The NamesforLife Semantic Index of Phenotypic and Genotypic Data: Phase I Final Technical Report

DOE Award Number: DE-SC0006191

FY 2011 SBIR/STTR Phase I Funding Opportunity (DE-FOA-0000413)

Topic 34 GTL and Related Biotechnologies

Subtopic b. Software Tools for the GTL Systems Biology Knowledgebase (SBK).

Date – May 31, 2012, revised November 28, 2013

PI - **George M. Garrity**^{1,2} (garrity@namesforlife.com)

¹NamesforLife, LLC, 333 Albert Ave., Suite 202 East Lansing, MI and

²Department of Microbiology and Molecular Genetics, 567 Wilson Rd., Michigan State University, East Lansing, MI

<http://www.namesforlife.com>

Phase I project goals: Predictive models depend on high quality input data. But not all data are of similar quality nor are all of the data amenable to computational analysis without extensive cleaning, interpretation and normalization. Key among those needed to make the 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 requires expertise in harvesting, modeling and interpreting these data. The *NamesforLife Semantic Index of Phenotypic and Genotypic Data* seeks to address this problem by taking the first steps toward building an ontology of bacterial and archaeal based on the taxonomic literature through the development of a draft vocabulary of phenotypic features.

Background

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 also complex and may be limited in taxonomic scope, requiring expert interpretation, as 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 phylum, class or order (e.g., *Cyanobacteria*, *Actinobacteria*) or a particular class of features (e.g., lipids, structural carbohydrates). In addition, phenotypic data must be viewed in a historical perspective (time when the data were collected) to understand what was measured and how it was measured. As was the case with microbial physiology, which had fallen out of fashion as a field of research, so too has “classical” or polyphasic microbial taxonomy resulting in a dwindling community of experts who can readily interpret the existing information in the literature and in various specialized databases. How might that expertise be captured and applied to developing a standardized language and ontology of microbial phenotype?

In 2003, Garrity and Lyons [1,2] proposed a novel approach to resolving ambiguity of biological nomenclature; a long-standing problem in the life sciences. Their approach, which loosely reflects the semiotic triangle of Ogden and Richards [3] with modifications from Sowa [4], provided a means of modeling the complex relationships that exist among names and the concepts and objects to which names apply (Figure 1). When coupled with Digital Object Identifiers (DOIs) [5], this provided a means by which names in digital content (e.g., journal articles, technical reports, web pages and databases) could be made actionable and resolve to an expertly curated record for the name, including its history of changes. They suggested that such an approach would empower end-users by providing access to literature and data using either current names or alternative names (synonyms). Their approach would also provide a means of keeping back-content and databases up-to-date, persistently and automatically, effectively *future-proofing* biological nomenclature.

Figure 1. The semiotic triangle (left) and its application to biological nomenclature (right). Ogden and Richards [3] show that uncertainty arises from a failure to recognize that names (symbols) that are assigned to objects (referents) have meaning to the agent that *interprets* them that may differ from the meaning intended by the agent that *transmits* them. With some adaptation, this model is applicable to biological nomenclature and addresses the unpredictable decay that occurs because the taxonomic concept to which a name refers changes as new members are recognized or other rearrangements occur.

NamesforLife, LLC (the Company) first reduced the NamesforLife (N4L) model to practice in 2007 by building a prototype that fully modeled all validly published names¹ of *Bacteria* and *Archaea*. The company subsequently developed a suite of web services and applications that are used to semantically enrich digital content in a web browser and to enhance digital content in XML or other formats so that each occurrence of a validly published bacterial or archaeal name becomes actionable and can return a wealth of information about a named organism or higher taxon, on demand. The web services provide direct access to expertly curated information served in a variety of formats (i.e., REST, JSON, XML, XHTML) [6-8].

NamesforLife has already created the necessary infrastructure to address the stated needs for a reference set of phenotypic data as a part of a previously funded DOE STTR program.

While the N4L Data Management Platform was originally conceived to provide persistent access to correct nomenclatural and biological data, it has subsequently evolved into a much richer resource and can now accommodate phenotypic, genotypic and genomic information, along with a record of all relevant literature in which each validly published

¹ A validly published name is defined as a name that conforms to *all* of the rules of the International Code of Nomenclature of Bacteria [11] and is available for use.

² SGML: Standard Generalized Markup Language, the parent language of XML and XHTML.; DTD:

name occurs in the scientific, technical, medical and patent literature [9]. Each reference is directly accessible using any of the names by which a particular organism or taxon may be known, in any given temporal context.

Early efforts to fully index the relevant literature provided a glimpse of what is possible with regard to extracting and associating a much richer data set with specific strains (e.g., those involved in patented process or products). Subsequently, we proposed building a semantic, (i.e., richly typed), index of the taxonomic literature by developing new methods to extract and classify relevant information (e.g., pathways, specific phenotypic traits) from that corpus and other digital resources, and to integrate normalized phenotypic data with the Company's index of genomes. We also proposed developing a registration service to capture strain information and related phenotypic and genotypic data in a manner that will make these data persistent and actionable (linked to related data). As part of our semantically-aware resources this could help to accelerate the discovery of new genes or the development of gene products and organisms that are relevant to the DOE mission. NamesforLife, LLC is in a unique position to accomplish this goal, based on the resources we have developed under our current and previous STTR awards and through our network of collaborators in the public and private sectors.

The N4L Data Model

The N4L data model and the associated semantic resolution technology are rooted in a long-standing collaboration between the PI and Catherine Lyons, a co-founder of the Company. Together, they developed and deployed the custom SGML DTD² used to publish *Bergey's Manual of Systematic Bacteriology, 2nd ed.*, for which the PI served as Editor-in-Chief. During the development work, it became apparent that names, which serve as the fundamental organizing principle of the *Manual*, also represented one of the major obstacles, not only in organizing the content, but also in using the content. The problem lies in a general misunderstanding of the principles of bacterial nomenclature among contemporary microbiologists [10] who equate names with identifiers and expect persistence. However, biological names are neither persistent nor unique. In fact, biological names are often semantically ambiguous.

The N4L data model is a model of semantic resolution in which we recognize that **names**, **taxa** (plural for taxon) and **exemplars** (representations of what is known about organisms) are separate classes with distinct properties (Figure 2). The exemplar concept is unique to NamesforLife [2] and is an index to the individual organism that may bear zero, one or more than one name and taxonomic concept. Exemplar objects are expressed in terms of phenotypic and genotypic data, culture collection deposits, links to gene and genome data in GenBank, and other relevant information, specific to a given organism. The exemplar provides a virtual representation of the referent in the Ogden-Richards semiotic triangle and persists beyond nomenclature and across taxonomic opinion. In ontological terms it is a reified relation associating all that is known about an individual organism. In the current N4L implementation exemplars hold the following information: culture collection and lab strain identifiers, 16S rRNA sequence identifiers, links to genome sequences where available, and links to key references. However, an exemplar object can

² SGML: Standard Generalized Markup Language, the parent language of XML and XHTML.; DTD: Document Type Definition, a formal syntax for structured languages (i.e., SGML and XML).

hold many types of information, gleaned from the literature or elsewhere. The exemplar model can also accommodate incomplete information (according to standard practice in taxonomy). Change in the state of knowledge about an organism is reflected by the exemplar. The N4L exemplar object is therefore a surrogate for the strain of interest, and can be used to monitor knowledge about that strain over time.

Figure 2. The N4L Semantic Model. Clarity of meaning requires precise and unambiguous definitions that can be directly referenced on demand. Biological names can be ambiguous because they are subject to change, are not guaranteed to be unique, and may exist in one-to-one, one-to-many, many-to-one and many-to-many relationships. Biological names (Names) and taxonomic concepts (Taxa) are precisely defined through the processes of typification of biological materials (type strains) at the species and subspecies level (upper figure) and type concepts (species, genera and orders) for higher taxa (lower figure) that are part of published circumscriptions.

Names, taxa and exemplars exist in a variety of one-to-one, one-to-many, and many-to-many relationships that arise over time, reflecting a gain in knowledge that triggers taxonomic rearrangements and accompanying nomenclatural events, as defined by the International Code of Nomenclature of Bacteria (ICNB) [11]. These taxonomic and nomenclatural events, which occur unpredictably and asynchronously, lead to a proliferation of synonymous names, and the problematic situation where the number of names is greater than the number of taxa and the number of strains to which those names apply (Figure 3). There is also a temporal component that must be addressed to resolve to the correct name. The N4L model provides a means of mapping the sequence of events to provide an accurate representation of the meaning of a name, at a given point in time in a given taxonomy (Figure 4).

While currently applied to biological names, the N4L data model is applicable to any well-defined subject language terminology. This is an area of considerable interest outside the life sciences and the subject of ongoing standards development for language resources. ISO Working Draft 24618 [12] independently arrived at an approach for using persistent identifiers to link terms and concepts, similar to that anticipated by Garrity and Lyons [2]. The current N4L implementation is 11 levels deep (domains (2), phyla (36), classes (79), subclasses (8), orders (154), suborders (25), families (384), genera (2,193), subgenera (5), species (11,658) and subspecies (548)). It provides full coverage of all ranks in common usage, including those outside the scope of the ICNB. In addition, the N4L model provides a mechanism for independent registration of organisms (by exemplar) and groups of organisms (by taxon) based on DOIs. The model is not constrained by conformance to the ICNB and supports registration of organisms that cannot yet be cultivated and deposited in a culture collection. The N4L model also accommodates trivial or colloquial names (e.g., lab

identifiers assigned by a LIMS or automated ID system). As all instances of these classes are identified and interlinked using DOIs, the information is directly and persistently addressable and formally cited as a micro-publication or in the literature.

Figure 3. Modeling of synonyms occurring in the taxonomy of *Bacteria* and *Archaea*. Taxonomic rearrangements and nomenclatural events lead to a number of complex relationships among names, taxa and exemplars that must be correctly modeled to resolve to the correct name, based on the rules of the ICNB. All such events are supported in the N4L model.

Figure 4. Assembly of N4L objects into a taxonomic hierarchy. In the N4L model, names, taxa, and exemplar objects are carefully mapped to provide an accurate representation of the precise meaning of a name at a given point in time. DOIs allow the information associated with these objects to be directly and persistently addressable on the web and formally referenced as micropublications.

N4L web services

The NamesforLife system is based on a multi-tier software architecture (Figure 5). Company databases include the nomenclatural database and a reference database holding all of the taxonomic literature relevant to *Bacteria* and *Archaea*. External databases include the Alexandria patent repository (IFI - Fairview) consisting of more than 80 million patent

documents from 70 countries, processed and indexed via *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 Company clients. DOIs provide the necessary infrastructure to ensure that intellectual property rights are protected through the policies of the International DOI Foundation. Other resources include the *N4L::Contextual Index*, which interrelates current scientific, technical, medical and patent literature using the Company's proprietary semiotic fingerprinting technology [13], and the *N4L::Abstracts*, a collection of approximately 50,000 citable micro-publications. The Company's current and planned product offerings are discussed in more detail in the commercialization plan. A unique aspect of the Company's software is that a single N4L-DOI provides direct access to information about a name, a taxonomic concept, the organism (at the species or subspecies level) and all of the associated web services. These services are designed to provide end-users of biological nomenclatures (or other dynamic terminologies such as biological, chemical or biochemical terminology, gene or enzyme nomenclature) with the appropriate information in the correct temporal and biological context, on demand. The same service can also be used by database owners, publishers, or other information providers to semantically enable their content, making it readily discoverable, even when the definition of a name or term has changed, with minimal modification to their existing workflows or data architecture.

Figure 5. The current N4L System Architecture

Degree to which Phase I demonstrated technical feasibility

In Phase I, we proposed two technical objectives, both of which served as prerequisites for achieving our principal objective in our currently funded Phase II project: a descriptive ontology of bacterial and archaeal phenotypes. Our first phase I objective was to extract a comprehensive list of terms from the taxonomic literature that have been used to describe phenotypes. We understood that meeting this objective would likely require the development of new software tools, as existing tools for indexing and tokenizing text might

be inadequate for our needs. Once the term list was compiled, the lexical and syntactic patterns of usage would be examined and disambiguated (synonymous terms identified and resolved to a single concept, preferred term selection governed by frequency of usage) by human curators. Upon completion, the term list would be made available for public comment.

We anticipated that each description of a type strain would include approximately 125-150 of these terms, and that overlap in term usage would be greatest for those taxa that were most closely related. Similarly, we anticipated that some of the terms would be unique to specific taxa or subsets of phenotypic features. We also anticipated a temporal pattern of usage, based on the methodology in use at the time a description published. We also knew at the outset of the project that the quality and nature of the digitized versions of the descriptions would vary significantly, based on the publishing technology in use at the time of initial publication and the digitization method applied to the back-content.

Phase I Results

Technical issues arising from the quality of the source material

During the Phase I project, we encountered a number of obstacles to the automated extraction and classification of terms from the descriptive text, which fell into the following broad categories.

Typographical errors - Approximately 5% of the extracted terms proved to be typographical errors. The frequency of errors was highest for chemical and biochemical terms, especially fatty acids. Most of these errors were found to occur at low frequency and could be manually resolved by curators. However in some cases the corrected terms proved to be ambiguous or meaningless when viewed in context. Erroneous terms have been included in the vocabulary as orthographic variants of known terms in order to create an electronically traversable path from the original literature to the descriptive ontology.

Formatting errors - Formatting errors led to ambiguous terms or unintentional synonymies. Some of these appear to be copy-and-paste errors (e.g., a portion of term is missing) created by authors and left uncorrected during the editing, copy editing and layout stages. Others appear to arise from ad-hoc normalization of terms appearing in the output of lab instruments (e.g., the MIDI Sherlock System used to identify fatty acids).

Uncorrected OCR - Between 1998 and 2003, the digital format of many journals changed from scanned images to text. For articles published prior to 2003, accurate extraction of text was hampered by the imprecision of optical character recognition (OCR) in use at the time (Figure 6). Relevant articles published prior to 1998 are limited to PDF format. These were created from scanned images of the original publication and subsequently processed by OCR. Diacritics, punctuation and white space are often mishandled by this process, resulting in incoherent text. Some of these errors are systematic and thus correctable (e.g., the lowercase 'L' in "et. al" is frequently interpreted as a numeric '1' or upper-case 'I'), but many errors were not.

Figure 6. Digital format of the corpus of taxonomic literature. The accumulated content of the *International Journal of Systematic and Evolutionary Microbiology*, (2000-present), the *International Journal of Systematic Bacteriology*, (published 1966-1999) and the *International Bulletin of Bacterial Nomenclature and Taxonomy* (published 1951-1965) exists in different formats of varying quality. Although the contemporary content is available as HTML/XML much of the historical content remains inaccessible, owing to shortcomings of uncorrected OCR text.

This limited our analysis to those articles that were created from a native text format, which covers 56% of the total amount of material available from the combined IJSEM/IJSB/IBBNT archive. Processing of additional historical material would require manual transcription of the scanned images, and in some cases would require access to an original print copy for those articles with poor scan quality because of the errors in the digitized copies (Figure 7). As a test, we manually corrected the OCR text from some of the older IBBNT articles and arrived at an average time of 20 minutes per page of text, resulting in an estimated 0.97 person-years of manual editing for the remaining 25,395 pages of material. We consider this to be outside the scope of Phase I project, except for those papers containing descriptions of genera and species that are still relevant.

Data utility - While many phenotypes can be inferred from the terms used in descriptions, establishing comparability, especially across long time periods, often requires an understanding of the underlying methods and growth conditions that were originally used. There is a connection between methodology and terminology that has a significant bearing on not only which phenotypes are described, but how those phenotypes are described.

Ambiguity of commonly used phrases - Many terms, (e.g., "optimal growth temperature", "non-cultivable") are inherently subjective and must be considered in the appropriate experimental context. Supplementary information typically found in the Materials and Methods section of an article, or in previously published articles must be referred to frequently as not all descriptions are self-contained. Indexing by traditional keyword-based methods or even by ontological methods without considering experimental conditions can distort the original meaning of a term.

Figure 7. Assessing the quality of scanned images as source documents. Poor scan quality and imprecise character recognition results in incoherent text. The original PDF document (left) contains many regions that were unsuccessfully converted to text (right), including the highlighted region “*Pseudomonas*”, which was interpreted as “f t!~aC/7il~'%(!-” owing to the faded quality of the original text. Although there are 25 occurrences of the term “pseudomonas” in the original article, not one instance of the term “pseudo” appears in the uncorrected OCR output.

Other limiting issues regarding PDF files - Early in the project we encountered significant problems with PDF text extraction:

- Diacritics were often added via overtyping (e.g., ‘O’, backspace, ‘umlaut’ [O’]) rather than using the correct Unicode character [Ö].
- Tables, figures, column delimiters and page breaks frequently occurred in the middle of the descriptive text.
- Existing PDF extraction libraries result in the loss of important formatting information used to convey technical information (italics, subscript and superscript), leading to tokenization errors.

We concluded that PDF documents would be unsuitable for indexing at the level of precision we required, without manual pre-processing to ensure the relevant text remains unbroken. Typically, these issues do not interfere with normal keyword-based indexing, since the text is often normalized, text order is not important and searches are assumed to be inexact. However, for utilization in tools such as a KWIC (KeyWord In Context) Interface (Figure 9), the text must be complete, ordered, and be a faithful representation of the original material. XML and HTML instances of the same articles did not suffer from these problems. To that end, we organized the article processing workflow so that the common point of entry into the document corpus was a minimal, normalized XHTML document. This provided a means of preserving the flow and formatting of the original text while also providing a straightforward way to render a human-readable version of the text in the indexing application.

Creating the corpus of bacterial and archaeal descriptions

In Phase I, our corpus comprised 4,061 articles published in the *IJSEM* (2003-2012), provided by our partner, the Society for General Microbiology. The process used to create the corpus is shown in Figure 8. Articles were first processed with *JTidy* [14] to produce

XHTML and then processed with *N4L::Scribe* to tag all bacterial and archaeal names. Formal taxonomic descriptions and emendations (which existed in one-to-one and one-to-many relationships with articles) were localized using regular expressions, extracted from the documents, and cast as 5,750 individual XHTML instances. Common tagging errors (e.g., improper Unicode characters, empty or incorrectly closed tags, inconsistent use of tags and whitespace) were identified and corrected using a combination of XSL stylesheet transformations and regular expressions. We expected to find a maximum of 6,089 descriptions and emendations based on queries of the N4L Nomenclature Database for IJSEM from 2003-2012, indicating that over 94% of the accessible literature was successfully processed. This far exceeded the expectations set in the Phase I work plan, as we expected to encounter difficulties when parsing descriptions from documents containing multiple novel taxa. The remaining descriptions were identified and were processed by the completion of the Phase I.

Indexing the corpus of bacterial and archaeal descriptions

The corpus was tokenized and indexed using *Apache Lucene* v3.5.0 [15]. Document indexing with *Lucene* is based on a series of tokens produced from a grammar, which are subsequently passed to an analyzer for inclusion/exclusion. Typically, the first step involves extracting plain text from a document (e.g., PDF, HTML), passing only the printable text to the tokenizer to produce a list of tokens (words). We found that the standard tokenizers used by *Lucene* were inadequate for identifying chemical compounds, fatty acids (routinely included in descriptions of whole-cell chemical composition) and measurements (e.g., pH, salinity, geographic location) because these terms typically contain characters that the default grammars consider to be token delimiters (spaces, periods, commas, hyphens, slashes, parentheses and brackets), as well as subscripted and superscripted elements, which are important for conveying meaning. These are all lost prior to the tokens being passed to the analyzer and there is insufficient information available after tokenizing to allow unambiguous reconstruction of the original term (Figure 8). To solve this problem, we developed several new lexical analyzers and tokenizers for *Lucene* using the *JFlex* lexical scanner generator [16].

We developed four distinct tokenizers to extract terms that were used to extract terms: (1) the standard *Lucene* tokenizer; (2) a newly developed chemical compound tokenizer; (3) a newly developed fatty acid tokenizer; (4) and a general measurement tokenizer.

Figure 8. UML Activity Diagram for Development and Indexing of Corpus.

Figure 9. The Extended KWIC Index (Java Swing application) developed by the Company utilizes several newly developed grammars to properly recognize fatty acid and other chemical terms while preserving formatting. In the above example, where the grammar embedded in the *Lucene* Standard Tokenizer identifies four tokens ("C19", "0", "cyclo", "ω8c"), the Fatty Acid Tokenizer developed by the Company correctly identifies a single token ("C19 : 0 cyclo ω8c"), which is mapped back to its position in the original, formatted document (left), producing the proper term "C₁₉:0 cyclo ω8c".

Unlike the default implementation of the *Lucene* tokenizer and analyzer, we stored the token types and term vectors and tied the position of each token back to its position in the original source document (XML in this case). This preserves the original formatting of any string (i.e., italics, subscript, and superscript). This is a critical step to ensure correct extraction and identification of terms pertaining to fatty acids and many other biochemical and chemical names.

The overlap between fatty acid nomenclature and other chemical and biochemical nomenclature resulted in a large number of false positives when attempting to combine these tokenizers into a single grammar. This became an early task to address in Phase II. We will attempt to solve this problem by chaining together multiple tokenizers, ordered by grammar specificity (e.g., fatty acids prior to other chemical terms). To *Lucene*, these now appear as the product of a single tokenizer. To our knowledge, this is a novel technique and no such indexing method yet exists elsewhere. Indeed, searches for complete fatty acid and chemical terms on Google indicate that only the standard delimiters are currently in use when tokenizing complex terms. We initially had considered GATE [17] as a possible solution, since it supports chemical names via a plug-in. However, we have found the plug-in only recognizes chemical names if they directly precede the full IUPAC version of the name, rendering it irrelevant for our purposes.

As an alternative to GATE, the Company developed a new Java client³ for the *WhatIzIt* streamed servlet [18], a freely available text mining service developed by the European Bioinformatics Institute. We processed the entire corpus of descriptive text using the *whatizitChemicals* pipeline, resulting in 250,570 tagged regions of text (6,144 unique terms). Although many of the complex chemical terms were properly tagged with ChEBI (Chemical Entities of Biological Interest [18]) identifiers, there were also a large number of false positives (e.g., terms tagged as chemicals that are common words, unrelated acronyms or symbolic representations of elements). This is a documented problem with the *WhatIzIt* service, and a symptom of the multiple-grammar problem described above. Although it has proved useful in identifying a large number of chemical terms (as well as their ontology provenance) that can be used for the developing the proposed descriptive ontology, and as part of the dictionary used for tagging documents with *N4L::Scribe*, we were unable to rely on this service alone.

Extracting and Classifying the High-Frequency Terms

Once the documents were properly tokenized and indexed, we used the *Apache Luke* [19] library to extract the stored term vectors from the *Lucene* index. The library performed to our expectations, allowing relatively easy extraction of the high-frequency terms (Figure 10). Classification of the terms proved more difficult than expected. As noted above, chemical and biochemical names, especially fatty acids, contained numerous spelling and formatting variations, confounding sorting and filtering. Within the corpus, we have found 749 distinct variations of fatty acid names. This number was reduced to a much smaller set after we resolved the synonymies among the terms (e.g., "10meC_{16:0}", "10-methylated C_{16:0}", "10-methyl C_{16:0}", "10-methyl hexadecanoic acid" and "10-methyl palmitic acid"). For many of these terms, the official (correct) spellings are defined in existing ontologies or other resources [21-23], but the variations in spelling and formatting needed to be associated with each in order to tie the literature to the correct term. *This is familiar ground to the Company and parallels the situation with biological nomenclature.* In Phase II, we are transforming these extracted term sets into a set of dictionaries that can be used in the *N4L::Scribe*, where synonymies and orthographic variations resolve to a single unambiguous identifier. These dictionaries also serve as a foundation on which the ontology is based.

The Extended KeyWord In Context (KWIC) Index

The Company developed a KWIC interface to the index to provide a (virtual) concordance of terms in the taxonomic literature. KWIC indices are used extensively by lexicographers and by indexers of patent literature. The KWIC index has proven useful in vocabulary development, as it provides curators with a means of localizing the immediate context and pattern of use of any term. We have designated this an "Extended KWIC Index" because it also provides search, source document editing and re-indexing capabilities, as well as a means of viewing the original formatted document with term highlighting.

³ EBI provides a Java client for their streamed service, but it has not been recently updated (2006) and no longer functions with the current version of their service.

The application was built using the Java Swing 1.6 framework, along with *Apache Lucene* and *Apache Luke* as described previously. Many of the visual components of the KWIC interface were designed and implemented *ab initio*, as no suitable components were available from existing libraries. A plug-in interface for filtering terms (with both inclusive and exclusive filtering modes) was developed to reduce the current working vocabulary. The currently implemented dictionary and ontology filters include: Bacterial Names, Authorities (author names), PubChem (CID-MeSH and MeSH-Pharm) terms [23], KEGG Lipids ontology [24], the American National Corpus, Webster's Dictionary, and Geographic Regions. There are also several filters for selecting/de-selecting the working terms based on numeric tokens and nucleotide sequences, as well as term length and frequency of occurrence in the corpus. A filter was also created for matching the classified terms.

Major Features Included in the NamesforLife Phenotypic Index

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

Figure 10. Major features to be included in the *NamesforLife Phenotypic Index*. Expanded descriptions of most of these elements can be found in Tindall et al. 2010 [20].

References Cited

1. Garrity GM, Lyons CM: **Future-proofing biological nomenclature**. *OMICS* 2003, **7**:31-31. doi:10.1089/153623103322006562
2. Garrity GM, Lyons CM: **U.S. Patent No. 7,925,444. Systems and Methods for Resolving Ambiguity Between Names and Entities**. 2011. Washington, DC: U.S. Patent and Trademark Office.
3. Ogden C, Richards I: *The Meaning of Meaning*. Harcourt, Brace, and World; 1923.
4. Sowa J: **Ontology, Metadata, and Semiotics** [Online]. (<http://users.bestweb.net/~sowa/peirce/ontometa.htm>) 2000.
5. Paskin, N: **The DOI Handbook. ed. 4.4.1**. International DOI Foundation. 2006. doi:10.1000/182

6. Parker CT, Roston GP, Lyons CM, Garrity GM: **Systems and methods for automatically identifying and linking names in digital resources**. European Patent Application EP20100729654; U.S. Patent Application 2010/0198841 A1; 2010.
7. Garrity GM: **NamesforLife BrowserTool takes expertise out of the database and puts it right in the browser**. *Microbiology Today* 2010, **37**:9.
8. Parker CT, Taylor, D, Mannor K, Wigley S, Osier N, Lyons CM, Garrity GM: **NamesforLife Semantic Resolution Services for the Life Sciences**. *Nature Precedings* 2010. doi:10.1038/npre.2010.5137.1
9. Garrity, GM, Parker, CT: **Recent trends in US patent grants and issues to be considered**. *Nature Precedings* 2010. doi:10.1038/npre.2010.4998.1
10. Tindall BJ: **Misunderstanding the Bacteriological Code**. *Int J Syst Bacteriol* 1999, **49**:1313-6.
11. Lapage SP, Sneath PHA, Lessel EF, Skerman VBD, Seeliger HPR, Clark WA (editors) (1992). *International Code of Nomenclature of Bacteria (1990 Revision)*. **Bacteriological Code**. Washington, DC: American Society for Microbiology.
12. ISO Technical Committee 37/SC 4 N 455: **ISO Working Draft 24618 - Citation of Electronic Resources**. International Standards Organization, 2008.
13. Garrity GM, Parker CT, Taylor D, Mannor K, Lyons CM: **Semantic Indexing of the Green Technology Patent Literature - An Application of the NamesforLife Contextual Index**. *Genomic Sciences Awardee Meeting IX*, 2011.
14. **HTML Tidy Library Project (JTidy r938 2009)** [Online]. (<http://tidy.sourceforge.net/>)
15. **Lucene Core; Java-based indexing and search technology v3.5.0**. [Online]. (<http://lucene.apache.org/>)
16. **JFlex - The Fast Scanner Generator for Java v1.5.0 (trunk)**. [Online]. (<http://jflex.de/>)
17. H. Cunningham, et al. **Text Processing with GATE (Version 6)**. University of Sheffield Department of Computer Science. April 2011. ISBN 0956599311.
18. Alberto L, Franck V, Mikael A, Rodrigo L: **Web Services at the European Bioinformatics Institute**. *Nucl. Acids Res.* 2007, **35** (suppl 2): W6-W11. doi: 10.1093/nar/gkm291
19. **Luke - Lucene Index Toolbox v3.5.0**. [Online]. (<http://code.google.com/p/luke/>)
20. **Chemical Abstracts Service**. American Chemical Society. [Online]. (<http://chemabs.cas.org/>)
21. O'Neil (Editor): **Merck Index, 14th Edition**. Cole-Parmer 2006. ISBN 978-0-911910-00-1
22. Bolton E, Wang Y, Thiessen PA, Bryant SH. **PubChem: Integrated Platform of Small Molecules and Biological Activities**. Chapter 12 in: Annual Reports in Computational Chemistry, Volume 4, American Chemical Society, Washington, DC, April 2008.
23. Cotter D, Maer A, Guda C, Saunders B, Subramaniam S: **LMPD: LIPID MAPS proteome database**. *Nucleic Acids Res.* 2006, **34**: D507-D510. doi:10.1093/nar/gkj122