Semantic Index of Phenotypic and Genotypic Data

Charles T. Parker[1]* (chuck.t.parker@namesforlife.com), Nenad Krdzavac,[2] Kevin Petersen,[2] Grace Rodriguez,[1] and **George M. Garrity**[1,2]

[1]NamesforLife, LLC; East Lansing, Michigan and [2]Michigan State University; East Lansing, Michigan

http://ontology.namesforlife.com

Project Goals: The goal of this project is to develop a semantic data resource that can serve as a basis for predictive modeling of microbial phenotype. The core technical objectives are twofold: (1) to build a database of normalized phenotypic descriptions (observational data) using the primary taxonomic literature of bacterial and archaeal type strains, and (2) to construct an ontology with reasoning capabilities to make accurate phenotypic and environmental inferences based on that data. This project is tightly coupled with ongoing DOE projects (the Genomic Encyclopedia of Bacteria and Archaea, the Microbial Earth Project, the Community Science Project) and with two key publications, *Standards in Genomic Sciences* (SIGS) and the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM).

The DOE Systems Biology Knowledgebase (KBase) was envisioned to provide a framework for modeling dynamic cellular processes of microorganisms, plants and metacommunities. The KBase will enable rapid iteration of experiments that draw on a wide variety of data and allow researchers to infer how cells and communities respond to natural or induced perturbations, and ultimately to predict outcomes.

Predictive models rely on high quality input data, but not all data are of similar quality nor are they amenable to computational analysis without extensive cleaning, interpretation and normalization. Key among those needed to make the KBase fully operational are phenotypic data, which are more complex than sequence data, occur in a wide variety of forms, often use complex and non-uniform descriptors and are scattered about specialized databases and scientific and technical literature. Incorporating phenotypic information into the KBase requires expertise in harvesting, modeling and interpreting these data.

The Semantic Index of Phenotypic and Genotypic Data will address this problem by providing a resource of reference phenotypic data for all validly published type strains of *Bacteria* and *Archaea*, based on concepts and observational data drawn from the primary taxonomic literature. In the Phase I project we developed software to construct and analyze a corpus of this literature and to extract putative feature domain vocabularies comprising approximately 40,000 candidate phenotypic terms used in new and emended descriptions of the 12,937 distinct type strains of *Bacteria* and *Archaea*. In Phase II, these vocabularies have served as the basis for developing a phenotypic ontology, a repository of phenotypic data that is undergoing normalization of

phenotypic descriptions for each species. We have found that many of the phenotypes applied to microbes describe a combination of quantitative environmental conditions and qualitative growth and metabolic capabilities. Such terms have proven difficult to implement in query systems because of their context-based interpretations and conceptual overlap across multiple feature domains. We have furthered our work on novel design patterns for ontology development [1] that address these problems and remove barriers to machine reasoning over these complex terms, while preserving the bi-directional mapping back to human interpretation at multiple levels of abstraction. A PCT patent application on this method was filed in Q3 2014 and the preliminary examination has found all claims to be novel, non-obvious and industrially applicable. Critical to implementing this system has been the adoption of a SPARQL Inferencing Notation reasoner coupled with a triplestore rule-based reasoner. This approach resolves ambiguity attributed to the semantic equivalence and imprecision of phenotypic terms arising in literature and in databases.

In order to better facilitate access to knowledge extracted from the literature and encoded in the ontology, we are implementing a special-purpose web portal to accommodate query and retrieval of biological resources by term or concept, with a multi-tier query platform conforming to current search standards and backed by Semantic Web and Linked Data query standards. In addition to linking to the primary literature, related ontologies and source data, we are also incorporating public data from NCBI, USDA and the Joint Genome Institute in order to provide researchers with immediate access to the appropriate resources for a set of strains, along with consistent, accurate interpretations of available knowledge about those strains that are usable for predictive modeling and in other research and commercial applications.

As part of our commercialization activities, we continue to develop several software components that resulted from this project into a commercial semantic search and document analysis platform with end products being used in *Standards in Genomic Sciences* and the *International Journal of Systematic and Evolutionary Microbiology*.

Publications

1. Parker, CT, Garrity, GM and Krdzavac, NB. Systems and Methods for Establishing Semantic Equivalence Between Concepts. U.S. Patent Application No. PCT/US2014/056808. Filed September 22, 2014. Washington, DC: *U.S. Patent and Trademark Office*.

Funding for this project was provided through the DOE SBIR/STTR program (DE-SC0006191). Public funding for development of the NamesforLife infrastructure was received from the DOE SBIR/STTR program (DE-FG02-07ER86321), the Michigan Small Business Technology Development Corporation, the Michigan Strategic Fund, and the Michigan Universities Commercialization Initiative.