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

NamesforLife
Bringing meaning to life...

Project Goals

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

Our core technical objectives are: (1) to build a database of normalized phenotypic descriptions (observational data) using the primary taxonomic literature of bacterial and archaeal type strains, and (2) to construct an ontology with reasoning capabilities to make accurate phenotypic and environmental inferences based on that data.

This project is tightly coupled with ongoing DOE projects (the 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).

The scope of this project covers areas of text-mining and data-mining, Natural Language Processing, terminology development, ontology development, machine reasoning and semantic analysis.

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*
- □Growth Curves
- ☑Cell Images

* 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. The KBase will enable rapid iteration of experiments that draw on a wide variety of data and allow researchers to infer how cells and communities respond to natural or induced perturbations, and ultimately to predict outcomes.

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

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 wide variety of forms, often use complex and non-uniform descriptors and are scattered about specialized databases and scientific and technical literature. Incorporating phenotypic information into the KBase requires expertise in harvesting, modeling and interpreting these data.

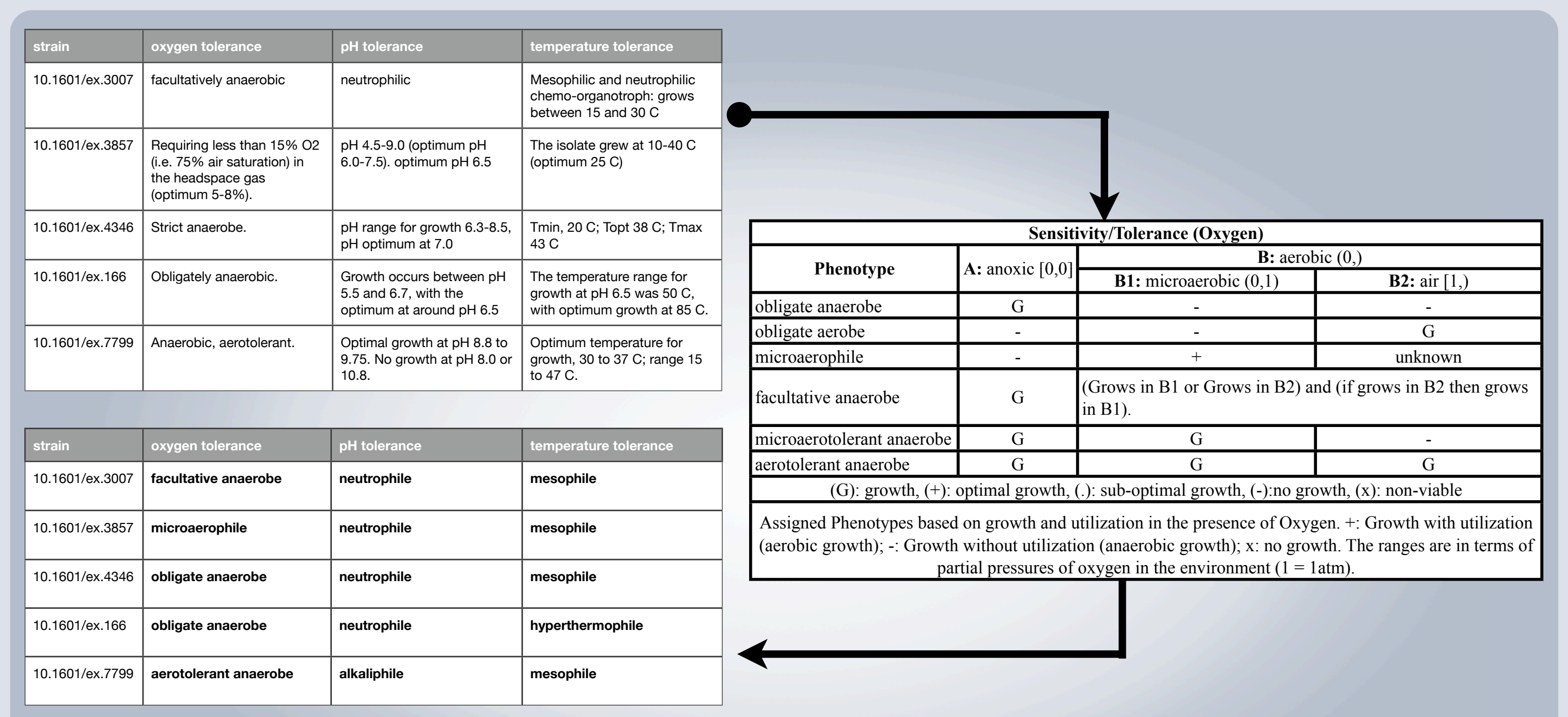
Our Solution

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 approximately 40,000 candidate phenotypic terms used in 5,750 (now expanded to 11,676 of 17,793 total) new and emended descriptions of the 11,492 distinct type strains of *Bacteria* and *Archaea*. In Phase II, 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 our past year of research, we have discovered novel design patterns for ontology development [1] that address these problems and remove barriers to machine reasoning over these complex terms.

Technical Challenges

This project has presented technical challenges that require creative solutions across several areas of information science.

Many ontologies consist of a large thesaurus of terms in a narrowly-defined domain and do not contain any reasoning capability beyond the taxonomic structure of the vocabulary and relations among concepts. Our objective is to develop an ontology that covers many broad feature domains and contains axioms encoded in first order logic that enable reasoning and inference over sparse phenotypic data, even in feature domains that contain partially-overlapping concepts and terms that map to undefined ranges of environmental conditions. In order to accomplish this, we have developed a core ontology model that maps between imprecise phenotypic features and precise environmental data.



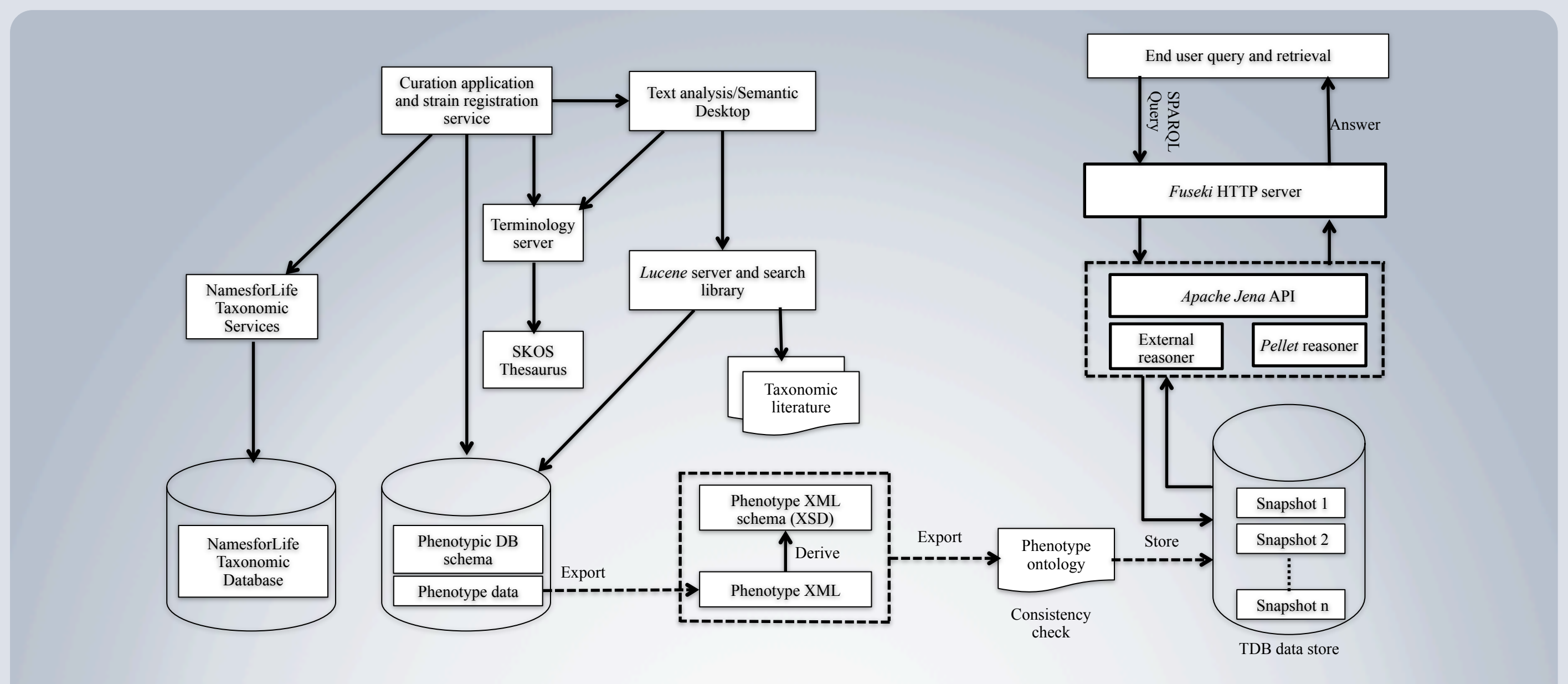
strain	oxygen tolerance	pH tolerance	temperature tolerance
10.1601/ex.3007	facultatively anaerobic	neutrophilic	Mesophilic and neutrophilic chemo-organotroph: grows between 15 and 30 C
10.1601/ex.3857	Requiring less than 15% O2 (i.e. 75% air saturation) in the headspace gas (optimum 5-8%).	pH 4.5-9.0 (optimum pH 6.0-7.5). optimum pH 6.5	The isolate grew at 10-40 C (optimum 25 C)
10.1601/ex.4346	Strict anaerobe.	pH range for growth 6.3-8.5, pH optimum at 7.0	Tmin, 20 C; Topt 38 C; Tmax 43 C
10.1601/ex.166	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 C, with optimum growth at 85 C.
10.1601/ex.7799	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.

Sensitivity/Tolerance (Oxygen)			
Phenotype	**A:** anoxic [0,0]	**B:** aerobic (0,)	
		B1: microaerobic (0,1)	**B2:** air [1,)
obligate anaerobe	G	-	-
obligate aerobe	-	-	G
microaerophile	-	+	unknown
facultative anaerobe	G	(Grows in B1 or Grows in B2) and (if grows in B2 then grows in B1).	
microaerotolerant anaerobe	G	G	-
aerotolerant anaerobe	G	G	G
(G): growth, (+): optimal growth, (.): sub-optimal growth, (-):no growth, (x): non-viable			
Assigned Phenotypes based on growth and utilization in the presence of Oxygen. +: Growth with utilization (aerobic growth); -: Growth without utilization (anaerobic growth); x: no growth. The ranges are in terms of partial pressures of oxygen in the environment (1 = 1atm).			

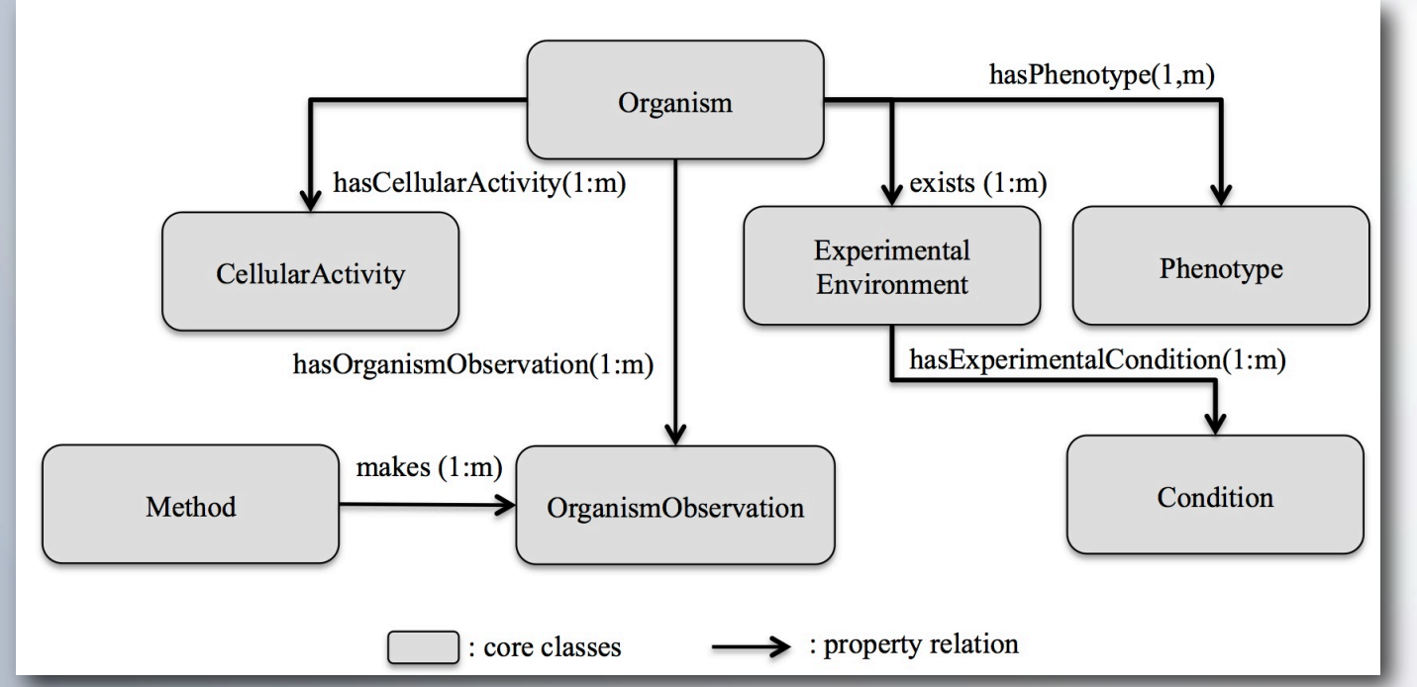
strain	oxygen tolerance	pH tolerance	temperature tolerance
10.1601/ex.3007	**facultative anaerobe**	**neutrophile**	**mesophile**
10.1601/ex.3857	**microaerophile**	**neutrophile**	**mesophile**
10.1601/ex.4346	**obligate anaerobe**	**neutrophile**	**mesophile**
10.1601/ex.166	**obligate anaerobe**	**neutrophile**	**hyperthermophile**
10.1601/ex.7799	**aerotolerant anaerobe**	**alkaliphile**	**mesophile**

The top left table shows some individual raw phrases extracted from the original literature using a custom grammar and classified into feature domains using a modified TF-IDF algorithm. When this text is normalized and given as input to a reasoner with our phenotypic ontology, the reasoner can infer appropriate mappings between each organism, its phenotype, and the environmental conditions it prefers, tolerates, or cannot survive in. The reasoning for Oxygen Sensitivity and Tolerance is depicted here as a table, but the logic is encoded in the ontology as axioms. More complex phenotypes overlap several feature domains, requiring extensive modeling and axiom development.

In our current work, we are applying these novel modeling techniques to encode *Tbox* axioms for automatically resolving ambiguity attributed to the semantic equivalence and imprecision of phenotypic terms arising in literature [2]. These axioms will enable reasoners (e.g., Pellet, Fact++) to make appropriate inferences over the ontology and phenotypic data. We are also developing a query and retrieval service linked to the ontology that will provide researchers with consistent, accurate interpretations of these data that are usable for predictive modeling and in other research and commercial applications.

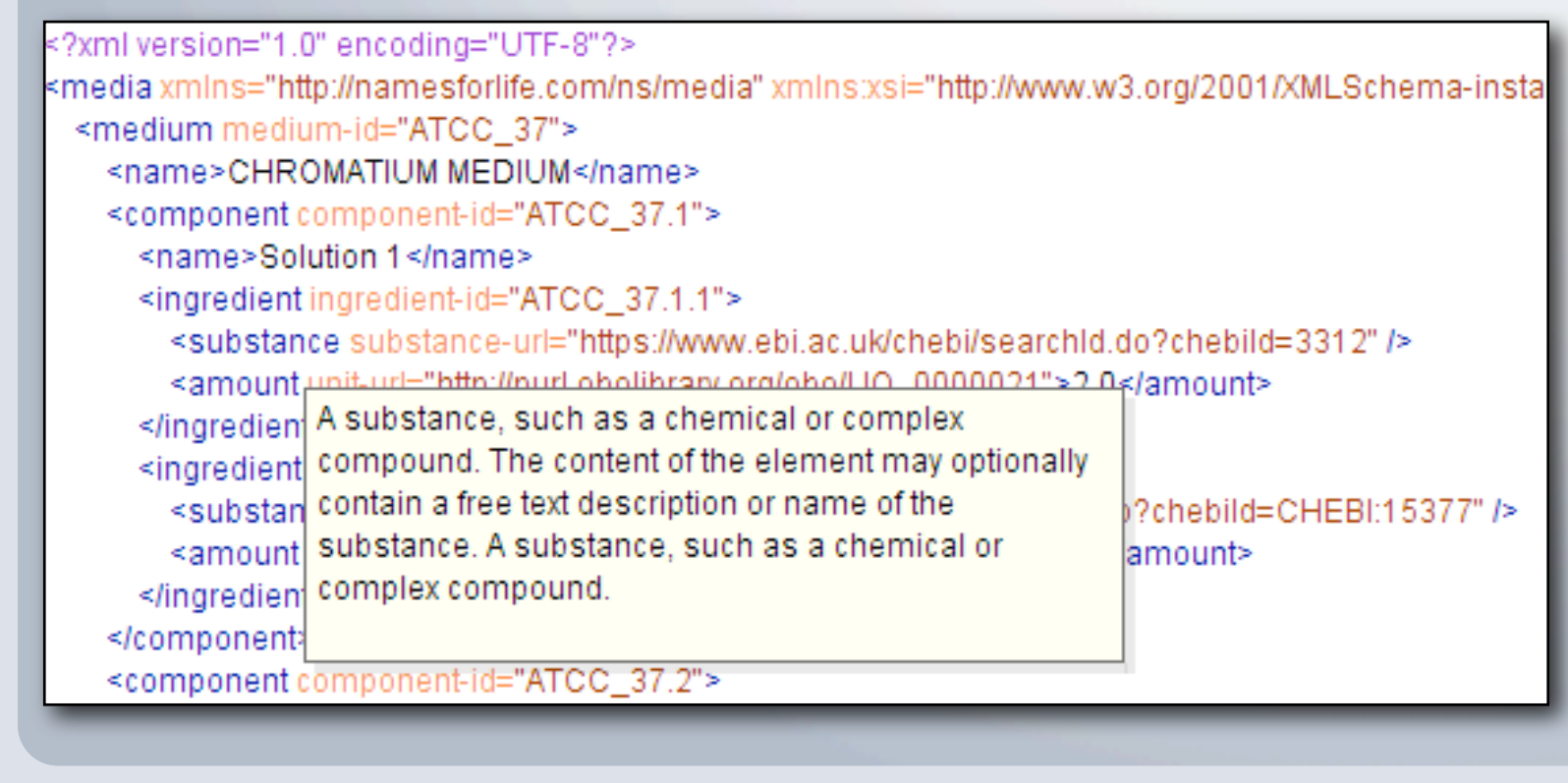


(Above). We have adopted a hybrid relational database (RDB) / ontology architecture in order to support curation, reasoning, search and query. The relational schema is mapped to the ontology via an intermediate XML schema. Nightly snapshots of the relational database will be loaded into the ontology data store and individual records will be checked for consistency over the ontology. We are currently using the Fuseki SPARQL query server, part of the Apache Jena ontology framework. The ultimate goal is to provide a set of end-user services for query and retrieval as well a strain registration service, which will offer researchers a way of describing strains prior to publication.



(Left). The core model is shown to the left and provides the high-level relationships between organism, cellular activity, environment and phenotype. The full model extends the classes in the core model and currently covers 115 narrowly-defined phenotypic feature classes in 16 major feature domains. The individual features are much more numerous, and are being extracted and curated from a candidate vocabulary of over 40,000 distinct terms.

The vocabulary is stored in a SKOS-XL, a W3C recommendation for thesaurus storage, based on the OWL and RDF standards. The ontology itself is encoded in OWL2 and the Tbox axioms are encoded in Description Logic, a general syntax for first order logic which is compatible with several commercial and Open Source reasoners.



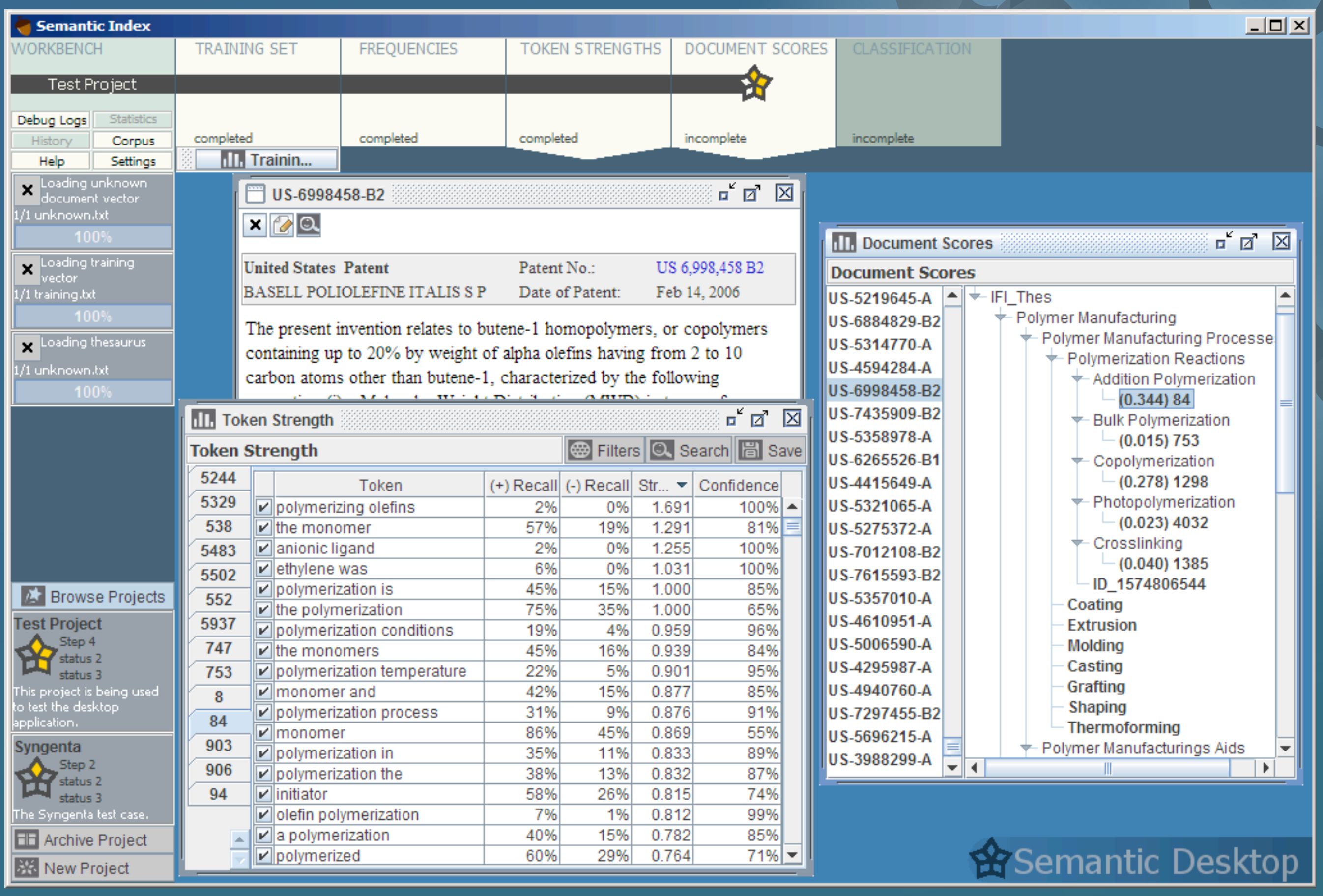
```
<?xml version="1.0" encoding="UTF-8"?>
<media xmlns="http://namesforlife.com/ns/media" xmlns:xsi="http://www.w3.org/2001/XMLSchema-insta
  <medium medium-id="ATCC_37">
    <name>CHROMATIUM MEDIUM</name>
    <component component-id="ATCC_37.1">
      <name>Solution 1</name>
      <ingredient ingredient-id="ATCC_37.1.1">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

(Left). 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 (in SKOS format) 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 between storage platforms.

At each interface between major architectural components, we take care to fully conform to existing standards for knowledge representation.

Several additional software components were developed to overcome technical barriers that arose during this project [3]. Originally implemented as command-line utilities for vocabulary extraction, annotation and document analysis, we are now developing these into a commercial semantic desktop application for document/corpus analysis and for bootstrapping terminology/ontology development.



completed | completed | completed | incomplete | incomplete

Loading unknown document vector 1/1 unknown.txt 100%
Loading training vector 1/1 training.txt 100%
Loading thesaurus 1/1 unknown.txt 100%

Test Project
Step 4
status 2
status 3
This project is being used to test the desktop application.

Syngenta
Step 2
status 2
status 3
The Syngenta test case.

US-6998458-B2

United States Patent	Patent No.:	US 6,998,458 B2
BASELL POLIOLEFINE ITALIS S P	Date of Patent:	Feb 14, 2006

The present invention relates to butene-1 homopolymers, or copolymers containing up to 20% by weight of alpha olefins having from 2 to 10 carbon atoms other than butene-1, characterized by the following

Token Strength

		Token	(+) Recall	(-) Recall	Str...	Confidence
5244	☑	polymerizing olefins	2%	0%	1.691	100%
5329	☑	the monomer	57%	19%	1.291	81%
538	☑	anionic ligand	2%	0%	1.255	100%
5483	☑	ethylene was	6%	0%	1.031	100%
5502	☑	polymerization is	45%	15%	1.000	85%
552	☑	the polymerization	75%	35%	1.000	65%
5937	☑	polymerization conditions	19%	4%	0.959	96%
747	☑	the monomers	45%	16%	0.939	84%
753	☑	polymerization temperature	22%	5%	0.901	95%
8	☑	monomer and	42%	15%	0.877	85%
84	☑	polymerization process	31%	9%	0.876	91%
903	☑	monomer	86%	45%	0.869	55%
906	☑	polymerization in	35%	11%	0.833	89%
94	☑	polymerization the	38%	13%	0.832	87%
	☑	initiator	58%	26%	0.815	74%
	☑	olefin polymerization	7%	1%	0.812	99%
	☑	a polymerization	40%	15%	0.782	85%
	☑	polymerized	60%	29%	0.764	71%

Document Scores

US-5219645-A
US-6884829-B2
US-5314770-A
US-4594284-A
US-6998458-B2
US-7435909-B2
US-5358978-A
US-6265526-B1
US-4415649-A
US-5321065-A
US-5275372-A
US-7012108-B2
US-7615593-B2
US-5357010-A
US-4610951-A
US-5006590-A
US-4295987-A
US-4940760-A
US-7297455-B2
US-5696215-A
US-3988299-A

- IFI_Thes
  - Polymer Manufacturing
    - Polymer Manufacturing Processe
      - Polymerization Reactions
        - Addition Polymerization
          - (0.344) 84
        - Bulk Polymerization
          - (0.015) 753
        - Copolymerization
          - (0.278) 1298
        - Photopolymerization
          - (0.023) 4032
        - Crosslinking
          - (0.040) 1385
        - ID_1574806544
      - Coating
      - Extrusion
      - Molding
      - Casting
      - Grafting
      - Shaping
      - Thermoforming
    - Polymer Manufacturings Aids

The individual software components developed to support ontology construction are being integrated into a single application, tentatively named the Semantic Desktop. The components may alternatively be deployed in a web service container or integrated with third party software. Our approach is language-independent and has been tested in several technical domains. 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. A user provides a thesaurus in SKOS Core format to visualize the results. The application includes a KWIC index (not shown above but demonstrated at GtL 2013) that allows an end user to view the context of any term across all documents in the corpus. Once complete, the application will be used to assist curators with phenotypic data extraction and supporting further ontology development. We plan to offer this as a commercial product in Q3 2014.

Current Work

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

Corpus construction is essentially complete (with the exception of monthly additions via newly published taxonomic literature, amounting to approximately 1,000 new type and reference strains per year). Our manual OCR correction effort has yielded descriptions for an additional 5,926 strains from historical literature. Our custom text analysis and annotation tools are operational and yielded excellent starting points for ontology development. 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 development staff is 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.

Future Work

As soon as the integration test is completed, work can begin on populating the phenotypic profiles from the semi-normalized protologues. 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. 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.

Publications

1. Parker, CT, Garrity, GM and Krdzavac, NB. *Systems and Methods for Inferring Properties of Objects in the Absence of Direct Observations.* U.S. Provisional Patent Application No. 61/880,244. Filed September 20, 2013. Washington, DC: U.S. Patent and Trademark Office.

2. Krdzavac, NB, Parker, CT, Garrity, GM. An Observation Ontology. *Bioinformatics*. 2014, Under review.

3. Garrity GM. *The NamesforLife Semantic Index of Phenotypic and Genotypic Data: Phase I Final Technical Report* [Internet]. 2012, doi:10.1601/report.sc0006191p1

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.

MICHIGAN STATE UNIVERSITY