2015 Disease Ontology update: DO's expanded curation activities to connect disease-related data

Elvira Mitraka[1] and Lynn M. Schriml[1,*]

[1] Institute for Genome Science, University of Maryland School of Medicine, Baltimore, MD, USA

ABSTRACT

The Human Disease Ontology is a widely used biomedical resource, which standardizes and classifies common and rare human diseases. Its latest iteration makes use of the OWL language to facilitate easier curation between a variety of working groups and to take advantage of the analyses available using OWL. The DO integrates disease concepts from ICD-9, ICD-10, the National Cancer Institute Thesaurus, SNOMED-CT, MeSH, OMIM, EFO and Orphanet. The DO Team is focused on enabling mapping and curation of large disease datasets for major Biomedical Resource Centers and integration of their disease terms into DO. Constant updates and additions to the ontology allow for coverage of the vast field of human diseases. By having close collaborations with a variety of research groups, such as MGD, EBI, NCI, the Disease Ontology has established itself as the go-to tool for human disease curation. Implementing a combination of informatic tools and manual curation DO ensures that it maintains the highest standard possible.

1 INTRODUCTION

The Disease Ontology (DO) (Schriml *et al.*, 2012) is the core disease data resource for the biomedical community. Human disease data is a cornerstone of biomedical research for identifying drug targets, connecting genetic variations to phenotypes, understanding molecular pathways relevant to novel treatments and coupling clinical care and biomedical research. Consequently, across the multitude of biomedical resources there is a significant need for a standardized representation of human disease to map disease concepts across resources, to connect gene variation to phenotypes and drug targets and to support development of computational tools that will enable robust data analysis and integration.

2 CURRENT STATUS

DO has proven to be an invaluable genomics and genetic disease data resource used for evaluating and connecting diverse sets of data, used by diverse curation groups to connect human disease to animal models and genomic resources and used to informatically identify representative phenotype sets (Köhler *et al.,* 2014; Schofield *et al.*, 2010), functionally similar genes (Fang and Gough, 2013; Singleton *et al.*, 2014), human gene and genome annotations (Peng *et al.*, 2013; Osborne *et al.*, 2009), pathways, cancer variants (Wu *et al.,* 2014) and immune epitopes (Vita *et al.,* 2014). The DO website (http://www.disease-ontology.org), is a web-based application that allows users to query, browse, and visualize the Disease Ontology and disease concept data.

The latest version of DO has close to 9,000 disease terms, more than 16,000 synonyms and almost 39,000 cross-references to other biomedical resources. Those resources include the ICD-9 and ICD-10, the National Cancer Institute (NCI) Thesaurus (Sioutos *et al.,* 2007), SNOMED-CT (Donnelly, 2006) and MeSH (https://www.nlm.nih.gov/mesh/MBrowser.html) extracted from the Unified Medical Language System (UMLS) (Bodenreider, 2004) based on the UMLS Concept Unique Identifiers for each disease term. DO also includes disease terms extracted directly from Online Mendelian Inheritance in Man (OMIM) (Ambereger *et al.,* 2011), the Experimental Factor Ontology (EFO, http://www.ebi.ac.uk/efo/) and Orphanet (Maiella *et al.,* 2013).

The DO files are available in both OBO and OWL format from DO's SourceForge site (http://sourceforge.net/p/ diseaseontology/code/HEAD/tree/trunk) and can be found at http://purl.obolibrary.org/obo/doid.obo and http: //purl.obolibrary.org/obo/doid.owl. DO's OBO and OWL files are also available from the OBO Foundry (http://www.obofoundry.org/cgi-bin/detail. cgi?id=disease ontology) and GitHub (https://github.com/obophenotype/human-disease-ontology/tree/master/src/ontology).

3 CURRENT WORK

Due to the huge amount of data generated at an increasingly rapid pace, the genomics community is trying to streamline its data processing efforts. Ontologies are an avenue that lead to this, but even they can become too big and unwieldy in their effort to capture all available data. There are instances where the multitude of information captured is not needed.

The Gene Ontology is one the most widely used ontology and one of the most comprehensive. It covers the molecular functions, biological processes and location in cellular components of gene products, containing more than 40,000 terms. In order to make it more accessible and less resource intensive the Gene Ontology Consortium has created slim version of the ontology. These "GO slims" are smaller versions of GO that contain only a subset of the terms, repre-

[*] To whom correspondence should be addressed: lschriml@som.umaryland.edu

senting a general knowledge of a specific field, without going too deep into the hierarchy.

Due to the breadth of its user base the DO team decided to create its own slim files, the DO Cancer Slim (Wu *et al.,* 2015) being the most prominent, containing terms needed by the pan-cancer community. In the same vein a DO MGI slim is being created. It contains all the terms that were modified or created during an intensive curatorial effort to map concepts between DO and OMIM. It will give insight into the overlap between DO and OMIM, as well as which disease types are more heavily featured in the MGD. Meaning it can give even more information about which diseases do not have a mouse model to represent them.

4 UPCOMING WORK

Future plans include the definition of all disease terms in DO and the creation of DO slims for every major curation project of all the MODs. These slims will enable DO users to review the representation and classification of MOD associated diseases, to compare the diseases represented between MODs and to compare the different animal models associated with a particular disease or types of diseases across species.

ACKNOWLEDGEMENTS

This work was supported in part by the National Institute of Health – National Center for Research Resources (R01RR025341) and NIH/NIGMS (R01 GM 089820-06).

REFERENCES

Alexandrescu,A. (2001) *Modern C++ Design: Generic Programming and Design Patterens Applied.* Addision Wesley Professional, Boston.

Amberger,J., Bocchini,C. and Hamosh,A. (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.*, **32**, 564–567

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

Donnelly,K. (2006) SNOMED-CT: the advanced terminology and coding system for eHealth. *Stud. Health Technol. Inform.*, **121**, 279–290.

Fang, H. and Gough,J. (2013) DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res.* **41**:D536-D544.

Köhler,S., Doelken,S.C., Mungall,C.J., Bauer,S., Firth,H.V., Bailleul-Forestier,I., Black,G.C., Brown,D.L., Brudno,M., Campbell,J., FitzPatrick,D.R., Eppig,J.T., Jackson,A.P., Freson,K., Girdea,M., Helbig,I., Hurst,J.A., Jähn,J., Jackson,L.G., Kelly,A.M., Ledbetter,D.H., Mansour,S., Martin,C.L., Moss,C., Mumford,A., Ouwehand,W.H., Park,S.M., Riggs,E.R., Scott,R.H., Sisodiya,S., Van Vooren,S., Wapner,R.J., Wilkie,A.O., Wright,C.F., Vulto-van Silfhout,A.T., de Leeuw,N., de Vries,B.B., Washingthon,N.L., Smith,C.L., Westerfield,M., Schofield,P., Ruef,B.J., Gkoutos,G.V., Haendel,M., Smedle, D., Lewis,S.E. and Robinson,P.N. (2014) The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **4**:D966-D974.

Maiella,S., Rath,A., Angin,C., Mousson,F. and Kremp,O. (2013) Orphanet and its consortium: where to find expert-validated information on rare diseases. *Rev. Neurol. (Paris)*, **169**, S3–S8.

Osborne,J.D., Flatow,J., Holko,M,. Lin,S.M., Kibbe,W.A., Zhu,L.J., Danila,M.I., Feng,G. and Chisholm,R.L. (2009) Annotating the human genome with Disease Ontology. *BMC Genomics*. **10** Suppl 1:S6.

Peng,K., Xu,W., Zheng,J., Huang,K., Wang,H., Tong,J., Lin,Z., Liu,J., Cheng,W., Fu,D., Du,P., Kibbe,W.A., Lin,S.M. and Xia,T. (2013) The Disease and Gene Annotations (DGA): an annotation resource for human disease. *Nucleic Acids Res.* **41**:D553-D560.

Schofield,P.N., Gkoutos,G.V., Gruenberger,M., Sundberg,J.P. and Hancock, J.M. (2010) Phenotype ontologies for mouse and man: bridging the semantic gap. *Dis. Model Mech.,* **3**:281–289.

Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.

Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, Margraf RL, Durtschi J, Eilbeck K, Reese MG, Jorde LB, Huff CD, Yandell M (2014) Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.,* **94**:599-610.

Sioutos,N., de Coronado,S., Ha.ber,M.W., Hartel,F.W., Shaiu,W.L. and Wright,L.W. (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.*, **40**, 30–43.

Vita,R., Overton,J.A., Greenbaum,J.A., Ponomarenko,J., Clark,J.D., Cantrell,J.R., Wheeler,D.K., Gabbard,J.L., Hix,D., Sette,A. and Peters,B (2014) The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43**:D405-D412

Wu,T.J., Shamsaddini,A., Pan,Y., Smith,K., Crichton,D.J., Simonyan,V. and Mazumder,R. (2014) A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE). *Database (Oxford)*, bau:022.

Wu,T.J., Schriml,L.M., Chen, Q.R., Colbert,M., Crichton,D.J., Finney,R., Hu,Y., Kibbe,W.A., Kincaid,H., Meerzaman,D., Mitraka,E., Pan,Y., Smith,K.M., Srivastava,S., Ward,S., Yan,C. and Mazumder,R. (2015) Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. *Database (Oxford),* bav:032.