Project Goals
Predictive models depend on high quality input data, but not all data are of similar quality and are all amenable to computational analysis without extensive cleaning, interpretation, and normalization. Key among these needed to make 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 will require expertise in harvesting, modeling and interpreting the data. The NamesforLife Semantic Index of Phenotypic and Genotypic Data will be built on an ontology of biological and archival phenotypes based on the taxonomic literature. This project aims to achieve its first objective: a draft vocabulary for the phenotypic features of the taxonomic type strains.

Background
To manage dynamic terminologies Garrity and Lyons developed a semantic model (the N4L data model) that represents names, taxa (global for taxon), and exemplars (representations of what is known about a name) with a rich set of features and relationships. The N4L data model provides a way to represent all of the complex relationships that exist among names and the concepts and objects to which names refer. Each such identification is pivotally linked to a portal which allows for placement of forward-pointing links in the published literature and in databases and provides a mechanism for resolving ambiguities ("future proofing" a name). The NamesforLife Semantic Index of Phenotypic and Genotypic Data will require expertise in harvesting, modeling and interpreting the data. The project will require access to information while reading content in HTML form. This will allow for the placement of forward-pointing links in the published literature and in databases and provides a mechanism for resolving ambiguities ("future proofing" a name). The NamesforLife Semantic Index of Phenotypic and Genotypic Data will be built on an ontology of biological and archival phenotypes based on the taxonomic literature. This project aims to achieve its first objective: a draft vocabulary for the phenotypic features of the taxonomic type strains.

The NamesforLife Semantic Index of Phenotypic Data

Our approach towards developing a draft vocabulary of bacterial and archaeal phenotypes is based on the long-term objective of this STTR project is to develop a semantic index of bacterial and archaeal phenotype that can be used to augment annotation efforts and to provide a basis for predictive modeling of microbial phenotypes. The index is based on published descriptions of taxonomic type and non-type strains that have been the subject of ongoing genome sequencing efforts as this will provide a mechanism whereby hypothetical sequences can be tested and reproducibility verified. This project is supported by the Michigan State University, East Lansing, MI, the University of Michigan, Ann Arbor, MI, and the University of California, Berkeley, CA.

Figures and Diagrams

Figure 1. Creation of a nomenclatural taxonomy from N4L knowledge objects is accomplished by mapping the linkages via DOIs. This ensures that N4L::Exemplar objects are properly and persistently identified using the DOIs, which are citable micro-publications that are directly accessible via EPO Green technology patent collection from Fairview Research (n=380,000 patents) reveal pathogen and environmental profile of 126 bacterial and archaeal strains.

Figure 2. The N4L Data Architecture
N4L::Guide uses a layered architecture. Company databases include the nomenclatural database and a reference database holding all of the taxonomic literature that is relevant to the names tracked by the company. The Alexandria patent repository (IP - Fairview) contains more than 80 million patent documents from 70 countries and is processed and indexed with N4L::Scribe. We also provide a means of accessing public repositories (e.g., the Genomes Online Database) or private databases and provide the necessary infrastructure to ensure that intellectual property rights are protected. NamesforLife tools and methods provide access to these repositories and enrich web content by rendering contextually correct captions based on biological names. Other custom solutions include the N4L::Native Index, which integrates current scientific, technical and medical and patent literature using our proprietary semantic fingerprinting technology, and the N4L::Taxonomy Abstracts, a collection of approximately 50,000 citable micro-publications.

Figure 3. End user access to NamesforLife content - N4L::Guide is designed to provide readers with full access to information while reading content in HTML form. (left panel) Each instance of a validly published bacterial or archival name is converted into a link by the N4L::Scribe client side using the Reader Edition or server side using the Publisher or Developer Editions. (right panel) N4L::Guide text is citable via DOIs. Each species and subspecies is represented by three separate documents (Name, Taxon, and Abstract). (left panel) The dynamic model with a NamesforLife model. Higher taxa are represented by two documents (Name and Taxon Abstracts).

Figure 4. NamesforLife web services for other content types. (left panel) Semantic enrichment of the Microbial Earth Project tree by the N4L::Guide. (right panel) N4L::PatentScrobe is designed to annotate US and non-US patents and is available to clients with a need for access to these documents. The NamesforLife DOIs are a persistent index of patented names and provides end-users with not only annotation services, but also for indexing, searching and analyzing this corpus of literature using the Company’s Semantic Fingerprinting. N4L::PatentScrobe is a special case of vector space models. Ours is distinct in that it uses an externally managed terminology in which the synonyms and other semantic ambiguities are automatically resolved by the N4L data model. This allows end users to fine-tune the level of taxonomic or phylogenetic granularity to meet specific needs. In our approach to mining text, the meaning of a term is known a priori and defines the scope of the search space. The combination of terms allows as to deduce the likely meaning of a document based on the properties of the ontologies that are referenced. The more complex the fingerprint, the greater the resolving power.

Figure 5. Clustering of patents by organism and technology classification. Preliminary experiments using the EPO Green technology patent collection from Fairview Research (n=380,000 patents) reveal a meaningful similarity matrix was clustered, visualized as a heat map, and output as an ordered list of patent IDs.

Figure 6. N4L::KWIC Index A view of the curatorial environment used for manual review and editing of terms used in the NamesforLife Index of Phenotypic Terms of Archaea and Bacteria.

Lessons learned
Fatty acids are described using dozens of variations and so many deviations from the standard nomenclature as to render them useless as keywords without creating mappings to resolve synonyms. Delimitation breaks the down fatty acid terms in ways that show history (and meanings). For instance, C18:1ω9c, which is not a proper fatty acid, is split into C18 and C18:1, which is not a proper conceptual break within the term. Similarly, when C16:1ω7c:11c. is converted to plain text (as required by Lucene and other search platforms), the lexical structure is lost when the HTML tags are stripped, resulting in the nonsensical string C16:1ω7c:11c. Subsequently, the default tokenization breaks on punctuation and white space, resulting in the tokens C16 and 11c, further losing the original meaning of the term.

To work around these issues, we have developed custom grammars in FIsys (Java Fast Lexical Analyzer, used in Lexire for tokenizing). Our custom grammars properly tag nearly all variations of fatty acids in the corpus, as well as extract measurements, chemical compounds and strain identifiers. A drawback of creating such complex grammars is that they interface with each other when used in serial tokenization. Therefore, we ran them independently, creating multiple indices.

Given the number of variations in fatty acid and other compound names in the literature, it is difficult to navigate the term lists. Therefore, we developed a second level of indexing against the term lists to manageable working sets. Our approach towards developing a draft vocabulary of bacterial and archaeal phenotypes is based on the long-term objective of this STTR project is to develop a semantic index of bacterial and archaeal phenotypes that can be used to augment annotation efforts and to provide a basis for predictive modeling of microbial phenotypes. The index is based on published descriptions of taxonomic type and non-type strains that have been the subject of ongoing genome sequencing efforts as this will provide a mechanism whereby hypothetical sequences can be tested and reproducibility verified. This project is supported by the Michigan State University, East Lansing, MI, the University of Michigan, Ann Arbor, MI, and the University of California, Berkeley, CA.

NamesforLife, LLC, East Lansing, Michigan, US and Edinburgh, UK

Michigan State University, East Lansing, MI

Acknowledgments
Funding for this project was provided through the DOE SBIR/STTR program (DE-FG02-07ER86321), the Michigan Small Business Technology Development Corporation, the Michigan Strategic Fund, and the Michigan Universities Commercialization Initiative.