NamesforLife Semantic Resolution Services for the Life Sciences

George M. Garrity[1,2], Catherine M. Lyons[1], Charles T. Parker[1] and James R. Cole[1,2]
NamesforLife, LLC, East Lansing, MI and
Michigan State University, East Lansing, MI

Abstract

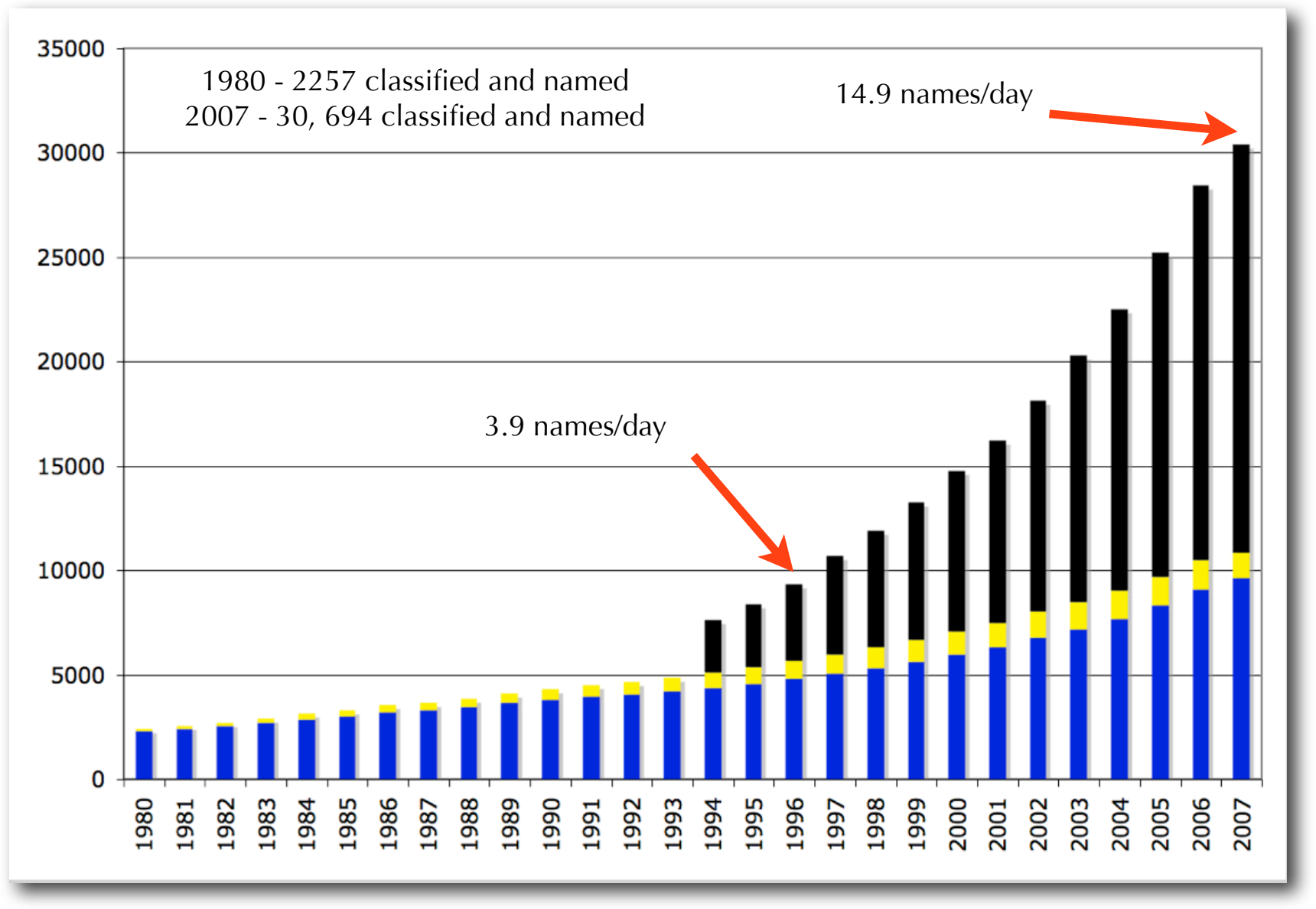
Within the Genomes-to-Life Roadmap, the DOE states that a significant barrier to effective communication in the life sciences is a lack of standardized semantics that accurately describe data objects and persistently express knowledge change over time. As research methods and biological concepts evolve, 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. NamesforLife (N4L) is a novel technology designed to solve this problem.

Background

With our rapid understanding of the depth and breadth of bacterial diversity, the list of names investigators must deal with not only grows, it also undergoes incremental redefinition on a daily basis. While these name changes are of considerable interest to a relatively small number of experts engrossed in bacterial classification, they present a significant problem to both end-users and information providers, who must invest a significant and increasing effort to map from new names to names in publications that predate any change. Failure to maintain name currency may prevent scientists and non-scientists alike from tracking important developments in their field and can trigger inappropriate or life-threatening responses.

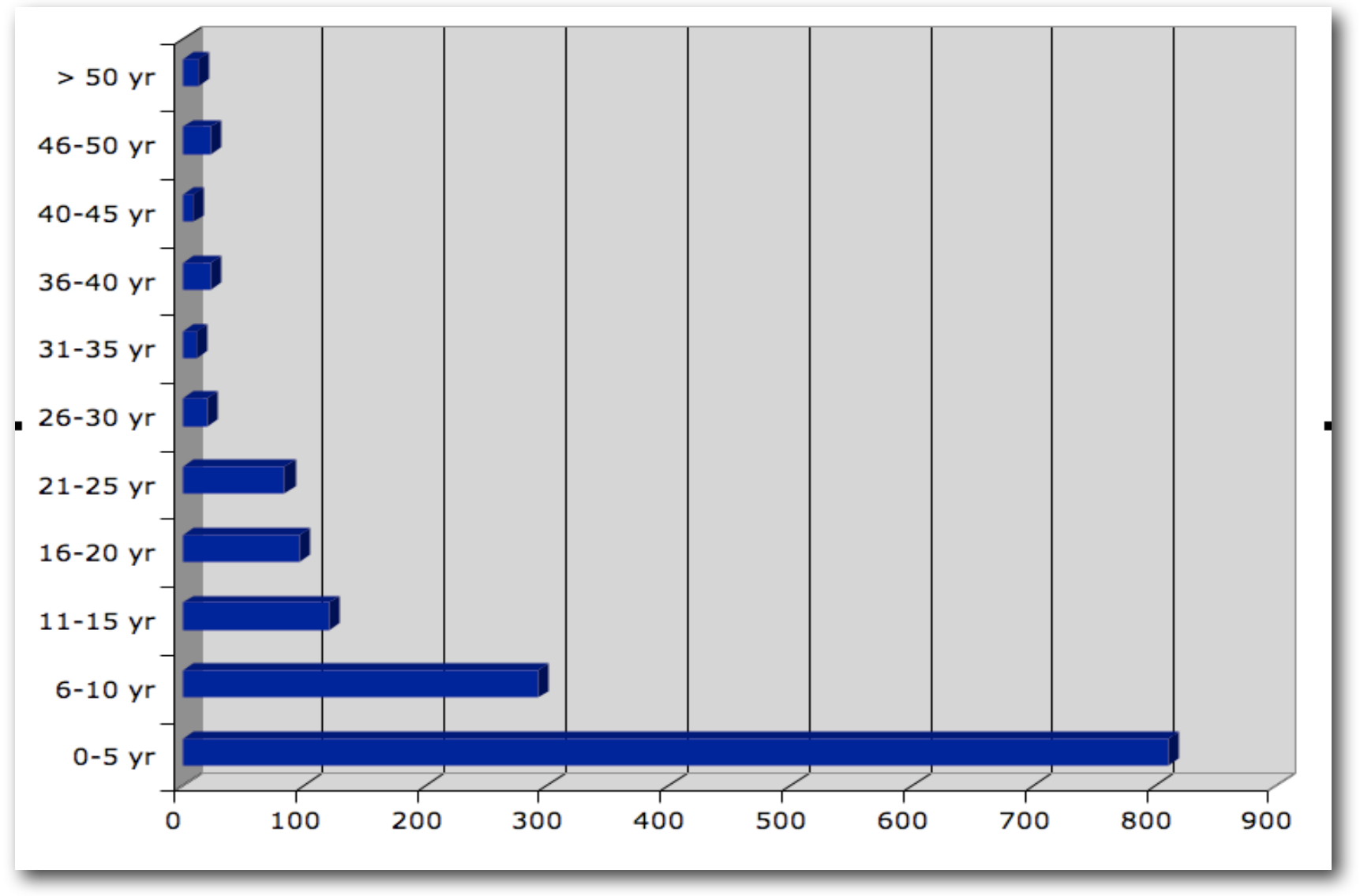
Biological names are the oldest and most highly evolved form of controlled vocabulary and follow precise rules with regard to formation and usage. Two concepts that are not widely understood, but highly useful are typification and priority. Before a name can receive "official" standing in the literature, it must formally be defined in a proposal in which the named taxonomic group to which the name applies is circumscribed and a type exemplar declared for diagnostic purpose. The author must also verify that no other name that applies to the same group has priority in the literature. Names in botany and zoology use preserved (non-living) type exemplars and date from May 1, 1753. Names of *Bacteria* and *Archaea* use viable types and date from January 1, 1980 (publication of the Approved Lists of Bacterial Names). Names of *Bacteria* and *Archaea* also differ from those of plants and animals in that they undergo a formal registration process and are not considered to be validly published until they appear in print in the International Journal of Systematic and Evolutionary Microbiology (IJSEM), or its predecessor, the International Journal of Systematic Bacteriology (IJSB).

The figure below shows the annual growth in the number of validly published names since 1980. At the beginning of 4Q 2007, there were 9662 validly published names (subspecies through classes) and 1495 synonyms for which there were type exemplars. The rate at which new validly published names appear in the literature has accelerated considerably in the recent past and currently stands at 3.9 names/day (including taxonomic rearrangements and emendations). This number is however significantly lower than the number of names that appear in INSDC records and GenBank taxonomy, that have no standing in the literature (14.9 names/day). Trivial names appearing on INSDC records add further confusion to the process and occur at a rate more than five-fold higher.



Changing names and meanings

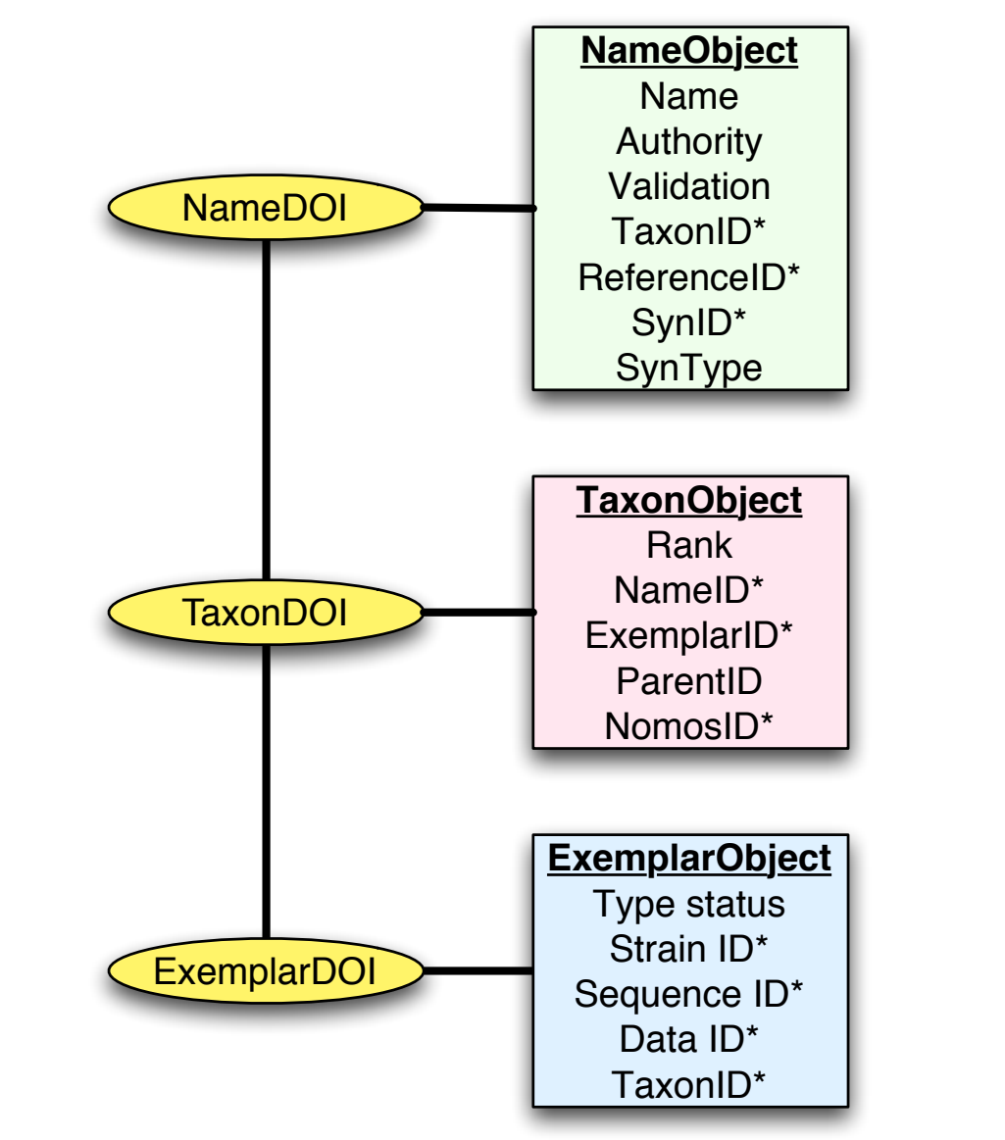
Once a name comes into usage, it begins to undergo a gradual shift in "meaning" as others practitioners apply the published definition to identify unknown isolates or sequences. While "lumping and splitting" are reasonably well understood, as these lead to changes in nomenclature, emendations are less well understood as these lead to a refinement of the circumscription and naming authority, but not a change in name. The figure below reveals that the rate at which explicit (published) emendations occur varies significantly. All taxa undergo frequent implicit emendations (especially higher taxa) as each taxonomic proposal results in upward and downward cascades in group membership, tied to the monthly publication of the IJSEM.



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, rules-based, and carefully maintained. It is also a problem if importance to the Genomes-to-Life program because of the significant investment by the DOE to understand *Bacteria* and *Archaea* at the systems level.

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

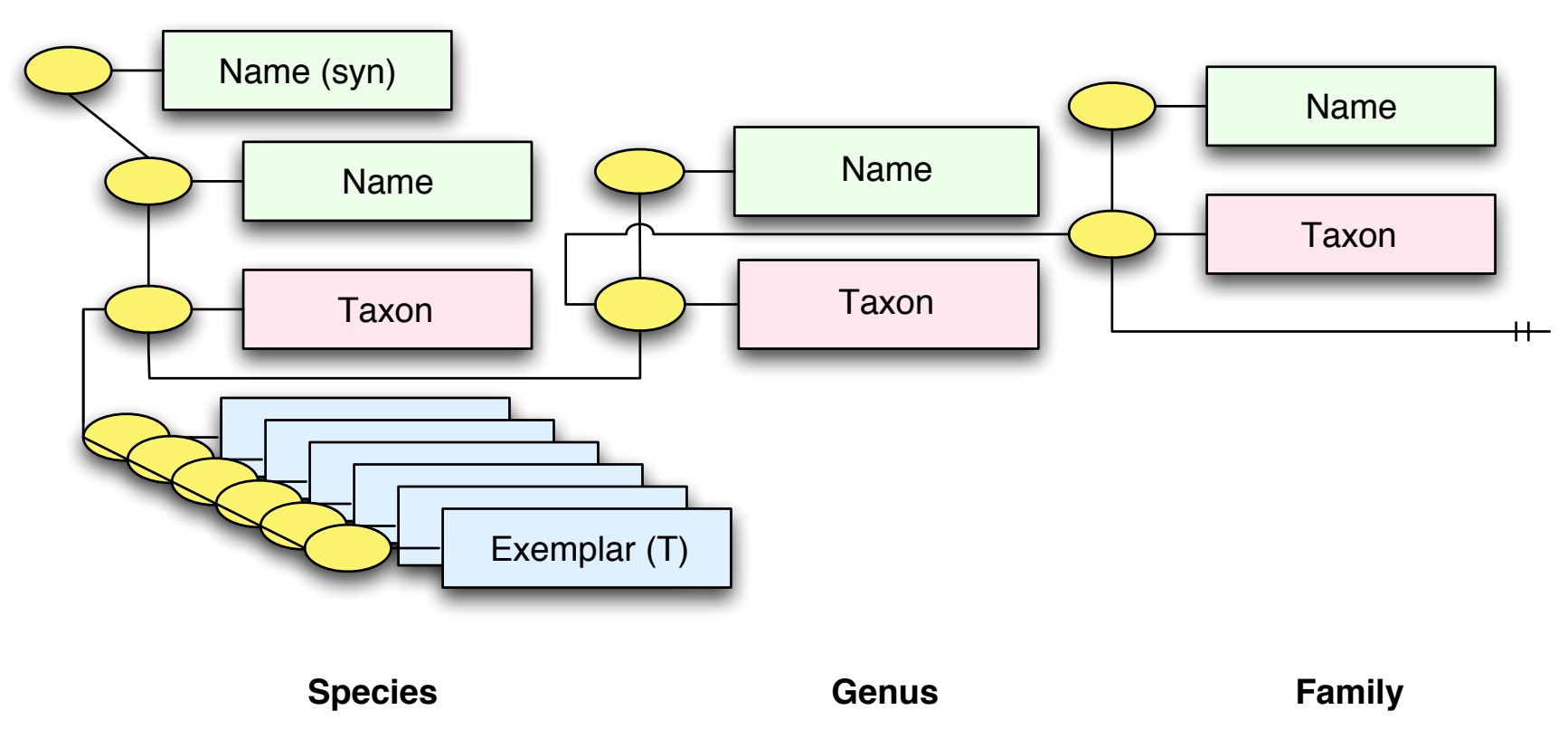


Data model

The figure above depicts the N4L data model. In its simplest form, it consists of three first-class objects (which we refer to as information objects) representing names, taxa, and exemplars. The objects are interlinked to each other by DOIs and each can be accessed directly by its given DOI. Name objects contain information that is specific to nomenclatural events, as defined by the International Code of Nomenclature Bacteria. In addition to the name string, each object contains information about the naming authority, date of effective and/or valid publication, name status, the reference in which the naming event(s) occurred, links to synonyms, and synonymy status, and a link to the taxon object to which the name is assigned.

Taxon objects serve as aggregating elements above the rank of genus and provide the path through a given hierarchy. Although the model is designed to support multiple taxonomic views, through the use of a separate first-class object called a Nomos, we have not yet implemented this feature. At the species and subspecies level a separate aggregating object holds information about strain deposits and corresponding sequence identifiers. The model is general and can accommodate any other type of information that exits for a given strain. Separate exemplars exist for type and non-type strains.

A simple view of the manner in which the objects link together in a taxonomic hierarchy is depicted below. This is an idealized view of a properly formed hierarchy, free from anomalies. When restricted to an error-free taxonomy of type material there are apparent one-to-one relationships between names and taxa, however, many-to-one relationships arise when orthographic corrections and variants are factored into the taxonomy. One-to-one, many-to-one, and one-to-many relationships occur between taxa and exemplars, and one-to-one and many-to-one relationships occur between names and type-exemplars. The addition of non-type exemplars adds many-to-many relationships at this level.



As proof of principle, a working model of the N4L technology has been built. It allowed us to validate our concepts and gain new insights into previously unaddressed complexities of dynamic vocabularies. In this project, we are reducing the working model to a service that can automatically annotate occurrences of names in the scientific literature and databases. The initial target is the *International Journal of Systematic and Evolutionary Microbiology*, the publication of record for all nomenclatural changes for *Bacteria* and *Archaea*. To accomplish this objective, we are addressing several technical problems, including transfer of the current data into a more suitable environment to simplify updating and generation of N4L information objects; development of tagging rules to embed links to N4L information objects into on-line content; enabling multiple resolution through the Handle server; development of mini-monographs as an improved human interface to N4L; and development of additional infrastructure to support on-the-fly translation of N4L tagged data in published content.

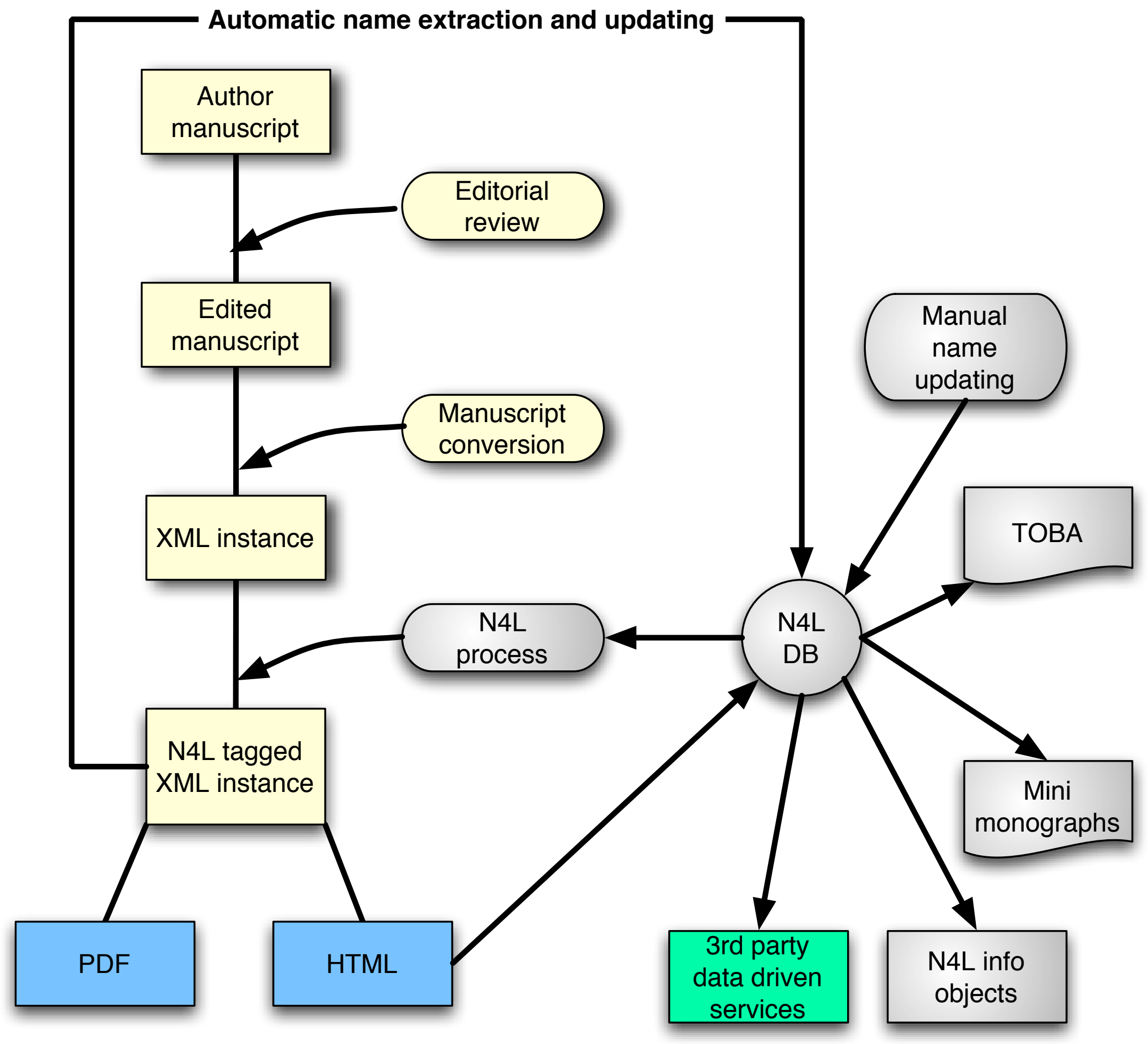
Current efforts

Our partner on this project is the Society for Microbiology (Reading, UK), the publisher of the IJSEM. The journal is a contemporary blended publication that is distributed in print and on the web. The SGM workflow is depicted below, in the yellow boxes. Author manuscripts are received in the form of MS Word document files through an online submission system. The manuscripts undergo a traditional peer and editorial review process. On acceptance, the Word files are brought into compliance with the journal styles sheet, references and other facts check and linked, key meta-data added to the file and then converted into XML instances using eXstyles and the NLM DTD. In the current work flow, those files are submitted to a compositor for layout and production of PDF files for typesetting and printing. XML instances are also provided to HighWire Press for hosting the journal on the Web.

We have two main goals. The first is to be able to embed N4L DOIs in the IJSEM content to enable readers to obtain key information regarding the current and past "meaning" of any bacterial or archaeal name. Ideally, we seek to do this as early in the pre-production phase as practical, with the least amount of impact on the editorial or publishing activities of the SGM.

A prerequisite for accomplishing this goal is that we must maintain the currency of N4L information objects so that links embedded in the literature always point to information that is current at the time of reading. With N4L, once the link is embedded, the paper becomes immune to future changes in that name so long as the N4L database is maintained. Historically, this has been a time consuming task requiring significant curatorial input. This task, however, may become much simpler in the future as the name capture routines used in the N4L process can be used to identifying names that likely represent new taxonomic proposals that are not yet in our data base.

A second goal of this project is to migrate our collection of N4L information objects into a more stable and easily maintained environment. In addition to providing the N4L processor with a means of embedding N4L information into prepublication documents in a completely non-obtrusive way, this database will also provide a means of automatically generating material that is relevant to the community, such as updates of the *Taxonomic Outline of Bacteria and Archaea* (TOBA) or custom mini-monographs. Direct access to N4L information objects in HTML or XML form provides a way for third-party data providers to link to the corresponding information objects and retrieving a detailed history of past and the present "meaning" of each name, any and all synonyms, associated core data, the record of past nomenclatural changes, and a complete bibliographic record of corresponding events that are associated with the name. It also provides an indication of the current and past placement of the strain within a given taxonomy.

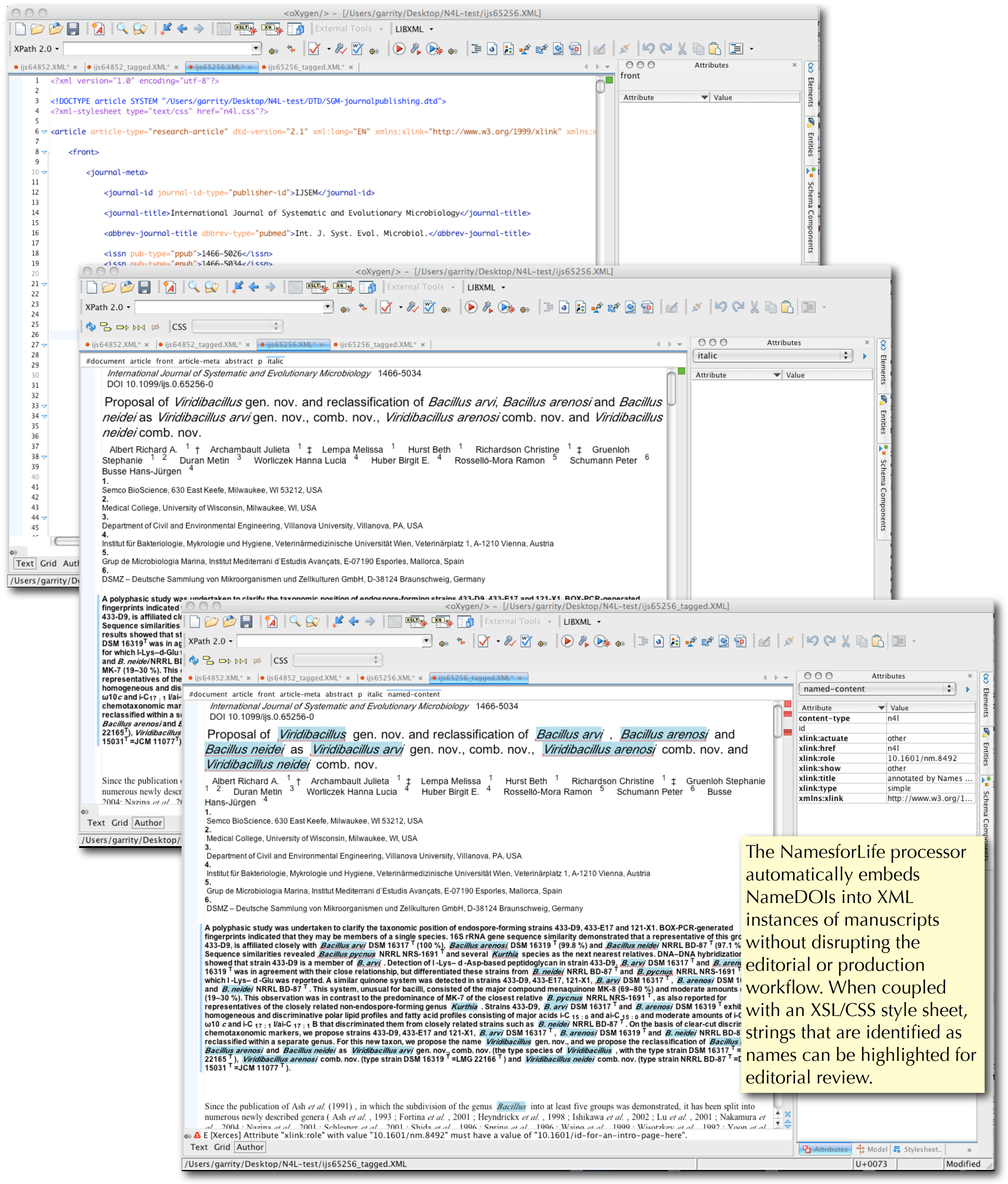


The advantages of the NamesforLife approach are the cost savings and improved quality of the nomenclatural information that can be provided as a service to the community. The production version of the NamesforLife semantic resolution service will provide a direct feed information that is part of the official nomenclatural record and is persistently linked to the published chain of events. Because NamesforLife provides a mechanism to link any occurrence of a name that appears in digital content, N4L-enabled papers that published in the past are automatically refreshed in lock-step with the field. This will ensuring that investigators can easily recover essential information whether they use the most recently published name or an earlier variant.

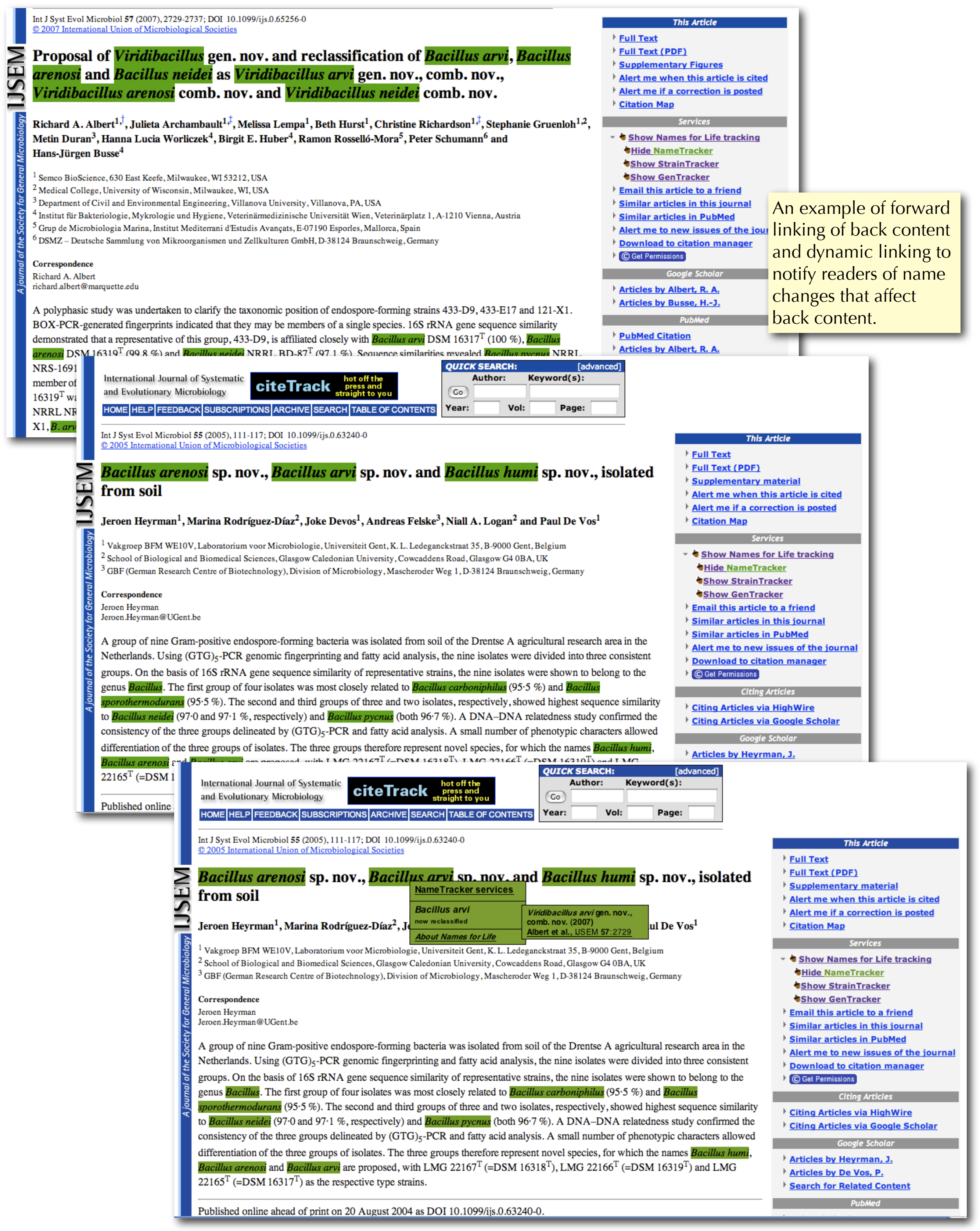
A prerequisite for this activity is the availability of a current global taxonomy and associated information that can be accessed either by batch or interactive applications. We are addressing this in a companion project, the *Taxomatic*, which is a system for examining and correcting taxonomies based on large-scale examination of supporting phylogenetic data.

Acknowledgments

We wish to thank B.J. Tindall (DSMZ, Braunschweig) and J. Euzeby (Ecole 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.



The NamesforLife processor automatically embeds NameDOIs into XML instances of manuscripts without disrupting the editorial or production workflow. When coupled with an XSL/CSS style sheet, strings that are identified as names can be highlighted for editorial review.



An example of forward linking of back content and dynamic linking to notify readers of name changes that affect back content.