Names are key to the big new biology

Patterson, D. J.[1], Cooper, J.[2], Kirk, P. M.[3] , Pyle, R. L.[4], Remsen, D. P.[5]

[1.] Biodiversity Informatics, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA. dpatterson@eol.org

[2.] Landcare Research, PO Box 40, Lincoln 7640, New Zealand. CooperJ@landcareresearch.co.nz

[3.] CABI UK, Bakeham Lane, Egham, Surrey, TW20 9TY, United Kingdom. p.kirk@CABI.ORG

[4.] Department of Natural Sciences, Bishop Museum, 1525 Bernice St., Honolulu, Hawaii 96817, USA. deepreef@bishopmuseum.org

[5.] GBIF, Universitetsparken 15, Copenhagen ø, DK 2100, Denmark. dremsen@gbif.org.

Abstract

Those who seek answers to big, broad questions about biology, especially questions emphasizing the organism (taxonomy, evolution, ecology), will soon benefit from an emerging names-based infrastructure. It will draw on the almost universal association of organism names with biological information to index and interconnect information distributed across the Internet. The result will be a virtual data commons, expanding as further data are shared, allowing biology to become more of a "big science". Informatics devices will exploit this 'big new biology', revitalizing comparative biology with a broad perspective to reveal previously inaccessible trends and discontinuities, so helping us to reveal unfamiliar biological truths. Here, we review the first components of this freely available, participatory, and semantic Global Names Architecture.

The value of taxonomy to a biology that is changing

"New Biology" is a vision [1] of a discipline evolving to become considerably more data-intensive as it accommodates increasing amounts of under-analysed data from high-throughput molecular and environmental technologies, and from large-scale digitization programs such as the Biodiversity Heritage Library (BHL, http://www.biodiversitylibrary.org/). In addition, there is pressure on scientists to share their results. More of the biological community will have access to on-line resources. Biology will shift towards the data-intensive 'big sciences' [2, 3]. Web services that use names to index and organize information about organisms will be a critical part of this 'big new biology'.

This change will require an organizational framework that is able to manage billions of pieces of information about our current catalogue of 2,200,000 or so living and expired species. The information will be distributed across thousands of Web sites. Three devices have the potential to organize information on all species. The first might use information from the molecular machinery that is common to all organisms. The second, phyloinformatics [4], would call on hypotheses through which the ancestor-descendent relationships within all life are explored. While the logics are appealing, neither phylogenetic nor genetic analyses have been applied to the majority of species, let alone all species. Today, they would fail as comprehensive information management devices. Fortunately, the third option, taxonomy, extends to all formally described species and so offers a life-wide axis by which all biological information might be organized [5, 6].

Taxonomy is supported by 5,000 to 10,000 professional taxonomists worldwide [7; http://www.gti-kontaktstelle.de/taxonomy_E.html]. This 'team' [8] is united by principles founded in the codes of nomenclature. Taxonomists discover and describe biodiversity, arrange species into classifications with sensitivity to phylogenetic insights, are aware of all of the literature that bears on the identity of the taxa, and provide services to those who rely on authoritative information. However, many taxonomists feel unable to meet the expectations of the discipline, home institutions, or exasperated users [9, 10], and even believe that taxonomy as a scientific discipline is in danger of extinction [5, 11, 12]. Others argue that the "information age" offers new opportunities to serve those who depend on taxonomic knowledge [6, 13-15], and that using taxonomy to manage on-line biological information can reinvigorate the discipline [16]. A small community of innovative taxonomists, computer scientists, and data managers (collectively "biodiversity informaticians") are pursuing this vision and are building data standards, information exchange protocols, resources, and services that can bring distributed data together as a virtual pool. Taxonomists use their expertise to add taxonomic principles, practices and knowledge as 'Taxonomic Intelligence', ensuring that the products are sensitive to the character of biology [17-18].

Taxonomy has two special features that suit it for re-use in biodiversity informatics. The first is the system of scientific names. Their almost universal use allows them to be treated as metadata to index biodiversity-related information, much as names are used in the index of a book. Secondly, classification schemes transform lists of names into organizational structures (ontologies) that group data, permit generalizing statements, allow users to infer properties of taxa, expand or focus searches, or to browse information in a biologically relevant fashion. The value of names as metadata and classifications as ontologies led to the vision of a names-based infrastructure to serve

biology [19]. This approach is used in major life-wide projects such as the Encyclopedia of Life [http://www.eol.org, 20, 21]. Now this approach is being transformed into a 'Global Names Architecture' (GNA) that aims to make the informatics potential of names and hierarchies freely available through the Internet.

To be effective in information management, GNA must overcome an array of challenges. It must index all references to organisms. It must bring together information on the same taxon even when different names are used to refer to it, and it will need to 'know' when the same name refers to more than one taxon. The system must be dynamic, adapting to changes in nomenclature, phylogeny or taxonomy [18]. To scale to the task, it must automatically draw on new information as it is published in authoritative on-line sources, a process that will be made possible by the widespread adoption of agreed protocols, standards and identifiers [22].

GNA will initially serve three areas of biology with interests in names. The first is taxonomy. Taxonomists use 'names' as tokens for concepts of species (and other taxa) and compile lists of names to catalogue and discriminate all approximately 1,900,000 named extant species and 250,000 named extinct species [23, 24]. Species are indefinite objects and taxonomists necessarily dispute where their boundaries lie. Their views are referred to as taxonomic concepts [22, 25-27]. The architecture must be able to discriminate competing concepts, and link all of them to specimens, georeferenced data, publications, and other usages that inform the concepts.

The second area deals with names from the perspective of the Codes of Nomenclature. In this context, the meaning of a name derives from 'nomenclatural acts' that begin with the creation of a new name and include subsequent actions that refine or change it. The results are compiled as nomenclators: definitive listings of code-governed names, their orthography, and bibliographic citations.

The third area, biodiversity informatics, is broader than taxonomy and nomenclature. Informaticians need to keep track of any string of alphanumeric characters that was used to refer to taxa. The strings include scientific names, vernacular names (which in some contexts are the formally preferred names (e.g., the Australian Standard Fish Names http://www.fishnames.com.au/) and surrogates for names. Surrogates include provisional names and specimen, culture or strain numbers which refer to a taxon. "SAR-11" ('SAR' refers to the Sargasso Sea) was a surrogate name given in 1990 to an important member of the marine plankton. Only a decade later did it become known as *Pelagibacter ubique* [28].

The names problems

The needs of taxonomists and nomenclators can be satisfied with relatively minor modifications of traditional practices. But, the biodiversity informaticians are encountering unfamiliar problems that confound the merger of distributed data. They require a more innovative system.

The largest problem is that of 'many-names-for-one-species', where data on the same species have been indexed with different names. Until addressed, it prevents all information about the same species being brought together. This problem has many sources, such as when new research leads to the relocation of a species to a different genus. For example, a proposal to break up the genus that contains *Drosophila melanogaster* would lead to the species epithet

'*melanogaster*' being combined with a different genus name to create a new binomial, *Sophophora melanogaster* [29, 30]. Such taxonomic revisions create indisputable homotypic synonyms (*Drosophila melanogaster* and *Sophophora melanogaster* refer to the same species). The names infrastructure must bring together information that was published using either name. A second type of synonymy, heterotypic synonymy, occurs when a taxonomist opines that taxa previously considered distinct are the same. Again, the challenge is to bring information labeled under different names together. The solution to this problem must also manage vernacular names and surrogates.

Figure 1: Lexical variants of scientific names. A few of the valid alternative spellings of *Cyclotrachelus sodalis*, image from Canadian Biodiversity Information Facility (http://www.cbif.gc.ca/), used with permission.

C. sodalis (LeC)
C. sodalis (LeC.)
C. (*E.*) *sodalis* (LeC)
C. (*E.*) *sodalis* (LeC.)
C. sodalis (Le Conte)
C. sodalis (LeC. 1848)
C. sodalis (LeC., 1848)
C. (*E.*) *sodalis* (LeConte)
C. (*E.*) *sodalis* (Le Conte)
C. (*E.*) *sodalis* (LeC. 1848)
C. (*Evarthrus*) *sodalis* (LeC)
Cyclotrachelus sodalis (LeC)
C. (*Evarthrus*) *sodalis* (LeC.)
Cyclotrachelus sodalis (LeC.)
C. (*E.*) *sodalis* (Le Conte 1848)
C. (*E.*) *sodalis* (Le Conte, 1848)
C. (*Evarthrus*) *sodalis* (LeConte)
C. (*Evarthrus*) *sodalis*(Le Conte)
Cyclotrachelus (y.) *sodalis* (LeC)
Cyclotrachelus sodalis (LeConte)
Cyclotrachelus sodalis (Le Conte)
Cyclotrachelus (*E.*) *sodalis* (LeC.)
C. (*Evarthrus*) *sodalis* (LeC. 1848)
C. (*Evarthrus*) *sodalis* (LeC., 1848)
Cyclotrachelus sodalis (LeC. 1848)
Cyclotrachelus sodalis (LeC., 1848)

Cyclotrachelus (*E.*) *sodalis* (LeConte)
Cyclotrachelus (*E.*) *sodalis* (Le Conte)
C. (*Evarthrus*) *sodalis* (Le Conte 1848)
C. (*Evarthrus*) *sodalis* (Le Conte, 1848)
Cyclotrachelus sodalis (Le Conte 1848)
Cyclotrachelus sodalis (Le Conte, 1848)
Cyclotrachelus (*E.*) *sodalis* (LeC. 1848)
Cyclotrachelus (*E.*) *sodalis* (LeC., 1848)
Cyclotrachelus (*Evarthrus*) *sodalis* (LeC)
Cyclotrachelus (*Evarthrus*) *sodalis* (LeC.)
Cyclotrachelus (*E.*) *sodalis* (Le Conte 1848)
Cyclotrachelus (*E.*) *sodalis* (Le Conte, 1848)
Cyclotrachelus (*Evarthrus*) *sodalis* (LeConte)
Cyclotrachelus (*Evarthrus*) *sodalis* (Le Conte)
Cyclotrachelus (*Evarthrus*) *sodalis* (LeC. 1848)
Cyclotrachelus (*Evarthrus*) *sodalis* (LeC., 1848)
Cyclotrachelus (*Evarthrus*) *sodalis* (Le Conte 1848)
Cyclotrachelus (*Evarthrus*) *sodalis* (Le Conte, 1848)

Most of the alternative names for species come from different ways in which names are represented (Fig. 1). Variants are caused by different styles of citing authors, how names are abbreviated, unintended errors, truncations, or concatenations. As each string, right or wrong, is associated with one or more usages, all variants must be included within the indexing structure.

Two solutions address the "many-names-for-one-species" problem. The first standardizes on a 'correct' name and seeks to apply that name universally. This is not viable because the chosen name will be arbitrary when, as is common, there is disagreement about the number of species or how each species should be defined. This solution cannot be applied retrospectively (at least not without the second solution); is costly to maintain, and does not cover vernacular or surrogate names. The second solution is to link together (reconcile) all known names for a given taxonomic concept (Fig. 2). Reconciliation can be applied to any name, and preferred names can be 'flagged' to meet the needs of the first solution. With reconciliation, queries initiated with one name are transformed into actions involving all names.

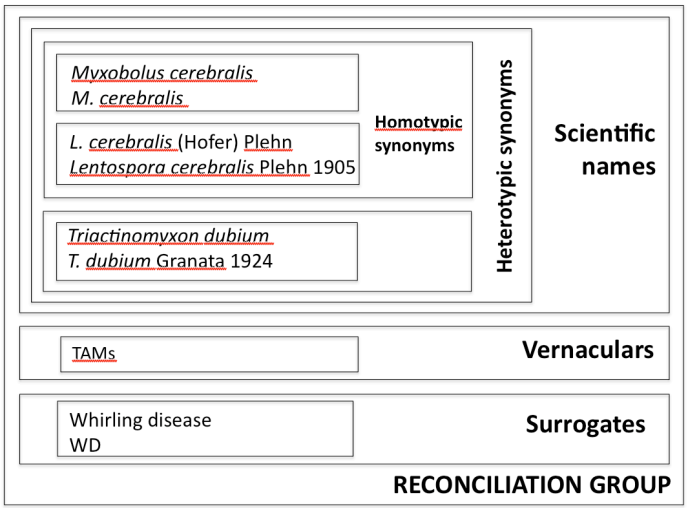

Figure 2: Reconciliation groups. A reconciliation group is an aggregate of all names used to refer to a taxon. It is comprised of one or more scientific names, with or without vernacular names or surrogates. Every name may be written out in one or more ways because the names of genera may or may not be abbreviated, and information about authorities may or may not be included. These lexical variants of names are included in the smallest boxes of the diagram. Homotypic synonyms include names that have the same type material – and they will have the same species element in the binomial name – *Myxobolus cerebralis* and *Lentospora cerebralis* are homotypic synonyms. The heterotypic synonymies are subjective, and emerge from a judgment by one or more taxonomists that *Triactinomyxon dubium* is the same species as was described as *Myxobolus cerebralis.* The vernacular names are non-scientific names that refer to the organisms. Surrogates are terms that also identify the taxon - in this case through the symptoms of the disease.

A second names problem arises when one name is used for more than one taxon. *Bacillus* is a genus of stick insects and of bacteria, *Aotus,* a type of legume and a monkey. This problem risks bringing together information on different organisms, leading to incorrect outcomes. This problem increases as biological research becomes 'bigger' expanding from narrower taxonomic territories to include all taxa. Now the 14% of plant genera that have homonyms elsewhere shifts from an amusing anecdote to a serious problem for data integration [31]. The solution will register homonyms and apply disambiguating devices. Generic names, the most abundant source of homonyms, can be disambiguated with reference to taxonomic context, species names, authorship, or by the included taxa.

Components of GNA

GNA is being developed as a modular structure that can expand and adapt as opportunities and needs emerge. The initial elements (Fig. 3) form a virtual layer that integrates information and services from sources (providers) to serve users (consumers).

At the core of GNA is a "Usage Bank" (GNUB) that is designed to index all published statements about life on Earth. The occurrence of a name on one or more occasions within a source constitutes a 'usage'. Usages occur in publications, field notes, databases, and classifications, on web pages, specimen labels in museums, and herbarium sheets. Initially, the usage bank will emphasize usages that bear on nomenclature [32, 33]. It will interconnect with prospective Web-based registry systems that will be used to formally establish new species instead of continuing the tradition of erecting new species in scientific publications (33). Through its association with nomenclators, the usage bank will inform the names architecture of correct scientific names and their spellings, will link to taxonomic treatments and specimens to provide insights into synonymies and taxonomic concepts. The first iteration of the usage bank is ZooBank (http://zoobank.org/), the ICZN registry for names of animals [34, 35]. Efforts are underway to incorporate nomenclators for fungi.

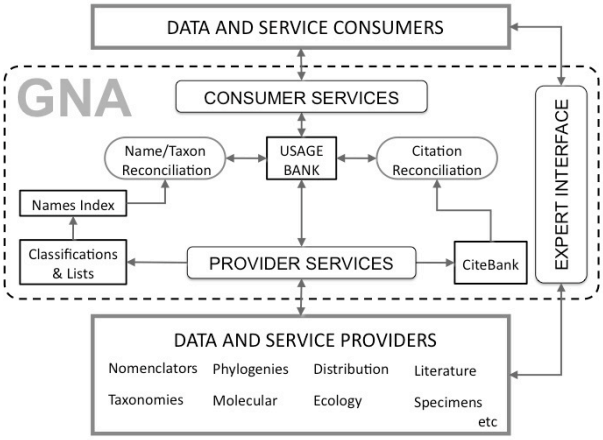

Figure 3: Initial elements of the Global Names Architecture. The virtual layer (dashed) is provided with services to access data and services from providers and to serve the needs of consumers. Given that many providers are also consumers, many interactions will be bidirectional. The central database is the Usage Bank. It indexes and points to occurrence of names in sources. Names, lists, and classifications from users are first compiled in a names index and in a classification repository, and citations are captured in CiteBank. As the content comes from many sources, the form and quality will vary. Reconciliation services seek to normalize the content of these databases, contributing consistent, de-duplicated, standardized content to the Usage Bank. An interface allows users to add, comment or correct the underlying infrastructure or annotate records in the databases.

The names index (GNI) is a simple index of all unique forms of name strings (i.e. correctly and incorrectly spelled scientific names with or without author information, or nomenclatural annotations, or vernacular names, or surrogates for names). The index (http://gni.globalnames.org/) currently includes about 19,000,000 names. The index links to data held by contributors and provides a simple discipline-specific means of linking distributed information (Fig. 4, model 'c').

NameLink, a prototype tool (http://labs.eol.org/?q=node/10/), recognizes names in documents and inserts anchors to which links known to GNI or to other digital objects can be attached. The names index is being enhanced with services to reconcile different versions of names and to disambiguate homonyms.

Biologists express their understanding of evolutionary relationships as classifications and trees. Both can be represented as parent-child structures, and are therefore interchangeable means of grouping or navigating data. Many catalogues of species, such as lists of marine species (http://www.marinespecies.org/), place lists within hierarchies of convenience. When the names are extracted to form simple lists, they can quickly filter data sets, instantly converting, for example, an encyclopedia of all life into an encyclopedia of marine life. Hierarchies can communicate insights into evolutionary history, and can be used to infer the distributions of attributes and test phylogenetic hypotheses. By accessing list and hierarchy repositories such as the GBIF ChecklistBank (http://names.gbif.org/), GNA can exploit the informatics and biological value inherent in parent-child structures and lists.

Citebank (http://citebank.org/) is an open repository for bibliographic citations relating to biodiversity. It fosters collaboration to build definitive reference lists. With content coming from many sources, the styles of citation vary and CiteBank must provide reconciliation services to map variant forms together. CiteBank will include a document submission module to allow sharing of documents while complying with the "safe harbor" principles of the Digital Millenium Copyright Act (http://www.copyright.gov/legislation/dmca.pdf). The early version of Citebank contains bibliographies of the BHL, other digital libraries, publishers, institutional repositories, and contributed bibliographies from specialist groups. CiteBank will have tools, like those in use by BHL, to find names in documents and automatically provide taxonomic indices.

Reconciliation and disambiguation services are being included to overcome the problems that accompany the federation of distributed but non-standardised information. Variously formed names and citations will have to be rendered into standard forms. This is achieved through reconciliation. First generation "fuzzy" matching algorithms (http://www.cmar.csiro.au/datacentre/taxamatch.htm) applied to names discover lexical variants and have reduced almost 19,000,000 names to about 6,000,000 reconciliation groups. Fuzzy matching is supplemented with parsing algorithms that reveal that, for example, '*Mycosphaerella eryngii* (Fr. ex Duby) Johanson ex Oudem. 1897', '*Mycosphaerella eryngii* (Duby) ex Oudem. 1897', and '*Mycosphaerella eryngii* (Fr. ex Duby) ex Oudem. 1897', all contain the same canonical binomial, *Mycosphaerella eryngii*, allowing all these strings along with their fuzzily matched variants to be placed in the same reconciliation group. With time, reconciliation services will bring together homotypic synonyms. Homonym discovery tools that flag homonyms and their children will minimize the risks of linking data on different taxa that have the same name.

The scale of the challenge to manage billions of data objects about millions of species arising on thousands of Web sites can be addressed through algorithms and by promoting information exchange with machine-readable standards and protocols. Yet, the properties of the species include a myriad of idiosyncrasies and are defined by complex interactions that defy rule-based analysis and organization [36]. The automated processes will not serve biology perfectly. The names architecture compensates with an interface that allows experts to identify gaps, correct

errors, disambiguate homonyms and help build reconciliation groups. Elements of the interface will allow names to be added, edited, 'deleted', or commented on; other functions will enable editing, merger, or division of reconciliation groups, as well as the integration of vernacular and surrogate names. Flagging tools can be used to annotate names and their relationships, and finally, classification tools will allow users to build or improve classifications.

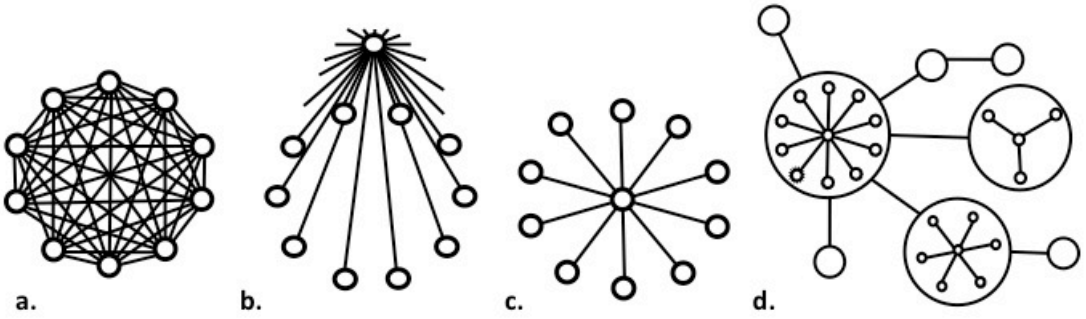

Figure 4. Models of interconnections among information at Web-sites. The oldest system (a) relies on hyperlinks between pages at each site. Such systems are costly to establish and maintain. Remote indexing of sites (b) is exemplified by search engines that allow users to visit one site and from it gain access to many other sites; the approach is very efficient but, as a 'one-size-fits-all' solution, may fail to accommodate the nuances of a discipline. A hub and spoke model (c) relies on specialist knowledge and agreed communications protocols to show content from a distributed array of sites through a unified specialist portal; structures like this for biology are made possible through a names-based infrastructure. The future is likely witness flexible integrated linked systems (d) that flexibly interconnect differing arrays of single web sites, data in semantic data clouds, and indexed systems. The appeal is that such systems can evolve into increasingly complex and customizable structures as more indexing and management services become available.

The big new biology needs to be readied to participate in newer trends of data integration, such as semantic data linking (Figs 4, 5). As biologists digitize data and make them available through web services, they have relied on search engines and hyperlinks to make content discoverable and to draw attention to related data. More automated data federation has been made possible through the adoption of web services, data standards, universally unique identifiers, and atomization of content. Now common keywords can foment a rich digital world of linked-data able to generate unsuspected insights [37]. A little time spent with Google Earth reveals how information generated for quite different purposes can be integrated using common denominators, such as georeferences, to deliver rich new services. The resulting semantic web has an almost anarchic quality, but it has enormous potential ([38], http://richard.cyganiak.de/2007/10/lod/). Semantic data-linking can be improved with services that manage discipline-specific data, metadata and ontologies. Data-linking for biology will benefit from rich services associated with taxonomic names, such as those that address the names problems (Fig. 5). To fulfill this role, the GNA will emphasize web services that broadcast and collate new knowledge in forms that are readily

understood by other machines.

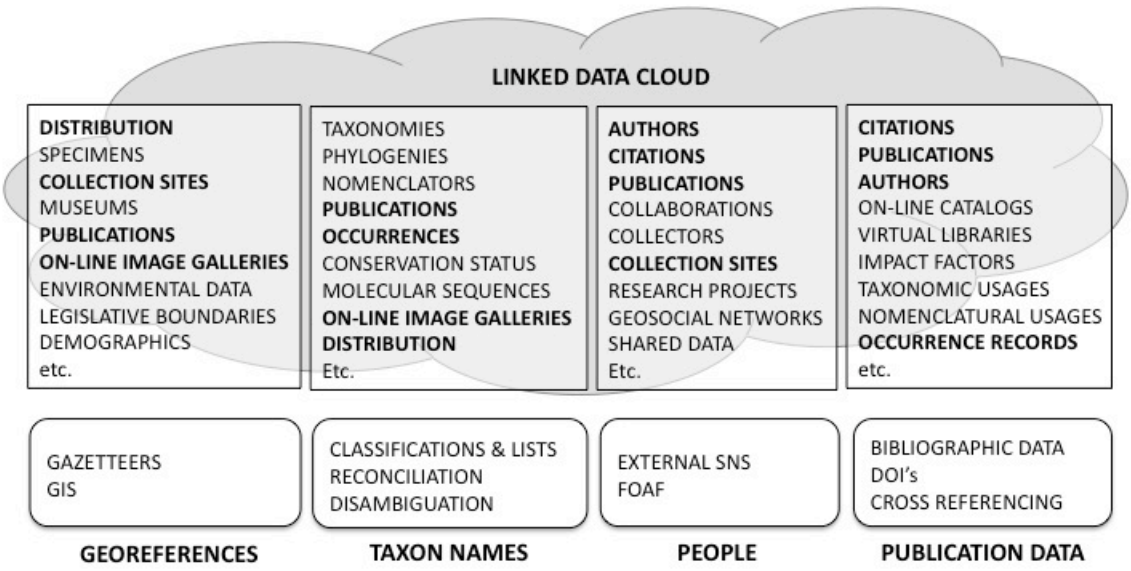

Figure 5. Linked data. Four examples illustrate the potential of vocabularies to interconnect a rich array of data (listed in square boxes) hosted in machine-readable environments, and provide the basis of cross-walks among knowledge areas (bolded terms). Georeferencing, publications and social networks (people) are already enhanced by services (rounded boxes), such as FOAF (Friend of a friend) that can be used to describe or build networks (e.g. http://network.nature.com/) or through technologies which interconnect the scientific literature (such as serviced by CrossRef http://www.crossref.org/ with almost half a billion links). Other links may identify where individuals have collected specimens and, being attached to location, provide for potential cross-walks to other data linked by georeferences such as maps, images, habitat type, environmental conditions, local institutions, and experts. Names of organisms, enhanced with reconciliation and disambiguation services, open up almost any data relating to organisms. As an illustration, a 16S sequence of a rare spider in GenBank, through lat.long. metadata, allows access to maps showing where in Argentina this organism was found, integrates the result with distributional data from the Global Biodiversity Information Facility that show the known world distribution, lists all publications in the Biodiversity Heritage Library that refer to the organism, all publications by the data depositor and of her colleagues, her obituary, and a description of the species emphasizing venomousness caused by a toxin that blocks sodium channels and has an LD50 of 0.0005 mg/kg.

And where is all this heading?

One reflection of the big new biology will be a biologically informed Internet. Users of search engines will find all information about a species irrespective of which name was used; no longer will biologists need to unpack nomenclatural history, but can expect systems to know that much of the information about *Pneumocystis pneumoniae* can be found under the name *Pneumocystis carinii*. We can expect electronic documents to be automatically brought up to date in matters nomenclatural and taxonomic, and for names in documents viewed with browsers to automatically link to other resources of our choosing.

The first beneficiaries of GNA will be the communities from which its architects and

engineers are drawn. Nomenclaturalists will have access to on-line reference information cross-linked to searchable page images from on-line virtual libraries. Taxonomists will be able to check on all previously used names and will not create new homonyms. They will register new species quickly and easily, linking them to descriptions on-line without the delays associated with representing knowledge with ink on paper. Taxonomists will adapt on-line classifications to suit their own needs, and the parent-child statements they create will be captured and drawn together to assemble an editable and dynamic catalogue of all life. Ecologists will find services to ensure that they identify components of their ecosystems correctly, and text-editing programs will prompt authors with the correct names for their objects of study. 'Normalizing' names-services will correct names in databases and data-linking projects will use common identifiers to merge complementary data.

The potential of data-linking is evident from mapping applications. Biologists of the future, assisted by GNA, can expect services to keep them abreast of new information about clades or taxa of interest. Users will have access to bigger and broader arrays of data, with valuable datasets identified with automated pointers that inform us that, for example, other ecologists and molecular biologists who used this data set also used those other data sets. Through their availability, suspicious data can be flagged for cautious treatment and the quality of data will improve. The capacity of this "crowd sourcing" to be creative as well as critical was powerfully demonstrated with Open Mapping that produced the most useful maps in the immediate aftermath of the recent Haitian earthquake (http://haiti.openstreetmap.nl/). With a virtual data commons, data become part of a dialog, and we can expect more tools to allow users to annotate data, or for nature lovers to confirm, deny, or track the spread of invasive species or to register biological responses to climate change. Connections among previously unassociated data will provide a fertile pasture to nourish new hybrid scientists who combine biology and computer sciences. From them and those working at the boundaries of the different subdisciplines of biology we can expect a flush of new services, analytical tools, and visualizations to reveal trends, patterns and discontinuities in data. They will take an unfamiliar, distant view of the knowledge landscape that is biology to reveal patterns not evident from reductionist approaches, while directing our attention to features of the underlying biology that deserve study. As a reinvention of comparative biology, such tools will become the 'Macroscope' [39, 40] able to extract new insights from the big new biology.

Acknowledgements

Many people have reacted, advised, and otherwise assisted in the development of ideas laid out here; they include Stan Blum, Patrick Leary, Dimitry Mozzherin, Rod Page, and Tony Rees. Our thanks are due to them. DJP thanks the NSF for support through the Data Conservancy project and the Alfred P. Sloan and John D. and Catherine T. MacArthur foundations for their support.

REFERENCES

1. National Research Council of the National Academies (2009) *A new biology for the 21st century.* The National Academies Press

2. Hey, T. et al. (eds) (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research (http://research.microsoft.com/en-us/collaboration/fourthparadigm/)
3. Kelling, S. et al. (2009) Data-intensive science: a new paradigm for biodiversity studies. *BioScience* 59, 613-620
4. Page, R. D. (2010) Phyloinformatics in the age of Wikipedia. Nature Precedings: 28 June 2010, http://precedings.nature.com/documents/4583/version/1, hdl:10101/npre.2010.4583.1
5. Agnarsson, I. and Kuntner, M. (2007) Taxonomy in a changing world: seeking solutions for a science in crisis. *Systematic Biology* 56, 531–539
6. Wheeler. Q.D. (ed.). (2008) *New Taxonomy*. (Systematics Association, Special Volume 76), CRC Press
7. Hopkins, G. W., and R. P. Freckleton. (2002) Declines in the numbers of amateur and professional taxonomists: implications for conservation. *Animal Conservation* 5, 245-249
8. Knapp, S. (2008) Taxonomy as a team sport. In *The New Taxonomy* (Systematics Association, Special Volume 76), (Wheeler, Q. ed.), pp 33-53. CRC Press
9. Godfray C. H. J. (2002). Challenges for taxonomy. *Nature* 417, 17–19
10. Bortulus, A. (2008) Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *Ambio* 37, 114–118
11. De Carvalho, M. R. et al. (2008a) Taxonomic impediment or impediment to taxonomy? A commentary on systematics and the cybertaxonomic-automation paradigm. *Evolutionary Biology* 34, 140–143
12. De Carvalho, M. R. et al. (2008b) Systematics must embrace comparative biology and evolution, not speed and automation. *Evolutionary Biology* 35, 150–157
13. Stein L. (2002) Creating a bioinformatics nation. *Nature* 418, 125
14. Mace G. M. et al. (2003) Preserving the tree of Life. *Science* 300, 1707-1709
15. Beach, J. H. et al. (1993) Hierarchical taxonomic databases. In *Advances in Computer Methods for Systematic Biology: Artificial Intelligence, Databases, Computer Vision,* (Fortuner R, ed.), pp 241 – 256, John Hopkins University Press
16. Patterson, D. J. (2009) Future Taxonomy. In *Systema Naturae 250 - the Linnaean Ark*" (Polaszek, A. ed.) pp 115-124 CRC Press
17. Patterson, D. J. et al. (2006). Taxonomic Indexing - extending what taxonomy is. Systematic Biology. *Syst. Biol.* 55, 367-373
18. Franz, N. M. and Thau, D. (2010) Biological taxonomy and ontology development: scope and limitations. *Biodiversity informatics* 7, 45-66
19. Patterson D.J. et al. (2008) Principles for a names-based cyberinfrastructure to serve all of biology. *Zootaxa* 1950, 153-163
20. Wilson, E. O. (2003) *The encyclopedia of life*. *TREE* 18, 77-80
21. Patterson, D. J. 2010. Encyclopedia of Life. In *Information and communication technologies for biodiversity conservation and agriculture* (Maurer, L, and Tochterman, K. eds), pp 67-81, Shaker Verlag
22. Kennedy, J. et al. (2006) Standard data model representation for taxonomic information. *Omics* 10, 220-230
23. Chapman, A. D. (2009) *Numbers of Living Species in Australia and the World (*2nd edn) Australian Biological Resources Study
24. Raup, D. (1991) *Extinction: bad genes or bad luck?* Norton
25. Berendsohn, W. G. (1995) The concept of potential taxa in databases. *Taxon* 44, 207-212
26. Berendsohn, W. G. and Geoffroy, M. (2007) Networking taxonomic concepts – uniting without "unitary-ism". In *Biodiversity Databases: From Cottage Industry to Industrial Networks* (Systematics Association Special Volume, 73), (Curry, G. and Humphries, C., eds.), pp 13-22, Taylor & Francis
27. Chavan, V. et al. (2005) Resolving taxonomic discrepancies: role of electronic catalogues of known organisms. *Biodiversity informatics* 2, 70-78

28. Rappe, M. S. et al. (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* 418: 630-633
29. van der Linde, K. and Houle, D. (2008) A supertree analysis and literature review of the genus *Drosophila* and related genera. Insect Systematics and Evolution 39, 241-267 1.
30. ICZN (2010) Opinion 2245 (Case 3407) *Drosophila* Fallén, 1823 (Insecta, Diptera): *Drosophila funebris* Fabricius, 1787 is maintained as the type species. *Bull. Zool. Nomenc.* 67, 106-115
31. McNeill, J. (1997) Key issues to be addressed. In *The new bionomenclature: The BioCode debate*. (Biology International Special Issue 34), (Hawksworth, D. L. ed). 17–40, IUBS
32. Crous, P.W. et al. (2004) MycoBank: an online initiative to launch mycology into the 21st century. *Studies in Mycology* 50, 19–22
33. Polaszek, A. et al. (2005) A universal register for animal names. *Nature* 437, 477
34. Pyle, R. and Michel, E. (2008) Zoobank: Developing a nomenclatural tool for unifying 250 years of biological information. *Zootaxa* 1950, 39–50
35. Pyle R., Michel, E. (2009) Unifying nomenclature: ZooBank and Global Names Usage Bank. *Bull. Zool. Nomenc.* 66, 298
36. Mayr, E. (2004) *What makes biology unique?* Cambridge University Press
37. Berners-Lee, T. et al. (2006) Tabulator: Exploring and Analyzing linked data on the Semantic Web. Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI06) workshop, Athens, Georgia, http://swui.semanticweb.org/swui06/papers/Berners-Lee/Berners-Lee.pdf
38. Page, R. D. M. (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers *Briefings in Bioinformatics* 9, 345-354
39. De Rosnay, J. (1975) *Le macroscope: vers une vision globale.* Seuil, Paris.
40. Ausubel, J. H. (2009) A botanical macroscope. *Proc. Natl. Acad. Sci.* 106, 12569