NamesforLife Semantic Resolution Services for the Life Sciences:

Moving Towards an Extensible and Interoperable System of Nomenclature

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...

Abstract

A major challenge in bioinformatics, the life sciences, medicine, and law is using correct and informative biological names. For a variety of reasons, proper names are not always used in scientific, technical, medical or legal literature, leading to accumulated semantic ambiguity that readers of the literature and end users of databases are left to resolve on their own. To assist those confronted with ambiguous names, we developed a generalizable semantic model (N4L) that represents names, taxonomic concepts, and exemplars (representations of biological entities) as distinct objects. By assigning each object a unique Digital Object Identifier (DOI), it becomes possible to place forward-pointing links into published literature, databases, and vector graphics that can be used as part of a mechanism for resolving ambiguity, thereby "future proofing" a nomenclature or terminology. A complete and up-to-date implementation of the N4L model for the *Bacteria* and *Archaea* is now available online. The system is professionally curated and fully backed by literature references. A variety of tools and web services are available to readers, publishers, database owners, software developers, and others. We are currently adding phenotypic, genotypic, and genomic information to the exemplars to add greater value to end users.

Background

To assist those confronted with ambiguous names (which not only includes researchers but clinicians, manufacturers, patent attorneys, and others who use biological data in their routine work), we developed a generalizable semantic model that represents names, concepts, and exemplars (representations of biological entities) as distinct objects (Figure 1). By identifying each object with a Digital Object Identifier (DOI) (Figures 2-4), it becomes possible to place forward-pointing links in the published literature, in databases, and vector graphics that can be used as part of a mechanism for resolving ambiguities, thereby "future proofing" a nomenclature or terminology. A full implementation of the N4L model for the *Bacteria* and *Archaea* was released in April, 2010. The system is professionally curated and represents a Tier III resource in Parkhill's view of informatic services (Parkhill et al. Genome Biology 2010, 11:402, Figure 7). A variety of tools and web services have been developed for readers, publishers, and others (***N4L Guide***, ***N4L Autotagger***, ***N4L Semantic Search***, ***N4L Taxonomic Abstracts***) and we are incorporating other taxonomies into the N4L data model, as well as adding additional phenotypic, genotypic, and genomic information to the existing exemplars to add greater value to end users (Figures 8-9).

The N4L Model

To manage dynamic terminologies, we have developed a semantic model (*the N4L data 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.

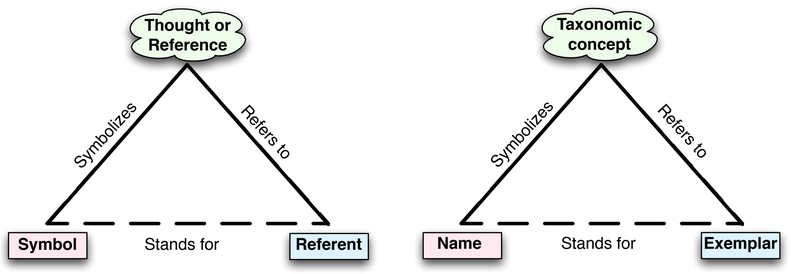

Figure 1. The semiotic triangle (left) and its application to biological nomenclature (right). Ogden and Richards (1923) and Sowa (2000) show that uncertainty arises from a failure to recognize that names (symbols) that are assigned to objects (referents) have meaning to the agent that interprets them 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.

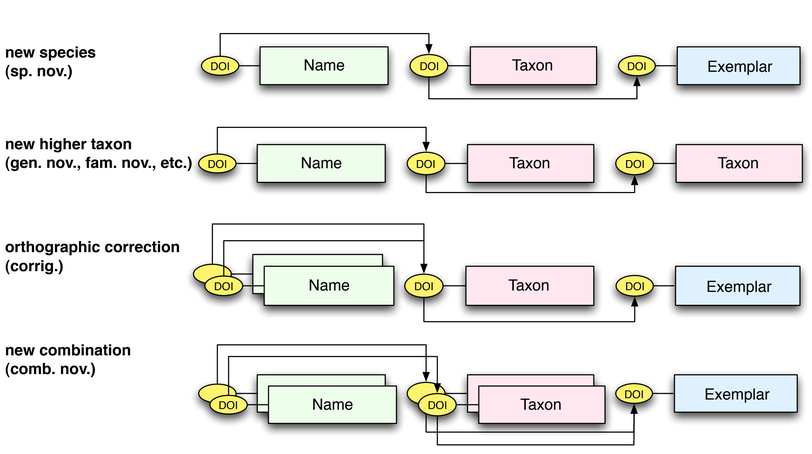

Figure 2. Some common nomenclatural events depicted using the *NamesforLife* semantic model. Clarity of meaning requires precise and unambiguous definitions that can be directly referenced, on demand. Biological names can be ambiguous because they are subject to change, are not guaranteed to be unique, and may exist in one-to-one, one-to-many, many-to-one and many-to-many relationships. Biological names (***Names***) and taxonomic concepts (***Taxa***) are precisely defined through the processes of typification of biological materials (***type strains***) at the species and subspecies level (top of figure) and type concepts (species, genera and orders) for higher taxa (second from top) that are part of published circumscriptions.

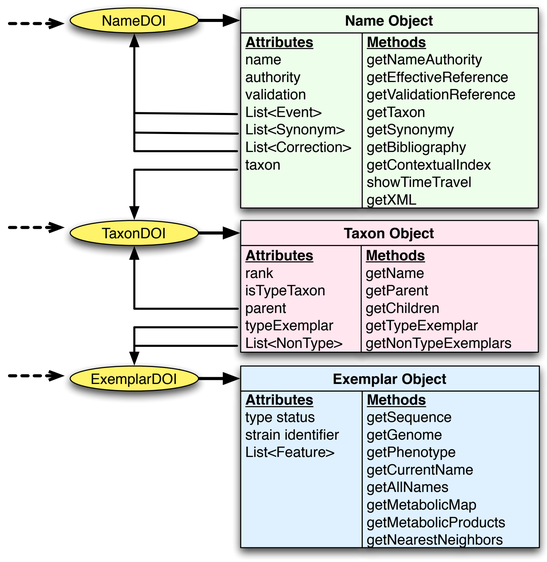

Figure 3. The N4L data model. A single N4L-DOI provides direct access to information about a name, a taxonomic concept, the organism (at the species or subspecies level) and associated web services. The N4L service provides end-users of biological nomenclatures or other dynamic terminologies with the appropriate information, in the correct temporal and taxonomic context, on demand. The same service can also be used by database owners, publishers, or other information providers to semantically enable their offerings, making them readily discoverable, even when the definition of a name or term has changed.

A Meaningful Taxonomy

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.

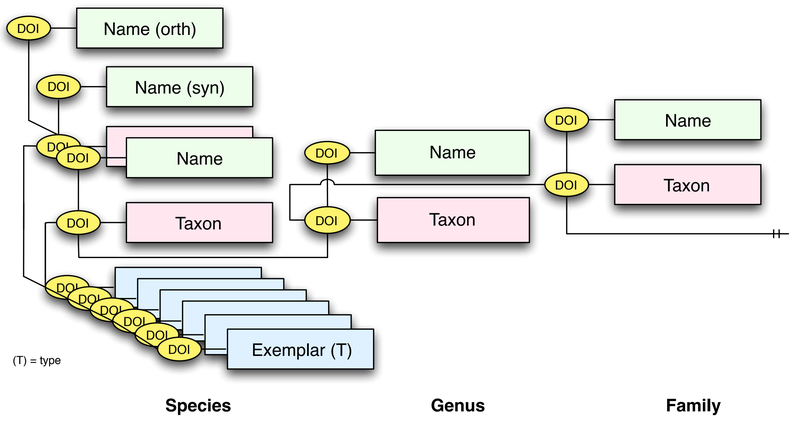

Figure 4. Assembly of N4L objects into a taxonomic hierarchy. In the N4L model, names, taxa, and exemplar objects are carefully mapped to provide an accurate representation of the precise meaning of a name at a given point in time. DOIs allow the information associated with these objects to be directly and persistently addressable on the web and formally referenced as micropublications (***N4L Taxonomic Abstracts***).

Database Statistics

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

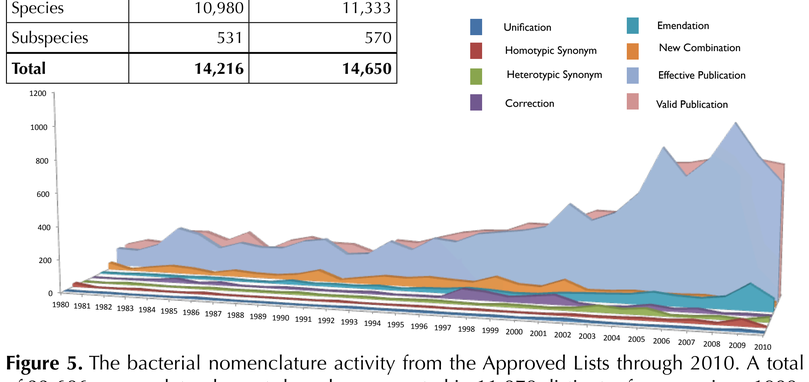

Figure 5. The bacterial nomenclature activity from the Approved Lists through 2010. A total of 33,606 nomenclatural events have been reported in 11,870 distinct references since 1980.

Figure 6a (below left). Cataloging of taxonomic information. GenBank began cataloging taxonomic information found in sequence metadata in 1993. By 2004, the number of Not Validly Published (NVP) names appearing on GenBank records exceeded the number of validly published names and now outnumber Validly Published (VP) names (1.4:1). This adds significantly to the burden of those attempting to use this information for tracking and identification, since it adds an additional 6.8 names/day, none of which have any formal standing in the literature. This divergence is increasing every year at a linear rate.

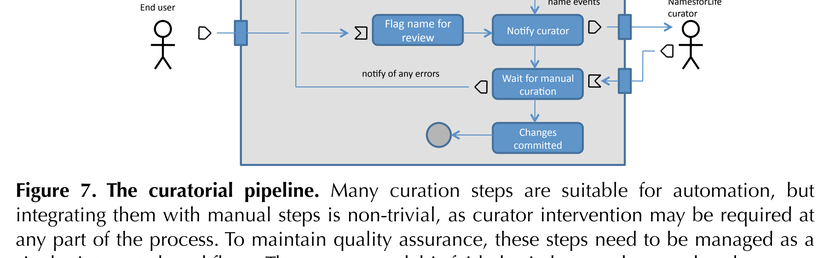

Figure 6b (above right). Cumulative growth in the number of validly published names since 1980. At the beginning of 1Q 2011, there were 13,833 validly published names (subspecies through classes) and 1,909 synonyms for which there were publicly available type strains. The rate at which new validly published names appear in the literature plateaued in 2006, at around 800 names per year (excluding taxonomic rearrangements and emendations). The linear increase in the rate of divergence translates to an exponential increase in total names.

Figure 7. The curatorial pipeline. Many curation steps are suitable for automation, but integrating them with manual steps is non-trivial, as curator intervention may be required at any part of the process. To maintain quality assurance, these steps need to be managed as a single, integrated workflow. The current model is fairly basic but can be tuned to the nature of the data to improve individual automated steps and support adjustment of the process as needed. The goal is to minimize the number of un-handled process exceptions that require curator assistance, while maintaining the overall quality of the curation process.

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*** (Figure 8b), 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

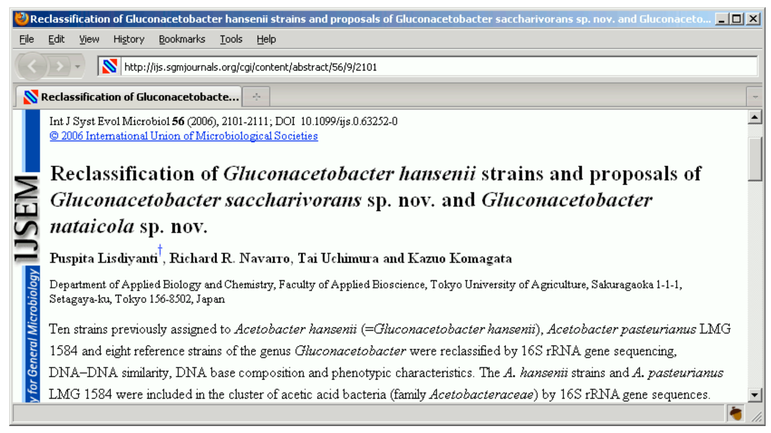

Figure 8a (above). An sample article prior to being semantically enabled by the ***N4L Guide***.

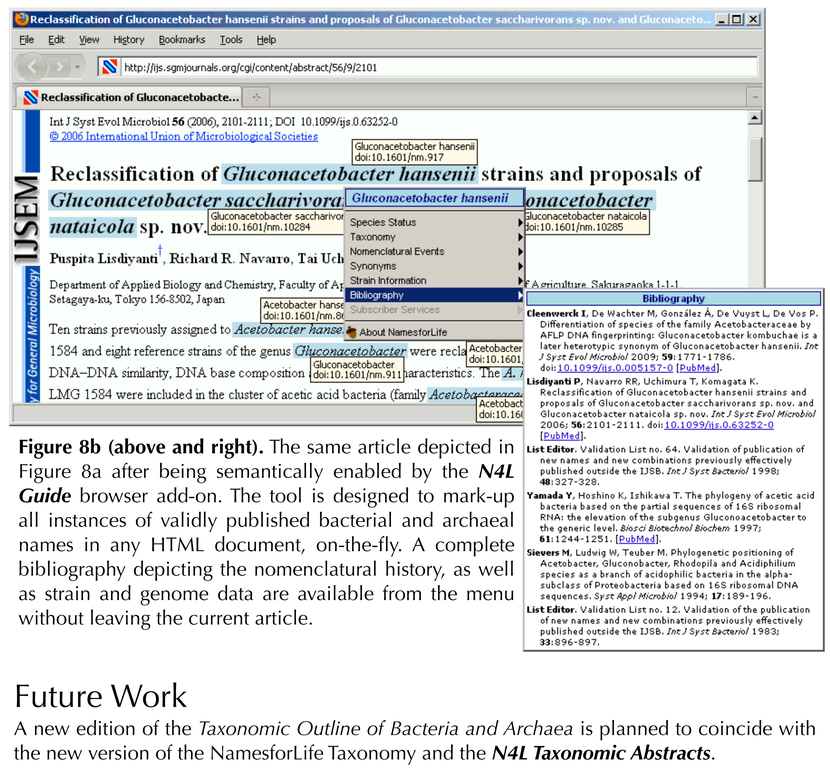

Figure 8b (above and right). The same article depicted in Figure 8a after being semantically enabled by the ***N4L Guide*** browser add-on. The tool is designed to mark-up all instances of validly published bacterial and archaeal names in any HTML document, on-the-fly. A complete bibliography depicting the nomenclatural history, as well as strain and genome data are available from the menu without leaving the current article.

Future 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***.

The development of the ***N4L Contextual Index*** is nearing completion. The first implementation of this system creates a semantic path from bacterial nomenclature into US Patents and Patent Applications. It is available online at (free *NamesforLife* account required):

http://services.namesforlife.com/patent/search

Additional Contextual Index services will become available throughout the year, including the PubMed Central Open Archives, USPTO and WIPO patents.

The ***N4L Taxonomic Abstracts*** are currently in development, and are scheduled for release in Q1 2011. 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*.

An interactive taxonomically-aware web application is currently under development in cooperation with the Department of Energy Joint Genome Institute (DOE JGI). The current prototype is shown below in Figure 9, and is freely available at:

http://microbial-earth.namesforlife.com

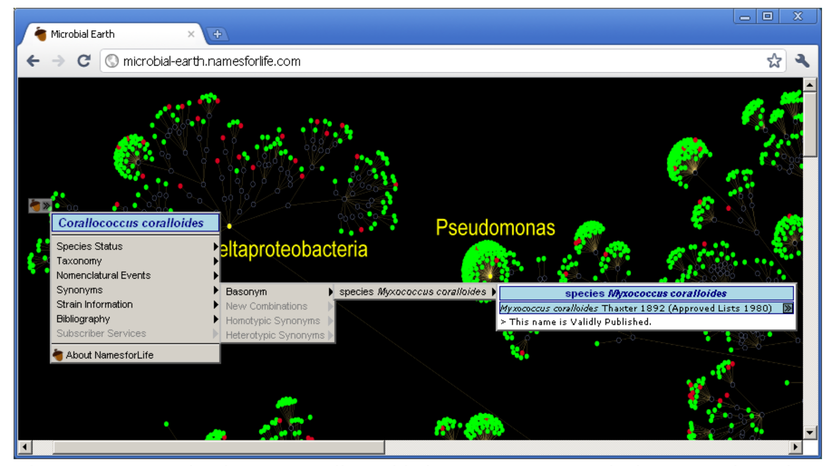

Figure 9. An example of a semantically enabled vector image using the browser-independent server version of ***N4L Guide*** in *Chrome 8*. The tree is from the forthcoming Microbial Earth project of Kyrpides et al.

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.