BRIEFINGS IN BIOINFORMATICS. VOL 8. NO 5. 347–357
Advance Access publication August 17, 2007
doi:10.1093/bib/bbm037

Biodiversity informatics: organizing and linking information across the spectrum of life

Indra Neil Sarkar

Submitted: 8th April 2007; Received (in revised form): 22nd July 2007

Abstract

Biological knowledge can be inferred from three major levels of information: molecules, organisms and ecologies. Bioinformatics is an established field that has made significant advances in the development of systems and techniques to organize contemporary molecular data; biodiversity informatics is an emerging discipline that strives to develop methods to organize knowledge at the organismal level extending back to the earliest dates of recorded natural history. Furthermore, while bioinformatics studies generally focus on detailed examinations of key 'model' organisms, biodiversity informatics aims to develop over-arching hypotheses that span the entire tree of life. Biodiversity informatics is presented here as a discipline that unifies biological information from a range of contemporary and historical sources across the spectrum of life using organisms as the linking thread. The present review primarily focuses on the use of organism names as a universal metadata element to link and integrate biodiversity data across a range of data sources.

Keywords: *biodiversity informatics; taxonomic intelligence; taxonomic name reconciliation; taxonomic name recognition; federated search engines; knowledge integration; encyclopedia of life*

INTRODUCTION

Continual improvements in technology enable us to generate more types of data across a wider spectrum of life than ever before imaginable. Among the grandest challenges in biology are transforming volumes of raw data into usable knowledge about our world and its inhabitants. This transformation poses significant challenges that necessitate the assistance of automated methods. Informatics strategies have been shown to facilitate the organization of biological data towards the development of testable hypotheses. To date, much of informatics in the biological domain (generally termed 'bioinformatics') focuses on studying molecular aspects of life across a number of key 'model' organisms and systems. Biodiversity informatics has emerged as a suite of informatics techniques that can augment traditional bioinformatics approaches by linking information at the organism level across a wide spectrum of data types and organisms, often times within a historical context. While 'biodiversity informatics' can be considered a new discipline, the use of automated techniques is not entirely new to the biodiversity domain. Indeed, the development of systematic techniques (often involving computers) has proven to be essential in the study and cataloguing of the

Corresponding author. Indra Neil Sarkar, MBLWHOI Library, Marine Biological Laboratory, Woods Hole, MA 02543, USA. Tel: +1-508-289-7632; Fax: +1-508-540-6902; E-mail: sarkar@mbl.edu

Indra Neil Sarkar, PhD, is the Informatics Manager in the MBLWHOI Library at the Marine Biological Laboratory in Woods Hole, Massachusetts, USA. Dr Sarkar's research involves the development and use of a range of computational techniques (including novel knowledge gathering and discovery methods, phylogenetics, information theory and natural language processing) in the study of the evolutionary history of infectious diseases. His research also involves the integration of disparate knowledge sources through the development of ontologies, information retrieval methods and indexing technologies within the greater context of biodiversity. To this end, he has been working towards the integration of historical documents with contemporary documents in an effort to identify information that may guide the development of new theories. Collectively, Dr Sarkar strives to identify knowledge that can aid in the understanding of disease and in the development of therapies and/or treatments to eradicate or control disease. He received his Bachelor of Science degree in Microbiology, with concentrations in Computer Science and Philosophy of Science, from the Lyman Briggs School of Science at Michigan State University and his Doctorate in Biomedical Informatics from the College of Physicians and Surgeons at Columbia University in the city of New York.

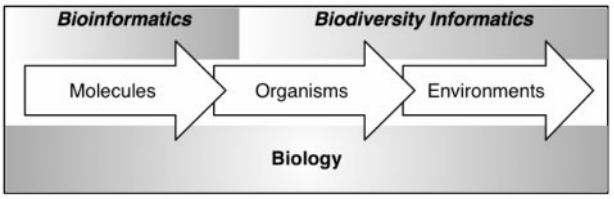

Figure 1: Biodiversity Informatics. The spectrum of biological knowledge can be binned into three categories: molecular, organismal and ecological. Bioinformatics studies primarily focus on the development of hypotheses at the molecular level. Biodiversity informatics primarily emphasizes the organization of knowledge at the higher levels, at the level of the organism.

speciation, distribution and evolution of life on Earth. However, to date, there has been limited integration and harmonization of biodiversity information within the context of molecular studies that are the focus of bioinformatics.

'Biodiversity' refers to the holistic study of life on Earth in light of its inherent variation [1]. In this context, diversity of biology is considered at three levels [2]: (i) molecular, (ii) organismal and (iii) ecological. Bioinformatics studies primarily focus on examining biological hypotheses from the molecular perspective. Information at the organism level, including scientific description and distribution information, may be available for many species that are not yet (or, in the case of extinct species, never) associated with molecular information. Biodiversity informatics thus aims to identify linkages within and across all three levels of biological data relative to the organism. The relationship between bioinformatics and biodiversity informatics, relative to the spectrum of biology is shown in Figure 1. The fundamental tenet in biodiversity informatics is that biological information can be linked through the organism towards the development of new, testable hypotheses. Additionally, because much insightful data may currently be locked away in historical archives that are becoming available through digitization movements, biodiversity informatics promises to complement contemporary knowledge with archival information.

The creation of methods and systems to consolidate, organize and categorize available information, regardless of its available form, is an essential step for large-scale comparative biological studies. The organization of biological information from an array of resources into consolidated knowledge bases for subsequent archival and research purposes is a significant informatics task. This centralization of a range of data types into a single resource can enable a range of comparative studies. In the bioinformatics community, centralized systems like the *Entrez* system at the National Center for Biotechnology Information (NCBI) provide access to biomedical information across many resources [3, 4]. This information can be retrieved and organized using standardized ontologies or terminologies [5–8]. Knowledge-based systems have also emerged to capture and organize biological information relative to particular domains (e.g. molecular interactions [9–11]). Finally, the magnitude of natural language biomedical literature has given rise to a range of natural language processing (NLP) systems to capture relevant information from biomedical literature [12–16]. Recent discussions in both the scientific [17–20] and popular media [21, 22] have described the need for similar frameworks to organize existing biodiversity knowledge both within the context of existing data and the modernization and incorporation of archival knowledge.

This brief review begins with a discussion of how organisms can be used to link information across disparate resources. The use of organisms within an information retrieval framework, termed 'taxonomically intelligent information retrieval', is then described within the context of biomedical literature retrieval using Medline. Finally, the use of organisms as a universal metadata element is explored towards the development of Web-based, federated search tools that enable retrieval and subsequent linkage of information across a range of resources.

THE ORGANISM AS A UNIFYING THREAD FOR KNOWLEDGE ABOUT LIFE ON EARTH

All information pertaining to life, including molecular data, is associated with at least one organism. Biodiversity informatics is thus a species-centric discipline [18, 23]. The research community is familiar with—even takes granted of the fact that—organism identifiers are used to annotate and organize almost all biological data. In current practice, for example, every entry in GenBank is associated with an organism (denoted by the 'TaxId' field, which links to NCBI Taxonomy). To this end, organism identifiers (which can include an organism's name or a name surrogate, like a strain number

or concept within an ontology) can be used to link data across a wide range of biologically relevant databases. Organism identifiers can thus be used to identify and link information from data sources that might contain information on gene expression, ecology, conservation and distribution. They can also be used to identify historical documents that might contain the original description of the organism, which often includes in-depth morphological character descriptions. This aspect of using the name as a metadata anchor to link biological data across resources reflects the point that 'All accumulated information of a species is tied to a scientific name, a name that serves as the link between what has been learned in the past and what we today add to the body of knowledge'. [24] To date, most efforts have focused on the (mostly manual) organization of information associated with a single organism or a group of related organisms (as in the case of the above quote, entomology). To keep pace with the increasing rate of new biological data being made available, there is a significant need to develop resources and techniques for organizing biological information across a wider spectrum of life. Informatics solutions can then be developed to further the exploration of comparative biology hypotheses across a wide range of organisms within the context of multiple axes of information (e.g. morphological features and geographic distribution).

There are an estimated two million organisms that are associated with taxonomic treatments [2]. A taxonomic treatment generally includes an organism name, morphological description, distributional information and other related (e.g. phylogenetic) information. These taxonomic treatments can be used to supplement contemporary molecular data, especially in the context of identifying 'genotype–phenotype' correlations (e.g. molecular patterns can be associated with morphological descriptions between taxonomic groups). However, using organism names as identifiers to link information can be problematic, especially in a historical context. Organism names change over time—e.g. before 1919, data associated with *Escherichia coli* were labeled with *Bacillus coli* or *Bacterium coli*. Issues remain even in light of an array of regulatory bodies that strive to develop systematic rules to stabilize names and minimize ambiguity [25–27]. Reconciliation techniques are thus needed to interconnect multiple names, either objectively (e.g. *Doryteuthis pealeii* and *Loligo pealeii* are names that refer to the common squid) or subjectively (e.g. *Brucella abortus* and *Brucella suis* are names that refer to the causative agent for Brucellosis, which affects a range of hosts). Disambiguation methods are also needed to distinguish different organism concepts associated with the same names that refer to more than one species (e.g. *Peranema* refers to a genus of both a fern and a euglena). The successful development of comprehensive scientific name indices can be used to identify relevant data across a wide range of resources. A centralized index might also foster the development of applications that can be used to infer linkages between organisms across heterogeneous data sources.

The cataloguing of scientific names into a single, publicly accessible resource is a paramount first step to develop a framework for organizing biological knowledge [28, 29]. Such an endeavor is not a new concept. The Unified Medical Language System® (UMLS®) began development in the mid-1980s as a means to create a standard language for biomedicine that could be used by computer-based clinical information systems [30]. The UMLS includes terms from over 100 biomedical terminologies and ontologies organized into over one million concepts [31]. The UMLS does contain some scientific name terminologies, most notably NCBI Taxonomy. However, in addition to NCBI Taxonomy, there are a number of other resources that maintain lists of scientific names. These include Species2000 (http://www.sp2000.org/) and the Integrated Taxonomic Information System (ITIS; http://www.itis.usda.gov/), both of which are associated with the Catalogue of Life project [32]. Organism names are also maintained by groups of researchers focused on a particular taxonomic group—e.g. IndexFungorum (Fungi; http://www.bioone.org/), AlgaeBase (Algae; [33]), CephBase (Cephalopods; [34]), Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ; Microorganisms derived from Euzeby's List [35]), and FishBase (Fish; [36]). To organize these different lists from multiple sources, the Universal Biological Indexer and Organizer (uBio; http://www.ubio.org) project has been working towards the integration of scientific and vernacular names. The uBio databases are designed to function much in the same way as the UMLS, as an aggregator of lists of concepts and hierarchies into a single resource. Currently, uBio contains over 10 million organism name strings, which have been collected from a range of existing

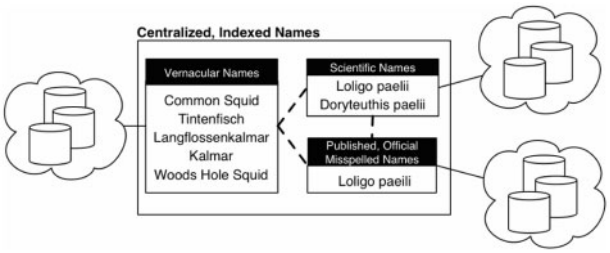

Figure 2: Organism name reconciliation. Much of biological content is associated with organisms, and thus annotated with organism names. However, organism names can be represented by vernacular, scientific and published misspellings. Centralized name-based indices are designed to relate multiple variants for the same organism name (shown by dashed lines) such that content can be related across multiple databases (linked by solid lines).

scientific name resources, including all of the previously mentioned. The general structure of names in uBio is to organize vernacular, scientific and published name variants (e.g. misspellings) into reconciliation groups (Figure 2). This structure enables the linkage of information across various resources for the same organism, regardless of the name used for annotation. Because taxonomic opinions can also be subjective in nature, the structure is also designed to accommodate competing viewpoints.

Beyond organism names contained in catalogs and databases, a significant number of organism names are embedded in natural language texts, which include many historical texts that are becoming increasingly available through various digitization efforts [37]. Perhaps the most significant, relevant digitization effort currently underway is the Biodiversity Heritage Library (BHL; http://www.bhl.si.edu/). Ultimately, the BHL aims to digitize all recorded natural history, starting with the 10 core BHL member institutions (American Museum of Natural History, The Field Museum, Harvard University Botany Libraries, Harvard University Ernst Mayr Library of the Museum of Comparative Zoology, Marine Biological Laboratory/Woods Hole Oceanographic Institution, Missouri Botanical Garden, Natural History Museum, The New York Botanical Garden, Royal Botanic Gardens and Smithsonian Institution). Initially, the BHL digitization efforts will focus on volumes that are out-of-copyright or for which permissions have been obtained. There have already been discussions about expanding the scope of digitized literature beyond pre-1923 (the current accepted criteria for determining copyright restriction in the United States). Nonetheless, the pre-1923 literature represents millions of pages of natural science literature that have essentially been locked away until the conception of the BHL.

Because the BHL represents a significant corpus of historical literature that may not have been studied in decades, it will undoubtedly contain references to organisms that have long since been renamed, migrated or become extinct. It will thus become increasingly important to develop automated tools that can detect scientific names contained within digital text. Named Entity Recognition (NER) algorithms, which have been shown effective for identifying gene and protein names from biomedical literature [38], can be applied to this task. NER algorithms that are specific for identification of organism names, called Taxonomic Name Recognition (TNR), have been developed to exploit the linguistic and contextual nature of taxonomic names, as dictated by Linnaean rules used for most organism scientific names since 1754 (e.g. Latinized form of words in 'Genus species' format) [39, 40]. The basic TNR approach posits that taxonomic names can be identified from natural language text using a combination of taxonomic nomenclature rules and/or a lexicon of non-taxonomic terms. TNR methods need to be flexible enough to process 'dirty' OCR documents—given the scope of the BHL, it is difficult to predict the accuracy of how good the OCR quality will ultimately be.

In addition to literature and molecular data, organism names are associated with specimen collections. Most of these data are becoming increasingly available in one of two structured XML schemas—DarwinCore (http://digir.net/schema/conceptual/darwin/2003/1.0/darwin2.xsd) or ABCD (http://www.bgbm.org/TDWG/CODATA/Schema). The systematic representation of data in museum and herbaria collections can ultimately lead to the linkage of knowledge from molecules (e.g. in repositories like GenBank), to contemporary literature (e.g. indexed in Medline), to historical literature (e.g. indexed in the BHL), to raw specimen data that are available in museums and herbaria. The importance of museum specimens in the study of zoonoses has been demonstrated—e.g. rodent skins preserved in a museum collection were instrumental in elucidation of the disease etiology of hantavirus [41].

TAXONOMICALLY INTELLIGENT INFORMATION RETRIEVAL

Identifying pertinent information within large databases may be facilitated through the use of controlled vocabularies to represent concepts that are associated with each item of information. The Medline literature citation database indexes its content using a controlled vocabulary of indexing terms, the Medical Subject Headings (MeSH) [42]. The use of a controlled vocabulary to facilitate information retrieval queries has been positively demonstrated when using MeSH to retrieve information from Medline [42, 43]. Both the retrieval and information linking processes are dependent on the quality of the keyword controlled vocabulary as well as their consistent application to information objects (e.g. MeSH terms that are used to annotate Medline citations) [44]. A significant challenge for the success of this approach is the identification of pertinent terms for a given query [44]. One way to identify terms is to browse through a hierarchical organization of terms. For example, it has been shown that a browsing system that provides a 'real-time' list of available terms can help guide searching through a corpus of documents [45]. The Medline search interface enables users to take advantage of the hierarchical structure of MeSH. Organism names can also be placed into hierarchies that can assist in the navigation of knowledge sources [46]. This is naturally done within biology through taxonomies, which reflect hierarchies that are often used to organize organisms according to relatedness. The term 'Taxonomic Intelligence' was introduced in connection with the 'Logic-based Integration of Taxonomic Conflicts in Heterogeneous Information Systems' (LITCHI) initiative [20]. Taxonomic intelligence may also be incorporated into existing information retrieval paradigms to identify organism knowledge, such as represented in literature.

The ability of an information retrieval system to reliably and accurately return results is measured according to two metrics: 'recall' and 'precision'. Recall, or sensitivity, assesses the ability to retrieve expected results; Precision, or the positive predictive value, is determined based on an assessment of number of correct results relative to the results that were retrieved. A taxonomic information retrieval tool with perfect recall would, therefore, need to identify all the variant forms of a given organism that exist within a knowledge base. The first task towards this goal will be the comprehensive collection of scientific names into a biological names register that may readily accommodate new classifications or nomenclature standards [47]. For example, if seeking all the literature in Medline associated with *Escherichia coli*, the query '*Escherichia coli*' should return articles that contain reference to its known variants (e.g. '*Bacterium coli*' and '*Bacillus coli*') or its misspellings (e.g. '*Escheria coli*'). However, performing Medline queries for '*Escherichia coli*,' '*Bacterium coli*,' '*Bacillus coli*,' or '*Escheria coli*' reveals that none of these retrieve the same number of results (at the time of this writing, they respectively return 233 339; 182 939; 182 907; and 22 citations). The only query that reliably retrieves all relevant results is '*Escherichia coli* OR *Bacterium coli* OR *Bacillus coli* OR *Escheria coli*' (233 384 citations). The respective recall values for retrieving documents from Medline using only *Escherichia coli*, *Bacterium coli*, *Bacillus coli* or *Escheria coli* are thus 99%, 78%, 78% and <1%. This finding implies that individuals seeking information from Medline would need to know all synonyms (including misspellings) for a given organism before querying PubMed.

Because Medline is manually (or semi-automatically) curated and indexed [48], one can assume that relevant articles are annotated with relevant MeSH terms. Querying Medline for those articles that have been annotated with the MeSH term '*Escherichia coli*' results in 178 461 citations. Based on this value, the precision for previous query for '*Escherichia coli*' is 76% (the fraction of those articles known to be associated with *Escherichia coli* relative to the total number of results returned, or 178 461 out of 233 384). While '*Escherichia coli*' is a MeSH term, neither '*Escheria coli*' nor '*Bacterium coli*' are. Although the PubMed interface generally expands queries to relevant terms (e.g. *Bacterium coli* for *Escherichia coli*), it is difficult to assess the absolute accuracy without manual examination. However, manual examination of the nine citations that contained the misspelled form, '*Escheria coli*', but not annotated with the MeSH term '*Escherichia coli*' reveals that 8/9 (89%) of them contain relevant knowledge. It is important to note that only through expert validation can articles be certified as containing relevant knowledge about a particular organism. Reliable validation will improve resources, and tools will emerge to curate knowledge such as that available through literature [49].

Table I: Current state of taxonomic information retrieval from medline

Domain	**Eukarya**				**Bacteria[M]**
	[349] (349)				[1 016 629] (201 741)
Kingdom	**Animalia[M]**			**Fungi[M]**	**Monera**
	[3 805 547] (51) {0}			[617 623] (42 192) {196 083}	[23] (23)
Phylum	**Chordata[M]**	**Arthropoda**	**Nematoda[M]**	**Ascomycota[M]**	**Eubacteria[M]**
	[11 664 676] (977) {11 664 654}	[176 348] (214)	[39 053] (6720) {38 820}	[86 266] (5885) {86 223}	[745 882] (1704) {0}
Class	**Mammalia**	**Insecta**	**Secernentea[M]**	**Saccharomycetes**	**Proteobacteria[M]**
	[11 441 596] (641)	[135 829] (842)	[31 250] (24) {31 245}	[35] (35)	[398 670] (2377) {397 754}
Order	**Rodentia[M]**	**Diptera[M]**	**Rhabditida[M]**	**Saccharomycetales[M]**	**Enterobacteriales**
	[1 931 565] (9 900) {1 931 476}	[81 574] (15 630) {81 038}	[9201] (212) {9156}	[67 171] (1714) {67 164}	[6] (6)
Family	**Muridae[M]**	**Drosophilidae**	**Rhabditidae**	**Saccharomycetaceae**	**Enterobacteriaceae**
	[1 836 137] (4231) {1 836 054}	[41 459] (166)	[41] (41)	[33] (33)	[252 917] (17 895)
Genus	***Mus*[M*]**	***Drosophila*[M]**	***Caenorhabditis*[M]**	***Saccharomyces*[M]**	***Escherichia*[M]**
	[26 311] (4292) {747 419}	[54 878] (54 882) {41 406}	[10 876] (10 876) {7626}	[76 757] (76 757) {62 990}	[231 562] (231 574) {177 469}
Species	***Mus musculus*[M*]**	***Drosophila melanogaster*[M]**	***Caenorhabditis elegans*[M]**	***Saccharomyces cerevisiae*[M]**	***Escherichia coli*[M]**
	[765 310] (2373) {747 419}	[25 476] (25 476) {21 475}	[10 507] (10 507)	[71 233] (71 233) {57 389}	[230 249] (230 249) {176 922}
Common Name(s)	***M. musculus***	***D. melanogaster***	***C. elegans***	***S. cerevisiae***	***E. coli***
	[26 311] (338)	[4039] (2872)	[8187] (4486)	[60 118] (26)	[199 314] (68 319)
	House Mouse	**Fruit Fly**	**Nematode Worm**	**Baker's yeast**	
	[765 099] (627)	[43 828] (371)	[43 169] (115)	[59 203] (1498)	
	Mouse	**Fruitfly**	**Worm**	**Yeast**	
	[811 316] (323 823)	[371] (371)	[120 671] (9605)	[140 803] (90 059)	
	Laboratory Mouse				
	[765 075] (618)				

[M]Term in MeSH (*translated by PubMed to "mice")
[] = Default PubMed Search
() = Quoted PubMed Search
{} = MeSH PubMed Search

The number of Medline articles retrieved via PubMed for five taxa at eight different taxonomic levels as well as common synonyms used to refer to the organisms (below heavy line). For each term used (shown in bold face with grey background), we performed three different queries: (i) 'Default'—the query term is entered into the PubMed interface as written; (ii) 'Quoted'—the query term is entered into the PubMed interface within quotes, thus preventing any expansions of the term and (iii) 'MeSH'—If the term is a MeSH term, specify to only search MeSH annotations within Medline.

Taxonomically intelligent tools that accommodate scientific name variation have been described in the context of harmonizing differing taxonomic hierarchies and species checklists [20]. The incorporation of taxonomic hierarchical categories is akin to how information retrieval searches can use the MeSH 'explode' feature to report content that is associated with a particular MeSH term and all of the granular terms that are inherited (e.g. if the MeSH term for 'Protozoan Infections, Animal' is chosen, the explode function will also include the more specific terms of 'Babesiosis', 'Cryptosporidiosis', 'Theileriasis', etc.). The same principle of 'hierarchical inclusion' should then naturally be extended to organize biological information for comparative studies—e.g. a developmental biologist working with a fruit fly (*Drosophila melanogaster*) may wish to identify literature pertaining to all *Drosophila*. However, taxonomic relationships between organisms tend to be less stable than traditional metadata keywords. To address this, synchronization methods are needed to interface between scientific names and their current taxonomic hierarchies. Exploring the retrieval of Medline literature for five 'model' organisms using a widely accepted taxonomic hierarchy reveals unexpected and inconsistent performance of hierarchical inclusion (Table 1). For example, more articles were recovered using the

default search of *Mus musculus* than with the more taxonomically inclusive term *Mus*.

There is, thus, a significant need for taxonomically intelligent information retrieval tools that enable one to identify literature at different levels of granularity according to taxonomic knowledge to address questions like, 'What parasites affect those organisms within the genus *Castor*?' The lack of taxonomic intelligence can limit the types of organisms that can be searched. For example, a *Giardia lamblia* (which is the organism associated with giardiasis, a common non-bacterial cause of diarrhea) researcher might want a list of other parasitic organisms that use the beaver as its host, since it is a mammal that spends much of its time in aquatic environments, which are a particular hotbed for a range of parasitic diseases. When using the search terms 'beaver' and 'parasites', PubMed performs a search for the MeSH term 'rodentia'. Such a general search term may impact the specificity of the returned citations (in this case, over 19 000 results are returned). To address this issue, PubMed does simultaneously perform a plain text word search of titles and abstracts. However, unless the author specifically uses the term 'beaver' in their title or abstract, PubMed will not retrieve the complete collection of relevant citations. Another approach to identify relevant articles on beavers and parasites would be to use the scientific name for the beaver. This presents another challenge, since there are two scientific names that are associated with beavers, depending on geographic location—the scientific name for the European beaver is *Castor fiber* (associated with 28 results pertinent to parasites), whereas in North America the name is *Castor canadensis* (associated with 10 results pertinent to parasites). Thus, PubMed does incorporate some taxonomic intelligence (e.g. it can reconcile 'beaver' to 'rodentia'); however, the lack of a complete taxonomy can preclude finer grained taxonomic searching.

PUTTING BIODIVERSITY INFORMATICS INTO ACTION—FEDERATED SEARCHES

There are billions of specimen records and observational data that exist in natural history collections worldwide and continue to grow, thanks to significant collection efforts [50, 51]. The Global Biodiversity Information Facility (GBIF; http://www.gbif.org/ [52]) and the Taxonomic Database Working Group (TDWG; http://www.tdwg.org) are organizations that strive to develop structured formats to represent and share biodiversity data. An overview of the emerging formats for biodiversity data have been recently reviewed elsewhere [23]. These structured data can be used to complement existing stores of genomic and biomedical knowledge (e.g. as stored in GenBank and Medline, respectively), leading towards the integration of knowledge across a range of biological resources.

The topic of knowledge integration in Biodiversity Informatics is rather timely—the recently funded Encyclopedia Of Life (EOL) project, which is inspired by E.O. Wilson [28], will depend on the development of the requisite informatics infrastructure to identify, validate and manage information such that they can be presented through a single portal. As EOL strives to create a Web site for all species known to be present on Earth, the scope of issues associated with organizing and linking data across the plethora of current and future repositories is immense. The ultimate goal of the EOL is to build a consumer-driven product that provides the most authoritative information on all species and the means to add, mine and analyze the information. The challenge is particularly acute, since biodiversity knowledge predominantly exists in collection institutions, especially natural history museums and herbaria. This knowledge includes studies on the evolution, speciation and distribution of life from around the globe.

The sheer volumes of data that are being produced across the entire spectrum of biology will undoubtedly make the traditional model of 'one-stop shopping' at centralized repositories a difficult proposition. Instead, a federated approach may become the only tractable alternative, where a single interface provides access to a number of repositories and other relevant data marts of knowledge. A number of resources might need to be consulted even for identification of relevant literature (i.e. not all relevant literature is indexed in Medline; some might be in more biology-centric indexing services, like BioONE; http://www.bioone.org/). Key to the development of federated searches is the accurate annotation of relevant content with controlled vocabularies or, even better, ontologies. Systematic annotation of conceptual entities within

a given database can facilitate the development of resources that can link knowledge across multiple databases [33].

Ultimately, if all data are represented or annotated in a systematic way, they can be automatically aggregated into portals that can serve as real-time collaboration environments for groups of experts from a range of disciplines with a common goal—for example, ecologists, taxonomists and other experts who are all interested in studying a single group of organisms. The identification of information across multiple resources requires universal anchors. As pointed out here, the organism and its name are central to almost all biological data. Geographic location information is also associated with much of biodiversity data. By linking relevant information associated with an organism and mapping it to geographic information (called 'georeferencing'), scientists can combine information from a range of data types (e.g. climatology and epidemiology) [34]. Such knowledge integration can further the understanding of the disease etiology and host epidemiology towards the development of prophylactic containment methods, vaccinations and treatments [53].

A number of prototype, federated search applications have been developed to bring together biodiversity content across a number of trusted resources. The Taxonomic Search Engine (TSE [36]), and the later ispecies (http://ispecies.org) applications demonstrate the ability to link information such as molecular data, phylogenetic trees and literature (including popular news feeds available through RSS). Like ispecies, the uBio Portal (http://portal.ubio.org) also brings together information such as images, in addition to providing links to relevant species-centric resources (e.g. for a fish like *Pomatomus saltatrix*, one might be pointed to FishBase for further information). Screen shots of the TSE and uBio Portal are shown in Figure 3. All of these federated resources store minimal content locally; instead, they rely on indices (which might be available through Web services, like SOAP) and annotated content to recover results. Perhaps the main advantage of developing federated services to link information across multiple resources is that no one organization or group has to maintain all information about all data. For example, rather than GenBank keeping track of all possible vernacular and scientific names for a given organism (including misspellings), these could be made available through a Web service (such as available to uBio NameBank, which provides a listing of both vernacular and scientific forms for nearly 2 million species). These types of federated interfaces could be customized for particular user needs (e.g. one may want to know what lethal organisms they might encounter on a forthcoming trip to a remote part of the world, or a conservationist might want to track trends of genomic data associated with species that are near extinction).

CLOSING THOUGHTS

Biodiversity informatics is starting to gain momentum as a scientific discipline. Biodiversity informatics is not meant to replace existing biological disciplines any more than bioinformatics is intended to replace 'wet-bench' work. Instead, biodiversity informatics aims to bring together relevant information into a form that can be used by biodiversity researchers. Furthermore, it strives to develop resources and services that may further the initiatives that can benefit from the use of biological data—from basic biology to biomedical science to general knowledge. The range of available data types and formats for biodiversity knowledge is humbling—e.g. climate, geographic and disease knowledge. There are a few anchoring metadata elements that can be used to link information across this array of knowledge resources. The organism and its name were presented here as one of the fundamental metadata elements that can be used to unify and link data alongside other metadata elements such as geographical information. The organization of content using organism names will facilitate the development of systems that are 'taxonomically intelligent', and may thus enable comparative biology inquiries at multiple granularities. On par with the scope of biodiversity, the promise of biodiversity informatics is immense. Through the development of federated search engines that enable the searching and navigation across multiple, disparate resources, biodiversity knowledge might be more readily put into multiple contexts—from the conservation of species on Earth to the health and well-being of our own species.

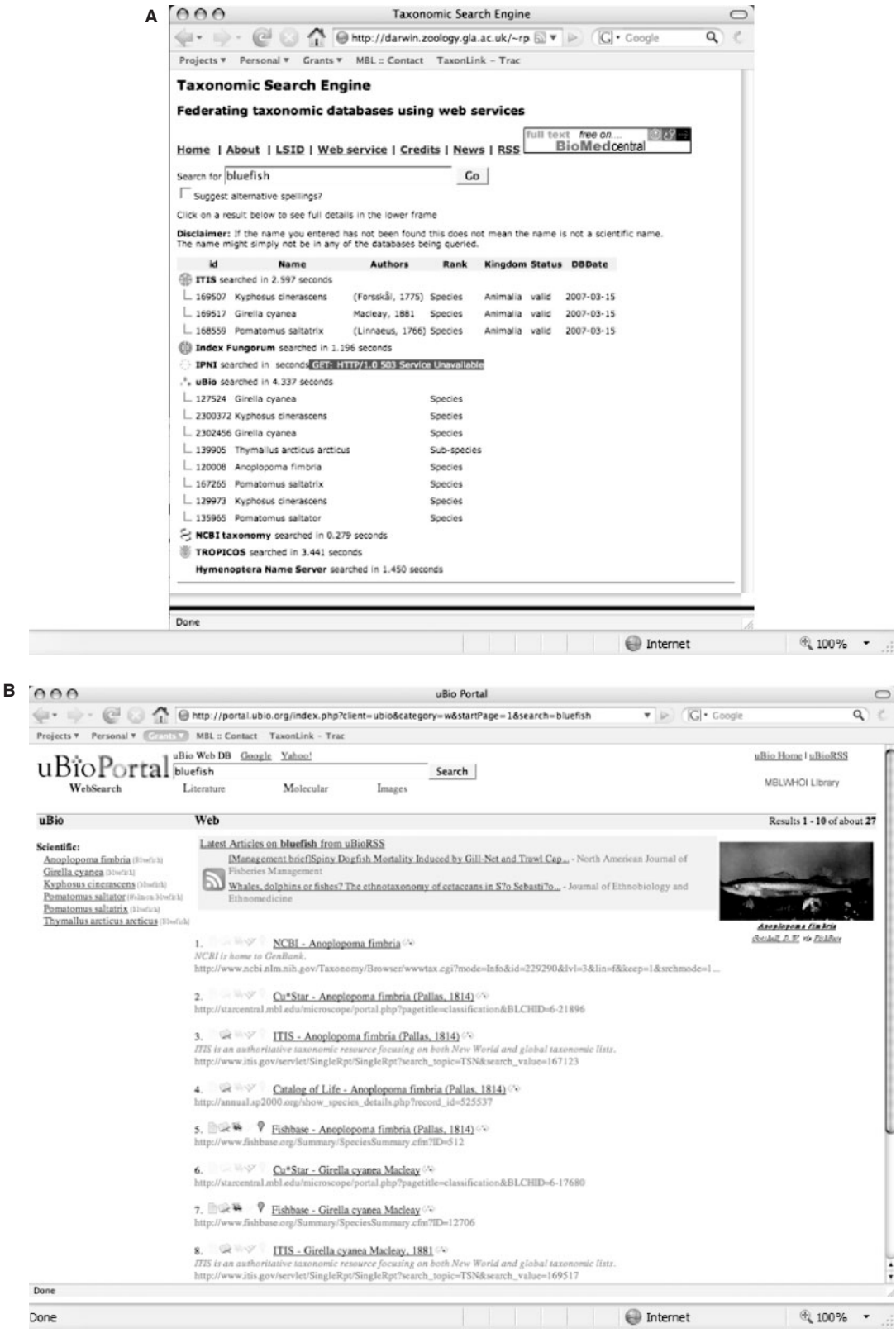

Figure 3: Biodiversity federated searches. Federated searches that dynamically identify and link relevant content from a range of resources have been prototyped by a number of interfaces. Shown here are the Taxonomic Search Engine (TSE) and the uBio Portal, both searching for content using the search term 'bluefish.'

Key Points

- Biological knowledge exists at three major levels: (i) molecular, (ii) organismal, and (iii) ecological. While bioinformatics focuses on facilitating molecular-based inquiries, biodiversity informatics aims to enhance inquiries at the organism level.
- The organism and its name are one of the few metadata elements that can link across a wide range of knowledge resources.
- The use of the organism name necessitates the development of name management systems to assist with disambiguation and reconciliation of name strings.
- The organization of knowledge according to organisms and their names can facilitate the development of 'taxonomically intelligent' information retrieval systems.
- The volume of biological data prohibits the easy development of aggregation databases to store all knowledge about all biology; federated interfaces may be a viable solution going forward.

References

1. Wilson EO, Peter FM. *'Biodiversity'*. Washington, D.C: National Academy Press, 1988.
2. Wilson EO. Systematics and the future of biology. *Proc Natl Acad Sci USA* 2005;**102** (Suppl 1):6520–1.
3. Kulikova T, Aldebert P, Althorpe N, *et al*. The EMBL nucleotide sequence database. *Nucleic Acids Res* 2004;**32**: 27–30.
4. Schuler GD, Epstein JA, Ohkawa H, *et al*. Entrez: molecular biology database and retrieval system. *Methods Enzymol* 1996;**266**:141–62.
5. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
6. Ashburner M, Lewis S. On ontologies for biologists: the Gene Ontology – untangling the web. *Novartis Found Symp* 2002;**247**:66–80; discussion 80–63, 84–90, 244–252.
7. Consortium GO. Creating the gene ontology resource: design and implementation. *Genome Res* 2001;**11**:1425–33.
8. Raychaudhuri S, Chang JT, Sutphin PD, *et al*. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res* 2002;**12**: 203–14.
9. Chen RO, Felciano R, Altman RB. RIBOWEB: linking structural computations to a knowledge base of published experimental data. *Proc Int Conf Intell Syst Mol Biol* 1997;**5**: 84–7.
10. Karp PD, Riley M, Saier M, *et al*. The EcoCyc and MetaCyc databases. *Nucleic Acids Res* 2000;**28**:56–9.
11. Selkov E, Galimova M, Goryanin I, *et al*. The metabolic pathway collection: an update. *Nucleic Acids Res* 1997;**25**:37–8.
12. Iliopoulos I, Enright AJ, Ouzounis CA. Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. *Pac Symp Biocomput* 2001;**6**:384–95.
13. Park JC, Kim HS, Kim JJ. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. *Pac Symp Biocomput* 2001;**6**: 396–407.
14. Yakushiji A, Tateisi Y, Miyao Y, Tsujii J. Event extraction from biomedical papers using a full parser. *Pac Symp Biocomput* 2001;**6**:408–19.
15. Yeh AS, Hirschman L, Morgan AA. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics* 2003;**19** (Suppl 1): i331–9.
16. Hirschman L, Park JC, Tsujii J, *et al*. Accomplishments and challenges in literature data mining for biology. *Bioinformatics* 2002;**18**:1553–61.
17. Agosti D. Biodiversity data are out of local taxonomists' reach. *Nature* 2006;**439**:392.
18. Soberon J, Peterson AT. Biodiversity informatics: managing and applying primary biodiversity data. *Philos Trans R Soc Lond B Biol Sci* 2004;**359**:689–98.
19. Blackmore S. Environment. Biodiversity update–progress in taxonomy. *Science* 2002;**298**:365.
20. Bisby FA. The quiet revolution: biodiversity informatics and the internet. *Science* 2000;**289**:2309–12.
21. Today we have naming of parts: a global registry of animal species could shake up taxonomy. *Economist*, 9 February, 2006.
22. McNeil DG. *'Hitting the Flu at Its Source, Before it Hits us'*. New York: New York Times, 2005.
23. Johnson NF. Biodiversity informatics. *Annu Rev Entomol* 2007;**52**:421–38.
24. Grimaldi DA, Engel MS. *Evolution of the Insects*. Cambridge: Cambridge University Press, 2005.
25. Greuter W, McNeill J, Barrie FR, *et al*. *International Code of Botanical Nomenclature (St Louis Code)*. Königstein: Koeltz Scientific Books, 2000.
26. Ride WDL, Cogger HG, Dupuis C, *et al*. *Internation Code of Zoological Nomenclature*. London: International Trust for Zoological Nomenclature, 1999.
27. Sneath PHA. *'International Code of Nomenclature for Bacteria'*. Washington, DC: American Society for Microbiology, 1992.
28. Wilson EO. The encyclopedia of life. *Trends Ecol Evol* 2003; **18**:77–80.
29. Polaszek A. A universal register for animal names. *Nature* 2005;**437**:477.
30. Hearings before the subcommittee on the Departments of Labor, Health and Human Services, Education, and Related Agencies of the House Committee on Appropriations, 99th Congress, 1st Session. Part 4B, (857), *Departments of Labor, Health and Human Services, Education, and Related Agencies of the House Committee on Appropriations*, 1985.
31. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
32. Gewin V. All living things, online. *Nature* 2002;**418**:362–3.
33. Schatz B, Mischo W, Cole T, *et al*. Federated search of scientific literature. *IEEE Computer* 1999;**32**:51–9.
34. Guralnick RP, Wieczorek J, Beaman R, *et al*. BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biol* 2006;**4**:e381.
35. Euzeby JP. List of Prokaryotic Names (http://www.bacterio.cict.fr/) 2006.
36. Page RD. A taxonomic search engine: federating taxonomic databases using web services. *BMC Bioinformatics* 2005;**6**:48.

37. Davidson SB, Overton C, Tannen V, *et al*. BioKleisli: a digital library for biomedical researchers. *Int J Digit Libr* 1997;**1**:36–53.
38. Cunningham H, Maynard D, Bontcheva K, *et al*. *'GATE: A framework and graphical development environment for robust NLP Tools and Applications'*. Philadelphia: Association for Computational Linguists, 2002.
39. Koning D, Sarkar IN, Moritz TD. TaxonGrab: Extracting taxonomic names from text. *J Biodiv Inform* 2005;**2**:79–82.
40. Sautter G, Boehm K, Agost D. A combining approach to final all taxon names (FAT). *J Biodiv Inform* 2006;**3**:46–58.
41. Yates TL, Mills JN, Parmenter CA, *et al*. The ecology and evolutionary history of an emergent disease: hantavirus pulmonary syndrome. *BioScience* 2002;**52**: 989–98.
42. Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA* 1994;**271**:1103–8.
43. O'Rourke A. Another fine MeSH: clinical medicine meets information science. *J Inform Sci* 1999;**25**:275–81.
44. Salton G. Another look at automatic text retrieval systems. *Commun ACM* 1986;**29**:648–56.
45. Srinivasan P. MeSHmap: a text mining tool for MEDLINE. *Proc AMIA Symp* 2001;642–6.
46. Patterson DJ, Remsen D, Marino WA, Norton C. Taxonomic indexing - extending the role of taxonomy. *Syst Biol* 2006;**55**:367-73.
47. Patterson DJ. Progressing towards a biological names register. *Nature* 2003;**422**:661.
48. Aronson AR, Mork JG, Gay CW, *et al*. The NLM indexing initiative's medical text indexer. *Medinfo* 2004;**11**: 268–72.
49. Eppig JT, Bult CJ, Kadin JA, *et al*. The Mouse genome database (MGD): from genes to mice – a community resource for mouse biology. *Nucleic Acids Res* 2005;**33**: D471–5.
50. Suarez AV, Tsutsui N. The value of museum collections for research and society. *BioScience* 2004;**54**:66–74.
51. Causey D, Janzen DH, Peterson AT, *et al*. Museum collections and taxonomy. *Science* 2004;**305**:1106–7.
52. Edwards JL, Lane MA, Nielsen ES. Interoperability of biodiversity databases: biodiversity information on every desktop. *Science* 2000;**289**:2312–4.
53. Daszak P, Cunningham AA, Hyatt AD. Emerging infectious diseases of wildlife–threats to biodiversity and human health. *Science* 2000;**287**:443–9.