Semantic Indexing of the Green Technology Patent Literature

An Application of the NamesforLife Contextual Index

George M. Garrity[1,2], Charles Parker[1], Dorothea Taylor[1], Kara Mannor[1], Catherine Lyons[1]
[1]NamesforLife, LLC, East Lansing, Michigan, US and Edinburgh, UK
[2]Michigan State University, East Lansing, MI

NamesforLife
Bringing meaning to life...

Background

As DOE research on biofuels, bioremediation and carbon sequestration moves from the laboratory into production or commercial environments, a number of important policy and business decisions must be made that demand correct information. These include establishing the patentability of a given technology, freedom to operate, and potential infringement of patents held by competitors, both in the U.S. and abroad. Failure to pay careful attention to these issues can have serious consequences beyond the payment of stiff penalties for infringement. These include lost opportunities arising for technology licensing, failure to detect and understand regional disparities, rapid growth in patent coverage of technologies by competitors and migration of technology across international borders. The scientific and technical literature provides an incomplete view of any field having commercial potential because the underlying technologies are typically not revealed in public until absolutely necessary, and then only after patent applications have been filed. While patents with corresponding papers are not uncommon as a means of announcing important new developments, they are not obligatory. Therefore, an awareness of developments in the field requires a thorough review of both bodies of literature. This approach integrates well with existing commercial, academic and USPTO data mining capabilities.

The N4L Nomenclature Model

To manage dynamic terminologies, we have developed a semantic model that represents **names**, **taxa** (plural for taxon), and **exemplars** (representations of organisms) as distinct objects. NamesforLife uses a context-driven model of semantic resolution that is based on the rules of biological nomenclature, specifically bacterial nomenclature, but is generally applicable.

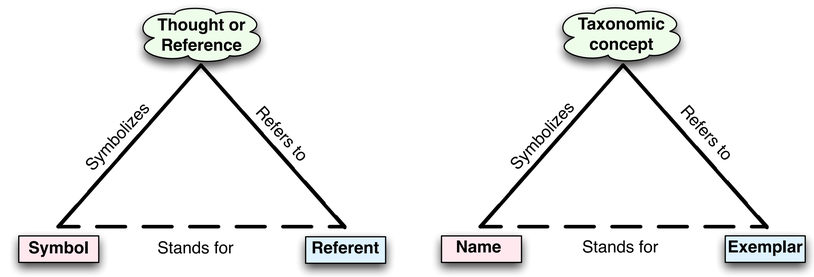

Figure 1. The semiotic triangle (left) and its application to biological nomenclature (right). Ogden and Richards (1923) and Sowa (2000) show that uncertainty arises from a failure to recognize that names (symbols) that are assigned to objects (referents) have meaning to the agent that interprets them 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 well-known problem of *name-rot*, the unpredictable decay that occurs because the taxonomic concept to which a name refers changes as new members are recognized or other rearrangements occur.

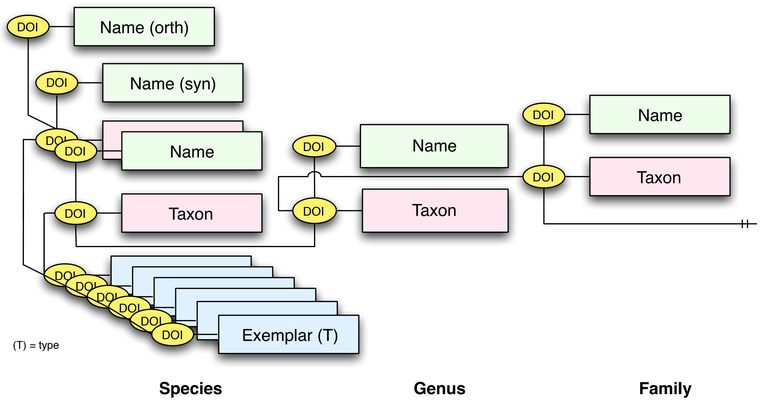

Figure 2. Assembly of N4L objects into a taxonomic hierarchy. In the N4L model, names, taxa, and exemplar objects are carefully mapped to provide an accurate representation of the precise meaning of a name at a given point in time. DOIs allow the information associated with these objects to be directly and persistently addressable on the web and formally referenced as micropublications (***N4L Taxonomic Abstracts***). The NamesforLife taxonomy is based on the published nomenclature and current taxonomic opinion, and is further refined through analysis of the best available 16S rRNA sequences for each type strain.

Nomenclature Project Status

At present, the *NamesforLife* Database (N4LDB) contains **14,650** distinct names, **13,883** of which are validly published, **119** *Candidatus*, and **47** that are illegitimate but relevant to the field. N4LDB also contains **14,939** exemplars (metadata representations of species/subspecies/strains), **9,461** of which represent distinct type strains for **11,511** taxa and **11,903** names, the remaining exemplars representing important non-type strains. The remaining **2,747** names are associated with higher taxa. The major classes of events that have occurred since publication of the Approved Lists in 1980, by event, are shown below. Less common events (Judicial Opinions, Revived Names, Rejected Names, Retractions, etc.) are not shown here.

Table 1. N4LDB Records by Rank

Rank	Taxa	Names
Domain	2	2
Phylum	35	36
Class	75	76
Subclass	7	7
Order	133	137
Suborder	24	25
Family	344	349
Subfamily	1	1
Genus	2,079	2,109
Subgenus	5	5
Species	10,980	11,333
Subspecies	531	570
Total	**14,216**	**14,650**

Table 2. Nomenclatural Events Recorded in N4LDB

Event	Count
Corrections	439
New Combinations	1,270
Heterotypic Synonyms	321
Homotypic Synonyms	163
Unifications	102
Automatically created names via rule 40d	53
Emendations	1,187
Validation List events	2,977
Valid Publication (excluding Validation Lists)	8,692

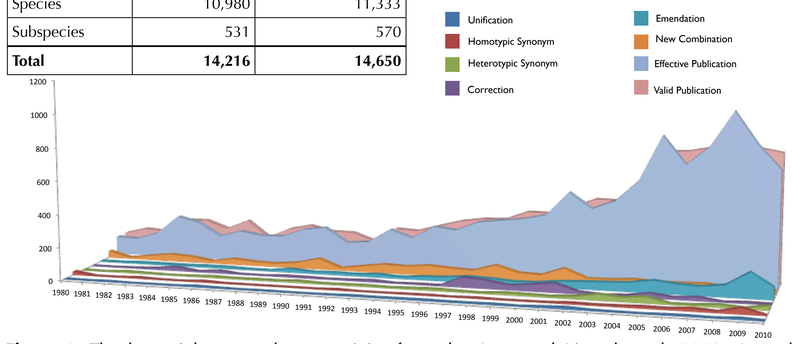

Figure 3. The bacterial nomenclature activity from the Approved Lists through 2010. A total of 33,606 nomenclatural events have been reported in 11,870 distinct references since 1980.

Indexing the Patent Space

NamesforLife, LLC has created a suite of software tools and techniques to manage dynamic terminologies using the underlying term set described in Tables 1 and 2, supplemented with links to the relevant taxonomic literature and key genetic and genomic information. The Company's N4L tools can automatically detect and tag bacterial and archaeal names in HTML and XML documents with a high degree of precision. An interactive browser-based application (***N4L Guide***) provides end users direct access to correct nomenclatural and supplementary information that is served-up on demand while reading the literature. N4L tools use ISO standard Digital Object Identifier (DOI) technology to create links at each occurrence of a validly published name in HTML documents. The company has also developed batch tools (***N4L Semantic Tagger***) that can embed N4L-DOIs into XML versions of scientific articles that are created as part of the contemporary publishing workflow and used to create human readable content in various forms (e.g., HTML, PDF, ink-on-paper). The company has also developed a unique way of tracking the occurrence of biological names in the literature, based on the usage of our tools (***N4L Contextual Index***).

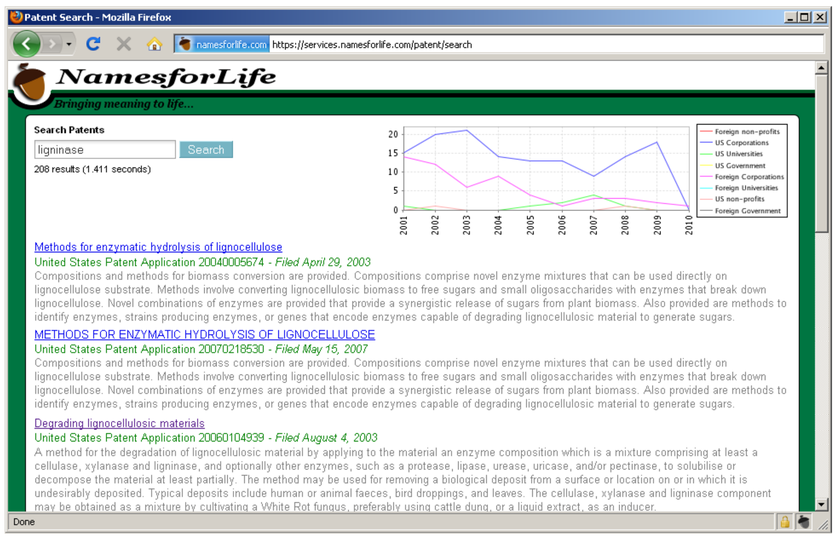

Figure 4. Searching U.S. Patents via the *N4L Contextual Index*. While initially intended as a tool for readers, authors, and publishers of scientific literature, the *N4L Contextual Index* can also be applied to other documents where bacterial names appear. As proof of principle, the company processed approximately 250,000 U.S. patents and patent applications with the *N4L Semantic Tagger* and then indexed the tagged documents using Apache Lucene to provide end users with additional search and retrieval capabilities. Simple graphical tools were added to support limited on-demand analyses of search results. These tools were designed to support data mining by non-commercial organizations and to highlight trends in commercialization for biodiversity research as part of ongoing discussion pertaining to the Convention on Biological Diversity. This work also led to the discovery of "terminological fingerprints" that could be used to classify patents and other documents using externally managed terms sets.

To validate the concept of "terminological fingerprinting", the company processed the European Patent Office (EPO) Green Technology Patent Collection, which consists of approximately 362,000 documents. In addition to detecting bacterial names, the *N4L Semantic Tagger* was modified to recover IPC and ECLA patent classifications, applicants, assignees, inventors, references, titles and other common patent landscape metadata.

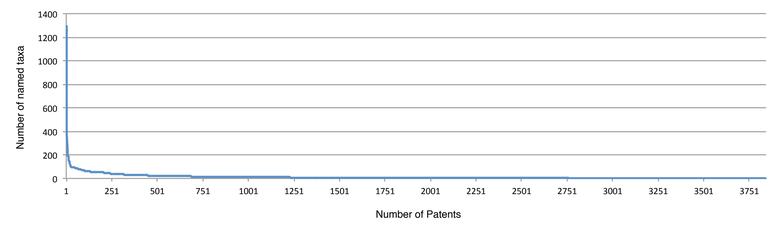

Figure 5. The long tail plot of bacterial and archaeal taxa referenced in the EPO Green Technology Collection. A total of 3,845 patents were found that made reference to 3,385 named bacteria and/or archaea held in the NamesforLife database. Of those, 626 names were unique to non-U.S. patents. The number of names per patent (name vectors) ranged from 1 - 1,290, with an average of 13 names and a median of 5 names. In addition to name occurrence, frequency data for each name occurrence per patent was tabulated. The resulting name vectors were then used to further examine the associations among the patents based on the IPC and ECLA patent classification systems. Simple associations could be derived directly from the captured data. However, more complex patterns involving multiple many-to-many relationships could only be ascertained from the cross-products of underlying contingency and frequency data.

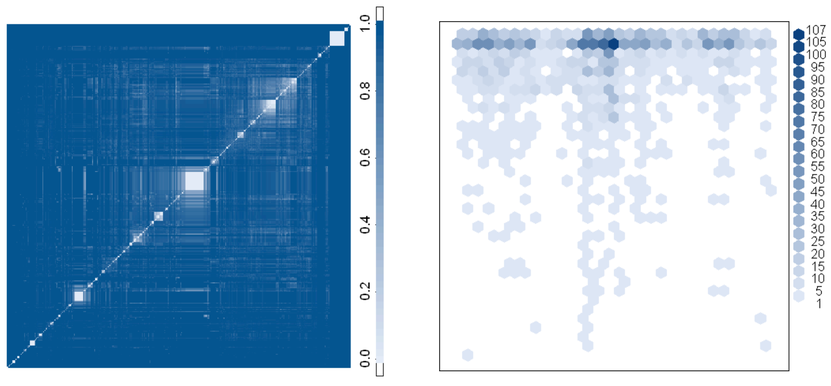

Figure 6. The Contextual Index was examined using routine approaches for exploratory data analysis and visualization (e.g., principal components analysis, robust clustering, 2D scatter plots, 3D spin plots and heatmaps). Each of these methods revealed strong evidence of terminological fingerprints in the patents. The heatmap on the left reveals the relationship among the Green Technology patents when classified using terminological fingerprints. While useful in revealing the underlying structure in the data, heatmaps are less useful as a component of a graphical user interface to interrogate large amounts of data. Hexagonal bin plots were found to be more suitable for large-scale applications, such as patent visualization as they scale well. The company is currently developing interactive hexagonal bin plots as a means of selecting subsets of patents that involve related technologies and microorganisms.

NamesforLife Semantic Services

A semantic tagging web service, ***N4L Scribe***, is now available. It tags bacterial names in any well-formed XML document with forward-linking Digital Object Identifiers. The service sits at the core of the server-side content enablement for ***N4L Guide*** (Figures 7a, 7b), and is intended for integration into existing publication workflows. Plug-ins are currently in development for several ubiquitous word processing and desktop publishing applications as well. The service can be tested out for free on our web site with a *NamesforLife* account.

The ***N4L Guide*** browser add-on detects and links bacterial names to the N4L database, providing up-to-date nomenclature, strain and genome information, and a full bibliography. The screenshots below demonstrate the use of this tool on an *IJSEM* article. Instructions for installing and using this tool can be found at the *NamesforLife* services website, located at:

https://services.namesforlife.com

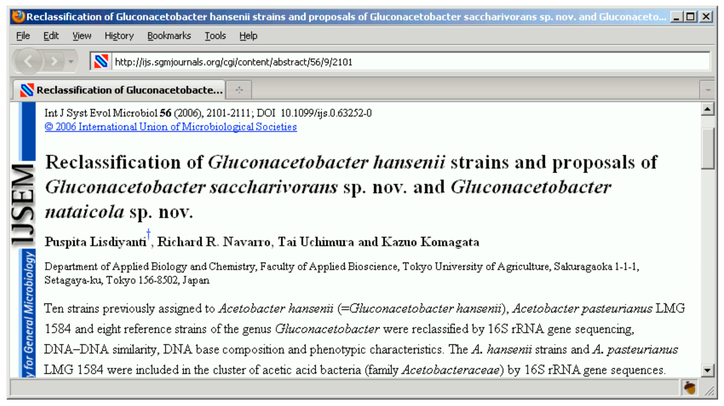

Figure 7a (above). An sample article prior to being semantically enabled by the ***N4L Guide***.

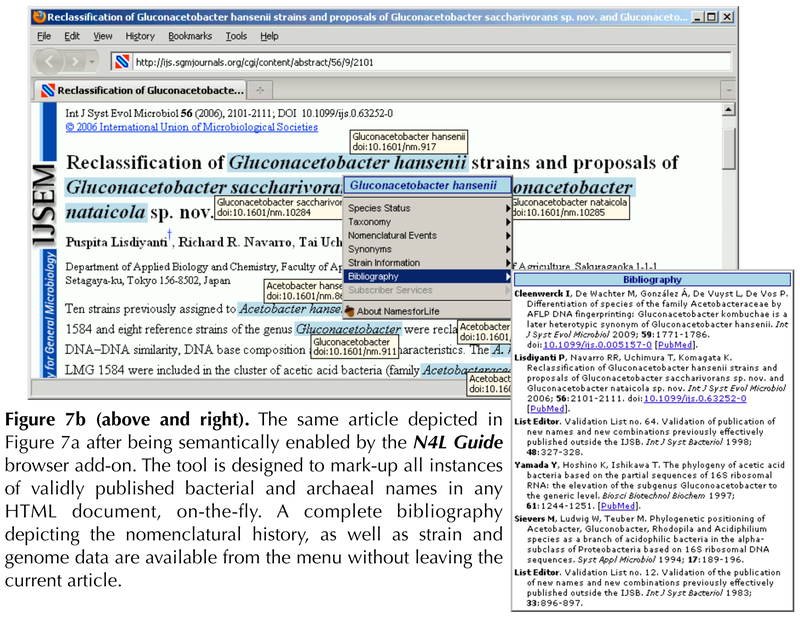

Figure 7b (above and right). The same article depicted in Figure 7a after being semantically enabled by the ***N4L Guide*** browser add-on. The tool is designed to mark-up all instances of validly published bacterial and archaeal names in any HTML document, on-the-fly. A complete bibliography depicting the nomenclatural history, as well as strain and genome data are available from the menu without leaving the current article.

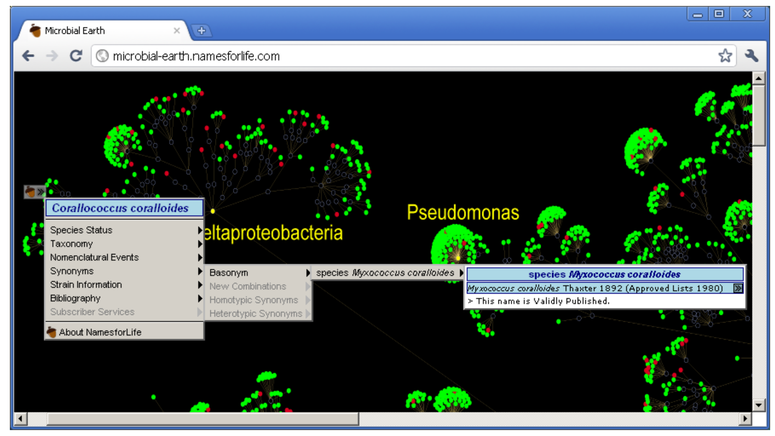

Figure 8. An interactive taxonomically-aware web application is currently under development in cooperation with the Department of Energy Joint Genome Institute (DOE JGI). The tree is from the forthcoming Microbial Earth project of Kyrpides et al. The prototype is freely available at http://microbial-earth.namesforlife.com.

Current Work

A new edition of the *Taxonomic Outline of Bacteria and Archaea* is planned to coincide with the new version of the NamesforLife Taxonomy and the ***N4L Taxonomic Abstracts*** (scheduled for release this quarter). These will provide a snapshot of Bacterial Nomenclature in the form of a citable micro-publication, and will serve to link existing literature to current nomenclature via CrossRef.

The *NamesforLife* database is kept in sync with the *Genomes OnLine Database* (GOLD), to provide curated links to metadata about all public genome sequencing projects, including non-type strains. We will soon deploy similar metadata for the *Human Microbiome Project*. We also plan to deploy a searchable database of phenotypic characteristics for the type strains of all *Bacteria* and *Archaea*.

Acknowledgments

We wish to thank B.J. Tindall (DSMZ, Braunschweig) and J. Euzéby (École Nationale Vétérinaire de Toulouse) for their helpful discussions regarding problematic nomenclature issues. We would also like to thank members of the International Committee on Prokaryotic Nomenclature for their support of these efforts, and Matt Winters, Denise Searles, Austin Kuo, Julia Bell, Judy Leventhal and Sheena Tapo for their assistance in curating the underlying taxonomic and nomenclatural information used in our models. This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Phase I and II STTR Awards DE-FG02-07ER86321 A001 - A005.