A Semantic Index of Phenotypic and Genotypic Data

Charles Parker[1], Nenad Krdzavac[2], Kevin Petersen[2], Amber Roberts[2], Grace Rodriguez[1] and George M. Garrity[1,2]
[1]NamesforLife, LLC, East Lansing, Michigan, US
[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 phenotypes. 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 threefold: (1) to construct an ontology of bacterial and archaeal phenotypes derived from the taxonomic literature, (2) to build a semantically-enabled database of phenotypic data using the ontology and primary taxonomic literature of bacterial and archaeal type strains, and (3) to develop commercially available services leveraging these unique resources.

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

While it is generally acknowledged that microbes are an invaluable source of commercially useful new products and processes, discovery is often based on finding not only the right strain but also the right growth conditions to achieve a desired outcome. Access to accumulated data and knowledge about the metabolic and genetic potential of the strains of interest are essential for success. But not all data are of similar quality nor are all data amenable to modern approaches of computational analysis without extensive cleaning, interpretation and normalization. Key among these 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.

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.

To address this need, we are constructing a Semantic Index of Phenotypic and Genotypic Data, built on an ontology of microbial phenotypes and growth conditions extracted from the taxonomic literature. In this project, we utilize a new class of Data as a Service (DaaS), based on reasoning over curated data. This approach provides a powerful query method that is tolerant of missing or ambiguous information. When combined with Description Logics (DL; a family of formal knowledge representation languages) it has the potential to unlock a wealth of hidden knowledge buried in observational data by supporting inferences about properties of interest. The resulting resource will serve as a useful bridge from discovery to production.

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 (Table 1). 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
- ☑N4L Exemplar DOI
- ☐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
- ☑% DNA-DNA similarity
- ☑% G+C composition
- ☑Whole genome
- ☐Other marker genes

Morphology

Micromorphology†
- ☑Cell size*
- ☑Cell shape*
- ☑Motility*
- ☑Sporulation*
- ☑Staining characteristics*
- ☐Intracellular inclusions*
- ☐Extracellular features*
- ☐Life cycle
- ☐Other characteristics

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

Physiological†
- ☐terminal e- acceptor
- ☐substrate utilization*
- ☐metabolic end-products
- ☐sensitivity/tolerance to chemical and physical agents*
- ☑optimal growth conditions*

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

A Note on 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 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).

Materials and Methods

In our previous work (a collaboration with the Society for General Microbiology), we developed a clean, complete and fully-referenced resource of nomenclature and taxonomy for Bacteria and Archaea. One major result from this work was the ability to collect and process the original and emended descriptions of all type strains of prokaryotes.

Creation of a clean corpus of primary taxonomic literature was the obvious first step of our current project. This involved several stages of document normalization, annotation, and manual cleanup. As a result of this effort we now have a collection of nearly 10,000 microbial descriptions in an XML/XHTML schema. For approximately 1,250 protologs 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 directly on the quality of the input.

In the second stage of our work, we needed to generate putative vocabularies for each of the major feature categories listed in Table 1. Traditional text-mining approaches proved inadequate early in our Phase I STTR, and as a result we developed a novel approach to generating cohesive putative vocabularies based on seed terms. As a follow-up to this work, we also developed a concept-mining application based on this approach (Figure 1). The *NamesforLife Semantic Index* is a Java application built on open source indexing libraries (*Apache Lucene*), our proprietary vocabulary generation algorithm and a flexible document/domain model backed by the NamesforLife nomenclature and taxonomy database.

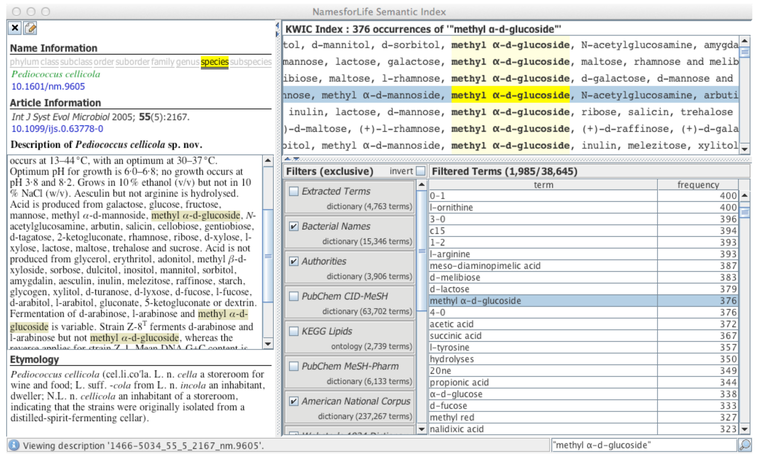

Figure 1. This *Extended KWIC* (*Key Word In Context*) *Index* incorporates several new software components developed during this project. This application is used to rapidly identify candidate terms for the ontology and investigate their usage in the taxonomic literature. In the above screenshot, we 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>>
```

Several iterative rounds of data modeling were performed, resulting in a core meta-model that is used as the main reference for conceptual and relational mapping. In order to leverage the best features of relational databases, structured text and ontology approaches, we have created a multi-faceted architecture that employs all three approaches (Figure 2). The relational database stores normalized data as interpreted by domain experts (biocurators). The XML markup is generated directly from the database and is used both to populate the ontology data store and to serve structured data to consumers of service end-points via our developer API (REST).

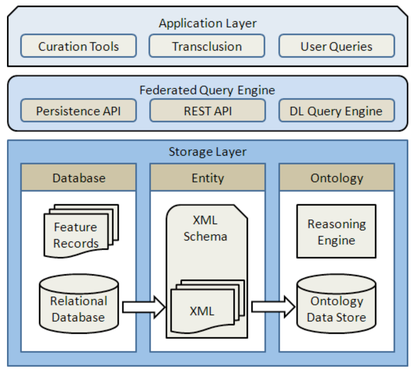

Figure 2. Our multi-faceted approach leverages the best features of three different approaches to Knowledge Storage. The relational database stores the individual phenotypic features as entered by domain experts, and answers queries based on that data, including statistical queries. The XML format is used in several ways: (1) as a long-term storage format that can be archived or delivered to subscribers; (2) as a transient data structure that can be delivered in multiple formats via a REST API; and (3) as a structured input to the ontology data store. A reasoning engine is leveraged to fulfill Description Logics queries that go well beyond what is possible with traditional relational database queries. In order to provide a consistent interface for end-users and data consumers, we are developing a federated query engine to abstract the queries for each system into a common API. DL queries to the ontology may be opened up as a direct service, with the federated engine leveraging named queries or cached queries.

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 a 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 a custom tokenizer to identify term vectors belonging to each of the major features in Table 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 (Table 1).

Many feature categories have terms that appear in common, some of which are context-dependent. In building out the phenotypic ontology, we have found a subset of terms (notably adjectives) that map into multiple ontological concepts, and concise definitions will be assigned for each distinct usage of the term based on 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 monthly additions via newly published literature, amounting to approximately 1,000 strains per year). Our manual OCR correction work has yielded descriptions for an additional 1,250 strains from historical literature. Our custom text analysis and annotation tools are operational and have already produced excellent starting points for ontology development (Figure 1). 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.

The core meta-model (Eclipse Ecore) is mostly complete and is refined as necessary. Several XML schemas and relational database schemas have been developed and will soon be ready for integration testing.

Ontology DL queries are currently available on a demonstration server (http://ontology.namesforlife.com) for a selected subset of the core ontology.

Our ontology experts are currently refining the core models of the ontology and developing axioms that can be employed by the reasoning engine to infer new knowledge from the accumulated, curated historical literature.

Current Work

Now that the core ontology development is finished (with *Cell Shape*, *Motility*, *Staining Characteristics* and *Fatty Acids* as the first descriptive feature domains completed), our ontology staff has moved on to axiom development, which is critical for reasoning over the data store and ontology. We expect much progress on this task during the next quarter, with the intent to perform a full integration test.

Our domain experts are currently mapping orthographic variants (common names and typographic errors) onto the ontology. Additionally, they are developing concise definitions for each term in the controlled vocabularies, with emphasis on precisely matching each concept as it is modeled in the ontology. These definitions will be available and applicable in every part of our system, including serving as the annotation for our XML schemas (Figure 3).

A REST API for data access is in early development phase, and will serve as a platform for application development and one part of the federated query engine (Figure 2).

```
<?xml version="1.0" encoding="UTF-8"?>
<media xmlns="http://namesforlife.com/ns/media" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  <medium medium-id="ATCC_37">
    <name>CHROMATIUM MEDIUM</name>
    <component component-id="ATCC_37.1">
      <name>Solution 1 </name>
      <ingredient ingredient-id="ATCC_37.1.1">
        <substance substance-uri="https://www.ebi.ac.uk/chebi/searchId.do?chebiId=3312" />
        <amount unit-uri="http://purl.obolibrary.org/obo/UO_0000021">2.0</amount>
      </ingredient>
      <ingredient
        <substan                                                    )?chebiId=CHEBI:15377" />
        <amount                                                     amount>
      </ingredient
    </component>
    <component component-id="ATCC_37.2">
```

A substance, such as a chemical or complex compound. The content of the element may optionally contain a free text description or name of the substance. A substance, such as a chemical or complex compound.

Figure 3. A sample of XML for descriptions of growth media. The underlying XML schema provides annotation of each element and attribute. Each annotation maps to a term in a controlled vocabulary that has a precise definition mapping directly to the ontological concept that it represents. The XML is generated automatically via the persistence layer for the relational database. The XML, relational database, and ontology all conform to the core meta-model, which ensures lossless conceptual mapping from one storage platform to another.

Future Work

As soon as the integration test is completed, work can begin on populating the phenotypic profiles from the semi-normalized protologs. 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 (Figure 4). We are also conducting a feasibility study on developing a discriminative feature engine based on DL reasoning. Our first commercial services are planned as subscription access to the federated query engine.

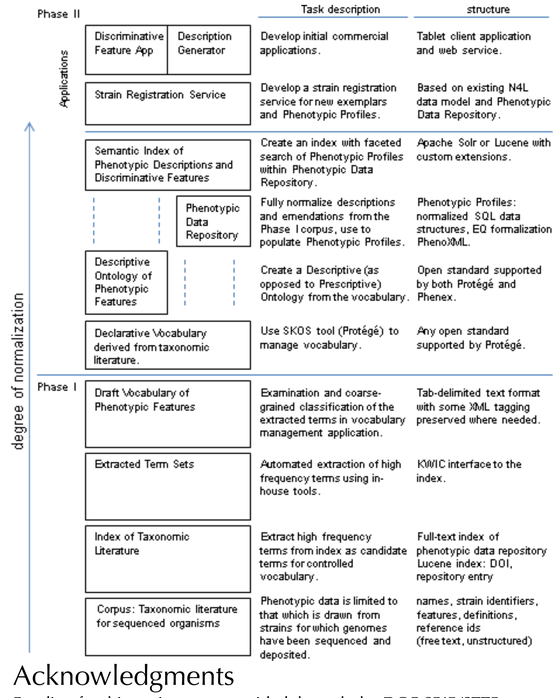

Figure 4. The scope of the project as viewed from a data normalization perspective. The quality and usefulness of each level of normalization depends directly on the quality of the next-lower level.

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.