Adapting Disease Vocabularies for Curation at the Rat Genome Database

Laulederkind SJ, Hayman GT, Wang SJ, Bolton ER, Smith JR, Tutaj M, De Pons J, Shimoyama M Department of Biomedical Engineering Medical College of Wisconsin and Marquette University Milwaukee, WI, USA

Abstract- The Rat Genome Database (RGD) has been annotating genes, QTLs, and strains to disease terms for over 15 years. During that time the controlled vocabulary used for disease curation has changed a few times. The changes were necessitated because no single vocabulary or ontology was freely accessible and complete enough to cover all of the disease states described in the biomedical literature.

The first disease vocabulary used at RGD was the "C" branch of the National Library of Medicine's Medical Subject Headings (MeSH). For at least a few years it was the most publicly accessible, complete, and useful vocabulary to describe diseases and disease processes. However, it still had many holes in its coverage of disease vocabulary and an improved vocabulary was much desired.

By 2011 RGD had switched disease curation to the use of MEDIC (MErged DIsease voCabulary), which was a combination of MeSH and OMIM (Online Mendelian Inheritance in Man) constructed by curators at the Comparative Toxicogenomics Database (CTD). MEDIC was an improvement over MeSH, because of the added coverage of OMIM terms, but it was not long before RGD curators saw the need for more disease terms. So within a couple of years, RGD began to add terms to MEDIC under the guise of the RGD Disease Ontology (RDO). Since RGD assigned a unique ID to every MEDIC term imported from CTD, it was easy to add specially coded IDs to indicate those additional terms from a separate, supplemental file.

Meanwhile, the human disease ontology (DO) had slowly been developing and expanding. As early as 2010, members of RGD were contributing to the development of DO. However, five years went by before MGD (mouse genome database) and RGD joined with DO in an organized attempt to make DO useful for the model organism community. From that collaboration came a large addition of OMIM-based terms, expansion of multiparentage of terms through axiomatic extension, and expansion of cross-references to clinical vocabularies. Based on the promise of those improvements, it was determined that the Alliance of Genome Resources could use the DO as a unifying disease vocabulary across model organism databases.

Despite the improvements in DO, RGD still had more than 1000 custom terms and 3800 MEDIC terms with annotations to deal with if RGD would convert to the use of DO. Those extra terms originated from OMIM, MeSH, and the biomedical literature. If RGD mapped those non-DO disease terms to DO, much granularity of meaning would be lost. For instance the

Dwinell MR Genomic Sciences and Precision Medicine Center Medical College of Wisconsin Milwaukee, WI, USA

1

term "osteoarthritis" in DO has no children terms, but the RGD version of DO has 11 children terms or variations of "osteoarthritis" (Figure 1). The extra details of those terms is lost to users of DO. To avoid the loss of granularity it was decided to extend the DO beyond the merged, axiomized DO file. After mapping/adding DO terms completely to the RGD version of MEDIC, a broader, deeper disease vocabulary has been achieved, by providing more term branches throughout the ontology and more child terms within those branches.

Keywords— Rat Genome Database, disease, vocabularies, online resource



Figure 1. RDO Children terms of Osteoarthritis.

ICBO 2018



Figure 2. Osteoarthritis with Mild Chondrodysplasia

2