Name: George M. Garrity, Sc.D.
Positions: Professor, Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI
Co-founder and Managing Member, NamesforLife, LLC, East Lansing, MI

Overview of current relevant efforts: Extraction of biological information from the primary literature is a non-trivial task. In addition to the hidden biases introduced by curators (who must interpret results described by authors in a variety of ways) the process is confounded by a "chicken or egg" dilemma. Proper storage of curated data requires a well-designed data model *a priori*, but the knowledge of how to properly model (normalize) the data is only available *posteriori*, once the structure of the underlying information is well understood. Initial assumptions about what information can be extracted and how that information should be stored must undergo near continual revision as additional resources are added to a project corpus. How might one maintain quality, consistency and usability of stored observational data over time, knowing that both the information and the underlying data are fluid and often inconsistent or even contradictory?

We have developed a generalized process to mitigate these challenges that includes a flexible data model, document analysis methods, and a workflow. We begin by properly defining a target corpus of literature and drawing a sample set of documents designed to maximize information density. The corpus is then subjected to a statistical analysis to uncover high-frequency topic-specific terms and phrases as well as recurring text and patterns that are amenable to parsing with regular expressions. The statistical analysis also serves as the basis for our initial data model and a definition of the subject domain(s) encompassed by the corpus. The results are reviewed by subject matter experts and data curators, extracted terms are flagged for relevancy, grouped into appropriate topical categories, and working definitions are developed for those that are deemed relevant. These results are then used to refine topical data models, which evolve into proper relational database schemas, XML schemas and ontologies in a relatively short time frame. Term sets are mapped into our N4L Data Model, which accommodates rearrangements of taxonomies and refinement of the contextual and temporal meaning of terms. As terms are added, DOIs are assigned and made available for use in NamesforLife semantic annotation, tagging and indexing services and for incorporation into our ontologies.

While text mining, natural language processing and machine reasoning are all thought of as computational problems, our experience teaches that the human element, provided by subject matter experts and data curators is crucial if one is to obtain useable and meaningful results. Subject language terminologies (SLTs) are dynamic and may contain terms that have many nuanced meanings. SLTs often contain rare terms that significantly alter meaning but are not easily uncovered by machine learning methods. The use of the right tools at the right time is important. Much of the early work of curators can be easily bootstrapped using familiar off-the-shelf tools, such as spreadsheets, for gathering preliminary data, manipulating the output of statistical analyses and quickly visualizing results. We have found that custom application development can be postponed until after databases are initially loaded and integrated into the NamesforLife environment. The other critical point that is often overlooked in text mining applications is the significant effect that changes in publishing technology can have on source documents, especially when older literature is included in the corpus. Minor differences in white space, formatting, character sets and typographical errors can have significant impact on precision, accuracy and recall.

Technology needs: Many of the core technologies and resources needed to successfully mine metagenomic literature is either already in hand or is actively being developed for other applications in both the public and private sectors. The most efficient use of resources is to adapt proven computational methods to solve current problems rather than attempting to reinvent existing tools. On the other hand, expertise in the development of language resources needed to drive text mining applications in a meaningful way are in much shorter supply. Considerable effort is needed to develop normalized subject language terminologies based on actual usage by domain experts, to provide a mechanism for sustainably delivering this information, and to persistently integrate this knowledge into the literature. Failure to address this problem in a meaningful way will significantly diminish the potential return of any text mining initiative.

Candidate challenge problems: (1) To define a test corpus of the metagenomic literature based on the total published output in the field to date. (2) To define the relevant data types (including overlap with other related fields). (3) To develop a baseline terminology for each of those data types, along with domain values that are reported in the literature.