-regulated in patients with COVID-19 will be added to the CIDO as new knowledge is acquired. Such systematic modeling and representation of the host-coronavirus interaction mechanisms would facilitate rational design of anti-coronavirus drugs and vaccines^{17,18}.

In pursuit of that aim, CIDO can logically define relations between drugs and roles or mechanisms of action – distinct hierarchies in CIDO – and so support advanced analysis of potential drugs used to treat COVID-19, as well as the quick query of drugs having specific roles or mechanisms of action potentially useful as treatments. Such application of CIDO for ontology-based integration and analysis of anti-coronavirus drugs is shown in our recent preprint paper¹⁷. Using literature mining we identified 72 chemical drugs and 27 monoclonal or polyclonal antibodies that have anti-coronavirus effects in experimental studies *in vivo* or *in vitro*. Many of these drugs were mapped to three ontologies: Chemical Entities of Biological Interest ontology (ChEBI)¹⁰, National Drug File – Reference Terminology (NDF-RT)¹⁹, and the Drug Ontology (DrON)²⁰. The subbranches of these ontologies that contain the mapped drugs and their related characteristics were extracted using the Ontofox tool²¹. Key information was identified by examining these subbranches. For example, based on their ChEBI annotations, many drug active ingredients are classified under the same chemical group: for example, chlorpromazine, dasatinib, terconazole, and chloroquine, all organochlorine compounds. Meanwhile, ChEBI classifies many drug chemicals having the same roles: chloroquine, conessine, lycorine, and mefloquine, all exhibit antimalarial activity. A ChEBI-based semantic similarity calculation method clustered 60 drugs into five major categories. The chemical information in ChEBI has also been imported to DrON. Developed by the U.S. Department of Veterans Affairs, Veterans Health Administration (VHA), NDF-RT organizes drugs by means of a formal representation of various drug characteristics such as mechanism of action (MoA), physiologic effect, and related diseases¹⁹. Using NDF-RT, we found that, of 35 drugs that have MoA annotations, 34 have MoAs of various inhibitors and antagonists. One shortcoming is that none of these ontologies covers all the needed information pertaining to our identified drugs. To study the anti-coronavirus drugs in a thorough manner we will need to identify and ontologically represent missing information of the sort that falls under the domain of the CIDO ontology. Thus, we plan to build logical relations linking drugs, coronaviruses, and the conditions under which the drugs work against the coronaviruses.

Another example of our ongoing work is the use of CIDO for the representation of vaccines against coronaviruses. We recently released another preprint paper on COVID-19 vaccine design using reverse vaccinology and machine learning¹⁸. Data pertaining to experimentally verified vaccine candidates in laboratory animal models

have also been collected and annotated¹⁸. We will systematically annotate these vaccine candidates, including their formulations and host responses, and work with the Vaccine Ontology (VO) development team to model, represent, and analyze these vaccines (<http://www.violinet.org/vaccineontology/>). Moreover, CIDO can be used alongside VO to support literature mining of vaccine-associated gene-gene interactions²².

In future research, we will use ontology-based approaches to investigate relevant host-coronavirus interactions to support the fundamental understanding of the disease and protective immune system mechanisms, precision medicine research, and rational vaccine design^{23,24}. More broadly, CIDO will provide community-based metadata standardization for interoperable and reproducible clinical and experimental studies, insofar as the coronavirus metadata ontology will be extracted from CIDO as a lightweight and relatively independent ontology to support data integration and knowledge discovery. We will use the standard to analyze clinical and basic research data and align the identified disease phenotype and transmission data with the underlying mechanisms introduced in other aims.

The extensive study of COVID-19 is generating new knowledge quickly. Given that our understanding of COVID-19 is changing rapidly, we need not only an ontology as wide in purview as CIDO, but also to be readily updated in a timely matter as more knowledge is generated about the disease, the virus, and the host response. To that end, we follow OBO Foundry guidelines for term requests and updating via issue tracking on the CIDO GitHub site (<https://github.com/CIDO-ontology/cido>). We welcome wide community participation in CIDO development and applications. CIDO is an open community; everyone is welcome. We are already collaborating with many groups and look forward to more collaborations with colleagues and people around the world. Starting with CIDO and its supported data integration, we expect that innovative computational and statistical algorithms and tools will be developed and applied to support basic studies of mechanisms, and translational applications such as predicting drugs and vaccines that hold promise for treating and preventing COVID-19.

Received: 15 April 2020; Accepted: 19 May 2020;

Published online: 12 June 2020

References

- Xu, R. Chance missed, but still there! Memoirs at the 10(th) anniversary of 2003 SARS outbreak. *J. Thorac. Dis.* **5**(Suppl 2), S90–93 (2013).
- Bernard-Stoecklin, S. *et al.* Comparative Analysis of Eleven Healthcare-Associated Outbreaks of Middle East Respiratory Syndrome Coronavirus (Mers-Cov) from 2015 to 2017. *Sci. Rep.* **9**, 7385 (2019).
- Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–995 (2013).
- Higdon, R. *et al.* Unraveling the Complexities of Life Sciences Data. *Big Data* **1**, 42–50 (2013).
- Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
- du Plessis, L., Skunca, N. & Dessimoz, C. The what, where, how and why of gene ontology—a primer for bioinformaticians. *Brief. Bioinform.* **12**, 723–735 (2011).
- Whetzel, P. L. *et al.* BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **39**, W541–545 (2011).
- Kibbe, W. A. *et al.* Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* **43**, D1071–1078 (2015).
- Groza, T. *et al.* The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *Am. J. Hum. Genet.* **97**, 111–124 (2015).
- Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–1219 (2016).
- Bandrowski, A. *et al.* The Ontology for Biomedical Investigations. *PLoS One* **11**, e0154556 (2016).
- Babcock, S., Beverley, J., Cowell, L. G. & Smith, B. The Infectious Disease Ontology in the Age of COVID-19. Preprint at, <https://doi.org/10.31219/osf.io/az6u5> (2020).
- Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
- Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251–1255 (2007).
- He, Y. *et al.* The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. *J. Biomed. Semant.* **9**, 3 (2018).
- Arp, R., Smith, B. & Spear, A. D. *Building Ontologies with Basic Formal Ontology*. (Cambridge, MA, USA, 2015).
- Liu, Y. *et al.* Ontological and bioinformatic analysis of anti-coronavirus drugs and their Implication for drug repurposing against COVID-19. Preprint at, <https://doi.org/10.20944/preprints202003.0413.v1> (2020).
- Ong, E., Wong, M. U., Huffman, A. & He, Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. Preprint at, <https://doi.org/10.1101/2020.03.20.000141> (2020).
- Carter, J. S. *et al.* Categorical information in pharmaceutical terminologies. In *AMIA Annu. Symp. Proc.* 116–120 (2006).
- Hogan, W. R. *et al.* Therapeutic indications and other use-case-driven updates in the drug ontology: anti-malarials, anti-hypertensives, opioid analgesics, and a large term request. *J. Biomed. Semant.* **8**, 10 (2017).
- Xiang, Z., Courtot, M., Brinkman, R. R., Ruttenberg, A. & He, Y. OntoFox: web-based support for ontology reuse. *BMC Res. Notes* **3**, 175 (2010).
- Ozgun, A., Xiang, Z., Radev, D. R. & He, Y. Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology. *J. Biomed. Semant.* **2**(Suppl 2), S8 (2011).
- Hoehndorf, R., Dumontier, M. & Gkoutos, G. V. Evaluation of research in biomedical ontologies. *Brief. Bioinform.* **14**, 696–712 (2013).
- Haendel, M. A., Chute, C. G. & Robinson, P. N. Classification, Ontology, and Precision. *Medicine. N. Engl. J. Med.* **379**, 1452–1462 (2018).

Acknowledgements

This project is supported by NIH grants U24CA210967 and P30ES017885 (to GSO); R01GM080646, 1UL1TR001412, 1U24CA199374, and 1T15LM012495 (to B.S.); 1UH2AI132931 (to Y.H.); the non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences 2019PT320003 (to H.Y.); and University of Michigan Medical School Global Reach award (to Y.H.).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020