# BioNames: surfacing the deep data of taxonomy

Roderic D. M. Page
rdmpage@gmail.com
http://iphylo.blogspot.com

## Summary

BioNames aims to create a biodiversity “dashboard” where at a glance we can see a summary of the taxonomic and phylogenetic information we have for a given taxon, and that information is seamlessly linked together in one place. It combines classifications from EOL with animal taxonomic names from ION, and bibliographic data from multiple sources including BHL, CrossRef, and Mendeley. The goal is to create a database where the user can drill down from a taxonomic name to see the original description, track the fate of that name through successive revisions, and see other related literature. Publications that are freely available will displayed in situ. If the taxon has been sequenced, the user can see one or more phylogenetic trees for those sequences, where each sequence is in turn linked to the publication that made those sequences available. For a biologist the site provides a quick answer to the basic question “what is this taxon?”, coupled with with graphical displays of the relevant bibliographic and genomic information.

## Background

If a goal of taxonomy is to make basic taxonomic information available to the wider biological community, then by some measures the discipline is doing well as evidenced by the existence of large online aggregations of taxonomic names, such as the Catalogue of Life and the Encyclopedia of Life. But on closer inspection these databases are often are little more than online collections of digitised 5x3 index cards (a technology Linnaeus himself pioneered; Müller-Wille and Charmantier, 2011). The lists of names may be digital, but taxonomic databases have failed to exploit two key developments in biological digitisation. The first is the massive increase in digitisation of scientific literature, including both content generated by scientific publishers scanning their back catalogues, and by digital libraries such as the Biodiversity Heritage Library (Pilsk, 2010). The second development is the rise in the "born digital" data of genomics (Benson et al. 2012) much of it linked to the primary literature.

These new digital developments have the potential to alter the way people interact with taxonomic data. Digitisation of the literature means that, in principle, anyone can read the original publication of a taxonomic name, and follow the subsequent taxonomic revisions. No longer do we need to rely on the word of an expert (or, more realistically, a cryptic entry in a database that has been gone through several filters before its appearance in a taxonomic database).

Similarly, the rise of genomic data removes the necessity for specific expertise in the morphology of a group. If you can download sequences from GenBank and build a tree (and there are readily available "how to" manuals for this task) then you can potentially investigate the taxonomy of any extant group of organisms. Whereas an in-depth knowledge of the morphology of one group does not readily extend to unrelated taxa (knowledge of wasp anatomy does not prepare you to study the mammalian skeleton), genomics data can span the tree of life. Furthermore, molecular phylogenetics is sufficiently scalable that we are entering the era of DNA metabarcoding (Taberlet et al. 2012) where bulk samples are being sequenced and the taxa contained are automatically identified.

The increasing use of sequence data has made taxonomic relationships computable (e.g., by building phylogenetic trees). Yet many DNA sequences are disconnected from taxonomy because they lack formal taxonomic names (Parr et al. 2011). Barcoding has been responsible for a massive influx of these "dark taxa" into the sequence databases http://iphylo.blogspot.co.uk/2011/04/dark-taxa-genbank-in-post-taxonomic.html. At the time of writing many of these unnamed barcode taxa are being suppressed by Genbank. But even without the barcoding sequences, dark taxa have been steadily increasing in number in recent years. This gap between names and genes is mirrored by biodiversity databases that reflect a long tradition of aggregating taxonomic names and data tagged with those names, but lack either links to the primary literature of taxonomy, or the genomic data that is flooding into GenBank.

Recently taxonomic journals have begun to mark up taxonomic names and descriptions (Penev et al. 2010), which is a precursor to linking names and data together. But these developments leave open the problem of what these links will point to. If we have a database of all taxonomic names and the associated literature, then such a database would provide an obvious destination for those links. Ultimately, we could envisage embedding new taxonomic publications within this system, so that each new publication becomes simply another document within the database. In the same way, we could use automated methods to extend the process of tagging names, specimens and literature cited to the legacy literature, so that the entire body of taxonomic knowledge becomes a single interwoven web of names, citations, publications, and data.

## BioNames

The BioNames project seeks to lay the foundations for the vision articulated above. Its goal is to create a database of taxonomic names linked to the primary literature and, wherever possible, to phylogenetic trees. Taxonomic names, concepts, publications and sequences will all identified using one or more globally unique identifiers. By using these identifiers we can create a network of linked information. Using identifiers rather than cryptic text strings (for example, abbreviated bibliographic citations) simplifies the task of linking - we can rely on exact matching of identifiers rather than approximate matching between names for what may or may not be the same entity. This is particularly relevant once we start to aggregate information from different databases, where the same literature reference may be represented by a different string. Furthermore, if we use existing identifiers we increase the potential to connect to other databases. For example, using a PubMed identifier for an article makes it straightforward to retrieve other forms of data associated with that article, such as sequences, controlled vocabularies, and citations. Using DOIs makes it easier to query databases such as Mendeley to retrieve information on the frequency of social bookmarking, or for users of Mendeley to discover their own publications in the database.

Identifiers by themselves help bind data together, but we also want access to the data itself. In the case of articles in BioStor we have access to the underlying OCR text, from which we can extract geographic localities and museum specimen codes (Page 2011). Many articles are freely available online as PDFs, these could also be processed using the same approach employed in BioStor. Long term this text mining could be extended to extract other information, such as literature cited, or ecological associations (e.g., articles that mention insect and mammal names together may contain information on host-parasite associations).

## Methods

As a starting point I will use the Index of Organism Names (ION) database http://www.organismnames.com which comprises names of taxa covered by the International Code of Zoological Nomenclature (i.e., animals and various other eukaryote groups). Each record in ION has a LSID with associated metadata in RDF, which can be harvested. Citations in ION are in the form of simple text strings and lack bibliographic identifiers. Several strategies will be used for locating digital identifiers (and their associated content) online. One is to parse the literature strings into their component parts (e.g., article title, journal, volume, pagination) and use OpenURL resolvers provided by CrossRef, BioStor, BioGUID, and elsewhere to locate digital identifiers. Another strategy is to harvest online bibliographies (such as lists of articles provided by publishers, or taxon-specific bibliographies), store these in Mendeley, then use approximate string matching to match citations in taxonomic databases to the records in Mendeley. To date I have some 200,000 references in Mendeley for this purpose.

A range of identifiers will be supported. The "gold standard" is the DOI, but other identifiers such as PubMed numbers, Handles, and URLs provided by digital archives such as JSTOR and CiNii will be used. To date I have mapped some 285,000 taxonomic citations to one or more identifiers, including 127,000 DOIs and 66,000 links to BioStor (see http://iphylo.org/~rpage/itaxon/?stats).

Taxa that are in the NCBI database will have associated sequence data. The PhyLoTA database http://phylota.net/ (Sanderson et al. 2008) has clustered GenBank into sets of similar sequences and constructed phylogenetic trees for each phylogenetically informative cluster. These clusters and trees are freely available for download. If a given taxon name being displayed is associated with a taxon in a phylogenetically informative cluster, BioNames will display the tree for that cluster. This will enable users to get a quick sense of the relationships of that taxon, and the extent to which the phylogeny and taxonomy correspond. For example, sequences may reveal considerable variation with a taxon regarded as a single species, or the tree may suggest an identification for a dark taxon.

### Services used

BioNames will make use of numerous web services, some of which I have developed, and some are provided by projects such as EOL, BHL, and Mendeley.

### *EOL*

To provide an easy way for users to navigate the collection of taxonomic names I will use the EOL API to retrieve and display taxonomic hierarchies that include the name. Images from EOL will be displayed to provide an "at a glance" sense of what taxon they are looking at.

### *Biodiversity Heritage Library*

The project uses a number of BHL services, either directly or indirectly via BioStor. The BioStor OpenURL resolver will be used to locate taxonomic articles in BHL. BioStor also provides a BioStor wrapper around the BHL search API that resembles the "discovered bibliography" currently shown on EOL but has improved date parsing, and groups BHL pages into articles (if the corresponding article is in BioStor), providing a more natural aggregation than the list of scanned items displayed by EOL.

### *Global Names Index*

The Global Names Index name parser API https://github.com/dimus/gni/wiki/api will be used to to convert names into their canonical form for use in searching.

### *Mendeley*

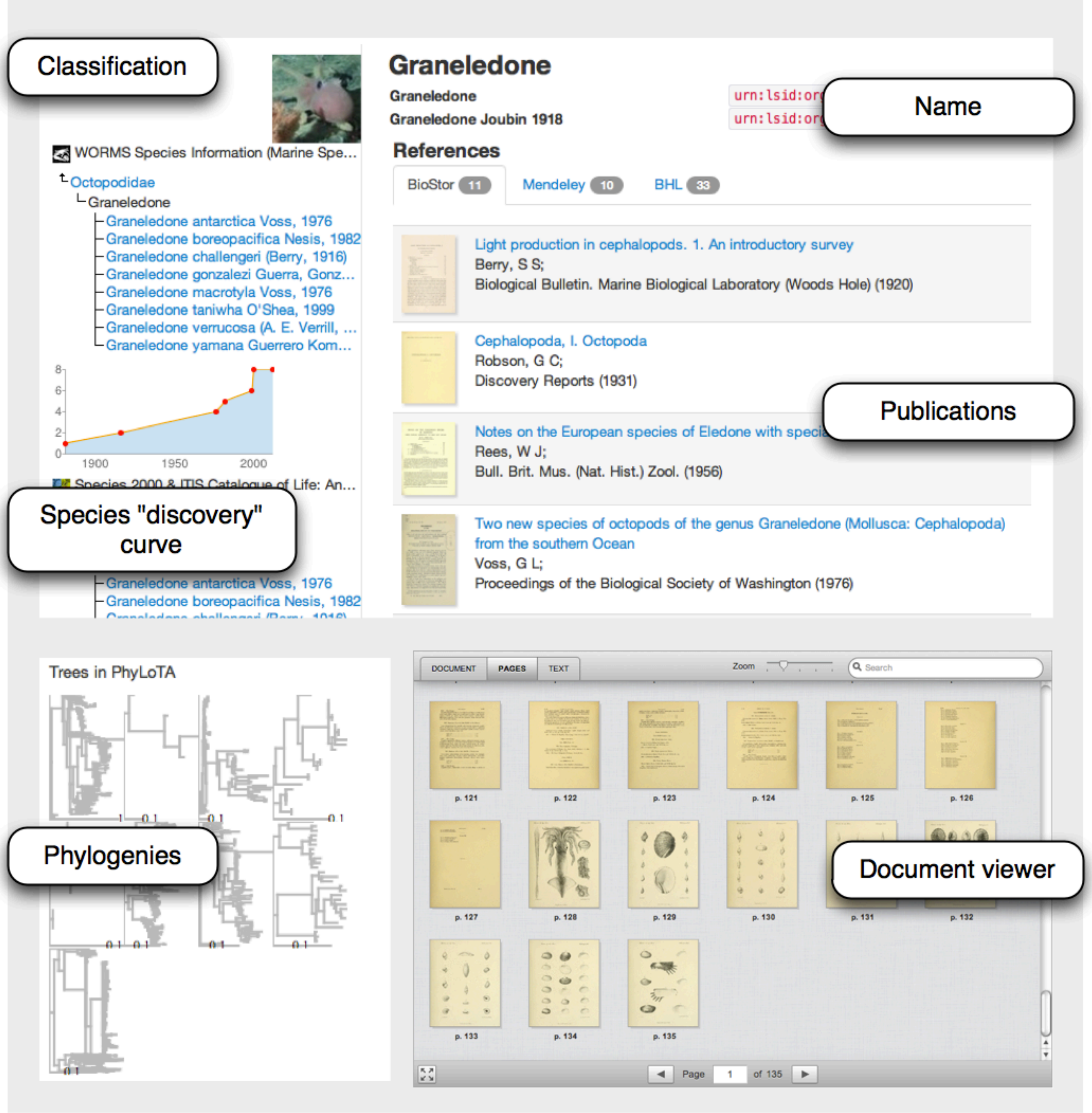
The Mendeley API will be used provide additional references for a name, as well as social bookmarking statistics. Mendeley's support for OAuth will also used to enable users to log in and discover their own publications in the database (for more details see http://iphylo.blogspot.co.uk/2011/12/these-are-my-species-finding-taxonomic.html).

### *PhyLoTA*

A local copy of the PhyLoTA database, coupled with a local copy of the EMBL sequence database will be used to provide a webservice that returns phylogenies and publications for a given NCBI taxon id (if phylogenetically informative clusters exist for that taxon).

## Interface

Part of the goal of BioNames is provide a seamless way to navigate between taxonomic names, classifications, phylogenies, and publications. Publications will be displayed either as thumbnails in lists, or as a browsable documents using the DocumentCloud viewer https://github.com/documentcloud/document-viewer. Classifications will be displayed along with species “discovery” curves summarising dates when the taxa were described (the dates will be derived from authority strings in the EOL classifications). Phylogenies will be displayed using SVG, either as thumbnails or as interactive trees. The figure below is a composite of some interface ideas currently being developed.



WORMS Species Information (Marine Spe…

Octopodidae
Graneledone
Graneledone antarctica Voss, 1976
Graneledone boreopacifica Nesis, 1982
Graneledone challengeri (Berry, 1916)
Graneledone gonzalezi Guerra, Gonz…
Graneledone macrotyla Voss, 1976
Graneledone taniwha O'Shea, 1999
Graneledone verrucosa (A. E. Verrill, …
Graneledone yamana Guerrero Kom…

Species 2000 & ITIS Catalogue of Life: An…

Graneledone antarctica Voss, 1976
Graneledone boreopacifica Nesis, 1982

**Graneledone**

Graneledone
Graneledone Joubin 1918

urn:lsid:or
urn:lsid:or

**References**

BioStor 11 | Mendeley 10 | BHL 33

Light production in cephalopods. 1. An introductory survey
Berry, S S;
Biological Bulletin. Marine Biological Laboratory (Woods Hole) (1920)

Cephalopoda, I. Octopoda
Robson, G C;
Discovery Reports (1931)

Notes on the European species of Eledone with speci
Rees, W J;
Bull. Brit. Mus. (Nat. Hist.) Zool. (1956)

Two new species of octopods of the genus Graneledone (Mollusca: Cephalopoda) from the southern Ocean
Voss, G L;
Proceedings of the Biological Society of Washington (1976)

Trees in PhyLoTA

## Relationship to ongoing work

This project is a natural extension of several projects I have been working on to link taxonomic names to the primary literature, as well as ways to visualisation phylogenetic trees. Ultimately it is a descendant of iSpecies http://ispecies.org which was a simple mashup that asked the question "what do we know about a species right now?" BioNames builds on previous experience with the document database CouchDB (e.g., http://iphylo.org/~rpage/afd/ and http://iphylo.org/~rpage/zoobank/), and the iTaxon project (http://iphylo.org/~rpage/itaxon/) to map literature strings in the ION database to digital identifiers for literature. It also uses my BioStor (http://biostor.org) project for extracting articles from BHL.

## Timeline and milestones

A prototype of BioNames is currently being developed. At present it is a mashup of queries to EOL, BioStor and BHL, as well as a local copy of the ION database linked to literature identifiers. CouchDB is being used to cache the results of API calls. The next step is to store data from ION, BioStor, and EOL directly in the database and develop views to query the data.

I anticipate releasing the initial prototype by the end of June. At the same time, source code will be released on GitHub. Project development will subsequently take place in the open, with regular releases for the duration of the project.

As the project unfolds I will be adding data from other taxonomic databases, such as the Catalogue of Life, ZooBank, and the Australian Faunal Directory.

## Dissemination plan

The primary output will be a website that will display taxonomic names, classifications, links to the primary literature, and evolutionary trees. Where literature is part of BioStor, or is otherwise freely available (i.e., open access PDFs or in digital archives such as Gallica) the publications will be displayed on the web site using the DocumentCloud viewer.

Progress on the project will be documented on my iPhylo blog, and code will be made available on GitHub.

An open access paper will be published that describes the project.

The mapping between names and identifiers for the primary literature will be made available on the project website as a data dump. In its simplest form this will be a simple pairing of taxon name identifier and literature identifier (e.g., taxon name LSID and DOI) but other formats will also be provided (e.g., RDF).

Discussions would be held with EOL about adding the taxon name - literature mapping to EOL web pages. This would require some thought about how to treat names and literature such that the digital identifiers are preserved, and links to online literature persist (for example, through caching). It may also require discussions with name sources (such as ION) about whether they will become EOL content providers and thus make their names available to EOL.

## Literature cited

Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. Nucleic acids research 40(Database issue): D48-53. doi: 10.1093/nar/gkr1202.

Müller-Wille S, Charmantier I (2011) Natural history and information overload: The case of Linnaeus. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43(1): 4-5. doi: 10.1016/j.shpsc.2011.10.021.

Page RDM (2011) Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. BMC bioinformatics 12: 187. doi: 10.1186/1471-2105-12-187.

Parr CS, Guralnick R, Cellinese N, Page RDM (2011) Evolutionary informatics: unifying knowledge about the diversity of life. Trends in ecology & evolution. doi: 10.1016/j.tree.2011.11.001.

Penev L, Agosti D, Georgiev T, Catapano T, Miller J, Blagoderov V, Roberts D, et al. (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. ZooKeys 50: 1-16. doi: 10.3897/zookeys.50.538.

Pilsk S, Person M, Deveer J, Furfey J, Kalfatovic M (2010) The Biodiversity Heritage Library: Advancing Metadata Practices in a Collaborative Digital Library. Journal of Library Metadata 10(2): 136-155. doi: 10.1080/19386389.2010.506400.

Sanderson MJ, Boss D, Chen D, Cranston KA, Wehe A (2008) The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. Systematic biology 57(3): 335-46. doi: 10.1080/10635150802158688.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology 21(8): 2045-50. doi: 10.1111/j.1365-294X.2012.05470.x.