Tailoring the NCI Thesaurus for use in the OBO Library

James P. Balhoff[1,*] Matthew Brush[2], Laura Christopherson[1], Sherri de Coronado[3], Gilberto Fragoso[3], Melissa A. Haendel[2], Christopher J. Mungall[4], Kimberly Robasky[1], Nicole Vasilevsky[2], and Lawrence W. Wright[3]

[1] Renaissance Computing Institute, University of North Carolina, 100 Europa Dr, Ste 540, Chapel Hill, NC, USA

[2] OHSU Library, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, Portland, OR, USA

[3] Center for Biomedical Informatics and Information Technology, National Cancer Institute, 9609 Medical Center Dr, Rockville, MD, USA

[4] Functional Genomics Dept., Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, MS977, Berkeley, CA, USA

ABSTRACT

The amount and diversity of knowledge and data being generated by the cancer research community is unprecedented in biomedicine, with data being collected on human samples, animals, and *in vitro* model systems. The data range from health records, to RNAseq, genomics, clinical trials, pathway analytics, modifier variant discovery, exposures, and pathology. To assist in standardizing these data, the National Cancer Institute (NCI) Thesaurus (NCIt) has been developed as a reference terminology and ontology that provides definitions, synonyms, and other information on nearly 10,000 cancers and related diseases, 8,000 single agents and combination therapies, and a wide range of other entities related to cancer and biomedical research. The NCIt is richly axiomatized with knowledge about tissues and cells of origin, causality, cancer grade, and much more.

NCIt is the *de facto* standard for clinical data and is utilized within Common Data Elements for clinical trial data collection, for submission to the NCI Genomic Data Commons, and as a pivotal component of the Clinical Data Interchange Standards Consortium (CDISC) for reporting to the FDA and other regulatory agencies internationally. However, NCIt is less well adopted within the basic research community. Most bioinformatics and basic researchers are familiar with the Gene Ontology, part of the Open Biomedical Ontologies (OBO) Library that includes a diversity of easy to use vocabulary standards developed within a community of best practice. However, the NCIt pre-dates most OBO ontologies and has evolved using a different set of design practices than current OBO ontologies. Currently, cancer is poorly represented across the OBO Library, and so there are efforts to try to expand cancer representation which are largely unaware of the extensive knowledge that is already encoded within the NCIt. Additionally, the knowledge contained within the NCIt is the outcome of years of clinical and molecular cancer classification by expert consultants, and is somewhat overwhelming to the non-expert (e.g. a basic science researcher or bioinformatician).

We aim to increase adoption of NCIt within the basic science bioinformatics community to support improved translation across the basic–clinical divide and data integration in projects such as the Monarch Initiative and the NCATS Data Translator. Here, we report on preliminary work and design plans. First, we are creating an OBO Library edition of NCIt, which will adhere to OBO conventions for identifiers, metadata, and accessibility. There are currently a number of experimental .obo format versions of NCIt, but none are official, and none have been created in collaboration with the NCI so as to ensure consistency, currency, and quality. We will also develop modules for specific subsets of NCIt, importable via distinct PURLs, for integration and reuse within other OBO ontologies, as well as equivalence axioms for bridging to existing terminologies such as those for anatomy (Uberon) and cell types (Cell Ontology, CL). A library of semantic queries will demonstrate use of the OBO Library edition of NCIt to answer questions in the cancer biology domain, both within its own terminology and via integration with other OBO ontologies and semantic web resources. We are also evaluating the consistency of cross-references to NCIt concepts found in related terminologies, with the goal of supporting use of NCIt in data integration and cross-dataset analyses. Finally, we utilize the NCIt for evidence modeling of cancer variant pathogenicity in accordance with the ACMG classification and in collaboration with ClinGen to support computational pathogenicity determination across diverse evidence sources.

By aligning with and supporting integration with OBO Library ontologies, the NCIt will be able to better support and benefit from work in the broader ontology and informatics community and support translational research.

ACKNOWLEDGMENTS

This work is supported by NCI/Leidos contract #17X118.