NamesforLife Semantic Resolution Services for the Life Sciences

Charles T. Parker[1], Dorothea K. Taylor[1], Kara Mannor[1], Sarah Wigley[1], Nicole Osier[1], Catherine Lyons[1], George M. Garrity[1, 2]*
[1]NamesforLife, LLC, East Lansing, MI and [2]Michigan State University, East Lansing, MI

NamesforLife
Bringing meaning to life ...

Abstract

A major challenge in bioinformatics, life sciences, and medicine is using correct and informative names. While this sounds simple enough, many different naming conventions exist in the life sciences and medicine that may be either complementary or competitive with other naming conventions. For a variety of reasons, proper names are not always used, leading to an accumulated semantic ambiguity that readers of the literature and end users of databases are left to resolve on their own. This ambiguity is a growing problem and the biocuration community is aware of its consequences.

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 (Figure 5). The system is professionally curated and represents a Tier III resource in Parkhill's view of bioinformatic services (Parkhill et al. Genome Biology 2010, 11:402, Figure 6) . 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 7-8).

The N4L Model

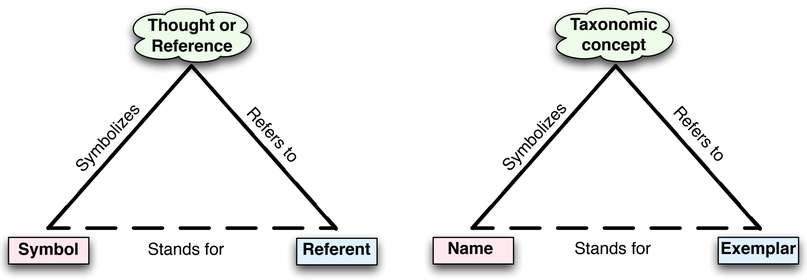


Figure 1- The semiotic triangle (left) and its application to biological nomenclature (right). Ogden and Richards (ref) show that uncertainty arises from a failure to recognize that names (symbols) that are assigned to objects (referents) have meaning to the agent that interprets them that may differ from the meaning intended by the agent that transmits them (9, 10). With some adaptation, this model is applicable to biological nomenclature and address 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 (4).

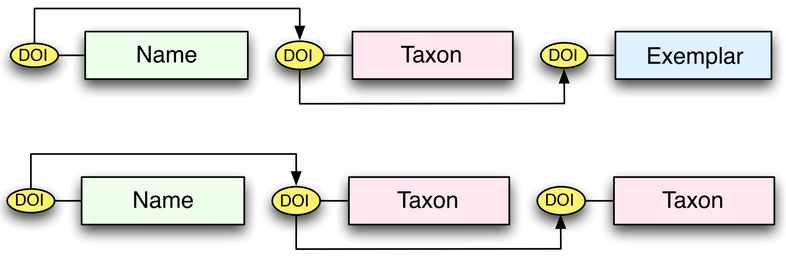


Figure 2. The NamesforLife semantic model. Clarity of meaning requires precise and unambiguous definitions that can be directly referenced, on demand. Biological names can be ambiguous because they are are subject to change, are not guaranteed to be unique, and may exist in one-to-one, one-to-many, many-to-one 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 (upper figure) and type concepts (species, genera and orders) for higher taxa (lower figure) that are part of published circumscriptions.

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. Each such object is identified with a Digital Object Identifier (DOI®) which enables placement of forward-pointing links in the published literature and in databases and provides a mechanism for resolving ambiguities, thereby "future proofing" a nomenclature. 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.

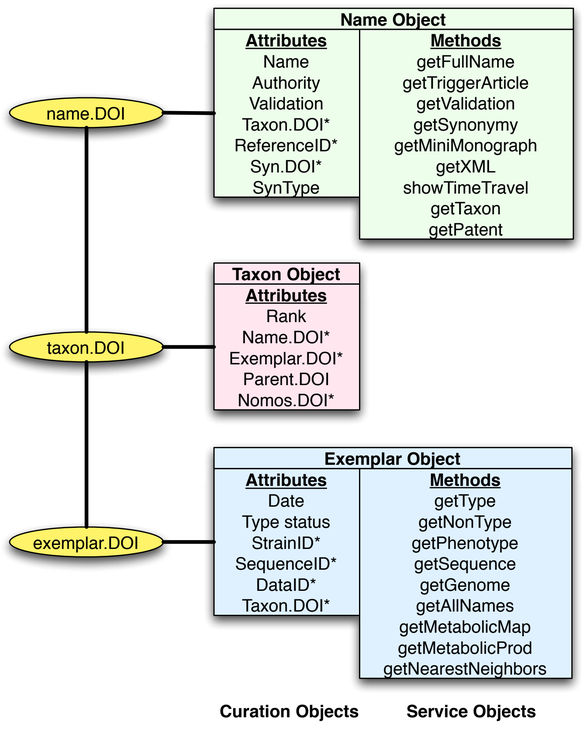


Figure 3 - The N4L data model. A single N4L-DOI provides direct access to information about a name, a taxonomic concept, the organism (at the species or subspecies level) and associated web services. The N4L service provides end-users of biological nomenclatures or other dynamic terminologies with the 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.

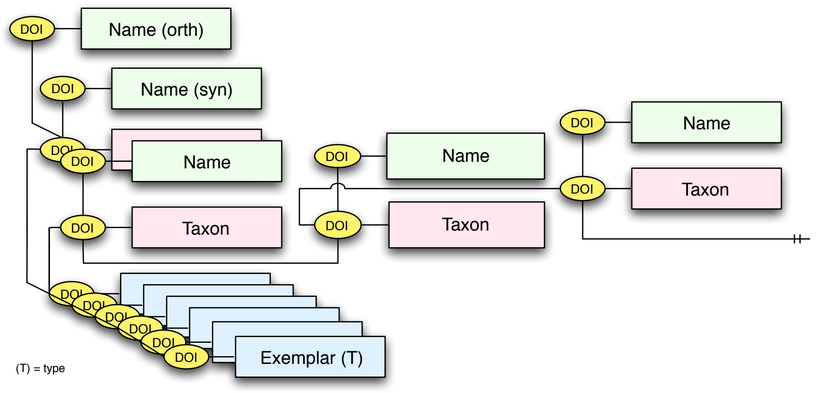


Figure 4. Assembly of N4L objects into a taxonomic hierarchy. In the N4L model, names, taxa, and exemplar objects are carefully mapped to provide an accurate representation of the precise meaning of a name at a given point in time. DOIs allow the information associated with these objects to be directly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

Database Statistics

At present, the *NamesforLife* Database (N4LDB) contains 14,331 distinct names, 13,601 of which are validly published, 119 *Candidatus*, and 47 that are illegitimate but relevant to the field. N4LDB also contains 13,944 exemplars (metadata representations of species/subspecies/strains), 9,171 of which represent distinct type strains for 10,698 taxa and 11,081 names. The remaining 3,250 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, etc) are not shown here.

N4LDB Records by Rank

Rank	Taxa	Names
Domain	2	2
Phylum	35	36
Class	75	76
Subclass	6	6
Order	132	136
Suborder	23	24
Family	341	346
Subfamily	1	1
Genus	2,038	2,068
Subgenus	5	5
Species	10,710	11,062
Subspecies	530	569
Total	**13,898**	**14,331**

Nomenclatural Events Recorded in N4LDB

Event	Count
Corrections	438
New Combinations	1,262
Heterotypic Synonyms	321
Homotypic Synonyms	163
Unifications	102
Automatically created names via rule 40d	53
Emendations	1,125
Validation List events	2,775
Valid Publication (excluding Validation Lists)	8,754

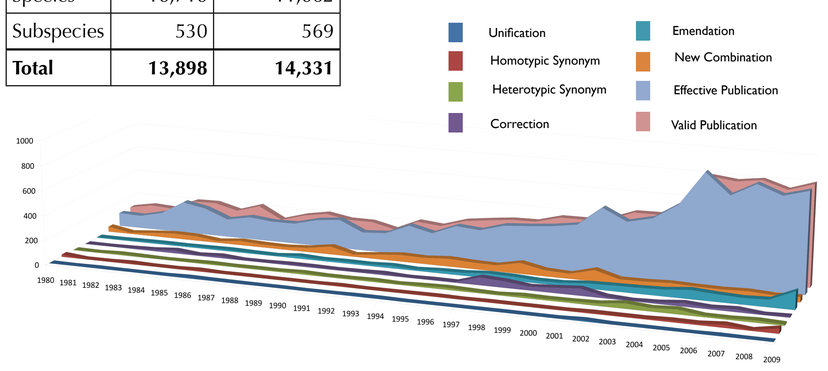


Figure 5. The bacterial nomenclature activity from the Approved Lists through 2010. A total of 30,534 nomenclatural events have been reported in 11,624 distinct references.

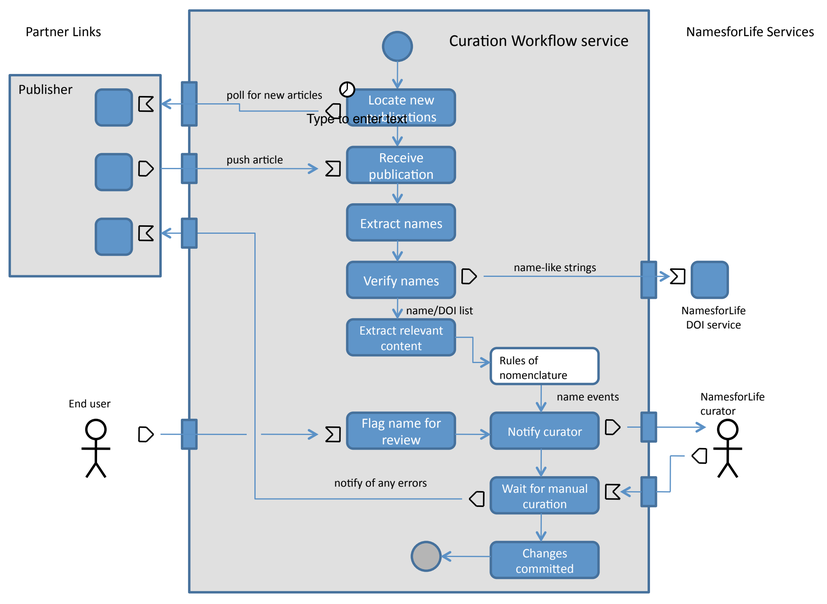


Figure 6. The curatorial pipeline - Many curation steps are suitable for automation, but integrating them with manual steps is non-trivial as curator intervention may be required at any part for the process. To maintain quality assurance, these steps need to be managed as a single, integrated workflow. 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 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.

The N4L Guide FireFox addon is available at no charge and can be downloaded at:
http://services.namesforlife.com

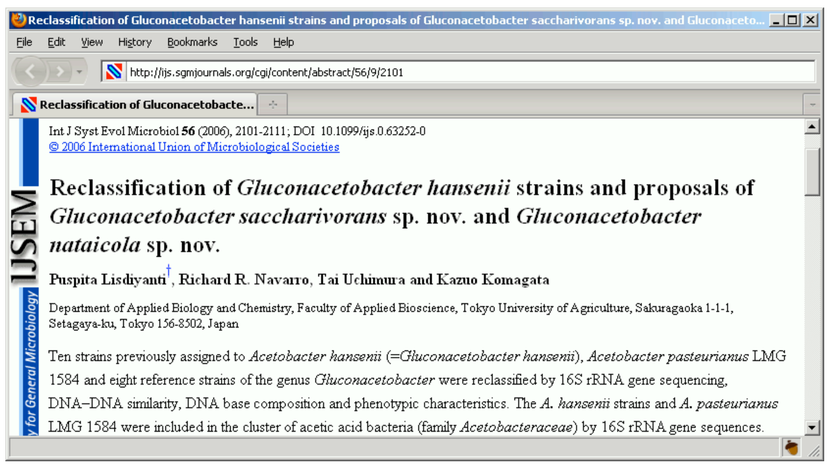
Semantic Services

The *NamesforLife* curators have recently completed linking the ongoing genome sequencing projects of *Bacteria* and *Archaea* listed in the Genomes Online Database, including all available non-type genomes. The N4L Nomenclature Database will continue to be updated as nomenclatural events occur are reported in the literature.

The ***N4LGuide*** Firefox 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 a Standards in Genomic Sciences article. Instructions for installing and using this tool can be found at the *NamesforLife* services website, located at https://services.namesforlife.com.

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 http://services.namesforlife.com/search/patents. Additional Contextual Index services will become available throughout the year.

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 nomenclature via CrossRef.



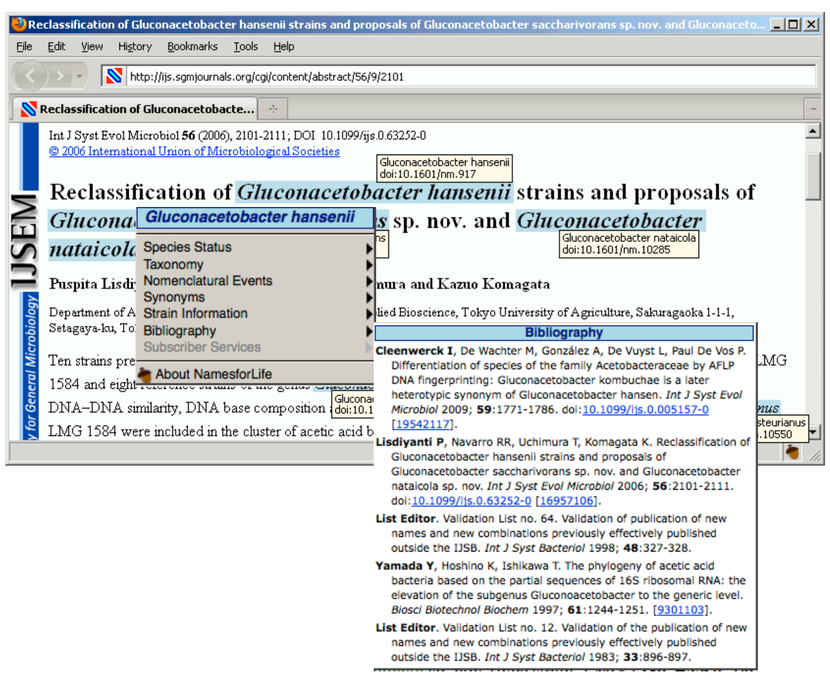


Figure 7. An example of an article, before and after being semantically enabled by the ***N4L Guide***. The tool is designed to mark-up all instances of validly published bacterial and archaeal names in any HTML document, on-the-fly.

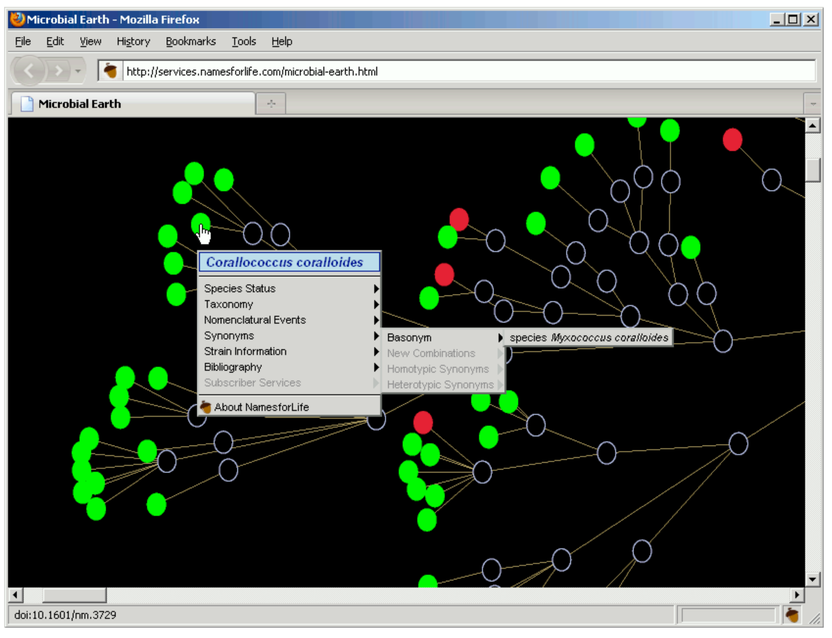


Figure 8. An example of a semantically enabled vector image using the ***N4L Guide***. This image is from a tree that is from the forthcoming Microbial Earth project of Kyrpides et al.

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.