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

The core objective of this STTR project is to develop a semantic index of bacterial and archaeal phenotypes that can be used to augment genome annotation efforts and provide a basis for predictive modeling of microbial phenotype. The index is built from published descriptions of strains that have been the subject of genome sequencing efforts in order to provide a foundation for hypothesis testing and validation. The goals of this project are twofold: (1) to construct an ontology of bacterial and archaeal phenotypes derived from the taxonomic literature, and (2) to build a semantically-enabled database of phenotypic data using the ontology and primary taxonomic literature of bacterial and archaeal type strains.

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).

Background

Predictive models depend on high quality input data, but not all data are of similar quality and not all are amenable to computational analysis without extensive cleaning, interpretation and normalization. Key among the types of information 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 these data.

Unlike sequence data, which are universally applicable, uniform and predictable, phenotypic data are complex, noisy and "taxonomically parochial". A trait may depend on growth conditions, vary during the life of a cell and/or change in response to environmental conditions. The descriptions of phenotype can be complex, and may be limited in taxonomic scope and an require expert for correct interpretation. There is currently no equivalent to BLAST to search for phenotypic data, and there is no central repository for such data. In some cases an entire terminology exists to describe the phenotypes 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 also needs to be viewed from an historical perspective to understand not only what was measured but how it was measured (growth on substrate *vs.* hydrolysis of indicator compound). It is also important to know which methods were applied and whether different methods within an array of data are measuring the same trait, and if so, whether the results are comparable.

The *Phenotypic Index* will address these issues by tying together observations under specific sets of growth conditions, supporting faceted search, retrieval and comparison of differentiating characteristics between (and within) taxonomic groups. Each phenotypic observation will be linked to a strain via a *NamesforLife Exemplar DOI* (*Digital Object Identifier*), which is directly linked to an actively maintained taxonomy and nomenclature.

Table 1. Major Features Included in the NamesforLife Phenotypic Index, by feature class. Some of these features (i.e., those marked as completed in the Strain Metadata and Genotypic feature categories) are already available via the NamesforLife Taxonomic Abstracts (http://services.namesforlife.com).

Strain Metadata	**Morphology**	**Chemotaxonomy**†
☑N4L Exemplar DOI	**Micromorphology**†	☑Fatty Acids*
☐Isolation source	☑Cell size*	☑Polar Lipids*
☐Isolation method†	☑Cell shape*	☐Mycolic Acids
☐Isolation substrate†	☑Motility*	☐Respiratory quinones
☐Geographic location	☑Sporulation*	☐Peptidoglycan composition
☐Environmental information	☑Staining characteristics*	☐Polyamines
☐Host	☐Intracellular inclusions	**Physiological**†
☑Strain Designation	☐Extracellular features	☐terminal e- acceptor
☑Collection ID(s)	☐Life cycle	☐substrate utilization
☑Taxon status (type/non-type)	☐Other characteristics	☐metabolic end-products
Genotypic	**Macromorphology**†	☐sensitivity/tolerance to chemical and physical agents
☑16S rRNA sequence	☐Growth on solid surfaces	☑optimal growth conditions*
☑% DNA-DNA similarity	☐Colony morphology	
☑% G+C composition	☐Growth in liquid	
☑Whole genome	☐Pigment production	
☐Other marker genes	☐Other features	

* features extracted but not yet curated
†features requiring normalization and ontological mapping

Corpus Construction

In previous work, *NamesforLife*, in cooperation with the *Society for General Microbiology*, created a digital library of the primary taxonomic literature for bacteria and archaea. This corpus serves as the source of both the ontology and the data. Since errors introduced at this point would cascade into the our final products, a number of quality control steps were introduced to ensure that clean XML representations of each taxonomic description were created.

Our text normalization workflow is a semi-automated process that required some analysis and development of heuristics, as well as manual intervention for text cleaning and transformation from the source articles. Direct text extraction from PDF documents is confounded by pagination, multi-column layouts, and figures or tables that interrupt the natural flow of text. In cases where only PDF documents were available, text was copied and pasted from the source into raw text files, restoring the natural flow of the text before being placed into a staging area for XML conversion (Figure 1). Transforming HTML to XML did not suffer from this problem, as the presentation and layout of HTML is (in general) separate from the underlying data. However, heuristics were employed to demarcate document sections (title, authors, keywords, publication metadata, abstract, material and methods, results and discussion, acknowledgements, references, copyright) and re-tag in XHTML format.

An HTML5-compliant document structure was chosen as the storage format for the corpus, as it supports the minimal requirements for preserving text formatting features (headers, paragraphs, italics, bold, subscript, superscript, lists, metadata, hyperlinks) along with document sections and subsections. HTML5 documents validate as XML, are directly usable with our existing tools, can be easily transformed to other representations and renders correctly in a browsers (allowing easy QA/QC without requiring supporting software or the use of a Content Management System).

Indexing and Advanced Parser Design

Indexing, document editing and text analysis were designed as a closed loop. Once a document is introduced into the system, edits and term extraction are performed using a single instance. Updates to the document are immediately reflected in the index. *NamesforLife* text analysis tools are built in *Java* using *Apache Lucene* with in-memory indices created on demand. A POJO (Plain Old Java Object) representation of the document structure is mapped to the in-memory DOM (Document Object Model) of the descriptions. All search and text analysis relies on the in-memory index. An intermediate business logic layer was developed to infer index fields and parameters via introspection of the POJO, obviating the need for index configuration.

Parsing the corpus presented a number of challenges. For instance, fatty acids are described using three different systems of nomenclature with numerous variations and deviations that render them useless as keywords in the absence of a mapping to resolve synonymies. Most indexing platforms break apart chemical terms in ways that destroy both their meaning and textual representation. For instance, when the formatted string `C<sub>18:1</sub>11c` is converted to plain text (as required by *Lucene* and other search platforms), the lexical structure is lost if the formatting is removed, resulting in a nonsensical string **`C18:111c`**. Also, the default *Lucene* tokenizer considers punctuation and white space as token boundaries, resulting in the tokens **`C18`** and **`111c`**, with concomitant loss of meaning.

We employ two strategies to solve this problem. First, we have developed custom grammars in *JFlex* (*Java Fast Lexical Analyzer*, used in *Lucene* for tokenizing). Our custom grammars properly tokenize all variations of fatty acids names in the corpus and correctly recognize measurements, other chemical compounds and strain identifiers. More importantly, our tokenizers recognize *NamesforLife Digital Object Identifiers*, which are inserted into XML documents by a business logic layer using *NamesforLife* annotation tools. These annotations reduce complex tokens to persistent identifiers, which are understood by our tools to have unambiguous meanings.

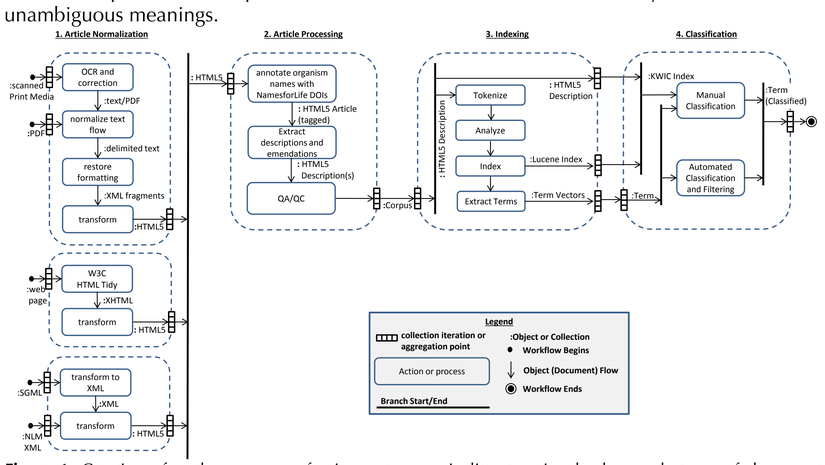

Figure 1. Creation of a clean corpus of primary taxonomic literature involved several stages of document normalization, annotation, and manual cleanup. As a result of this process we now have a collection of roughly 7,500 microbial descriptions in a minimal HTML5 format. For any documents that were available in multiple formats, the least lossy version was used (i.e., in cases where the XML instance was not available, we used SGML, HTML, or PDF versions as the source). For approximately 1,200 descriptions we resorted to scanning the original descriptions from books or journals to produce high quality PDF files that were then converted to text by OCR (Optical Character Recognition), followed by manual correction before conversion to XML. Manual OCR correction is a labor-intensive but necessary part of our approach since the quality of the resulting ontology and database depends entirely on the quality of the input.

Bootstrapping the Ontology

The concise nature of the phenotypic language lends our corpus to classification on a sentence by sentence basis. A survey of existing sentence splitters commonly used in text-mining applications failed to identify any that could correctly parse the complex terminologies found in our corpus, necessitating the development of a custom sentence splitter. We have also developed a text analysis method that employs an inverted index in combination with the TF-IDF coefficient (*Term Frequency - Inverse Document Frequency*) to provide an indication of *term strength* within predetermined document classes of a corpus. These have been used in combination with our custom tokenizer to identify, term vectors belonging to each of the major features in Listing 1, with a high degree of confidence.

The sentence classifier can be bootstrapped using a set of initial training terms to select appropriate subsets of the corpus and within a few iterations produce a set of high-scoring terms that are conceptually related and belong to topical category of interest. These lists are manually validated, and are being used to compile descriptive dictionaries for each major feature (Listing 1 and Figure 2).

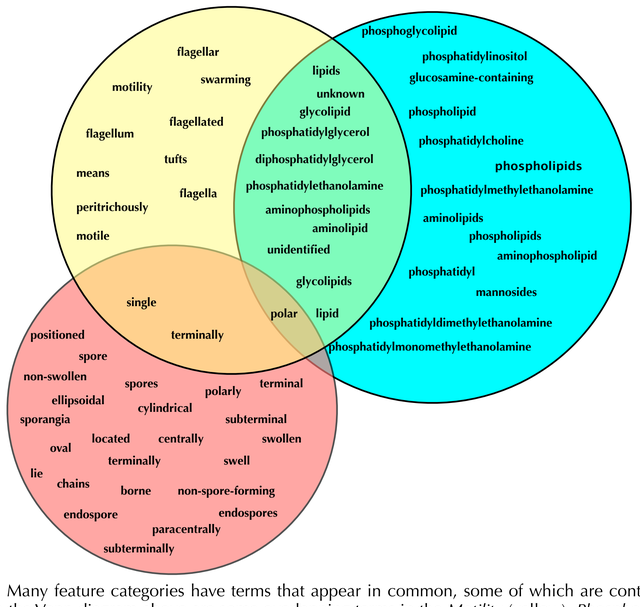

Figure 2. Many feature categories have terms that appear in common, some of which are context-dependent. Shown in the Venn diagram above are some overlapping terms in the *Motility* (yellow), *Phospholipid* (blue) and *Sporulation* (red) feature categories that have been extracted using the TF-IDF term strength algorithm described above. In building out the phenotypic ontology, we expect that a subset of terms (notably adjectives) will map into multiple groups, and concise definitions will be assigned for each distinct usage of the term in context.

Project Status

Much of the strain metadata (*N4L Exemplar DOI*, *Strain Designation*, *Collection Identifiers*, *Taxon Status*, *16S rRNA Accession*, and *Whole Genome Accession*) is already curated and is available in our *NamesforLife Taxonomic Abstracts*.

Corpus construction is essentially complete (with the exception of the ongoing OCR work, which has yielded descriptions for an additional 599 strains). Our custom text analysis and annotation tools are operational and have already produced excellent starting points for ontology development (Figures 2 and 3).

Feature categories consisting of mostly numeric data, such as *Cell Size*, *%G+C Composition* and *%DNA-DNA Similarity* can now be extracted from the corpus on demand (Figure 4).

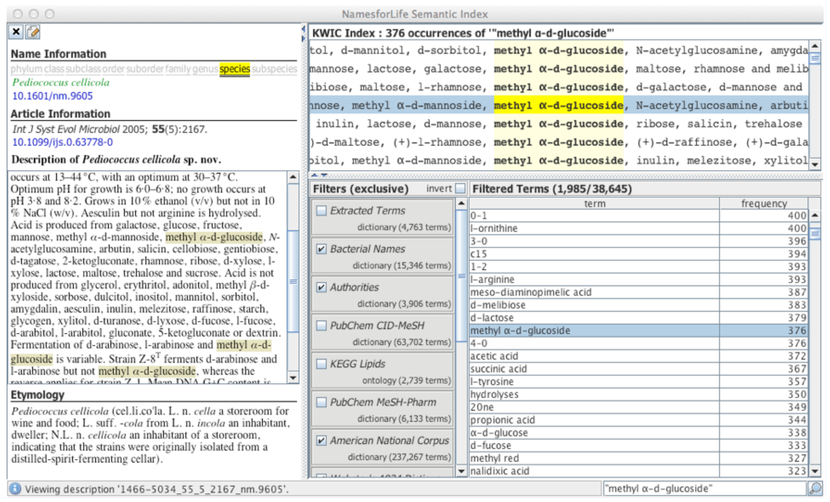

Figure 3. This *Extended KWIC* (*Key Word In Context*) *Index* incorporates several new software components developed during this project. This application is used to rapidly identify candidate terms for the ontology and investigate their usage in the taxonomic literature. In the above screenshot, we see that the descriptions of 376 type strains contain occurrences of "`methyl α-d-glucoside`". A curator can scan through each description in the taxonomic literature to collect examples that demonstrate every usage variation of that term (e.g. "acid production from", "no acid production from", "ferments", "does not ferment"). The ontology will contain entries for these metabolic processes as well as the chemical substrate. The phenotypic database will contain a mapping for this strain in a nested *EQ* (*Entity-Quality*) form:

`<Strain DOI, <Utilization, Substrate>>`

Current Work

The ontology development is starting to get underway, with *Cell Shape*, *Motility*, *Staining Characteristics* and *Fatty Acids* as the first descriptive feature domains targeted for completion. Orthographic variants (common names and typographic errors) will be mapped onto the ontology, as was done for names. Data modeling for storage of the *EQ* (*Entity-Quality*) relationships for the Phenotypic Database is underway. *Cell Size* in particular is problematic, as it is described in many ways depending on *Cell Shape*. Some preliminary data from the feature extraction is shown in Figure 4 and Tables 1a and 1b. A REST API for data access is in the design phase, and will serve as a platform for application development.

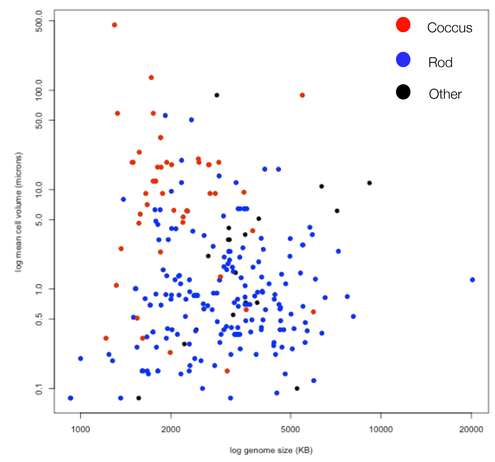

Figure 4. A log-log plot of microbial cell volume vs. genome size for those strains which have both data points available. Cell volumes are calculated directly from cell size measurements and morphologies that have been automatically extracted from the corpus. Genome size is calculated from sequence data downloaded from GenBank, via accessions in the *NamesforLife* database. A consistent, high quality data set with complete taxonomic coverage will provide answers to many questions about the relationships between physiological, taxonomic and genotypic characteristics.

Fatty Acid	Extracted
straight chain	399
unsaturated	
mono-	255
di-	12
tri-	4
{tetra-hexa}-	6
branched	315
hydroxy	141
DMA	13
cyclo-	4

Table 1a (left). A high-level classification of the distinct fatty acid terms extracted from the corpus. Many of these are orthographic variations and at least 5% are errors in the original publications. The usage of each term will be investigated in context using the *KWIC Index* (Figure 3). These will be mapped to identifiers in existing ontologies where available.

tokenized	733
reduced set	675
matched in *LipidMaps*	205
unidentified	529
normalized synonyms	412
synonyms (11-21)	6
synonyms (2-9)	76
unique	330

Table 1b (above right). A preliminary comparison of the fatty acid terms extracted from the taxonomic literature to those found in the *LipidMaps* database (http://lipidmaps.org/). The easily identifiable terms will serve as a starting point for ontology development.

Future Work

As soon as the ontology for any of the major features is completed, work can begin on populating the phenotypic profiles. Some of the major features (*Cell Size*, *%G+C*) will lend to bulk loading. Other features (e.g., *Isolation Method*, *Growth in Liquid*) will require some amount of curatorial interpretation. To minimize the amount of curation effort, we are investigating the use of Natural Language Processing to take advantage of the highly repetitive nature of the grammars used in sentence construction for microbial descriptions. Planned applications for this resource include faceted strain searching, a strain registration service, and the ability to recast existing strain descriptions in unambiguous language.

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, the Michigan Economic Development Corporation, and the Michigan Universities Commercialization Initiative.