The NamesforLife Semantic Index of Phenotypic and Genotypic Data

Charles Parker[1], Catherine Lyons[1] and George M. Garrity[1,2],
[1]NamesforLife, LLC, East Lansing, Michigan, US and Edinburgh, UK
[2]Michigan State University, East Lansing, MI

NamesforLife
Bringing meaning to life...

Project Goals

Predictive models depend on high quality input data, but not all data are of similar quality nor are all amenable to computational analysis without extensive cleaning, interpretation and normalization. Key among those needed to make projects such as the DOE Knowledgebase (Kbase) operational are phenotypic data, which are more complex than sequence data, occur in a wide variety of forms, use complex and non-uniform descriptors and are scattered about the literature and specialized databases. Incorporating these data into the Kbase will require expertise in harvesting, modeling and interpreting the data. The *NamesforLife Semantic Index of Phenotypic and Genotypic Data* will be built on an ontology of bacterial and archaeal phenotypes based on the taxonomic literature. This project aims to achieve its first objective: a draft vocabulary for the phenotypic features of the taxonomic type strains.

Background

To manage dynamic terminologies Garrity and Lyons developed a semantic model (the N4L data model) that represents **names**, **taxa** (plural for taxon), and **exemplars** (representations of what is known about organisms) as distinct information objects (US Patent 7,925,444). The model is rooted in semiotic theory and provides a way to represent all of the complex relationships that exist among names and the concepts and objects to which names apply. Each such object is identified with a Digital Object Identifier (DOI) which allows for placement of forward-pointing links in the published literature and in databases and provides a mechanism for resolving ambiguities ("future proofing" a nomenclature). The data model is a context-driven method of semantic resolution and has already been deployed for *Bacteria* and *Archaea* (prokaryotes).

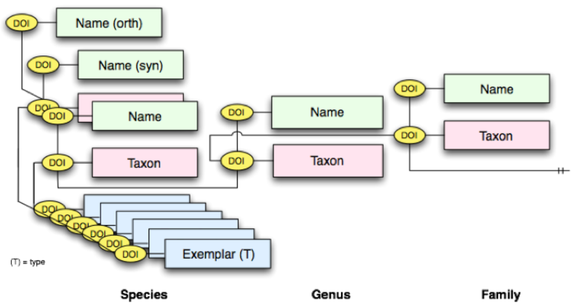


Figure 1. Creation of a nomenclatural taxonomy from N4L knowledge objects is accomplished by mapping the linkages via DOIs. This ensures that N4L::Exemplar objects are properly and persistently identified with any and all known synonyms, homonyms and orthographic variants in the correct manner. In addition to types, the NamesforLife taxonomy currently classifies all strains of *Bacteria* and *Archaea* for which either a sequenced 16S rRNA gene or genome exists in the public domain.

When coupled with Digital Object Identifiers (DOIs) this method provides a means by which names in digital content (e.g., journal articles, technical reports, web pages) and databases can be made actionable and directly linked to expertly curated information about the name, including its history of changes. NamesforLife, LLC has developed a suite of web services and applications based on this technology that can be used to semantically enrich or enhance digital content in a variety of formats.

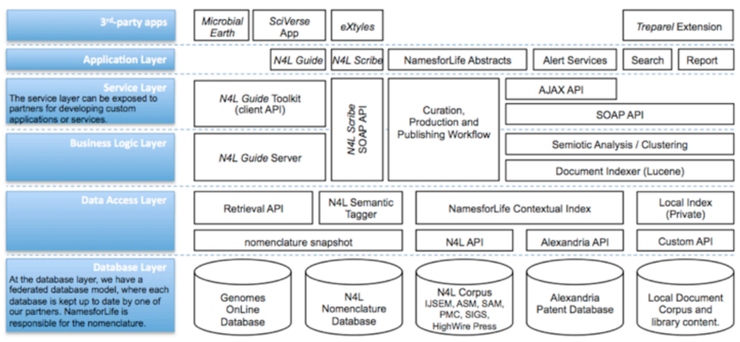


Figure 2. The N4L Data Architecture
NamesforLife uses a layered architecture. Company databases include the nomenclatural database and a reference database holding all of the taxonomic literature that is relevant to the names tracked by the company. The Alexandria patent repository (IFI - Fairview) contains more than 80 million patent documents from 70 countries and is processed and indexed with *N4L::Scribe*. We also provide a means of accessing public repositories (e.g., the Genomes OnLine Database) or private repositories that can be used by clients. DOIs provide the necessary infrastructure to ensure that intellectual property rights are protected. NamesforLife tools and methods provide access to these repositories and enrich web content by rendering contextually correct annotations based on biological names. Other custom solutions include the *N4L Contextual Index*, which interrelates current scientific, technical and medical and patent literature using our proprietary semiotic fingerprinting technology, and the *N4L::Taxonomic Abstracts*, a collection of approximately 50,000 citable micro-publications.

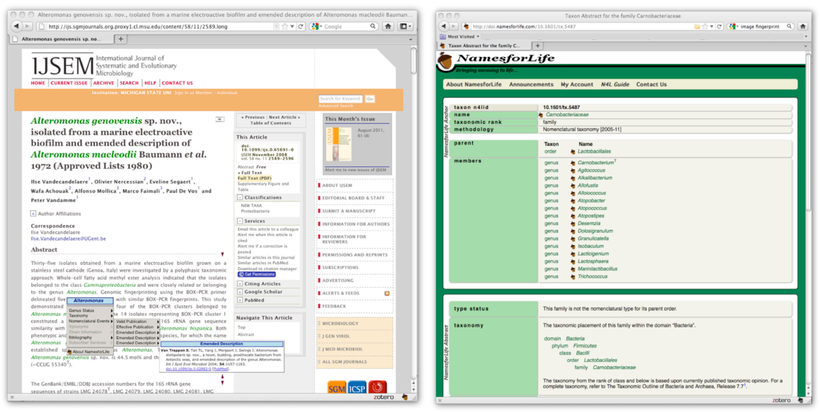


Figure 3. End user access to NamesforLife content - *N4L::Guide* is designed to provide readers with on-demand access to information while reading content in HTML form. (*Left panel*) Each instance of a validly published bacterial or archaeal name is converted into a link by the *N4L::Scribe* (either client side using the Reader Edition or server side using the Publisher or Developer Editions). (*Right panel*) *N4L::Taxonomic Abstracts* are citable micro-publications that are directly accessible via DOIs. Each species and subspecies is represented by three separate documents (Name, Taxon, and Exemplar Abstracts) in keeping with the NamesforLife model. Higher taxa are represented by two documents (Name and Taxon Abstracts).

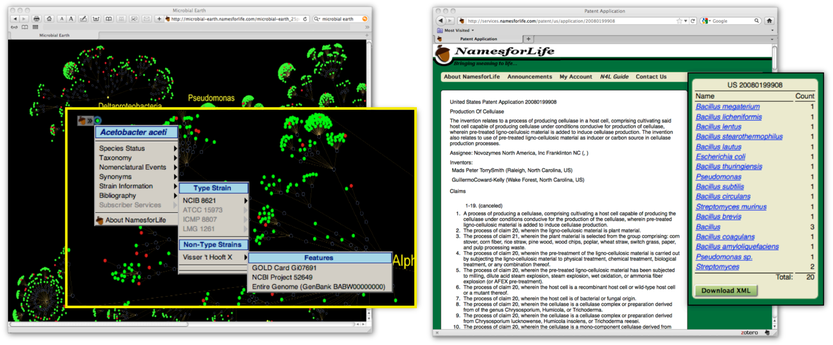


Figure 4. NamesforLife web services for other content types. (*Left panel*) Semantic enablement of the Microbial Earth Project tree by the *N4L::Guide*. (*Right Panel*) *N4L::PatentScribe* is designed to annotate US and non-US patents and is available to clients with a need for access to these documents. The *PatentScribe* embeds N4L::Name DOIs directly into XML instances of patents and provides end-users with not only annotation services, but a means for indexing, searching and analyzing this corpus of literature using the Company's Semiotic Fingerprinting.

Semiotic fingerprints are a special case of vector space models. Ours is distinct in that it uses an externally managed terminology in which the synonymies and other semantic ambiguities are automatically resolved by the N4L data model. This allows end users to fine-tune the level of taxonomic or phylogenetic granularity to meet specific needs. In our approach to mining text, the meaning of a term is known *a priori* and defines the scope of the search space. The combination of terms allows us to deduce the likely meaning of a document based on the properties of the organisms that are referenced. The more complex the fingerprint, the greater the resolving power.

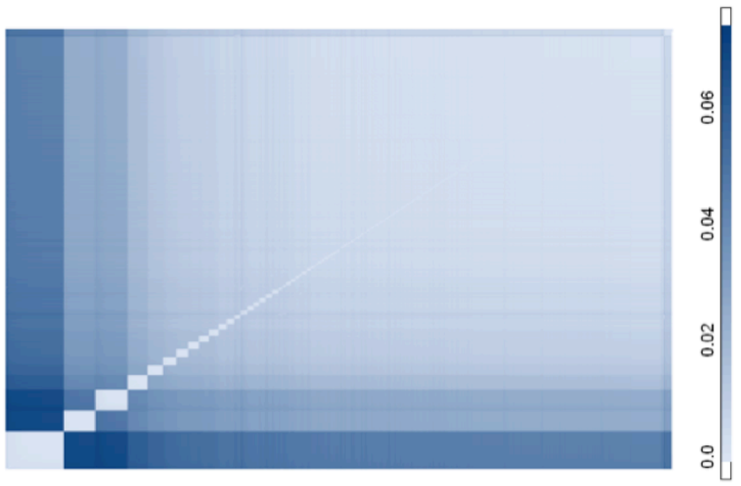


Figure 5. Clustering of patents by organism and technology classification. Preliminary experiments using the EPO Green technology patent collection from Fairview Research (n=380,000 patents) reveal the potential power of Semiotic Fingerprinting. A set of patents containing prokaryotic names (n=3,900) was produced using the *N4L:: PatentScribe,* which also extracts vectors of patent metadata (i.e., inventor, assignee, patent classification, patent authority, citations). The resulting similarity matrix was clustered, visualized as a heatmap, and output as an ordered list of patent IDs.

The NamesforLife Semantic Index of Phenotypic Data

Our methods and tools are not restricted to biological nomenclature and can be applied to terminologies of all types. Unlike sequence data, which are essentially universal, uniform and predictable, phenotypic data are inherently complex, noisy and "taxonomically parochial". The same trait may vary significantly under different conditions of growth, at different times during the life of a cell and under different environmental conditions. The language of phenotype is complex and may be limited in taxonomic scope and require expert interpretation. There is no equivalent to BLAST for searching for phenotypic data, and there is no central repository for such data. In some cases an entire language exists to describe the phenotypic features that apply to a single taxon (e.g., reproductive structures of *Cyanobacteria*, *Actinobacteria*, *Firmicutes*; complex life cycles of *Actinobacteria*, *Caulobacteria*) or a particular class of features (e.g., lipids). Phenotypic data must also be viewed from a historical perspective to understand what was measured and how it was measured (growth on substrate *vs.* hydrolysis of indicator compound). As such, it is critically important to know the methods that were applied and the comparability of the methods.

The long-term objective of this STTR project is to develop a semantic index of bacterial and archaeal phenotypes that can be used to augment annotation efforts and to provide a basis for predictive modeling of microbial phenotype. The index is based on published descriptions of taxonomic type and non-type strains that have been the subject of ongoing genome sequencing efforts as this will provide a mechanism whereby hypotheses can be tested and reproducibility verified. This project is tightly coupled with ongoing DOE projects (Genomic Encyclopedia of Bacteria and Archaea, the Microbial Earth Project, the Community Sequencing Project) and with two key publications, Standards in Genomic Sciences (SIGS) and the International Journal of Systematic and Evolutionary Microbiology (IJSEM). The first step towards accomplishing this goal, and the primary objective of this Phase I project is the development of a draft vocabulary.

Major Features Included in the NamesforLife Phenotypic Index

Strain metadata
- N4L Exemplar ID
- Isolation source
- Isolation method
- Isolation substrate
- Geographic location
- Environmental information
- Host
- Strain designation
- Collection ID(s)
- Taxon status (type/non-type)

Genotypic
- 16S rRNA sequence
- Other marker genes
- % DNA-DNA similarity
- % G+C composition
- Whole genome

Morphology

Micromorphology
- Cell size
- Cell shape
- Flagellation
- Sporulation
- Staining characteristics
- Other characteristics
- Intracellular inclusions
- Extracellular features
- Life cycle

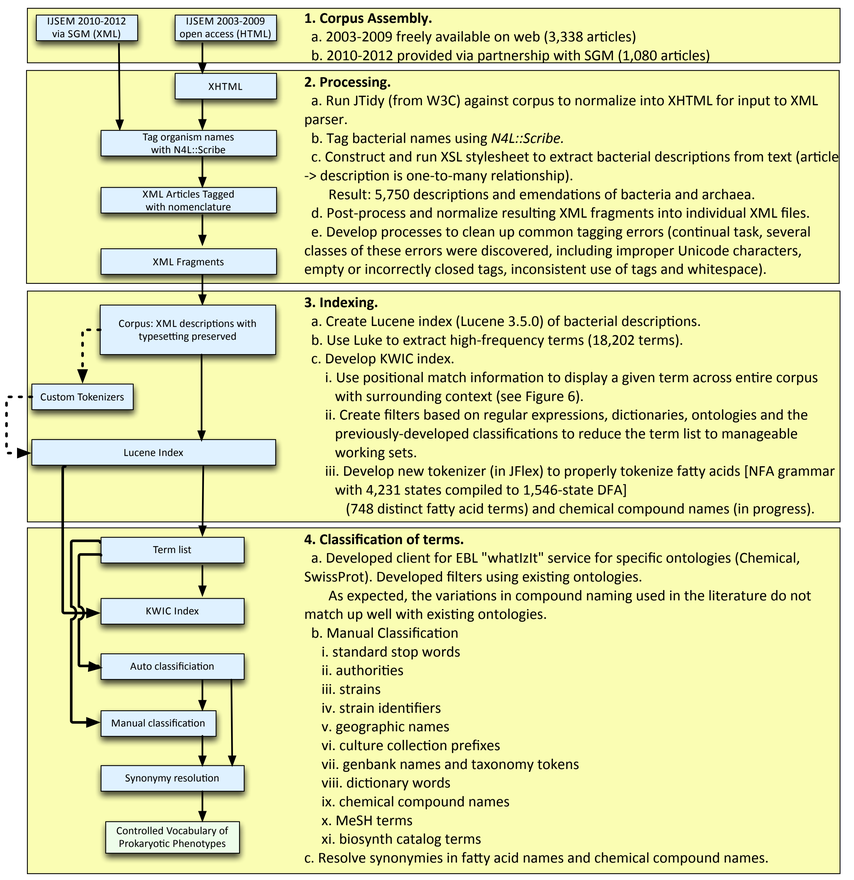
Macromorphology
- Growth on solid surfaces
  - Colony morphology
- Growth in liquid
- Pigment production
- Other features

Chemotaxonomy
- Fatty acids
- Polar Lipids
- Mycolic Acids
- Respiratory quinones
- Peptidoglycan composition
- Polyamines

Physiolgical
- terminal e- acceptor
- substrate utilization
- metabolic end-products
- sensitivity/tolerance to
  - chemical and physical agents

Our approach towards developing a draft vocabulary of bacterial and archaeal phenotype is based on a textual analysis of the richest source of descriptive information; the taxonomic literature. We follow a well-established path used for ontology construction based on derivation of domain-dependent hyponymy (is-a relationships) from a corpus and leverage tools, data resources and expertise that the Company has already developed. For this Phase I project, our target corpus consists of a subset of taxonomic literature of type strains from the IJSEM (2003-2012). The articles were indexed with Apache Lucene to produce two separate indices; one with the full articles and one with only the descriptions and emendations of organisms. We developed a KWIC (KeyWord In Context) interface to permit location and display of a given word in the corpus in its surrounding context to understand usage variations within and across different taxa. Selection of terms for analysis was initially done with Apache Luke, which provides facilities for determining usage frequency, coupled with curatorial review for relevance, categorization and synonymy.

Workflow



Lessons learned

Fatty acids are described using dozens of variations and so many deviations from the standard nomenclature as to render them useless as keywords without creating mappings to resolve synonymies. Default tokenization breaks apart chemical and fatty acid terms in ways that destroy terms (and meanings). For instance, C18:1ω7c is split into C18 and 1ω7c, which is not a proper conceptual break within the term. Similarly, when C_{18:1}11c is converted to plain text (as required by Lucene and other search platforms), the lexical structure is lost when the HTML tags are stripped, resulting in the nonsensical string C18:111c. Subsequently, the default tokenization breaks on punctuation and white space, resulting in the tokens C18 and 111c, further losing the original meaning of the text.

To work around these issues, we have developed custom grammars in JFlex (Java Fast Lexical Analyzer, used in Lucene for tokenizing). Our custom grammars properly tag nearly all variations of fatty acids in the corpus, as well as extract measurements, chemical compounds and strain identifiers. A drawback of creating such complex grammars is that they interfere with each other when used in serial tokenization. Therefore, we run them independently, creating multiple indices.

Given the number of variations in fatty acid and other compound names in the literature, it is difficult to navigate the term lists. Therefore, we have developed a second level of indexing against the terms themselves, an index of keywords, tokenized by the standard methods (punctuation and white space). The term list becomes searchable and manageable in this way. We have also retained the original token types (not normally available to an index) which allows us to hide entire groups of terms from the index (for instance, if we are working specifically on measurements or fatty acids).

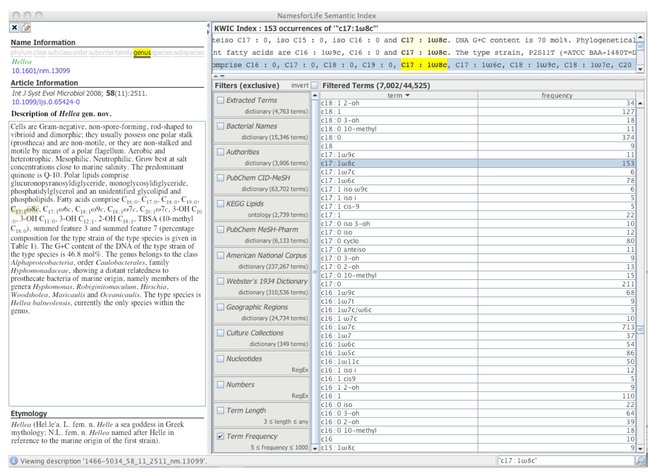


Figure 6. ***N4L::KWIC Index*** A view of the curatorial environment used for manual review and editing of terms used in the NamesforLife Index of Phenotypic Terms of Archaea and Bacteria.

Acknowledgments

Funding for this project was provided through the DOE SBIR/STTR program (DE-SC0006191). Funding for 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.