Making Phenotypic Data FAIR++ for Disease Diagnosis and Discovery

Melissa Haendel, Oregon Health & Science University; Jules Jacobsen, Queen Mary University; James Balhoff, RTI International; Jeremy Nguyen-Xuan, Lawrence Berkeley National Laboratory; Kent Shefchek, Oregon Health & Science University; Dan Keith, Oregon Health & Science University; Harry Hochheiser, University of Pittsburgh; Suzanna E. Lewis, Lawrence Berkeley National Laboratory; Sebastian Köhler, Charité – Universitätsmedizin Berlin; Peter Robinson, The Jackson Laboratory; Julie A. McMurry, Oregon Health & Science University; Tudor Groza, Garvan Institute of Medical Research, Sydney; Christopher J. Mungall, Lawrence Berkeley National Laboratory

While great strides have been made in exchange formats and data models for genomic sequence and variation data (e.g. Variant Call Format; VCF), the same is not true for phenotypic features. The heterogeneity of phenotype descriptions is a reflection of the different purposes for which they are collected (free-text clinical observations, QTLs, and newer high-throughput phenotype measurements) and the different contexts in which they are communicated (publications, databases, health records, registries, clinical trials, and even social media). Biocuration has effectively overcome these challenges in focused studies, but the field still challenged by the lack of broader phenotype standardization, accessibility, persistence, and computability. Consequently, it is extremely difficult to exchange, aggregate, and operate over phenotypic data in the same way that we do sequence data. We have therefore designed an exchange format standard for flexible, extensible, and expressive representation of a broad range of phenotypes in any species. The Phenotype eXchange Format (PXF) does not stop at just phenotypes; it accommodates any other evidence needed to make the very most of these phenotypes (eg. quantitative measurements, environments, and exposures). The goal is to enable PHI-free open exchange of phenotypic data in a way that can be automatically converted from existing registry/clinic data and easily shared outside paywalls for journals during publication of genotype-phenotype studies. In the context of the Global Alliance for Genomics and Health schemas and APIs, PXF will allow consumption of these phenotypic data for computational use by clinical labs for defining gene panels, for diagnostic pipelines, for rare disease patient matchmaking, and for deposition and aggregation in public knowledge bases such as the Monarch Initiative. Open exchange of phenotypic data promotes algorithmic innovation and sharing of our greater understanding of the correlations between genotype, environmental factors, and phenotypic outcomes in more holistic and translational manner. http://phenopackets.org