Towards a community recommended choice of ontologies to increase interoperability of genetic variant data

Annika Jacobsen[1], Friederike Ehrhart[2], Rajaram Kaliyaperumal[1], Mark Thomson[1], Erik A. Schultes[3], Jeroen F.J. Laros[1,4], Chris T. Evelo[2], Andra Waagmeester[5], and Marco Roos[1]

[1] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
{a.jacobsen,r.kaliyaperumal,m.thompson,j.f.j.laros,m.roos}@lumc.nl
[2] Department of Bioinformatics - BiGCaT, Maastricht University, Maastricht, The Netherlands
{friederike.ehrhart,chris.evelo}@maastrichtuniversity.nl
[3] Dutch Techcentre for Life Sciences, PO Box 8500, 3503 RM Utrecht, The Netherlands
erik.schultes@dtls.nl
[4] Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands
[5] Micelio, Antwerp - Ekeren, Belgium
andra@micelio.be

Keywords: Genetic variant, interoperability, ontology, semantic model

Abstract

In the field of rare diseases (RD), data describing genetic variants and associated phenotypes, so-called genotype-phenotype data, are crucial to uncover the pathogenic significance and role of these variants in disease development and progression. Since these data are sparse, distributed and isolated, data interoperability is crucial, both to link between different sources, but also to link relevant information from multiple sources to these variants. Consequently, efforts are underway to create guidelines for making these data FAIR (Findable, Accessible, Interoperable, and Reusable) for humans and computers [1]. The degree of interoperability is greatly impacted by the choice of ontologies by which data are described. An interoperable description of a genetic variant is in a computer-readable format such as by semantic models based on commonly used ontologies (application ontologies) [2, 3]. To reach a high level of efficiency in interoperability we require, that the same ontologies are used for describing the same concepts or, a proper mapping exists, when different ontologies are used. There are currently several models describing genetic variants, such as the sequence ontology, the genotype ontology, and the variation ontology. While semantic mapping efforts can technically resolve interoperability issues, a user community expects a FAIR data steward to present a clear and unified ontology

recommendation. A growing number of of community recommended ontologies such as the Human Phenotype Ontology and the Orphanet Rare Disease Ontology, are already recognized resources in the rare disease community and are being applied by rare disease researchers when describing phenotypic data and rare disease [4, 5].

Here, we offer a rationale for our initiative of undertaking the development of community recommended ontologies for genetic variants in the rare disease field. We envision that these recommendations will follow the generic guidelines for ontology selection, including application specific practices. The first step will be to survey existing ontologies according to these guidelines and propose a first version of the genetic variant ontology recommendations to the rare disease community. We hope to actively engage the community in further development of these ontology recommendations. After reaching a community agreement we will suggest these as official genetic variant ontology recommendations by the the rare disease community. We believe that these recommendation are very important for the non-ontology experts and that they will benefit genotype-phenotype interoperability.

Acknowledgements. The work leading to this poster is supported by grants from RD-Connect (FP7/20072013, grant agreement No. 305,444), Elixir infrastructure for life science data, Elixir-Excelerate (H2020-INFRADEV-1-2015-1), and BBMRI-NL2 (NWO National Roadmap for Large-Scale Research Facilities).

References

1. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 3: 160018. doi: 10.1038/sdata.2016.18 (2016).
2. Roos, M., Wilkinson, M.D., Kaliyaperumal, R., Thompson, M., Carta, C., Cornet, R., van Enckevort, D., Bonino, L.: Registries of domain-relevant semantic reference models help bootstrap interoperability in domains with fragmented data resources. SWAT4LS (2016).
3. Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., Parkinson, H.: Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 26:1112-8. doi: 10.1093/bioinformatics/btq099 (2010).
4. Lochmller, H., Cam, Y.L., Jonker, A.H., Lau L.P.L, Baynam, G., Kaufmann, P., Lasko, P., Dawkins H.J.S, Austin, C.P., Boycott, K.M.: IRDiRC Recognized Resources': a new mechanism to support scientists to conduct efficient, high-quality research for rare diseases. Eur J Hum Genet. 25: 162165 (2017).
5. IRDiRC webpage: Last updated on June 27, 2017. Retrieved from http://www.irdirc.org/activities/irdirc-recognized-resources/