The COVID-19 epidemiology and monitoring ontology

Núria Queralt-Rosinach, Paul Schofield, Robert Hoehndorf,
Claus Weiland, Erik Schultes, César Henrique Bernabé, Marco Roos

I. Motivation

One year ago, the novel COVID-19 infectious disease emerged and spread, causing high mortality and morbidity rates worldwide. In the OBO Foundry, there are more than one hundred ontologies to share and analyse large-scale datasets for biological and biomedical sciences. However, this pandemic revealed that we lack tools for an efficient and timely exchange of this epidemiological data which is necessary to assess the impact of disease outbreaks, the efficacy of mitigating interventions and to provide a rapid response [1]. Recently, several new COVID-19 ontologies have developed such as the IDO extension [2] or CIDO [3]. Hence, our research question was to determine if there was a good representation of epidemiological quantitative concepts in OBO ontologies. Our objectives were to identify missing COVID-19 epidemiological terms and implement axiom patterns for extensions to existing ontologies or to build a new, logically well-formed, and accurate ontology in OBO. In this study we present our findings and contributions for the bio-ontologies community.

II. Method

This work was conceived and mainly developed during open community hackathons [1],[2],[3]. Our approach was based on first, extracting a list of relevant epidemiological terms through manual curation of recent COVID-19 epidemiological studies published in peer-reviewed journals, medRxiv and public health surveillance websites, and mapping them to existing OBO ontologies. Curation was focused on quantitative data and indicators. Second, developing a minimal ontological representation of COVID-19 epidemiological quantitative information. And third, to refine and evaluate the model with domain expert input.

Our formal modeling followed a rationale already used in other studies: 1) determine the domain and scope of the ontology; 2) ontology reuse and addressing poor ontological coverage of COVID-19 epidemiology; and 3) development of a conceptual model [4, 5]. We extracted core domain knowledge concepts from [6, 7, 8]. We re-used ontological terms and models as much as possible using ontology search engines [4],[5]. To build an interoperable biomedical ontology, we decided to build an OBO ontology and use the OWL 2, a DL-based formalism and semantic web standard for knowledge representation to enable data sharing and formal reasoning. We used knowledge-engineering best practices following the OBO principles [6] and modularization guidelines [9] to achieve a logically well-formed model. Finally, we based our decisions on building a FAIR resource for health data and research following recent recommendations published by international data standard organizations [10, 11]. More information on the method, the list of sources used for curation and extracted terms, and the developed OWL ontology are open and publicly available for reproducibility and community re-use on GitHub.

[1] Virtual-biohackathon covid-19-bh20

[2] BioHackathon-EU-2020

[3] SWAT4HCLS 2021

[4] https://www.ebi.ac.uk/ols/index

[5] http://www.obofoundry.org/

[6] http://www.obofoundry.org/principles/fp-000-summary.html

III. Results

We provide a formal ontological model for COVID-19 epidemiology and monitoring (graphical and OWL representations are in our GitHub). With the rise of new variants of the virus that may challenge vaccine efficacy, a compatible logical model for quantities that enables researchers to represent and share machine-readable patient monitoring and epidemiology surveillance data for rapid analysis, modeling and response is an urgent need. In this work we re-used the SIO design pattern for measurements [7], a model already applied to patient health data for rare diseases in the EJP RD [8], clinical research data in the LUMC [12] and the measurements schema in the new GA4GH Phenopackets release [13]. The taxonomic structure is extended from IDO, a core ontology for infectious diseases. For domain concepts we re-used GFO [14] to formalize timelines concepts using the 'chronoid' and the GFO-based 'mortality' model approach [15]. To link patient-population is an RDA COVID-19 recommendation on data sharing, thus we checked common data models such as OMOP [9] and re-used the relationship used in Phenopackets based on *composition* semantics.

We filled the gap for epidemiological surveillance terms in OBO adding 100 new terms. From an initial set of 138 manually extracted terms, only 38 are covered by bio-ontologies, 21% (30 terms) IDO [16] and 24% STATO (33 terms) [17] (although including fallbacks this percentage could increase to 50%) and the rest by epidemiological-related ontologies such as APOLLO_SV [18] and GENEPIO [19]. We noticed that EPO [20] is not maintained since its publication and has been deprecated from OBO Foundry, and IDO is working towards epidemiological enrichment [21]. While interoperability within the OBO landscape is fostered by adopting the BFO backbone structure, the link with GFO can lead to incompatible temporal regions due to logical inconsistency [22]. Another issue that may be improved is the current absence of axioms and definition patterns that relate epidemiology (i.e., observations of a population) to clinical ontologies (i.e., observations on an individual) and allow reasoning for discovery. The re-use of the EQ model [23] or the adaptation of the REA model [24] will be evaluated. In the future, we will evaluate our ontology with domain experts and logical competency questions [25]. Moreover, we expect to use this model in FAIR-based projects such as TWOC [26] to publish epidemiological claims as nanopublications for trust [27]. We aim at FAIR reasoning and analytics of person-level real world observations over epidemiological surveillance information [28]. Therefore, checking common data models such as Phenopackets or OHDSI standards was done to enable the development of applications to discover patterns with ontology-guided machine learning algorithms and translational research.

IV. Conclusion

In the context of an infectious disease outbreak it is imperative to have these data as FAIR as possible to facilitate rapid analysis and support timely evidence-based decision making and trust. To enable the community to provide machine-readable epidemiological quantitative data and make it easier to share, we contributed with the development of an ontological representation, which was built based on ontology engineering best-practices such as reuse and ontology formalization through upper-level ontologies (i.e., GFO, SIO).

V. Acknowledgements

The authors would like to specially thank Dr. Birgit Meldal for her input and ideas. This initiative has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°825575 (the European Joint Programme Rare Diseases), and the Trusted World of Corona (TWOC; LSH

[7] https://github.com/MaastrichtU-IDS/semanticscience/wiki/DP-Measurements

[8] https://www.ejprarediseases.org/

[9] https://www.ohdsi.org/data-standardization/the-common-data-model/

Health Holland).

References

[1] Editorial. How epidemiology has shaped the covid pandemic. *Nature*, (589):491–492, 2021.

[2] IDO-COVID19 OWL ontology. http://purl.obolibrary.org/obo/2020-21-07/ido-covid-19.owl/. last accessed 2021/05/04.

[3] Y. He, H. Yu, and E. et al. Ong. Cido, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Sci Data*, 7(181), 2020.

[4] D. Sánchez and M. Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *J. Biomed. Inform*, (44):749–759, 2011.

[5] K.-M. Kouamé and H. Mcheick. Ontological approach for early detection of suspected covid-19 among copd patients. *Appl. Syst. Innov.*, (4):21, 2021.

[6] Ferran Martínez Navarro et al. *Vigilancia Epidemiológica*. McGraw-Hill Interamericana, 2004. ISBN:84-486-0245-5.

[7] B. MacMahon and D. Trichopoulos. Marbán SL, Harvard Medical School of Public Health, Boston, Massachussetts. ISBN:84-7101-317-7.

[8] Straif-Bourgeois S, Ratard R, and Kretzschmar M. Infectious disease epidemiology.y. *Handbook of Epidemiolog*, pages 2041–2119, 2014.

[9] Alan L Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including owl. http://www.cs.man.ac.uk/~rector/papers/rector-modularisation-kcap-2003-distrib.pdf.

[10] RDA COVID-19 Working Group. Rda covid-19 recommendations and guidelines on data sharing. https://zenodo.org/record/3932953#.YI7Dba4p5hH.

[11] Kees van Bochove, Emma Vos, Maxim Moinat, Sebastiaan van Sandijk, Tess Korthout, and Peyman Mohtashani. Ehden - d4.5 - roadmap for interoperability solutions. https://zenodo.org/record/4474373#.YI66yq4p5hH.

[12] Núria Queralt-Rosinach et al. Fair data management to access patient data. https://repository.publisso.de/resource/frl%3A6424232.

[13] GA4GH. Phenopackets v2. https://github.com/phenopackets/phenopacket-schema/issues/261.

[14] Heinrich Herre, Barbara Heller, Patryk Burek, Robert Hoehndorf, Frank Loebe, and Hannes Michalek. General formal ontology (gfo) - a foundational ontology integrating objects and processes [version 1.0]. Workingpaper, Unknown Publisher, July 2006.

[15] Santana F, Freitas F, Fernandes R, Medeiros Z, and Schober D. Towards an ontological representation of morbidity and mortality in description logics. *J Biomed Semantics*, Suppl 2(Suppl 2)(3):S7, 2012.

[16] Infectious Disease Ontology OWL ontology. http://purl.obolibrary.org/obo/ido/2017-11-03/ido.owl. last accessed 2021/05/04.

[17] STATO: the statistical methods ontology OWL ontology. http://purl.obolibrary.org/obo/stato.owl. last accessed 2021/05/04.

[18] Apollo structured vocabulary owl ontology.

[19] Genomic Epidemiology Ontology OWL ontology. http://purl.obolibrary.org/

obo/genepio/releases/2020-08-09/genepio.owl. last accessed 2021/05/04.

[20] Epidemiology Ontology OBO Foundry Website. http://www.obofoundry.org/ontology/epo.html. last accessed 2021/05/04.

[21] Shane Babcock, John Beverley, Lindsay G. Cowell, and Barry Smith. The infectious disease ontology in the age of covid-19. *Preprint*, 2021.

[22] Toward semantic interoperability with linked foundationalontologies in ROMULUS. https://researchspace.csir.co.za/dspace/bitstream/handle/10204/7042/Khan_2013.pdf?sequence=1&isAllowed=y. last accessed 2021/05/04.

[23] Mungall CJ, Bada M, Berardini TZ, Deegan J, Ireland A, Harris MA, Hill DP, and Lomax J. Cross-product extensions of the gene ontology. *J Biomed Inform*, 44(1):80–6, 2011.

[24] Dahdul WM Lapp H Mungall CJ Vision TJ. Mabee PM, Balhoff JP. A logical model of homology for comparative biology. *Syst Biol.*, 1(69):345–362, 2020.

[25] R. de Almeida Falbo. Sabio: Systematic approach for building ontologies. *FOIS*, 2014.

[26] Health-Holland funder. The trusted world of corona (twoc).

[27] Paul Groth, Andrew Gibson, and Jan. Velterop. The anatomy of a nanopublication'. *Information Services Use*, 30(1-2):51–56, 2010.

[28] V. et al Sherimon. Covid-19 ontology engineering-knowledge modeling of severe acute respiratory syndrome coronavirus 2 (sars-cov-2). *(IJACSA) International Journal of Advanced Computer Science and Applications*, 11(11), 2020.