NamesforLife
Bringing meaning to life...

Semantic Index of Phenotypic and Genotypic Data

Charles Parker[1], Nenad Krdzavac[2], Chuong Vo Phan[1], Kevin Petersen[2], Grace Rodriguez[1], Oishi Bagchi[2] and George M. Garrity[1,2]

[1]NamesforLife, LLC and [2]Michigan State University (East Lansing, Michigan)

MICHIGAN STATE UNIVERSITY

Project Goals

Our objectives are to:

(1) build a knowledge resource containing standardized phenotypic descriptions of prokaryotic type strains,

(2) develop a formal ontology capable of making accurate phenotypic and environmental inferences over this resource, and

(3) improve the visibility and accessibility of public-funded research projects that provide this data.

We are developing a standards-compliant semantic data resource to support predictive modeling of microbial phenotype.

This project is tightly coupled with ongoing DOE projects (*Genomic Encyclopedia of* Bacteria *and* Archaea, Microbial Earth Project, Community Science Program) and two key publications (*Standards in Genomic Sciences* and the *International Journal of Systematic and Evolutionary Microbiology*).

Background

The Problem

Despite significant improvements in genome annotation, many assertions are hypothetical and may lack experimental support. The taxonomic literature for prokaryotes contains a wealth of experimental phenotypic data, but that knowledge is currently in a form that does not lend itself to integration with databases or ontologies. Predictive models rely on high quality input data, but not all data are of similar quality nor are they amenable to computational analysis without extensive cleaning, interpretation and normalization. Key among the types of data needed to support current research are phenotypic data (Table 1), which are more complex than sequence data, occur in a variety of forms, often use complex and non-uniform descriptors, may be taxon-specific and are scattered throughout specialized databases and scientific, technical and medical literature. Integrating phenotypic data from such resources requires expertise in harvesting, modeling, interpreting, and validating these data, as well as a complete and actively maintained resource for all of the type strains.

Table 1. Feature classes included in the Prokaryote Knowledge Base, grouped by major feature domain. The features will be made available via the Taxonomic Abstracts (https://doi.org/10.1601/about) and several new services.

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

Our Solution

Our knowledge base is designed to address these problems by providing reference phenotypic data for nearly all type strains of *Bacteria* and *Archaea*, based on concepts and observational data drawn from the primary taxonomic literature (the corpus of literature that supports our up-to-date taxonomy and strain database). We developed software (*Semantic Desktop*) to extract putative feature domain vocabularies from this corpus, resulting in the discovery of over 40,000 candidate phenotypic terms used in new and emended descriptions of the 13,213 distinct type strains of *Bacteria* and *Archaea* (N4L Database, March 1, 2016). We have since developed this vocabulary into a precise thesaurus of phenotypic terms, which will ultimately conform to W3C SKOS-XL semantics, providing a link between microbial phenotype language, the semantic web and existing NamesforLife services (*N4L::Guide* and *N4L::Scribe*). Our use of existing standards and services, coupled with the broad coverage of prokaryotic taxa, will complement the MIGS and MIMS (MIxS) standards by providing a precise vocabulary to use when publishing descriptions of new taxa.

Our thesaurus complements MIxS by providing precise phenotypic language with broad taxonomic coverage.

Our ontology relates the environment and phenotype of an organism based on published observations.

Many of the phenotypes applied to microbes describe a combination of quantitative environmental conditions and qualitative growth and metabolic capabilities. Such terms are challenging to implement in query systems due to their context-based interpretations, imprecision and conceptual overlap across multiple feature domains.

To address this problem, the thesaurus was developed in parallel with a formal ontology that supports inference from observations of an organism under a set of environmental constraints, using meta-modeling techniques to implement rule and constraint templates using these complex terms. In developing a solution to this problem, we discovered a novel method for establishing semantic equivalence between concepts that enables precise, consistent, verifiable reasoning over imprecise terms at multiple levels of abstraction [1].

Challenges of Information Extraction (IE)

Extracting information from text is not an easy task. Prior to this phase of the project, we had already produced a curated taxonomy and strain database covering all prokaryotic type strains, and assembled a complete corpus of taxonomic literature, as well as a candidate vocabulary of phenotypic terms. Using these resources, some novel software methods and an extensive curation effort, we are coding raw text into phenotypic assertions based on our ontology and thesaurus. These assertions are interpreted by a reasoner to infer phenotype and other features based on all available information that has been reported about a strain. Our method is able to interpret these assertions at appropriate levels of abstraction to correctly answer queries and produce new knowledge.

strain	source	oxygen sensitivity (raw text)	pH sensitivity (raw text)	temperature sensitivity (raw text)
10.1601/ex.3007	rid.516	facultatively anaerobic	Mesophilic and neutrophilic chemoorganotroph: grows between 15 and 30 °C.	Mesophilic and neutrophilic chemoorganotroph: grows between 15 and 30 °C.
10.1601/ex.3857	rid.507	Requiring less than 15%O2 (i.e. 75% air saturation) in the headspace gas (optimum 5–8 %).	pH 4.5–9.0 (optimum pH 6.0–7.5). optimum pH 6.5	The isolate grew at 10–40 °C (optimum 25 °C)
10.1601/ex.4346	rid.500	Strict anaerobe.	pH range for growth 6.3-8.5, pH optimum at 7.0.	T_{min}, 20°C ; T_{opt}, 38°C ; T_{max}, 43°C ;
10.1601/ex.166	rid.490	Obligately anaerobic.	Growth occurs between pH 5.5 and 6.7, with the optimum at around pH 6.5	The temperature range for growth at pH 6.5 was 50–86 °C, with optimum growth at 85 °C.
10.1601/ex.7799	rid.301	Anaerobic, aerotolerant.	Optimal growth at pH 8.8 to 9.75. No growth at pH 8.0 or 10.8.	Optimum temperature for growth, 30 to 37°C; range, 15 to 47°C

strain	source	oxygen sensitivity (normalized text)	pH sensitivity (normalized text)	temperature sensitivity (normalized text)
10.1601/ex.3007	rid.516	facultative anaerobe	neutrophile	mesophile growth at 15 °C, growth at 30 °C
10.1601/ex.3857	rid.507	growth at 15%, optimal growth at 5%, optimal growth at 8%	optimal growth at pH 6.5	optimal growth at 25 °C
10.1601/ex.4346	rid.500	obligate anaerobe	optimal growth at pH 7.0	optimal growth at 38 °C
10.1601/ex.166	rid.490	obligate anaerobe	optimal growth at pH 6.5	optimal growth at 85 °C
10.1601/ex.7799	rid.301	aerotolerant anaerobe	optimal growth at pH 9.275	optimal growth at 33.5°C

strain	source	oxygen sensitivity (interpreted)	pH sensitivity (interpreted)	temperature sensitivity (interpreted)
10.1601/ex.3007	rid.516	facultative anaerobe	neutrophile	mesophile
10.1601/ex.3857	rid.507	microaerophilic	neutrophile	mesophile
10.1601/ex.4346	rid.500	obligate anaerobe	neutrophile	mesophile
10.1601/ex.166	rid.490	obligate anaerobe	neutrophile	hyperthermophile
10.1601/ex.7799	rid.301	aerotolerant anaerobe	alkalophile	mesophile

(Right) An Orthogonal Semantic Equivalence Map (OSEM) for sensitivity and tolerance to Oxygen. This provides the structure for implementing first order logic (rules and axioms) over three distinct concept taxonomies (SKOS-XL). Assertions may be supplied to an OSEM to infer semantically equivalent representations of phenotype over bi-directional (environment, observation) relations.

Phenotype		A: anoxic [0,0]	B: aerobic (0,)	
			B1: microaerobic (0,1)	B2: air [1,)
anaerobe	obligate anaerobe	+	x	x
	aerotolerant anaerobe	+	G	
aerobe	obligate aerobe	x	x	+
	microaerophilic	x	+	x

(G): growth; (x): no growth; (~): growth (suboptimal); (+): growth (optimal); ( ): don't care

A Hybrid Approach

We employ a hybrid relational database / thesaurus / formal ontology architecture in order to support curation, reasoning, search and query. The relational schema is mapped to the ontology via an intermediate XML Transfer Schema, which also serves as the basis for archiving complete strain records.

XML snapshots of the relational database (PostgreSQL) are loaded into the ontology data store (Virtuoso Open Source Edition). Assertions are validated (checked for consistency over the ontology) by the SPIN API (TopQuadrant, W3C Draft) and Jena API (HP, Apache License). The SPIN reasoner and optionally other OWL reasoners interpret rules and axioms encoded in the ontology to detect inconsistencies and constraint violations. Records that fail validation or Consistency and Constraint Checking are flagged for curatorial attention in the relational database.

The reasoner has its own curator account, which supplements and directs the activities of human curators.

As additional feature domains are modeled and new relations are discovered, rules, templates and axioms are developed and encoded in the SPIN and OWL Rule Language. A reasoner (the SPIN reasoner and Description Logics reasoners) may infer new knowledge about strains that was not directly reported in the literature.

(Right) An early working version of our faceted search for strains by phenotype.

The user interface is comprised of a collection of micro services that conform to the Open Search Description standard (where applicable), and allow query of the ontology, data and related resources in a variety of ways, including directly from a browser's native search bar.

Facet queries are converted into SPARQL queries and sent to our SPARQL endpoint, a custom Java Web Service built upon the Jena/ARQ API, which interprets the query and returns an answer (a matching set of strains, shown here by their Exemplar DOI).

Using our novel method [1] of Semantic Equivalence, we may construct representations of an environment as a set of environmental constraints to ultimately answer questions such as:

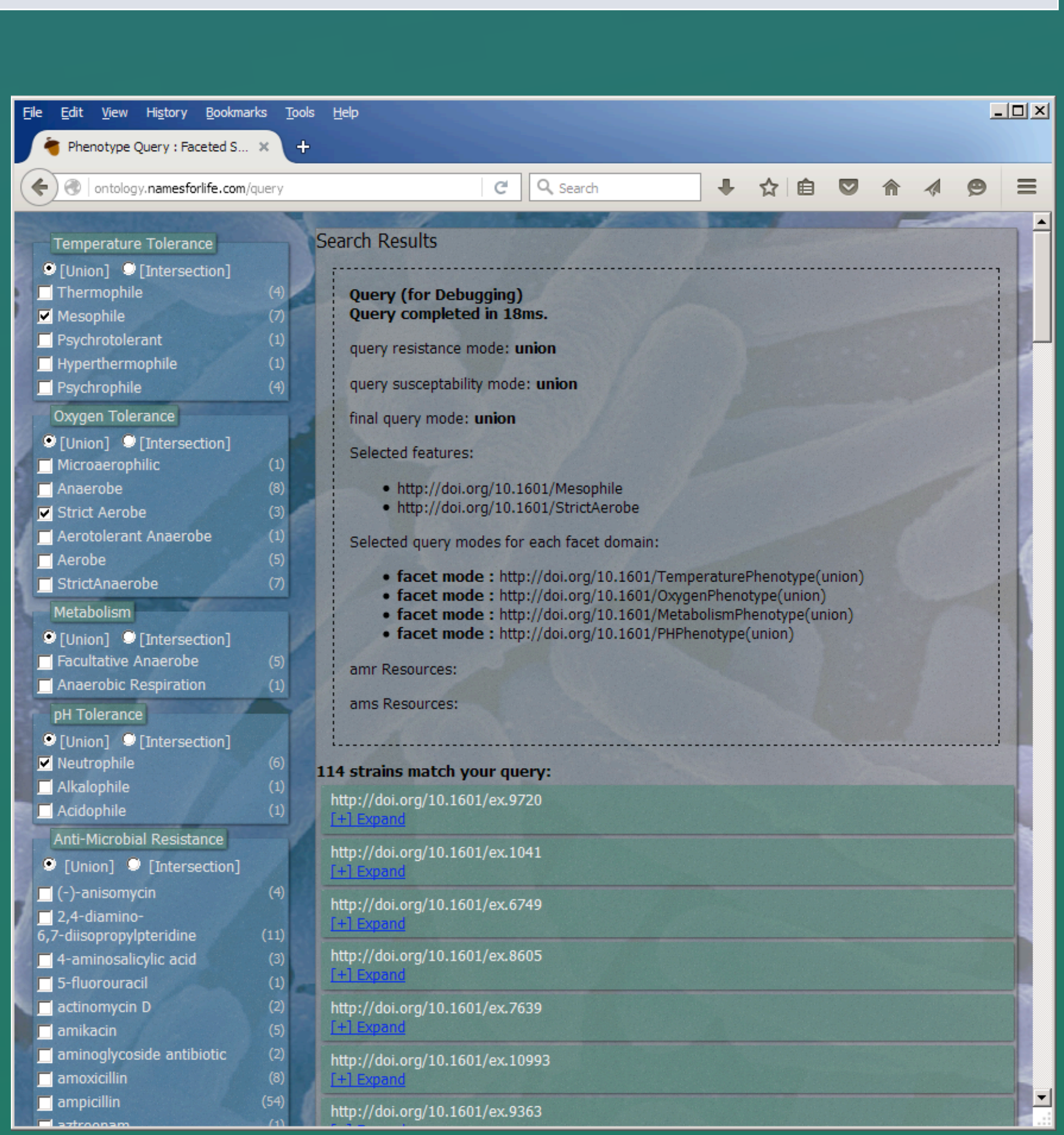
"What strains will grow in this environment?"

"How much phenotypic variation is there with the Streptomycetes?"

"What characteristics does Escherichia share with Salmonella?"

"Which type strains could be considered non-type strains of a different genus?"

"What strains are under-described for their taxa?"


File
Edit
View
History
Bookmarks
Tools
Help
Phenotype Query : Faceted S...
ontology.namesforlife.com/query
Search

Temperature Tolerance
○ [Union] ○ [Intersection]
☐ Thermophile (4)
☑ Mesophile (7)
☐ Psychrotolerant (1)
☐ Hyperthermophile (1)
☐ Psychrophile (4)

Oxygen Tolerance
○ [Union] ○ [Intersection]
☐ Microaerophilic (1)
☐ Anaerobe (8)
☑ Strict Aerobe (3)
☐ Aerotolerant Anaerobe (1)
☐ Aerobe (5)
☐ StrictAnaerobe (7)

Metabolism
○ [Union] ○ [Intersection]
☐ Facultative Anaerobe (5)
☐ Anaerobic Respiration (1)

pH Tolerance
○ [Union] ○ [Intersection]
☑ Neutrophile (6)
☐ Alkalophile (1)
☐ Acidophile (1)

Anti-Microbial Resistance
○ [Union] ○ [Intersection]
☐ (-)-anisomycin (4)
☐ 2,4-diamino-6,7-diisopropylpteridine (11)
☐ 4-aminosalicylic acid (3)
☐ 5-fluorouracil (1)
☐ actinomycin D (2)
☐ amikacin (5)
☐ aminoglycoside antibiotic (2)
☐ amoxicillin (8)
☐ ampicillin (54)

Search Results

Query (for Debugging)
Query completed in 18ms.

query resistance mode: union

query susceptability mode: union

final query mode: union

Selected features:

- http://doi.org/10.1601/Mesophile
- http://doi.org/10.1601/StrictAerobe

Selected query modes for each facet domain:

- facet mode : http://doi.org/10.1601/TemperaturePhenotype(union)
- facet mode : http://doi.org/10.1601/OxygenPhenotype(union)
- facet mode : http://doi.org/10.1601/MetabolismPhenotype(union)
- facet mode : http://doi.org/10.1601/PHPhenotype(union)

amr Resources:

ams Resources:

114 strains match your query:

http://doi.org/10.1601/ex.9720 [+] Expand
http://doi.org/10.1601/ex.1041 [+] Expand
http://doi.org/10.1601/ex.6749 [+] Expand
http://doi.org/10.1601/ex.8605 [+] Expand
http://doi.org/10.1601/ex.7639 [+] Expand
http://doi.org/10.1601/ex.10993 [+] Expand
http://doi.org/10.1601/ex.9363

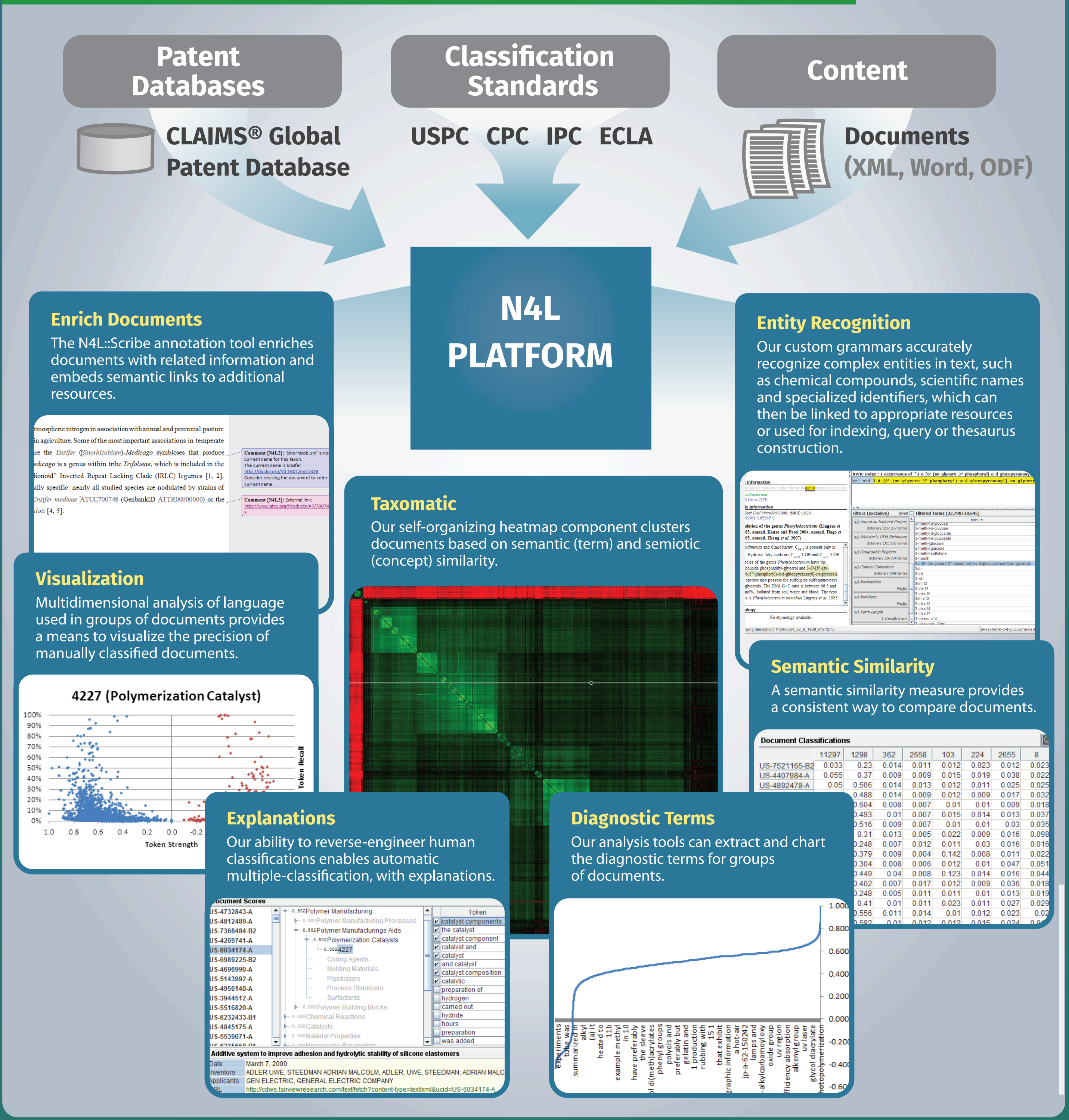
Current and Planned Products

During the course of this project we developed many software components that overcome specific technical barriers in terminology management, text mining, information extraction, knowledge transformation, entity recognition, document classification and annotation. The individual tools (N4L::Guide, N4L::Scribe, the Taxonomic Abstracts, Taxomatic, the KWIC Index and the Semantic Desktop) were implemented using W3C standards and recommendations (SPARQL, RDFS, RDF, OWL2, SKOS, SKOS-XL, XML, XSL, XSD, SPIN, OWL RL, DOI/CrossRef, CORS) and commercially-compatible FOS frameworks (Java, Apache, PostgreSQL, Virtuoso OSE, Jena/ARQ, SPIN Reasoner). We are integrating these components into a single software suite that can support a variety of document analysis needs.

Backed by the Fairview Research Alexandria platform (CLAIMS Global Patent Database), this analysis suite has access to the full text of the worldwide patent literature. We have demonstrated the ability to reverse-engineer the diagnostic phrases that human indexers use to classify large corpora of technical documents, and to measure both the quality of previously-annotated documents and the cohesion of individual document classifications. Our software provides a novel way to navigate and bridge multiple classification systems.

Our continued collaborations with the Joint Genome Institute, Fairview Research/IFI Claims and Oak Ridge National Laboratories provide excellent opportunities to test and refine the capabilities of this analysis suite while raising the visibility of other federal funded projects by completing the semantic linking between projects, entities and publications.

N4L Semantic Analysis Platform



Enrich Documents
The N4L::Scribe annotation tool enriches documents with related information and embeds semantic links to additional resources.

Entity Recognition
Our custom grammars accurately recognize complex entities in text, such as chemical compounds, scientific names and specialized identifiers, which can then be linked to appropriate resources or used for indexing, query or thesaurus construction.

Taxomatic
Our self-organizing heatmap component clusters documents based on semantic (term) and semiotic (concept) similarity.

Visualization
Multidimensional analysis of language used in groups of documents provides a means to visualize the precision of manually classified documents.

Semantic Similarity
A semantic similarity measure provides a consistent way to compare documents.

Explanations
Our ability to reverse-engineer human classifications enables automatic multiple-classification, with explanations.

Diagnostic Terms
Our analysis tools can extract and chart the diagnostic terms for groups of documents.

NamesforLife, LLC info@namesforlife.com namesforlife.com
(517) 214-8821 University Place, Suite 202 333 Albert Avenue East Lansing, MI 48823

Publications and Patents

1. Parker, CT, Garrity, GM and Krdzavac, NB. *Systems and Methods for Establishing Semantic Equivalence Between Concepts*. Provisional US Application No. 61/880,244 (Priority Date September 20, 2013) Washington, DC: U.S. Patent and Trademark Office. International Application No. PCT/US2014/056808 (Filed September 20, 2014); WIPO Application Number WO/2015/042536 (Published March 26, 2015) Geneva, Switzerland: World Intellectual Property Organization.

2. Sayood, K, Way, S, Ozkan, UN and Garrity, GM. *Classification of Nucleotide Sequences by Latent Semantic Analysis*. Provisional US Application No. 61/677,316 (Priority Date July 30, 2012); US Application No. 13/954,925 (Published May 1, 2014) Washington, DC: U.S. Patent and Trademark Office. International Application No. PCT/US2013/052797 (Filed July 30, 2013); WIPO Application No. WO/2014/022441 (Published February 6, 2014) Geneva, Switzerland: World Intellectual Property Organization.

3. Parker, CT and Garrity, GM. *Semiotic Indexing of Digital Resources*. US Patent No. 8,903,825 (Issued December 2, 2014); Provisional US Application No. 61/489,362 (Priority Date May 24, 2011); US Application No. 13/478,973 (Published January 10, 2013); Washington, DC: U.S. Patent and Trademark Office. International Application No. PCT/US2012/039168 (Filed May 23, 2012); WIPO Application No. WO/2014/022441 (Published November 29, 2012) Geneva, Switzerland: World Intellectual Property Organization. European Application No. EP 2012/0790213 (Published April 9, 2014) Munich, Germany: European Patent Office.

4. Garrity, GM, Parker, CT, Ussery, DW, Wanchai, V and Nookaew, I. Provisional US Application No. 62/232,925 (Priority Date September 25, 2015).

5. Parker, CT, Tindall, BJ and Garrity, GM. *International Code of Nomenclature of Prokaryotes (Prokaryotic Code 2008 Revision). Int J Syst Evol Microbiol* Published ahead of print November 20, 2015. PMID 26596770; doi:10.1099/ijsem.0.000778.

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. Additional support for developing the Semantic Analysis Platform was provided through a commercial partnership with Fairview Research/IFI Claims.