The COVID-19 Infectious Disease Ontology

John Beverley[1,2*], Shane Babcock[2,3], Lindsay Cowell[4], Barry Smith[2,5]

[1]Department of Philosophy, Northwestern University, Evanston, IL, USA
[2]National Center for Ontological Research
[3]Department of Philosophy, Niagara University, Lewiston, NY, USA
[4]University of Texas Southwestern Medical School, Dallas TX, USA
[5]Department of Philosophy, University at Buffalo, Buffalo, NY, USA

ABSTRACT

Summary: The Infectious Disease Ontology (IDO) Core - which represents terminological content common to investigations of infectious diseases – has long-provided a foundation for ontological extensions to specific infectious diseases. The growing array of virus specific extensions has created a need for the construction of a reference ontology comprised of content common to investigations of viral infectious diseases. The present pandemic, moreover, has created the need for a COVID-19 specific ontology that is semantically integrated with ontologies representing geography, healthcare policy, and biomedical domains. We report the development of two ontologies designed to meet these needs: The Virus Infectious Disease Ontology and the COVID-19 Infectious Disease Ontology.
Availability: Each ontology is available under the Creative Commons Attribution 4.0 license (https://creativecommons.org/licenses/by/4.0/) and up-to-date versions of each are found at the National Center for Biomedical Ontology Bioportal (https://bioportal.bioontology.org/), the Ontobee repository (http://www.ontobee.org/), and the Ontology Lookup Service (https://www.ebi.ac.uk/ols/index). Working versions available on GitHub (https://github.com/infectious-disease-ontology-extensions).
Contact: johnbeverley2021@u.northwestern.edu

1 INTRODUCTION

The COVID-19 pandemic prompted immense investigation of the SARS-CoV-2 virus. Rapidly, accurately, and easily interpreting generated data is of fundamental concern. This concern is a species of the general need for data interoperability during the contemporary era of Big Biomedical Data.

Ontologies – structured, controlled, vocabularies – support interoperability, and prevent the development of data silos [1] which undermine interoperability. The Open Biological and Biomedical Ontology (OBO) Foundry serves to ensure ontologies remain interoperable through adherence by its members to core ontology design principles [2]. OBO principles require ontologies have textual definitions for terms and relations, as well as logical axioms expressed in the OWL 2 Web Ontology Language (https://www.w3.org/TR/owl2-overview/), a World Wide Web Consortium (https://www.w3.org/) language developed for the semantic web. OBO ontologies extend from a common top-level architecture provided by the Basic Formal Ontology (BFO), an ISO/IEC approved standard 21838-2 [3], which provides general terms designed to serve as starting points for definitions in all scientific domains. Ontologies covering these specific domains extend from this starting point following a "hub and spoke" design approach; for example, the Infectious Disease Ontology (IDO) Core [4] is comprised of terminological content common to investigations of all infectious diseases. Ontologies covering more specific infectious diseases, in turn, extend from IDO, such as the Coronavirus Infectious Disease Ontology (CIDO) [5].

The growing list of virus specific IDO extensions has motivated construction of a hub covering content common to viral infectious disease investigations: the Virus Infectious Disease Ontology (VIDO) (https://bioportal.bioontology.org/ontologies/VIDO). Additionally, the present pandemic has motivated construction of a more specific extension of CIDO, covering terminological content specific to the pandemic: the COVID-19 Infectious Disease Ontology (IDO-COVID-19)

(https://bioportal.bioontology.org/ontologies/IDO-COVID-19). While IDO-COVID-19 is not the only COVID-19 specific ontology [6-9], existing ontologies are stand-alone initiatives, not obviously built in a principled manner, and not obviously interoperable with OBO library ontologies. Both VIDO and IDO-COVID-19 were developed in collaboration not only with relevant domain experts - such as immunologists and virologists – but also in collaboration with the developers of IDO. As such, each ontology aligns with principles outlined by the OBO Foundry, supporting interoperability with existing Foundry ontologies. **Figure 1** illustrates relationships among relevant OBO ontologies.

2 SPECIFICATIONS

VIDO resolves common ontological problems observed in ontologies representing viruses. For example, many OBO ontologies treat viruses as a subclass of *organism* yet define instances of *organism* as cellular. Viruses, however, are acellular. Some ontologists have suggested using the class *organism or virus or viroid* from the Common Anatomy Reference Ontology to avoid this issue [10]. Disjunctive classes, however, suggest class closure, and closure should be avoided in life sciences when possible. More worrisome, placing viruses in a class next to paradigmatic living entities, leads naturally to reflection on whether viruses are alive, an intractable question ontologists do not need to answer. To avoid these issues, VIDO and IDO developers collaborated to introduce:

acellular structure =def Object consisting of an arrangement of interrelated acellular parts forming an acellular biological unit.

To IDO as a sibling class of *organism*. In VIDO, this class is then treated as the parent of:

virus =def Acellular structure with RNA or DNA genetic material which uses host metabolic resources for RNA or DNA replication.

The above definition was developed to align with the Baltimore Classification [11], an exhaustive characterization of viruses in terms of genetic components. This stands in contrast to representing viruses using large, unwieldy, Linnean taxonomies found in some viral infectious disease ontologies, such as IDOSCHISTO [12-13]. These remarks provide a mere taste of what VIDO offers, further details of which are illustrated in **Figure 1** and can be found here [14].

Where VIDO is constructed as a hub from which spoke ontologies like CIDO extend, CIDO in turn acts as a hub for extension spokes like IDO-COVID-19. The latter ontology provides COVID-19 specific terms such as:

SARS-CoV-2 pathogenesis =def Coronavirus pathogenesis process which is the realization of an infectious disposition inhering in SARS-CoV-2, having at least process parts, SARS-CoV-2: (1) transmission, (2) localization in host, (3) establishment of infection, and (4) establishment of disorder.

Providing a bridge to virus replication stage terms imported from the widely-used Gene Ontology [15]. Clauses (1)-(4) in the preceding definition, moreover, reflect terms in IDO-COVID-19 representing sub-processes of pathogenesis, and so provide targets for researchers involved in rational drug design.

As another example, given the importance of asymptomatic carriers to infection spread, IDO-COVID-19 includes:

subclinical SARS-CoV-2 infection =def Infection by SARS-CoV-2 that is part of an asymptomatic host.

Where "asymptomatic" host is explicated by importing terms from OBO ontologies such as the Ontology of General Medical Science (OGMS). Terms such as the preceding are crucial in representing COVID-19 epidemiological and transmission data reflecting for example "super-spreader events".

3 APPLICATIONS

VIDO, CIDO, and IDO-COVID-19 are being used to annotate approximately 400 articles in the National Library of Medicine (https://www.nlm.nih.gov/) COVID corpus, which report COVID-19 clinical

trial, epidemiological, and pathogenesis data. The resulting gold standard will be used to train algorithms for automated annotating tasks, which will, in turn, be used to identify patterns in COVID-19 datasets. We illustrate with two examples, where colored words below correspond to colors representing OBO ontologies in **Figure 1**.

First, from an article in the *Lancet* [16]:

> Viral loads in throat and sputum specimens peaked 5-6 days after symptom onset, ranging from 10^4 to 10^7 copies per mL.

This color-coding illustrates VIDO's reuse of terms from OBO library ontologies, such as BFO, OGMS, IDO, the Ontology for Biomedical Investigations (OBI), the Uber-Anatomy Ontology (UBERON), the Information Artifact Ontology (IAO), and the Common Core Ontology (CCO) which supplies measurement and diagnostic terms. Also illustrated are terms that could not be imported, such as *viral load*, but were instead developed for VIDO. "Viral load" is a common measurement of the proportion of virions to fluid (often in milliliters), and is frequently measured from host sputum, hence:

> *viral load* =def Quality inhering in a portion of fluid that is the proportion of virions to volume of that portion of fluid.

Second, a recent description of disordered and healthy immune responses to SARS-CoV-2 [17]:

> SARS-CoV-2 replicates in a host cell, releasing virions through cell lysis. Host cells attract immune system cells to sites of infection, promoting inflammation. Typically, inflammation attracts T cells which neutralize SARS-CoV-2, while antibodies prevent further infection. Uncontrolled inflammation, however, results in organ disorders.

In addition to several ontologies mentioned above, this description illustrates IDO-COVID-19's reuse of terms from the Cell Ontology (CO), and interoperability across OBO Foundry ontologies.

VIDO and IDO-COVID-19 have been released for public use. Given how quickly our knowledge of COVID-19 changes, we encourage researchers in relevant disciplines to help developers further refine each ontology and apply them to COVID-19 datasets as we collectively work to stop the pandemic.

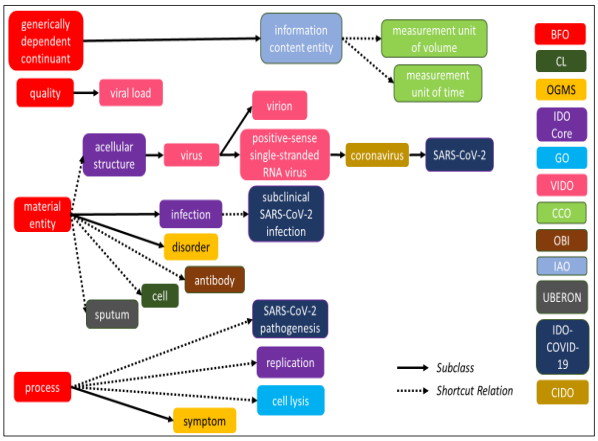

Figure 1: Ontological Relationships

ACKNOWLEDGEMENTS

Many thanks to Yongqun "Oliver" He and Asiyah Yu Lin, for assistance in VIDO and IDO-COVID-19 development. Darren Natale deserves thanks for critical feedback on a draft of IDO-COVID-19.

Funding: JB, SB supported by NIH / NLM T5 Biomedical Informatics and Data Science Research Training Programs. June 2017-May 2022. BS's contributions were supported by the NIH under NCATS 1UL1TR001412 (Buffalo Clinical and Translational Research Center).

REFERENCES

1. Arp R, Smith B, Spear A. *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press; 2015.
2. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007; 25:1251–1255. doi:10.1038/nbt1346.
3. https://www.iso.org/standard/74572.html; https://standards.iso.org/iso-iec/21838/-2/ed-1/en/
4. Babcock, S. Cowell, L. Beverley, J. Smith, B. (2020). *The Infectious Disease Ontology in the Age of COVID-19*. OSF Preprint. https://osf.io/2w865/
5. He, Y. *et al.* CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Scientific data*. 2020. (7):181.
6. *WHO COVID-19 Rapid Version CRF*. https:// bioportal.bioontology.org/ ontologies/ COVIDCRFRAPID. Accessed 27 Apr 2020.
7. *COVID-19 Surveillance Ontology*. https:// bioportal.bioontology.org/ontologies/COVID19. Accessed 27 Apr 2020.

8. *Linked COVID-19 Data Ontology*. https://github.com/Research-Squirrel-Engineers/COVID-19. Accessed 27 Apr 2020.
9. *COVID-19 Research Knowledge Graph*. https://github.com/nasa-jpl-cord-19/covid19-knowledge-graph. Accessed 27 Apr 2020.
10. Haendel M.A. et al. (2008) CARO – The Common Anatomy Reference Ontology. In: Burger A., Davidson D., Baldock R. (eds) Anatomy Ontologies for Bioinformatics. Computational Biology, vol 6. Springer, London. https://doi.org/10.1007/978-1-84628-885-2_16
11. Baltimore, D. (1971). *Expression of Animal Virus Genomes*. Bacteriological Reviews. 35, 235-41.
12. Babcock, S. Cowell, L. Beverley, J. Smith, B. (2020). *The Infectious Disease Ontology in the Age of COVID-19*. OSF Preprint. https://osf.io/2w865/
13. Babcock, S. Cowell, L. Beverley, J. Smith, B. (2020). Supplementary Documentation to [11].
14. Beverley, J. Babcock, Carvalho, G., S. Cowell, L., Duesing, S., Hurley, R., Smith, B. (2020). *Coordinating Coronavirus Research: The COVID-19 Infectious Disease Ontology*. OSF Preprint. https://osf.io/5bx8c/
15. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing. *Nucleic Acids Res.* 2019; 47:D330–D338. doi:10.1093/nar/gky1055.
16. Pan, Y. et. al. (2020). *Viral Load of SARS-CoV-2 in Clinical Samples*. The Lancet. 20(4):411-2. DOI:https://doi.org/10.1016/S1473-3099(20)30113-4.
17. Chousterman, B. G., Swirski, F. K. & Weber, G. F. Cytokine storm and sepsis disease pathogenesis. *Semin Immunopathol* **39**, 517-528, doi:10.1007/s00281017-0639-8 (2017).