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] and George M. Garrity[1,2]

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

MICHIGAN STATE UNIVERSITY

## Project Goals

**The goal of this project is to develop a semantic data resource to serve as a basis for predictive modeling of microbial phenotype.**

Our core technical objectives are to: (1) build a database of normalized phenotypic descriptions using the primary taxonomic literature of bacterial and archaeal type strains, (2) construct an ontology capable of making accurate phenotypic and environmental inferences based on that data, and (3) improve the visibility and accessibility of publicly-available research data.

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

The scope of this project covers many technical fields, including text-mining, Information Extraction, Natural Language Processing, indexing & search, terminology & ontology development, machine reasoning, semantic analysis, sequence analysis and taxonomic classification.

**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://doi.org/10.1601/about).

**Strain Metadata**
- ☑N4L Exemplar DOI
- ☑ Host
- ☑Strain Designation
- ☑Collection ID(s)
- ☑Taxon status (type/non-type)
- ☑Isolation substrate†
- □Isolation source
- □Isolation method†
- □Geographic location*
- □Environmental information

**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†**
- ☑optimal growth conditions
- ☑Cell Images
- ☑sensitivity/tolerance to chemical and physical agents*
- □substrate utilization*
- □terminal e- acceptor
- □metabolic end-products
- □Growth Curves

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

## Background

### The Problem

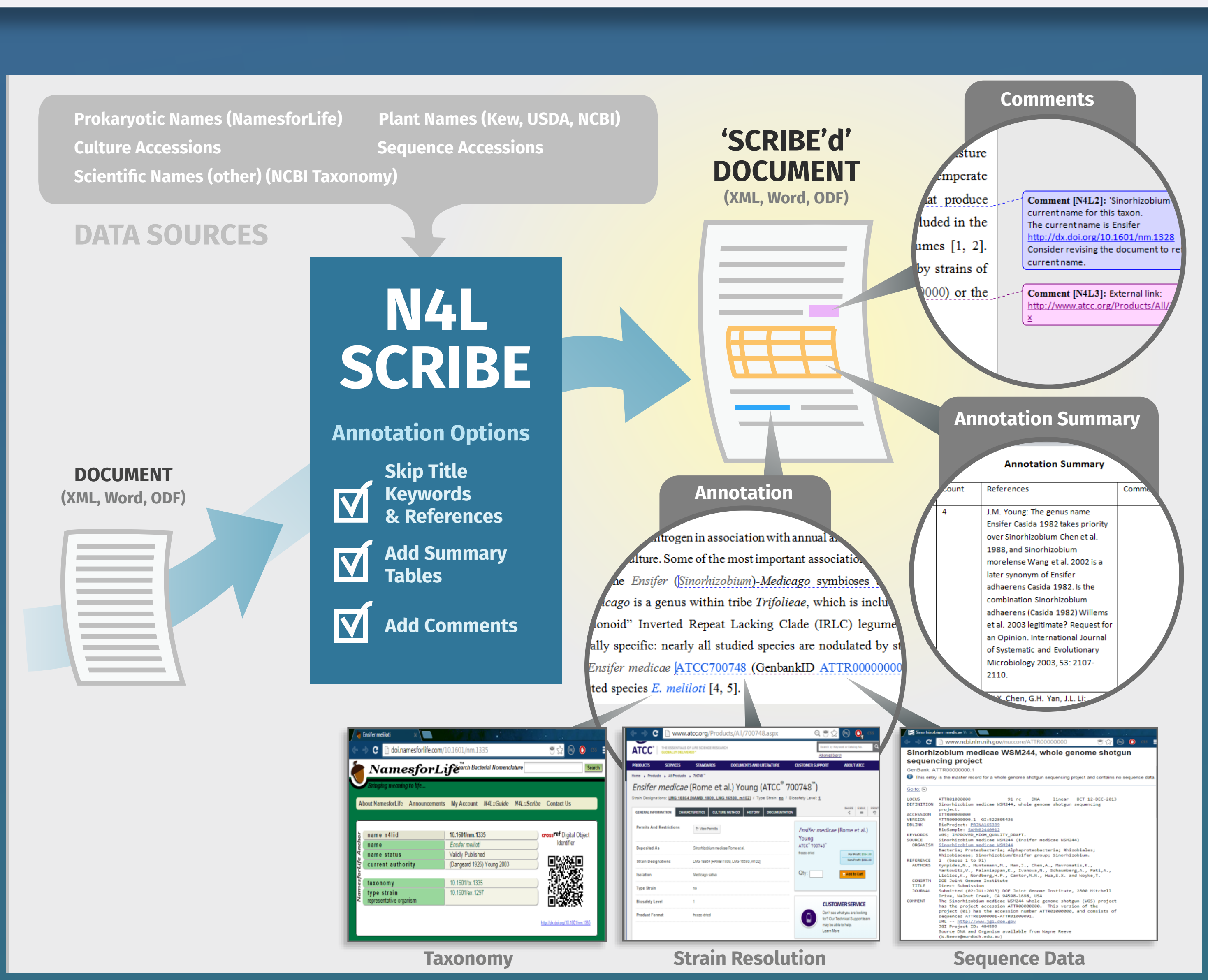
The DOE Systems Biology Knowledgebase (KBase) was envisioned to provide a framework for modeling dynamic cellular processes of microorganisms, plants and metacommunities. KBase will enable rapid iteration of experiments drawing on a variety of data to allow researchers to infer how cells and communities respond to natural/induced perturbations and ultimately predict outcomes.

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 those needed to make the KBase fully operational are phenotypic data, which are more complex than sequence data, occur in a variety of forms, often use complex and non-uniform descriptors and are scattered about specialized databases and scientific/technical/medical literature. Incorporating phenotypic data into the KBase requires expertise in harvesting, modeling interpreting and validating these data as well as a complete type strain dataset and taxonomy.

**This online resource complements KBase by providing a reference set of phenotypic data for nearly all published type strains of *Bacteria* and *Archaea*.**

### Our Solution

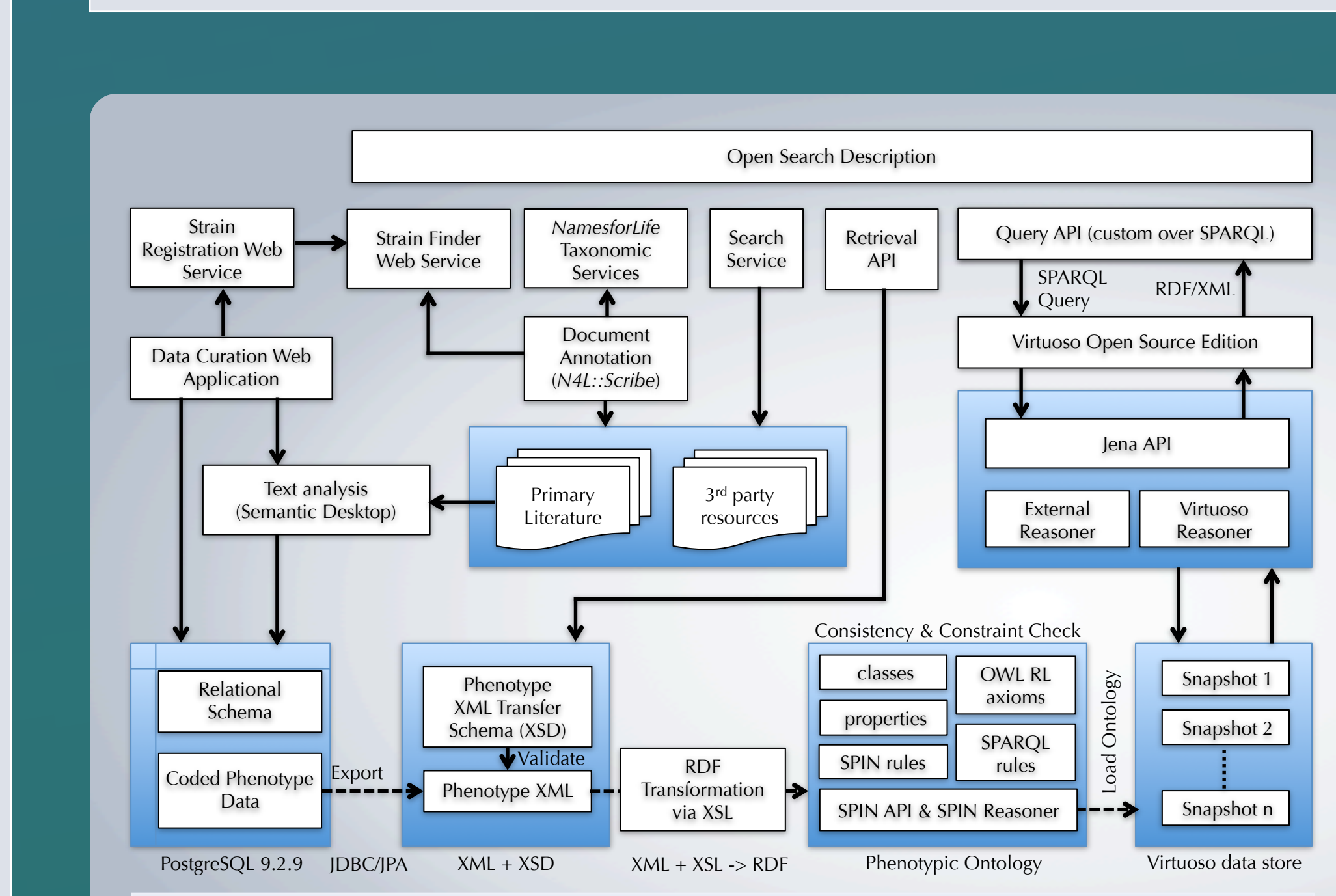
The Semantic Index of Phenotypic and Genotypic Data will address this problem by providing a resource of reference phenotypic data for all validly published type strains of *Bacteria* and *Archaea*, based on concepts and observational data drawn from the primary taxonomic literature. In the Phase I project we developed software to construct and analyze a corpus of this literature and to extract putative feature domain vocabularies comprising over 40,000 candidate phenotypic terms used in 20,224 new and emended descriptions of the 12,937 distinct type strains of *Bacteria* and *Archaea* (N4L Database, February 20, 2015). In Phase II/IIb, these vocabularies are serving as the basis for developing a phenotypic ontology, a repository of phenotypic data and normalized phenotypic descriptions for each species. 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 and conceptual overlap across multiple feature domains. In developing a solution to these problems, we discovered a novel method for establishing concept equivalence that enables precise, consistent, verifiable reasoning over these complex terms [1].


Prokaryotic Names (NamesforLife)
Plant Names (Kew, USDA, NCBI)
Culture Accessions
Sequence Accessions
Scientific Names (other) (NCBI Taxonomy)
DATA SOURCES
DOCUMENT
(XML, Word, ODF)
N4L SCRIBE
Annotation Options
Skip Title Keywords & References
Add Summary Tables
Add Comments
'SCRIBE'd' DOCUMENT
(XML, Word, ODF)
Comments
Comment [N4L2]: 'Sinorhizobium' current name for this taxon. The current name is Ensifer http://dx.doi.org/10.1601/nm.1328 Consider revising the document to re current name.
Comment [N4L3]: External link: http://www.atcc.org/Products/All/
Annotation Summary
Annotation
Ensifer medicae ATCC700748 (GenbankID ATTR00000000
Taxonomy
Strain Resolution
Sequence Data

## Document Annotation: The N4L Scribe

The Scribe document annotation service (http://scribe.namesforlife.com) has been significantly updated to recognize (in addition to bacterial and archaeal names), eukaryotic names, viral names, GenBank accessions and strain identifiers. This web service embeds links directly into documents (i.e., Microsoft Word [.DOC and .DOCX], Open/Libre Office [ODF], or any well-formed XML [including XHTML, NLM, JATS, etc.]) to the authoritative resources for any recognized names, identifiers or accessions. Additionally, summaries of nucleotide or protein sequences are generated so that authors, reviewers or editors may verify the accuracy of the identifiers used in the document.

Several new features are planned in the future, including enhancements to the Scribe SOAP API (https://ws.namesforlife.com/ws/scribe) as well as additional programming language support, support for external ontologies and user-supplied vocabularies. We are also testing delimited text and spreadsheet formats, which can provide nomenclature, taxonomy and strain resolution services for 3rd party databases.

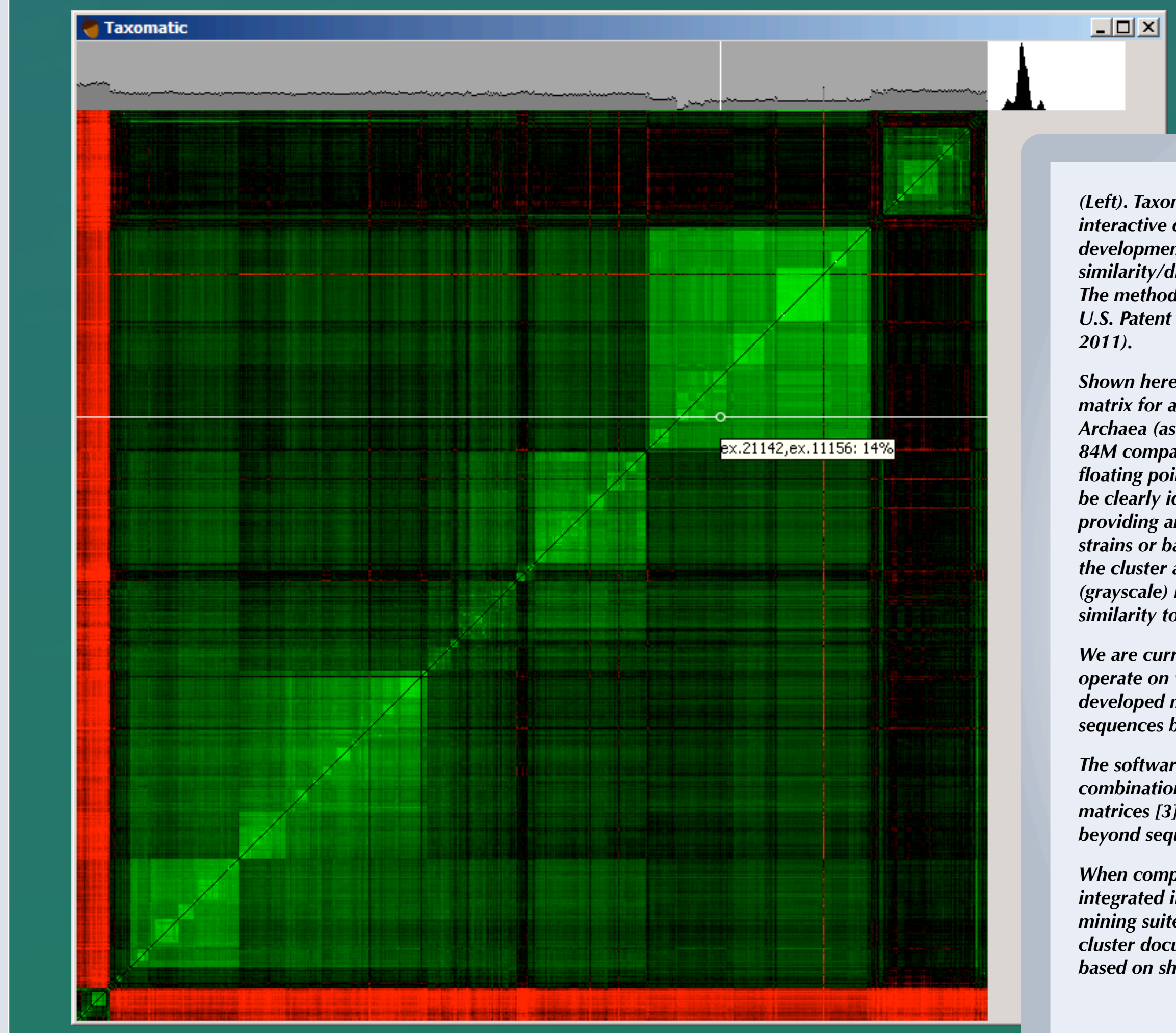


*Validation (consistency-checking over the ontology) of exported data from DB into phenotype ontology is handled by the SPIN API (TopQuadrant, W3C Draft), and Jena API (HP, Apache License). We use SPIN reasoner and optionally other OWL reasoners to detect inconsistencies and constraint violations. After that process, data are ready to be stored in a triple store and queried.*

*(Left). We adopted a hybrid relational database (PostgreSQL) / 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. Records that fail validation or Consistency and Constrain Checking are flagged for curatorial attention.*

*Snapshots of the relational database are loaded into the ontology data store and individual records are checked for consistency over the ontology. We have migrated away from the Fuseki SPARQL query server (part of the Apache Jena ontology framework) and are now using Virtuoso Open Source Edition as the data store and ontology framework, maintaining our goals of a complete Open Source framework that is free to both commercial and non-commercial use.*

*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.*



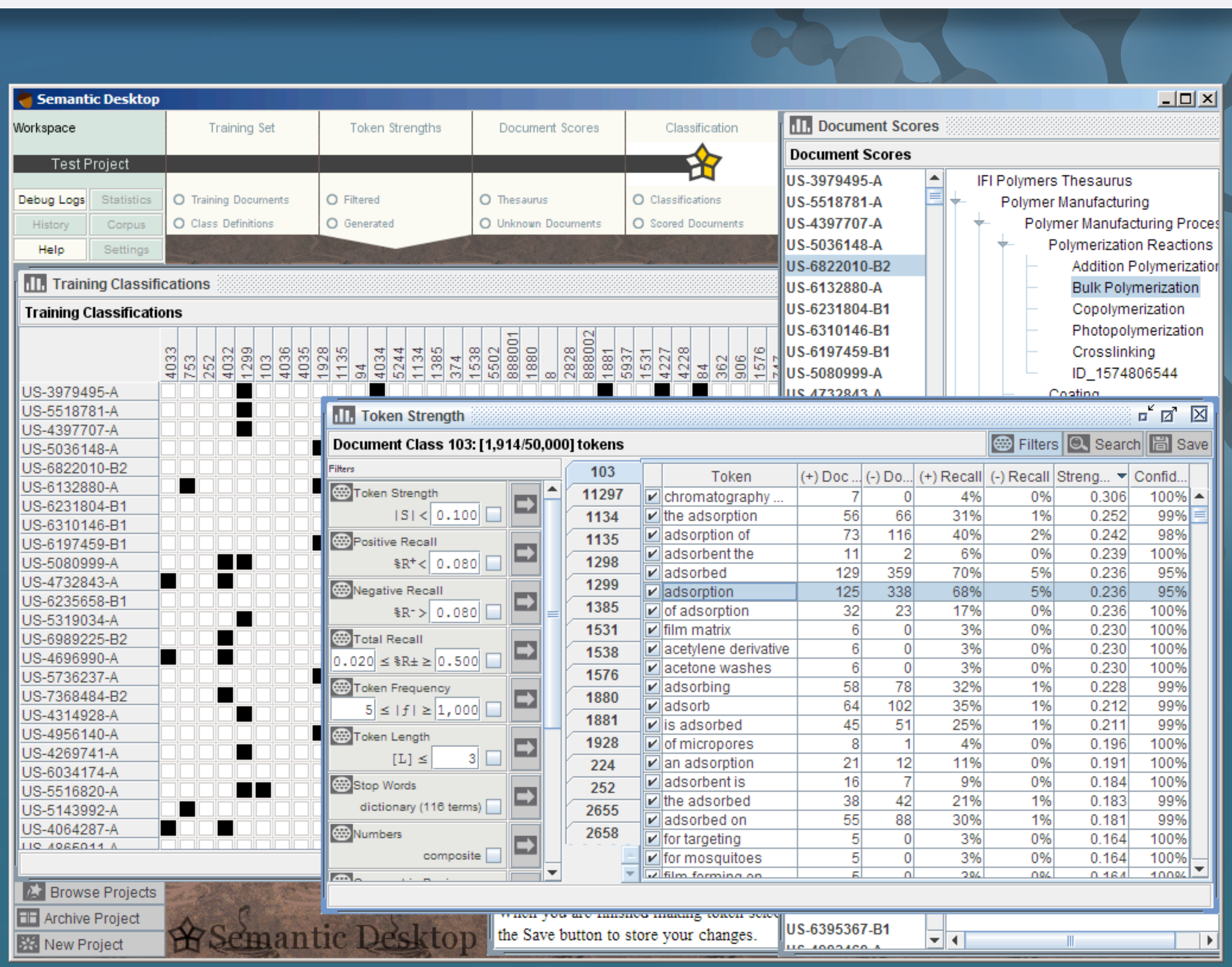
*(Left). Taxomatic is a Java component library and interactive desktop application under development for visualizing and ordering large similarity/distance matrices with a taxonomy. The methods of this approach are published as U.S. Patent 8,036,997 (Garrity and Lilburn 2011).*

*Shown here is the current 16S rRNA similarity matrix for all type strains of Bacteria and Archaea (as of February 2015, 12,937 ² / 2 = 84M comparisons arranged in a 168M cell floating point matrix). The taxonomic groups can be clearly identified, as well as any outliers, providing an easy way to detect misclassified strains or bad sequence data. The Archaea are the cluster at the far lower left. Along the top (grayscale) is a chart depicting a single strain's similarity to all other type strains.*

*We are currently extending this approach to operate on whole genomes using a recently developed method for classifying nucleotide sequences by Latent Semantic Analysis [2].*

*The software component can be used in combination with any matrix or collection of matrices [3], which extends its applicability beyond sequence analysis.*

*When complete, this component will be integrated into the Semantic Desktop text-mining suite (above right) to visualize and cluster documents (i.e., patents and articles) based on shared concepts [1,3].*



*Several additional software components were developed to overcome technical barriers that arose during this project. Originally implemented as command-line utilities for vocabulary extraction, annotation and document analysis, we have developed the individual software components into a set of libraries for text mining, information extraction, document classification and terminology development. The Semantic Desktop (above) is a Java Application based on those libraries, and the components may alternatively be deployed in a web service container or integrated with third party software. The above screenshot is part of a commercial case study using the Fairview Research Alexandria Patent Database, where we demonstrate the ability to reverse-engineer the logic 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.*

NamesforLife, LLC continues to develop its Intellectual Property based on technologies developed under the DOE STTR program. Our current patent portfolio is shown below grouped by patent family with priority dates, filing dates.

| Patent No. | Issued | Published | Priority | Status | Application No. | Provisional No. |
|---|---|---|---|---|---|---|
| US 7,925,444 | 4/12/2011 | 7/21/2005 | 1/16/2004 | Issued | US 10/759,817 | |
| US 8,036,997 | 10/11/2011 | 12/3/2009 | 6/16/2005 | Issued | US 11/922,273 | 60/690,969 |
| WO 2006/138502 | 4/2/2009 | 9/20/2007 | 6/16/2005 | Issued | PCT/US2006/023381 | |
| | | 8/5/2010 | 1/12/2009 | Pending | US 12/685,964 | 61/143,986 |
| WO 2010/081133 | | 7/15/2010 | 1/12/2009 | Pending | PCT/US2010/020734 | |
| EP 2386089 | | 1/16/2013 | 1/12/2009 | Pending | EP 2010/0729654 | |
| US 8,903,825 | 12/2/2014 | 2/23/2012 | 5/24/2011 | Issued | US 13/478,973 | 61/489,362 |
| WO 2012/162405 | | 11/29/2012 | 5/24/2011 | Pending | PCT/US2012/039168 | |
| EP 2715474 | | 4/9/2014 | 5/24/2011 | Pending | EP 2012/0790213 | |
| | | 5/1/2014 | 7/30/2012 | Pending | US 13/954,925 | 61/677,316 |
| WO 2014/022441 | | 2/6/2014 | 7/30/2012 | Pending | PCT/US2013/052797 | |
| | | 2015 | 9/20/2013 | Filed | | 61/880,244 |
| | | 2015 | 9/20/2013 | Filed | PCT/US2014/056808 | |

## Current and Planned Products

We recently deployed a strain finder service (http://doi.org/10.1601/strainfinder) that serves as a search interface, persistent landing page and forwarding service for strain identifiers (e.g., http://doi.org/10.1601/strainfinder?urlappend=%3fid=ATCC+23350). It integrates with the Taxonomic Abstracts and serves as a permanent, electronically traversable link from publications or 3rd party resources directly to specimens held in collections.

A Strain Registration database is being developed in collaboration with the Joint Genome Institute. NamesforLife will register Digital Object Identifiers and CrossRef metadata to all strains sequenced at JGI.

A faceted search engine over the phenotypic characters of prokaryotic strains is under development, which will be the main point of entry to the phenotypic knowledge base. Ontology specialists and developers may interact with the knowledge base using a SPARQL endpoint in addition to the faceted search.

## Publications

1. Parker, CT, Garrity, GM and Krdzavac, NB. *Systems and Methods for Establishing Semantic Equivalence Between Concepts*. International Application No. PCT/US2014/056808. Filed September 20, 2014. Washington, DC: U.S. Patent and Trademark Office and Geneva, Switzerland: World Intellectual Property Organization.
2. Sayood, K, Way, S, Ozkan UN and Garrity, GM. *Classification of Nucleotide Sequences by Latent Semantic Analysis*. International Application No. PCT US2013/052797. Published June 2, 2014. Washington, DC: U.S. Patent and Trademark Office and Geneva, Switzerland: World Intellectual Property Organization.
3. Parker, CT and Garrity, GM. *Semiotic Indexing of Digital Resources*. U.S. Patent 8,903,825. Issued December 2, 2014. Washington, DC: U.S. Patent and Trademark Office.

**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.*