NamesforLife Semantic Resolution Services for the Life Sciences

Charles T. Parker[1], Sarah Wigley[1], Nicole Osier[1], Jordan Fish[2], Qiong Wang[2], Donna McGarrell[2],
James R. Cole[1,2], Catherine Lyons[1], George M. Garrity[1,2]*(garrity@namesforlife.com)
[1]NamesforLife, LLC, East Lansing, MI and [2]Michigan State University, East Lansing, MI

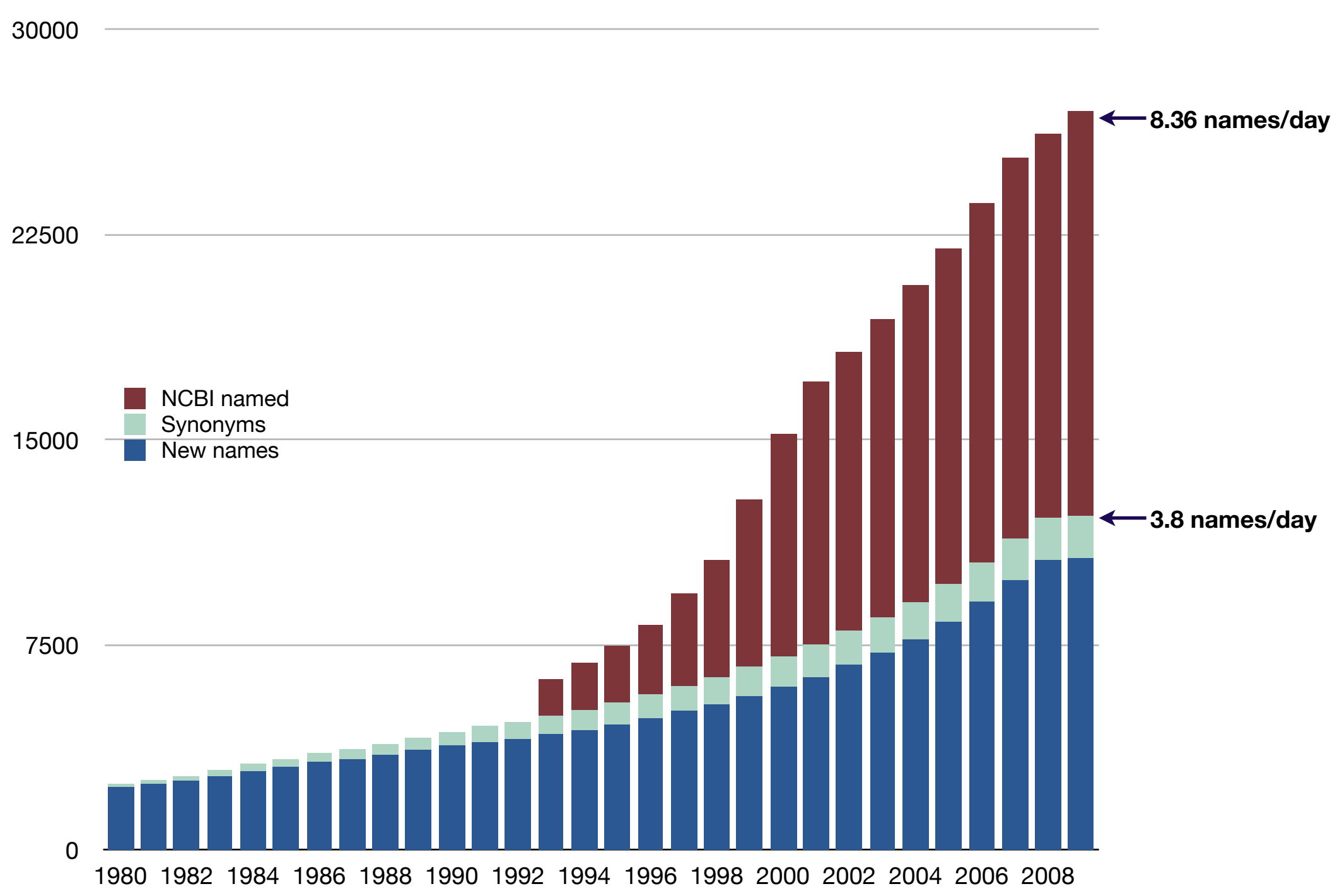
Abstract

Within the Genomes-to-Life Roadmap, the DOE states that a significant barrier to effective communication in the life sciences is a lack of standardized semantics that accurately describe data objects and persistently express knowledge change over time. As research methods and biological concepts evolve, certainty about correct interpretation of prior data and published results decreases because both become overloaded with synonymous and polysemous terms. Ambiguity in rapidly evolving terminology is a common and chronic problem in science and technology. N4L is a novel technology designed to solve this problem.

Background

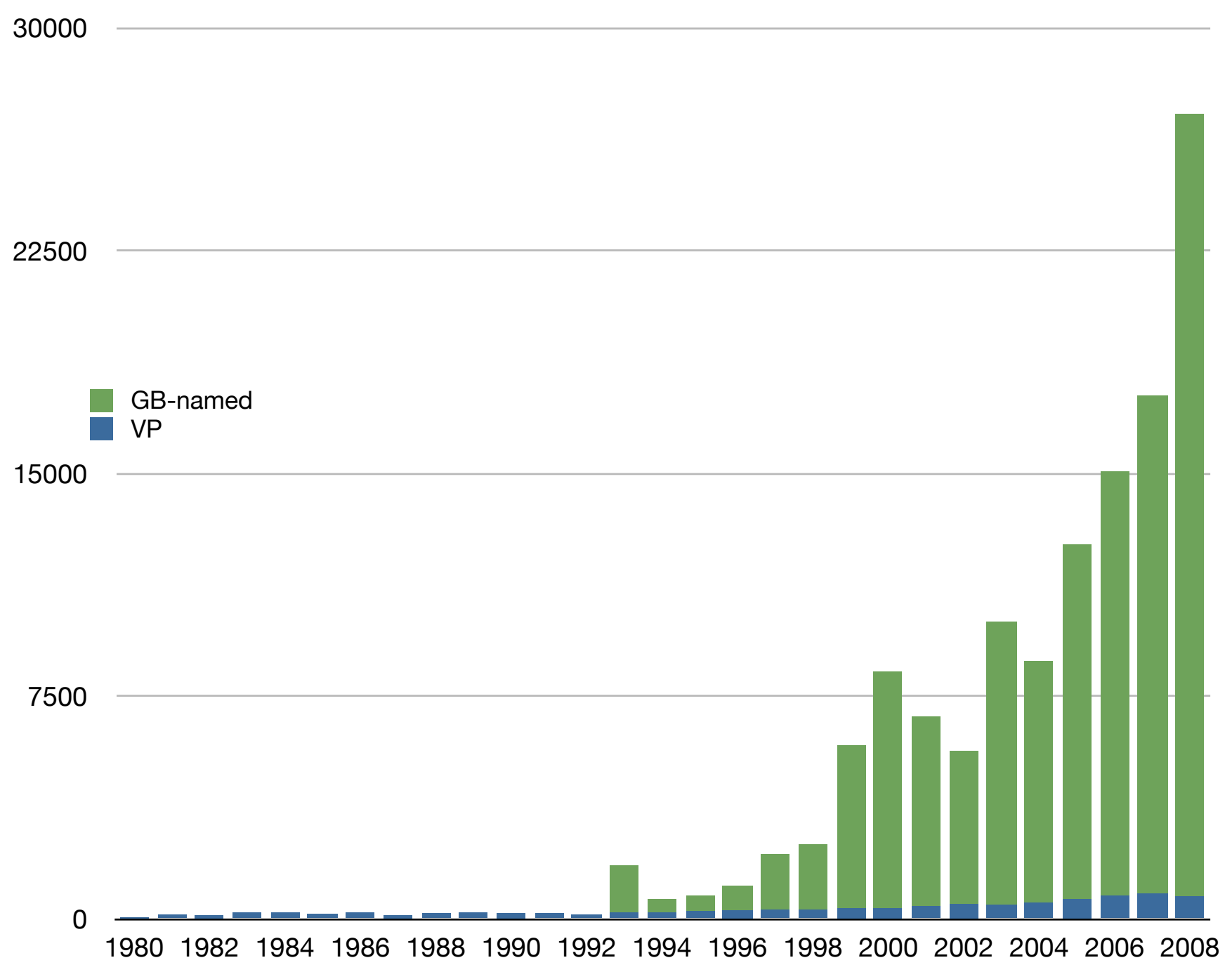
Understanding the correct meaning of a biological name, in the appropriate context, is important. It is not, however, a trivial task, and the number of individuals with expertise in biological nomenclature is limited. Such knowledge can, however, be accurately modeled and delivered through a networked semantic resolution service that provides end-users of biological nomenclatures or other dynamic terminologies with the appropriate information, in proper context, on demand. Such a service can also be used by database owners, publishers, or other information providers to semantically enable their offerings, making them more readily discoverable, even when the definition of a name or term has changed.

The figure below shows the cumulative growth in the number of validly published names since 1980. At the beginning of 1Q 2009, there were 11,940 *validly published names* (subspecies through classes) and 1453 synonyms for which there were publicly available type strains. The rate at which new validly published names appear in the literature plateaued in 2008, at 3.8 names/day (including taxonomic rearrangements and emendations). GenBank began cataloging taxonomic information for sequences derived from *Bacteria* and *Archaea* in 1993, and introduced the practice of tagging data with trivial or invalid names. By 2000, the number of invalid names associated with GenBank records exceeded the number of validly published names. At present, invalidly published names outnumber validly puiblished named (1.2:1). For both tracking and identification purposes, this adds significantly to the burden of those who must rely on this information for various purposes by adding an additional 4.5 names/day, none of which are governed by any rules or principles for assignment, usage, or uniqueness.



A looming problem

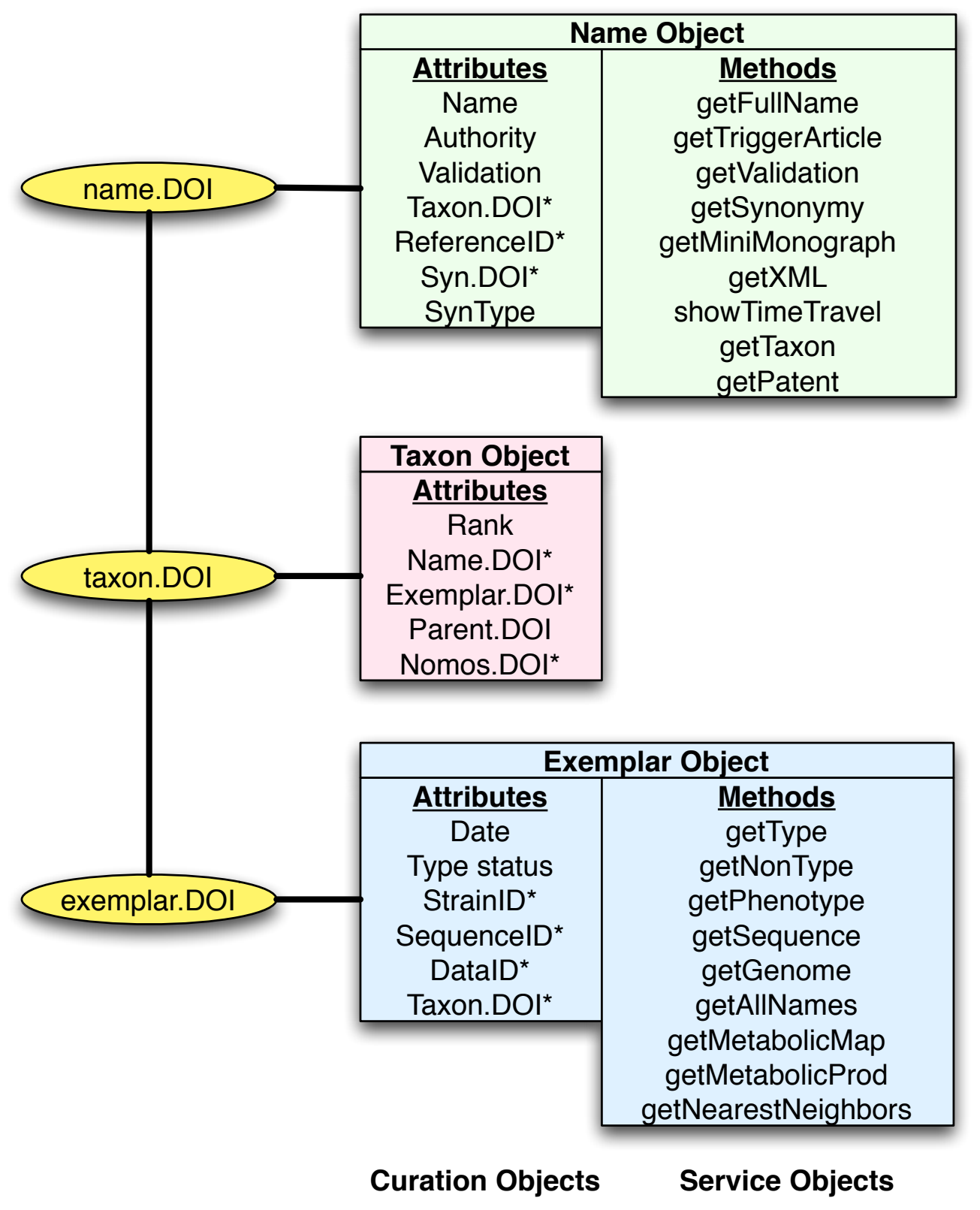
The adoption of DNA sequencing as the preferred method of rapidly characterizing *Bacteria* and *Archaea* has tremendously accelerated during the past five years, with the expected consequences. At present, the rate at which "named" sequences are added to the GenBank taxonomy exceeds the rate at which validly published names appear in the taxonomic record by a factor of approximately 35. This confounds the retrieval of related information from various databases and the scientific, technical and medical literature as many of these invalidly named species can not be readily tracked over time, nor can relationships be inferred to those species for which at least one genome sequence is available. This disconnect between the knowledge contained in the literature and the accumulated genomic data is likely to grow as faster and cheaper sequencing methods come into the market place.



N4L technology

Prokaryotic nomenclature represents an important opportunity to learn how we might best solve the problem of persistently expressing knowledge about biological systems that are described and defined by dynamic terminologies or names. It is a rich and complex vocabulary with numerous event types, yet is also tractable in size, is rules-based, and is carefully maintained. It is also a problem of importance to the Genomes-to-Life program because of the significant investment by the DOE in understanding *Bacteria* and *Archaea* at the systems level.

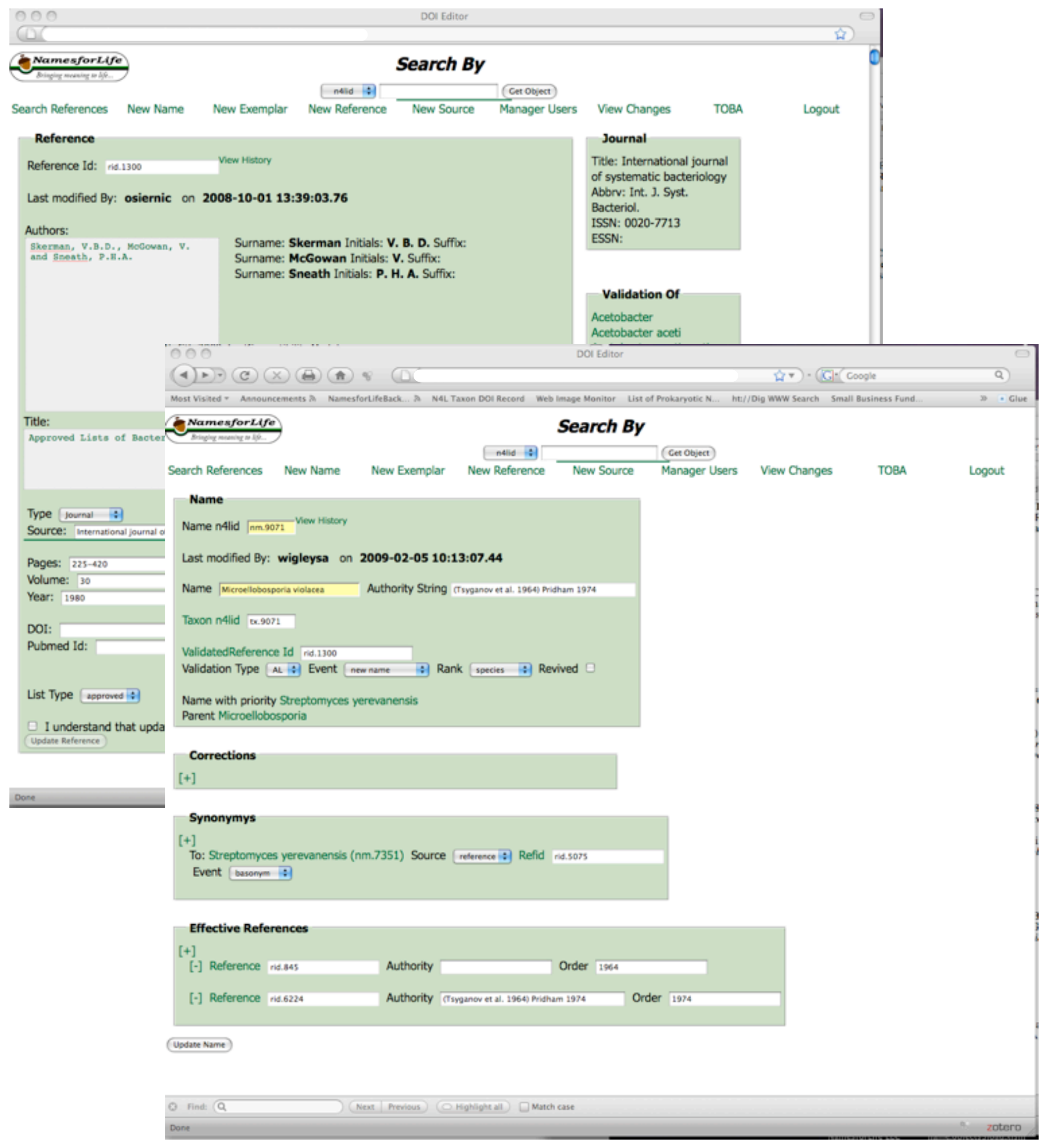
The core of N4L consists of a data model, an XML schema, and an expertly managed vocabulary coupled with Digital Object Identifiers (DOI®; a class of persistent identifiers based on the Handle System®) to form a transparent semantic resolution service that disambiguates terminologies, makes them actionable, and presents them to end-users in the correct temporal context. 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 is intended to provide 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.



Curation tools

Web-based tools have been developed to facilitate data entry and retrieval by NamesforLife curators. Some fields, such as the authority string are automatically populated based on the published rules of prokaryotic nomenclature (ICPN). There are several ways to query for information, including by name, and by accession number, and there is a citation matcher to find references currently in the database.

The curatorial tools connect to the database through a set of data validators to help curtail certain data entry errors such as ensuring exemplars are only connected to species/subspecies entities and correctly modeling various classes of taxonomic anomalies such as orthographic corrections, judicial opinions and automatically created coordinate subspecies under Rule 40d. As of February 1, 2009, the database held records on 11,407 named bacterial and archaeal taxa (8732 species, 491 subspecies, 2184 higher taxa), along with 7747 records identifying verified specimen holdings in biological resource centers and links to 7365 references in which the related taxonomic and nomenclatural acts were effectively and validly published.



Reference

Reference Id: rid.1300

Last modified By: **osiernic** on **2008-10-01 13:39:03.76**

Authors: Skerman, V.B.D., McGowan, V. and Sneath, P.H.A.

Surname: **Skerman** Initials: **V. B. D.** Suffix:
Surname: **McGowan** Initials: **V.** Suffix:
Surname: **Sneath** Initials: **P. H. A.** Suffix:

Journal

Title: International journal of systematic bacteriology
Abbrv: Int. J. Syst. Bacteriol.
ISSN: 0020-7713
ESSN:

Validation Of

Acetobacter
Acetobacter aceti

Title: Approved Lists of Bacter

Type: Journal
Source: International journal o

Pages: 225-420
Volume: 30
Year: 1980
DOI:
Pubmed Id:

List Type: approved

□ I understand that upda

Name

Name n4lid: nm.9071

Last modified By: **wigleysa** on **2009-02-05 10:13:07.44**

Name: Microellobosporia violacea   Authority String: (Tsyganov et al. 1964) Pridham 1974

Taxon n4lid: tx.9071

ValidatedReference Id: rid.1300

Validation Type: AL   Event: new name   Rank: species   Revived □

Name with priority Streptomyces yerevanensis
Parent Microellobosporia

Corrections

[+]

Synonymys

[+]
To: Streptomyces yerevanensis (nm.7351) Source: reference Refid: rid.5075
Event: basonym

Effective References

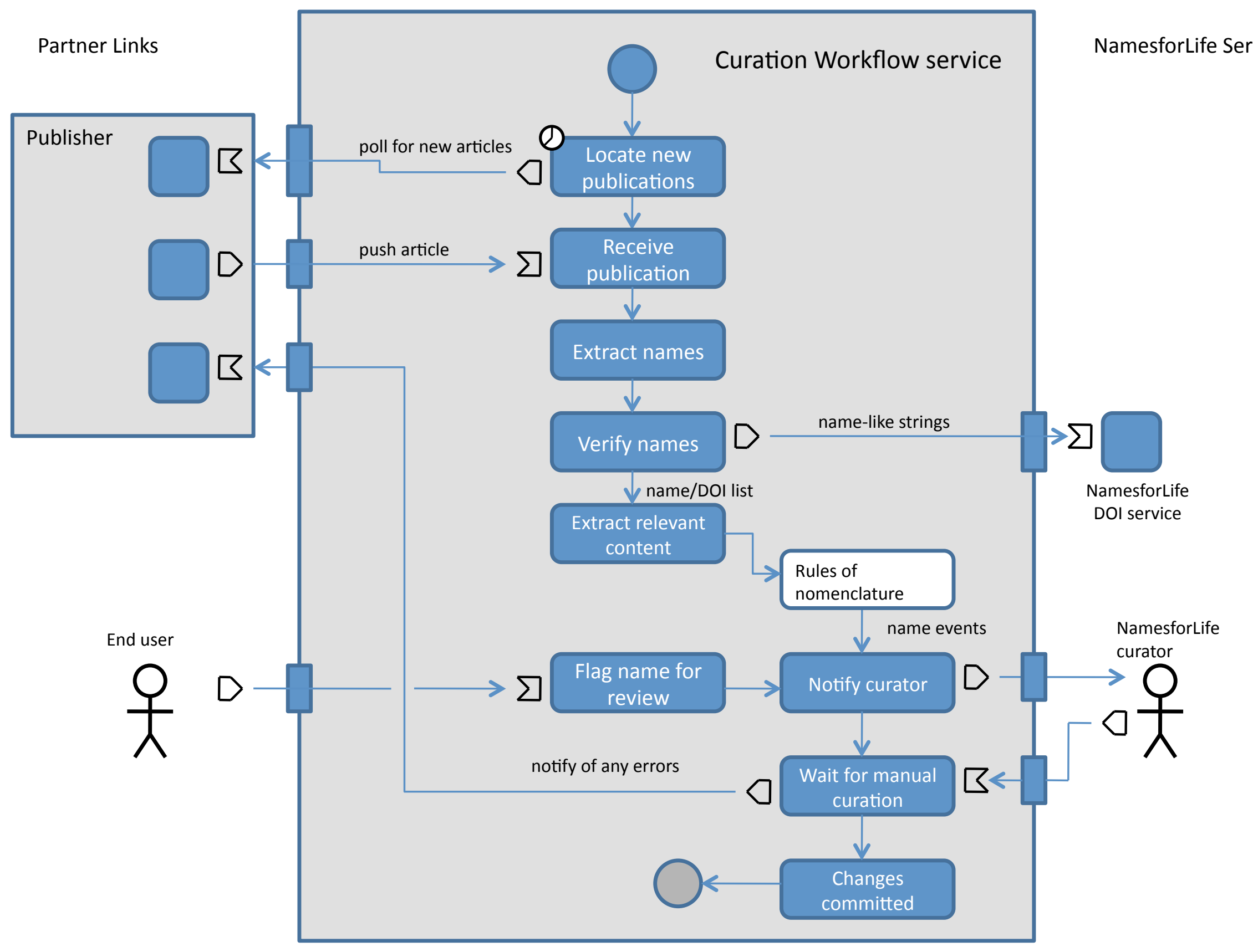
[+]
[-] Reference: rid.845   Authority: Order: 1964
[-] Reference: rid.6224   Authority: (Tsyganov et al. 1964) Pridham 1974   Order: 1974

Current efforts

Many curation steps are suitable for automation, but integrating them with manual steps is non-trivial, as curator assistance could be required at any part for the process. If these deviations from normal procedure are not well documented, they could call the quality of the data into question.

In order to maintain quality assurance, these steps need to be managed as a single, integrated workflow. One approach that we are taking to solve this problem is to break down the data curation process into tasks that are suitable for modeling in Business Process Execution Language (BPEL). The goal is to have a self-documenting workflow that is adhered to by both the automated tasks and the curators.

When a name event occurs, the business process is triggered and flows through the automated steps until an exception is raised, indicating that curator assistance is needed. This could potentially show up as a task in a curation tool, or on a list of names that require attention. The manual step is then performed (shown as a message being passed from the curator back into the service; or if preferred, the curator is modeled as a service with the curation tool as his/her API) and the process continues from that point. Using a BPEL engine, it is possible to replay the entire series of events that led to a name requiring curator discretion. The model can start out in a fairly basic form that is gradually tuned to the nature of the data by improving individual automated steps and adjusting the process as needed. The goal is to minimize the number of unhandled process exceptions that require curator assistance, while maintaining the overall quality of the curation process.

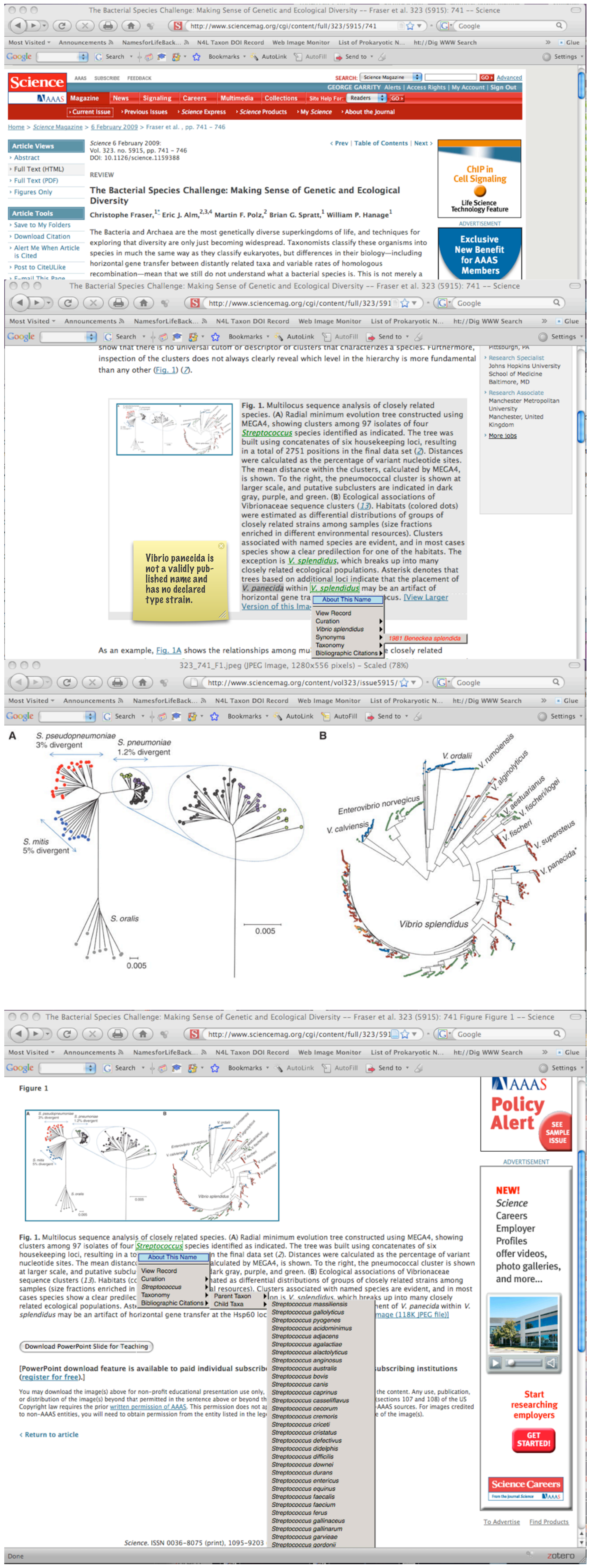


The target audience of N4L services is the broad scientific community and others who may need to know the precise meaning of biological names or other terms, in correct temporal context as they are encountered in other digital content (scientific or technical literature, regulatory literature, databases, etc). The dynamic, yet asynchronous nature of biological nomenclature and similar terminology poses a significant burden on information providers, as they must either invest in constantly maintaining their offerings to keep current or shift that burden to their end-users. If the former, the costs can be significant, and, in the absence of a means to synchronize updates across an entire domain of knowledge, end users are still confronted with apparent discrepancies across data sources and content providers. If the burden is shifted to end-users, they must then locate alternative information sources, typically hosted through a web portal, that must be queried separately. This makes utilization of content cumbersome and can lead to considerable ambiguity.

The NamesforLife approach is to semantically enable content in a manner that is transparent to end-users at two points in the value chain: at the source (the data provider or publisher) and at the client side (the end-user). In either case, the end-user experience is the same. At each occurrence of a validly published bacterial or archaeal name, they can have access to precise authoritative information by simply clicking on the name. Tools to enable publishers' content at the pre-publishing stage that embed persistent N4L identifiers in inline text ensures that their readers will always have access to the correct meaning of the name (as well as additional information), even if the name has changed since publication. Our web-based client supports semantic enablement of other digital content, one-the-fly, providing similar seamless access to NamesforLife content at each point where a validly published name occurs. This provides the reader with *direct access* to a wealth of information to aid in the interpretation of each enabled article as is shown in the figures to the right.

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, R. Dunford and R. Fraser (SGM, Reading) for their assistance in making IJSEM content available for this study and their time, energy and willingness to make this project a reality. We would also like to thank members of the International Commission 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 II STTR Award DE-FG02-07ER86321 A001.



The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity -- Fraser et al. 323 (5915): 741 -- Science

http://www.sciencemag.org/cgi/content/full/323/5915/741

GEORGE GARRITY Alerts | Access Rights | My Account | Sign Out

Home > Science Magazine > 6 February 2009 > Fraser et al. , pp. 741 - 746

Science 6 February 2009:
Vol. 323. no. 5915, pp. 741 - 746
DOI: 10.1126/science.1159388

REVIEW

The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity

Christophe Fraser,[1*] Eric J. Alm,[2,3,4] Martin F. Polz,[2] Brian G. Spratt,[1] William P. Hanage[1]

The Bacteria and Archaea are the most genetically diverse superkingdoms of life, and techniques for exploring that diversity are only just becoming widespread. Taxonomists classify these organisms into species in much the same way as they classify eukaryotes, but differences in their biology—including horizontal gene transfer between distantly related taxa and variable rates of homologous recombination—mean that we still do not understand what a bacterial species is. This is not merely a

show that there is no universal cutoff or descriptor of clusters that characterizes a species. Furthermore, inspection of the clusters does not always clearly reveal which level in the hierarchy is more fundamental than any other (Fig. 1) (7).

Fig. 1. Multilocus sequence analysis of closely related species. (**A**) Radial minimum evolution tree constructed using MEGA4, showing clusters among 97 isolates of four *Streptococcus* species identified as indicated. The tree was built using concatenates of six housekeeping loci, resulting in a total of 2751 positions in the final data set (2). Distances were calculated as the percentage of variant nucleotide sites. The mean distance within the clusters, calculated by MEGA4, is shown. To the right, the pneumococcal cluster is shown at larger scale, and putative subclusters are indicated in dark gray, purple, and green. (**B**) Ecological associations of Vibrionaceae sequence clusters (13). Habitats (colored dots) were estimated as differential distributions of groups of closely related strains among samples (size fractions enriched in different environmental resources). Clusters associated with named species are evident, and in most cases species show a clear predilection for one of the habitats. The exception is *V. splendidus*, which breaks up into many closely related ecological populations. Asterisk denotes that trees based on additional loci indicate that the placement of *V. panecida* within *V. splendidus* may be an artifact of horizontal gene tra[...]cus. [View Larger Version of this Ima

Vibrio panecida is not a validly published name and has no declared type strain.

As an example, Fig. 1A shows the relationships among mu[...]e closely related

Pittsburgh, PA
Research Specialist
Johns Hopkins University School of Medicine
Baltimore, MD
Research Associate
Manchester Metropolitan University
Manchester, United Kingdom
More jobs

323_741_F1.jpeg (JPEG Image, 1280x556 pixels) - Scaled (78%)

http://www.sciencemag.org/content/vol323/issue5915/

The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity -- Fraser et al. 323 (5915): 741 Figure Figure 1 -- Science

Figure 1

Fig. 1. Multilocus sequence analysis of closely related species. (**A**) Radial minimum evolution tree constructed using MEGA4, showing clusters among 97 isolates of four *Streptococcus* species identified as indicated. The tree was built using concatenates of six housekeeping loci, resulting in a to[...]n the final data set (2). Distances were calculated as the percentage of variant nucleotide sites. The mean distanc[...]alculated by MEGA4, is shown. To the right, the pneumococcal cluster is shown at larger scale, and putative subclu[...]dark gray, purple, and green. (**B**) Ecological associations of Vibrionaceae sequence clusters (13). Habitats (cc[...]ated as differential distributions of groups of closely related strains among samples (size fractions enriched in[...]al resources). Clusters associated with named species are evident, and in most cases species show a clear predilec[...]on is *V. splendidus*, which breaks up into many closely related ecological populations. Aste[...]ent of *V. panecida* within *V. splendidus* may be an artifact of horizontal gene transfer at the Hsp60 loc[...]mage (118K JPEG file)]

Streptococcus massiliensis
Streptococcus gallolyticus
Streptococcus pyogenes
Streptococcus acidominimus
Streptococcus adjacens
Streptococcus agalactiae
Streptococcus alactolyticus
Streptococcus anginosus
Streptococcus australis
Streptococcus bovis
Streptococcus canis
Streptococcus caprinus
Streptococcus casseliflavus
Streptococcus cecorum
Streptococcus cremoris
Streptococcus criceti
Streptococcus cristatus
Streptococcus defectivus
Streptococcus didelphis
Streptococcus difficilis
Streptococcus downei
Streptococcus durans
Streptococcus entericus
Streptococcus equinus
Streptococcus faecalis
Streptococcus faecium
Streptococcus ferus
Streptococcus gallinaceus
Streptococcus gallinarum
Streptococcus garvieae
Streptococcus gordonii

Download PowerPoint Slide for Teaching

[PowerPoint download feature is available to paid individual subscrib[...]ubscribing institutions (register for free).]

You may download the image(s) above for non-profit educational presentation use only, [...]the content. Any use, publication, or distribution of the image(s) beyond that permitted in the sentence above or beyond th[...](sections 107 and 108) of the US Copyright law requires the prior written permission of AAAS. This permission does not ap[...]AAAS sources. For images credited to non-AAAS entities, you will need to obtain permission from the entity listed in the leg[...]e of the image(s).

‹ Return to article

Science. ISSN 0036-8075 (print), 1095-9203

Saturday, February 7, 2009