

A Community Effort for COVID-19 Ontology Harmonization

Asiyah Yu Lin¹, Yuki Yamagata², William D. Duncan³, Leigh C. Carmody⁴, Tatsuya Kushida², Hiroshi Masuya², John Beverley⁵, Biswanath Dutta⁶, Michael DeBellis⁷, Zoë May Pendlington⁸, Paola Roncaglia⁸, Yongqun He⁹

¹ National Human Genome Research Institute, NIH, Bethesda, MD, USA

² RIKEN, Japan

³ Lawrence Berkeley National Laboratory, Berkeley, CA, USA

⁴ The Jackson Laboratory, Bar Harbor, ME, USA

⁵ Northwest University, Evanston, IL, USA

⁶ Indian Statistical Institute Bangalore Centre, India

⁷ Individual Consultant and Researcher, San Francisco, CA, USA

⁸ European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

⁹ University of Michigan Medical School, Ann Arbor, MI, USA.

Abstract

Ontologies have emerged to become critical to support data and knowledge representation, standardization, integration, and analysis. The SARS-CoV-2 pandemic led to the rapid proliferation of COVID-19 data, as well as the development of many COVID-19 ontologies. In the interest of supporting data interoperability, we initiated a community-based effort to harmonize COVID-19 ontologies. Our effort involves the collaborative discussion among developers of seven COVID-19 related ontologies, and the merging of four ontologies. This effort demonstrates the feasibility of harmonizing these ontologies in an interoperable framework to support integrative representation and analysis of COVID-19 related data and knowledge.

Keywords

Knowledge integration, COVID-19, SARS-CoV-2, ontology, harmonization

1. Introduction

Despite the development and distribution of effective COVID-19 vaccines, COVID-19 pandemic remains a challenge to overcome. The sheer volume of data collected by researchers, the speed at which it is generated, range of its sources, quality, accuracy, and need for assessment of usefulness, results in complex, multidimensional datasets [1], often annotated in specific terminologies and coding systems by researchers in distinct disciplines. The value of cross-discipline meta-data analysis is obvious, and evident in the present pandemic. However, with the extensive COVID-19 research, we face a big challenge of *data silos*, which significantly undermine interoperability, meta-data analysis, reproducibility, pattern identification, and discovery and reusability across disciplines [2].

Ontologies - interoperable, logically well-defined, controlled vocabularies representing common entities and relations across disciplines - is a well-known solution to data silo problems. Ontologies are widely used in bioinformatics and biomedical data standardization, supporting data integration, sharing, reproducibility, and automated reasoning. To meet different needs for COVID-19 studies, different groups of ontology developers have worked separately since the start of the pandemic, resulting in the

International Conference on Biomedical Ontologies 2021, September 16–18, 2021, Bozen-Bolzano, Italy

EMAIL: asiyah.lin@nih.gov (A. 1); yuki.yamagata@riken.jp (A. 2); yongqunh@med.umich.edu (A. 9)

ORCID: 0000-0003-2620-0345 (A. 1); 0000-0002-9673-1283 (A. 2); 0000-0001-9189-9661 (A. 9)



© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

development of several COVID-19 ontologies. A lack of coordination among these groups would risk the proliferation of COVID-19 ontologies using distinct, potentially non-interoperable, vocabularies.

The Workshop on COVID-19 Ontologies (WCO-2020) held on Oct. 23 and Oct. 30, 2020 brought the developers from international groups to report their efforts on building COVID-19 related ontologies. To harmonize heterogeneous knowledge and data for better COVID-19 study, the workshop attendees formed a COVID-19 Ontology Harmonization Working Group (WG) and discussed the ways to harmonize these related ontologies. This paper reports the current results of our harmonization effort.

2. Scope and Methods

In this study, the following seven COVID-19 related ontologies were covered in the ontology harmonization process by the COVID-19 Ontology Harmonization Working Group:

1. Virus Infectious Disease Ontology (VIDO) [3]
2. Ontology of Coronavirus Infectious Disease (CIDO) [4]
3. COVID-19 Infectious Disease Ontology (IDO-COVID-19) [5]
4. Controlled Vocabulary for COVID-19 (COVoc)
5. Homeostasis imbalance process ontology (HOIP) [6]
6. Medical Action Ontology (MAxO)
7. Ontology for collection and analysis of COviD-19 data (CODO) [7]

Each of the above ontologies has their own scope and purpose. Three ontologies: Virus Infectious Disease Ontology (VIDO), Coronavirus Infectious Disease Ontology (CIDO), and COVID-19 Infectious Disease Ontology (COVID-19-IDO) all extend the Infectious Disease Ontology (IDO) [5].

The mission statement of the COVID-19 Ontology Harmonization WG is to harmonize different COVID-19 related ontologies to support COVID-19 related data and knowledge interoperability. To achieve the mission, WG members held regular virtual Zoom meetings and communicated through emails. We identified overlapping domains or subdomains from different ontology groups and built consensus on ontology terms needed to characterize specific COVID-19 related entities.

2.1 VIDO

VIDO (<https://bioportal.bioontology.org/ontologies/VIDO>) is an extension of the IDO designed to bridge IDO - which is composed of terms common to any scientific investigation of infectious disease - to virus-specific ontologies. As such, VIDO follows OBO Foundry guidelines closely. VIDO is composed of terms common to any investigation of viral infectious diseases, including virus classification, virus infection epidemiology, pathogenesis, and treatment. For example, VIDO defines terms such as virus, prion, viricide, virus infection incidence, and so on.

2.2 CIDO

By extending IDO and other OBO ontologies including the Ontology for Biomedical Investigations (OBI), CIDO (<https://github.com/cido-ontology/cido>) is developed to cover coronavirus infectious diseases including their etiology, transmission, epidemiology, host-coronavirus interaction, pathogenesis, diagnosis, prevention, and treatment. CIDO covers SARS-CoV, SARS-CoV-2, and MERS-CoV, and other coronavirus strains that cause common human cold.

2.3 COVID-19-IDO

COVID-19-IDO (<https://bioportal.bioontology.org/ontologies/IDO-COVID-19>), which was created by the developers of VIDO, is a direct extension of VIDO. As such, IDO-COVID-19 covers the epidemiology, classification, pathogenesis, and treatment of terms used to represent infection by the SARS-CoV-2 virus strain and the associated COVID-19 disease.

2.4 COVoc

Controlled Vocabulary for COVID-19 (COVoc) (<https://github.com/EBISPOT/covoc>) is an application ontology created in collaboration between the European Bioinformatics Institute (EMBL-EBI) and the Swiss Institute of Bioinformatics (SIB) in March 2020. Its primary use case is to enable seamless annotation of biomedical literature to core databases and ELIXIR tools (ELIXIR is a European-wide intergovernmental organization for life sciences). The ontology covers 9 axes related to the COVID-19 pandemic (biomedical vocabulary, cell lines, chemical entities, clinical trials, conceptual entities, diseases and syndromes, geographic locations, organisms, and proteins and genomes). COVoc utilizes existing OBO ontologies where possible to augment connections to other useful resources such as the COVID-19 Data Portal (<https://www.covid19dataportal.org/>).

2.5 CODO

Ontology for Collection and Analysis of COviD-19 Data (CODO) (<https://w3id.org/codo>, <https://github.com/biswanathdutta/CODO>) is a formal Ontology for collection and analysis of COVID-19 data [8]. The goal of the ontology was to collect data about the pandemic so that researchers could answer questions, for example about infection paths based on information about relations between patients, clusters, geography, time, comorbidities, etc. The current CODO 1.3 primarily provides the terms and relations for representing COVID-19 data and information, such as epidemiology, clinical findings, etiology, diagnosis, treatment facility, comorbidity, including the statistical data on disease spread and casualty by space and time, and resource requirements. The developed ontology can be used by the various agencies, namely doctors, hospitals, policy-makers, government agencies, application developers, etc. for various purposes, such as for developing applications, like search, question-answering systems, risk detection systems; for document annotation; for developing knowledge graph, etc. The ontology was designed by analysing disparate COVID-19 data sources such as datasets, literature, services, government published COVID-19 guidelines, WHO literature, etc.

2.6 HOIP

Homeostasis imbalance process ontology (HOIP) (<https://bioportal.bioontology.org/ontologies/HOIP>) focuses on homeostatic imbalances between virus action and innate defense processes and covers the causal relationship of organelle/cellular/organ processes from early stage to clinical manifestation in COVID-19. The design patterns between CIDO and HOIP have now been aligned after shared discussion and communication.

2.7 MAXO

Medical Action Ontology (MAXO), launched in the spring of 2020, is a broad ontology that provides a structured vocabulary to medical procedures, interventions, therapies, treatments, or clinical recommendations. MAXO was designed to provide a thorough resource for annotating medical actions to diseases, particularly rare diseases. Given the broad nature of MAXO and the timing of the ontology development, much of the hierarchy was added with a keen awareness of the diagnostics and treatment of SARS-CoV-2. While there are no COVID-19-specific terms, terms like ‘ventilation with proning’ (MAXO:0000619) and ‘clinical RNA detection testing’ (MAXO:0000592) were added to annotate COVID-19 clinical data sets. To capture the relationship between treatments and diseases, a new tool, Phenotypic Observation Explication Tool (POET), was developed to establish a relationship between MAXO, Human Phenotype Ontology (HPO), and Mondo Disease Ontology (Mondo) terms. This tool will allow researchers to actively participate in annotating COVID-19 data sets or other diseases in their expertise. MAXO annotations and the POET tool will be available on the HPO website (hpo.jax.org) by 2022.

3. Ontology Overlapping and Term Reuse

The ontology harmonization is started by identifying the scopes and development methods by different ontologies covered in this work. We found that instead of reinventing the wheel, each ontology has imported and reused many terms from other ontologies where possible (Table 1). The top 1 reused ontologies (reused in six out of the seven ontologies) are: OBI, UBERON, CL, GO Biological process, ChEBI, PRO, and RO. The top 2 reused ontologies (reused in five out of the seven ontologies) are BFO, NCBI taxon, symptom ontology and Vaccine Ontology. Many of these reused ontologies are Open Biomedical and Biological Ontologies (OBO) Foundry [8] ontologies.

Table 1.

Ontology term reuse by COVID-19 related ontologies

Ontology	Domain	VIDO	CIDO	COVID-19-IDO	HoIP	CODO	MAxO	COVoc
BFO	Upper ontology	Yes	Yes	Yes	Yes	Yes		
IAO	Information content		Yes	Yes	Yes			Yes
OBI	Data item		Yes	Yes	Yes	Yes	Yes	Yes
NCBI taxon	Taxonomy	Yes	Yes	Yes	Yes			Yes
UBERON	Anatomical structure	Yes	Yes	Yes	Yes		Yes	Yes
CL	Cell	Yes	Yes	Yes	Yes		Yes	Yes
GO	Biological process	Yes	Yes	Yes	Yes		Yes	Yes
PATO	Phenotype		Yes		Yes		Yes	
HPO	Phenotype		Yes		Yes		Yes	Yes
ChEBI	Chemical compound	Yes	Yes	Yes	Yes		Yes	Yes
PRO	Protein	Yes	Yes	Yes	Yes		Yes	Yes
HGNC	Gene		Yes					
OGG	Gene				Yes			
DO	Disease		Yes		Yes	Yes		
MONDO	Disease						Yes	Yes
SNOMED CT	Disease					Yes		
NDF-RT	Disease/Finding		Yes					
Symptom	Symptom	Yes	Yes	Yes	Yes	Yes		
Vaccine Ontology	Vaccine	Yes	Yes	Yes	Yes	Yes		
RO	Relational ontology	Yes	Yes	Yes	Yes		Yes	Yes

4. Ontology Alignment and Harmonization

Given that most of the 7 ontologies follow the OBO Foundry ontology development principles, such as reusing terms defined in OBO foundry ontologies, Our harmonization exercise found that these ontologies can be aligned under the Basic Formal Ontology (BFO) upper level ontology (Figure 1). Figure 1 below shows how VIDO, CIDO, IDO-COVID-19, MAxO and HoIP can fit into BFO's structure.

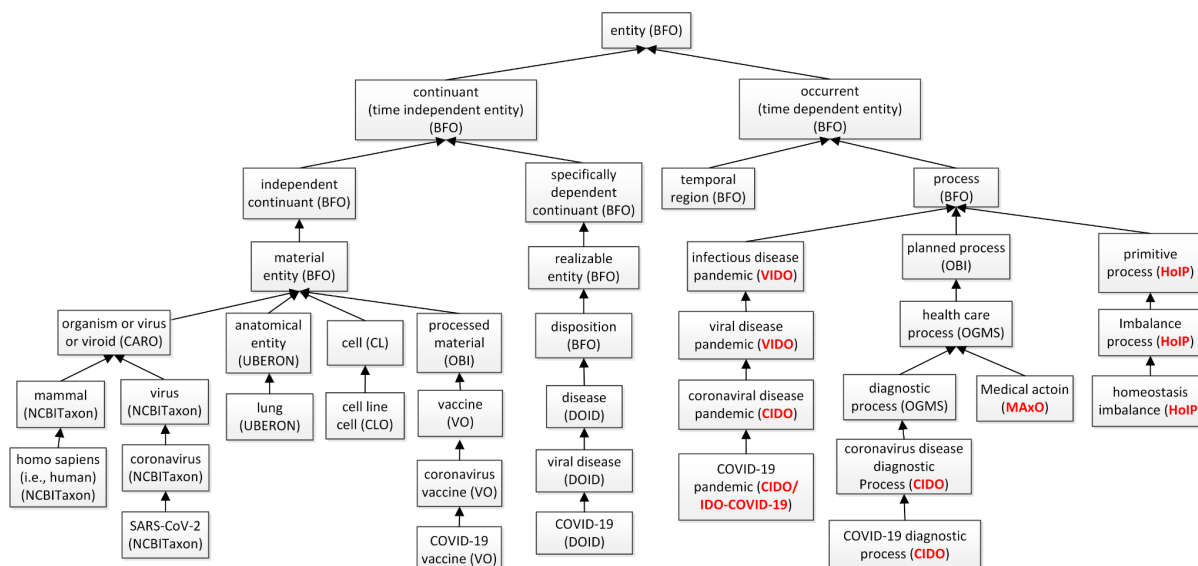


Figure 1: Hierarchical representation of selected terms from different ontologies that are harmonized under the BFO upper level ontology. The red colors represent ontologies focused in this ontology harmonization study. Terms from many ontologies such as BFO, NCBITaxon, and VO have been used by our ontologies as well.

The relationship between CIDO and IDO-COVID-19 provides an example of precisely the sort of distinct overlapping ontology development efforts our working group was designed to address. Via this alignment exercise and observing the scope of CIDO appears broad enough to include IDO-COVID-19, our working group has decided to incorporate the latter ontology into CIDO. Incorporation of terms from IDO-COVID-19 into CIDO will, moreover, strengthen the logical relationship between CIDO and VIDO, given how closely related VIDO and IDO-COVID-19 are.

The HoIP developers are working on mapping and aligning with all GO process terms. Concerning harmonization, HoIP ontology has started to compare their processual entities to those in CIDO. For example, although the labels of 'SARS-CoV-2 entry to cell' (CIDO:0000088) and 'viral entry into host cell [COVID-19]' (HoIP:0037063) are different, as the HoIP entity is described using object property restriction ('has agent' some SARS-CoV2), it can be mapped to correspondent CIDO term. As an application ontology, the COVoc developers rely on CIDO developers to create new terms, and COVoc imports and reuses CIDO for their application purpose. At the time of writing, CODO developers started to align the current build to BFO as its upper ontology, which increases the future possibilities of better alignment.

5. Discussions

While ontology creates a common language and reduces the work of mapping, the emergence of multiple ontologies may form individual silos by themselves. Given the report of many COVID-19 related ontologies, our COVID-19 Ontology Harmonization WG provided a timely effort to collaboratively identify the overlapping between different ontologies and achieve the harmonization of seven ontologies. Currently, seven ontologies have very different perspectives due to their use cases.

Entities within these seven ontologies are defined heterogeneously and described in various ways with various granularities. We should align not only the same URIs but also the meaning (semantics) of the entities. Therefore, it is necessary to investigate and compare entities among ontologies carefully, such as definition, superclass, logical restrictions, and related entities. Towards the formal alignment of these ontologies, we plan to clarify and make explicit the relationships such as equivalent class among the ontologies.

Members of the COVID-19 Ontology Harmonization WG made substantial efforts to characterize SARS-CoV-2 and COVID-19 data in a collaborative, computationally tractable, responsible manner. These ontologies are also being used in different use case studies, supporting productive and interoperable COVID-19 research.

Our working group has also recognized many future challenges such as funding, resource and time commitment, and challenging infrastructure development. We are pleased to find that the willingness to join the harmonization work is high, and more interested parties are joining the effort. We aim to continue this collaborative effort to further support our active COVID-19, leading to enhanced public health.

6. Acknowledgements

We acknowledge the organizers and attendees of the 2020 Workshop on COVID-19 Ontologies (WCO-2020), which initiated our get-together and collaboration on the ontology harmonization effort.

The Office of Data Science Strategy, NIH, provided funding for AYL as a Data and Technology Advancement (DATA) National Service Scholar. JB was supported by NIH / NLM T5 Biomedical Informatics and Data Science Research Training Programs (5T15LM012495–03) during development of VIDO and IDO-COVID-19. CODO work has been supported by Indian Statistical Institute through internal project grant. PR's work of COVoc has been made possible in part by a grant from Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation. YY was supported by the RIKEN Open Life Science Platform Project during knowledge systematization of COVID-19 infectious processes and development of HoIP ontology.

7. References

- [1] Y. He et al. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Scientific data*. 2020. (7):181.
- [2] R. Arp, B. Smith, A. Spear. *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press; 2015.
- [3] J. Beverley, S. Babcock, G. Carvalho, L. G. Cowell, S. Duesing, R. Hurley, B. Smith (2020). *Coordinating Coronavirus Research: The COVID-19 Infectious Disease Ontology*. OSF Preprint. <https://osf.io/5bx8c/>
- [4] Y. Liu, J. Hur, W. K. B. Chan, Z. Wang, J. Xie, D. Sun, S. Handelman, J. Sexton, H. Yu, Y. He. Ontological modeling and analysis of experimentally or clinically verified drugs against coronavirus infection. *Sci Data*. 2021 Jan 13;8(1):16. doi: 10.1038/s41597-021-00799-w.
- [5] S. Babcock, J. Beverley, L. G. Cowell, B. Smith. *The Infectious Disease Ontology in the Age of COVID-19*. 2021, June 10. <https://doi.org/10.31219/osf.io/az6u5>
- [6] Y. Yamagata et al. *Ontology development for building a knowledge base in the life science and structuring knowledge for elucidating the COVID-19 mechanism*. The 31th Annual Conference of the Japanese Society for Artificial Intelligence, 2021.
- [7] B. Dutta, M. DeBellis (2020). CODO: an ontology for collection and analysis of COVID-19 data. In *Proc. of 12th Int. Conf. on Knowledge Engineering and Ontology Development (KEOD)*, Lisboa, Portugal, 2-4 November 2020, vol. 2, pp. 76-85 (DOI: <https://doi.org/10.5220/0010112500760085>).
- [8] The Open Biomedical Ontologies Foundry. <http://obofoundry.org/>. Accessed 10 June. 2021.