BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 5. 345–354
Advance Access publication April 29, 2008
doi:10.1093/bib/bbn022

Biodiversity informatics: the challenge of linking data and the role of shared identifiers

Roderic D. M. Page

Submitted: 26th February 2008; Received (in revised form): 3rd April 2008

Abstract

A major challenge facing biodiversity informatics is integrating data stored in widely distributed databases. Initial efforts have relied on taxonomic names as the shared identifier linking records in different databases. However, taxonomic names have limitations as identifiers, being neither stable nor globally unique, and the pace of molecular taxonomic and phylogenetic research means that a lot of information in public sequence databases is not linked to formal taxonomic names. This review explores the use of other identifiers, such as specimen codes and GenBank accession numbers, to link otherwise disconnected facts in different databases. The structure of these links can also be exploited using the PageRank algorithm to rank the results of searches on biodiversity databases. The key to rich integration is a commitment to deploy and reuse globally unique, shared identifiers [such as Digital Object Identifiers (DOIs) and Life Science Identifiers (LSIDs)], and the implementation of services that link those identifiers.

Keywords: *biodiversity informatics; DOI; identifiers; knowledge integration; LSID; Semantic Web; taxonomy*

INTRODUCTION

Integrating diverse sources of digital information is a major challenge facing biodiversity informatics. Not only are we faced with numerous, disparate data providers, each with their own specific user communities, but also the information in which we are interested is diverse, and includes taxonomic names and concepts, specimens in museum collections, scientific publications, genomic and phenotypic data and images. Of course, the problem posed by integration is not unique to biodiversity informatics—the wider bioinformatics community is keenly aware of this challenge [1]. However, most bioinformatics integration efforts link together relatively few databases built upon similar data (e.g. macromolecular sequences and their annotations). At the time of writing the Global Biodiversity Information Facility (GBIF: http://www.gbif.org) lists some 217 different biodiversity data providers, serving a total of 145 660 886 records, mostly (but not limited to) museum specimens. The Catalogue of Life (http://www.catalogueoflife.org) contains over a million names contributed by 47 sources. If we add the contents of the 'traditional' bioinformatics databases GenBank and PubMed, along with the taxonomic literature accumulated since 1758 (much of it yet to be digitized), then the magnitude of the challenge facing biodiversity informatics becomes readily apparent. My purpose in this review is to outline the role that shared globally unique identifiers [2] might play in integrating these diverse sources of data.

Integration requires shared identifiers, in other words, a way to determine whether two items of data refer to the same entity or not. The obvious candidate for shared identifier is the taxonomic name of an organism [3]. It is a natural link between different databases that store information about that organism [4], and the basis of current tools that aggregate information from multiple sources, such as

Roderic D. M. Page, Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK. Tel: +44 141 330 4778; Fax: +44 141 330 2792; E-mail: r.page@bio.gla.ac.uk

Roderic D. M. Page is Professor of Taxonomy at the University of Glasgow. Perhaps best known as the author of phylogenetic computer programs such as TreeView, TreeMap, GeneTree, and COMPONENT, his recent work focuses on integrating phylogenetic, taxonomic, and specimen databases. He is the developer of the http://ispecies.org search engine and maintains a blog on phyloinformatics at http://iphylo.blogspot.com. From 2004 until the end of 2007, he was the editor of *Systematic Biology*.

iSpecies (http://ispecies.org) (Figure 1), and the Encyclopedia of Life (http://www.eol.org).

However, taxonomic names have serious limitations as identifiers in databases [5] as they are not completely stable, nor are they globally unique. Names may change due to taxonomic revision, there may be multiple names (synonyms) for the same taxon, and the same name may refer to different taxa (homonyms).

Some of the complexities of relying on taxonomic names to links records in different databases can be illustrated using the TbMap project [6] which links 52 778 names in the phylogenetic database TreeBASE (http://www.treebase.org) to names in a number of other databases, including the NCBI Taxonomy that underlies GenBank. Less than half the names in TreeBASE correspond exactly to a name in the NCBI Taxonomy database, and not all of those names that do match refer to the same taxon: *Loricaria* in TreeBASE is a plant, whereas *Loricaria* in GenBank is a catfish. Conversely, completely different names may refer to the same taxon. The plant name *Gastrolobium ebracteolatum* in TreeBASE has no match in GenBank, despite the authors of the TreeBASE study [7] listing two sequences from that plant. In GenBank these DNA sequences (AY015102 and AY015219) are listed under the name *Oxylobium lineare*. This is not an error—the two names are synonyms as a consequence of the nomenclatural contortions that result from moving *O. lineare* to the genus *Gastrolobium* [8].

Some of the problems encountered when using taxonomic names are a consequence of phylogenetic research outpacing taxonomic description, hence in sequence databases it is not uncommon to find taxa identified only to genus level or higher (e.g. *Drosophila* sp.). In poorly known taxonomic groups there will be many such taxa, consequently, it may not be clear which undescribed species is being referred to. This lack of formal names can hamper progress by leading to an unwitting duplication of effort. As an example, three different publications on ant phylogeny in the period 2004–06 have each submitted a 28S rRNA sequence obtained from California Academy of Sciences specimen casent0500379 (http://www.antweb.org/specimen.do?name=casent0500379) to GenBank (Figure 2), and each time the sequence has been recorded under a different taxonomic name, namely *Proceratium* sp. CS-2003-1 [9]; *Proceratium* sp. 1 CSM-2006 [10] and, *Proceratium* sp. Ma02 [11]. While the two papers published in 2006 appeared within four months of each other and so the authors may have been unaware of each others' work, the earlier 2004 paper [9] was cited by Moreau *et al.* [10].

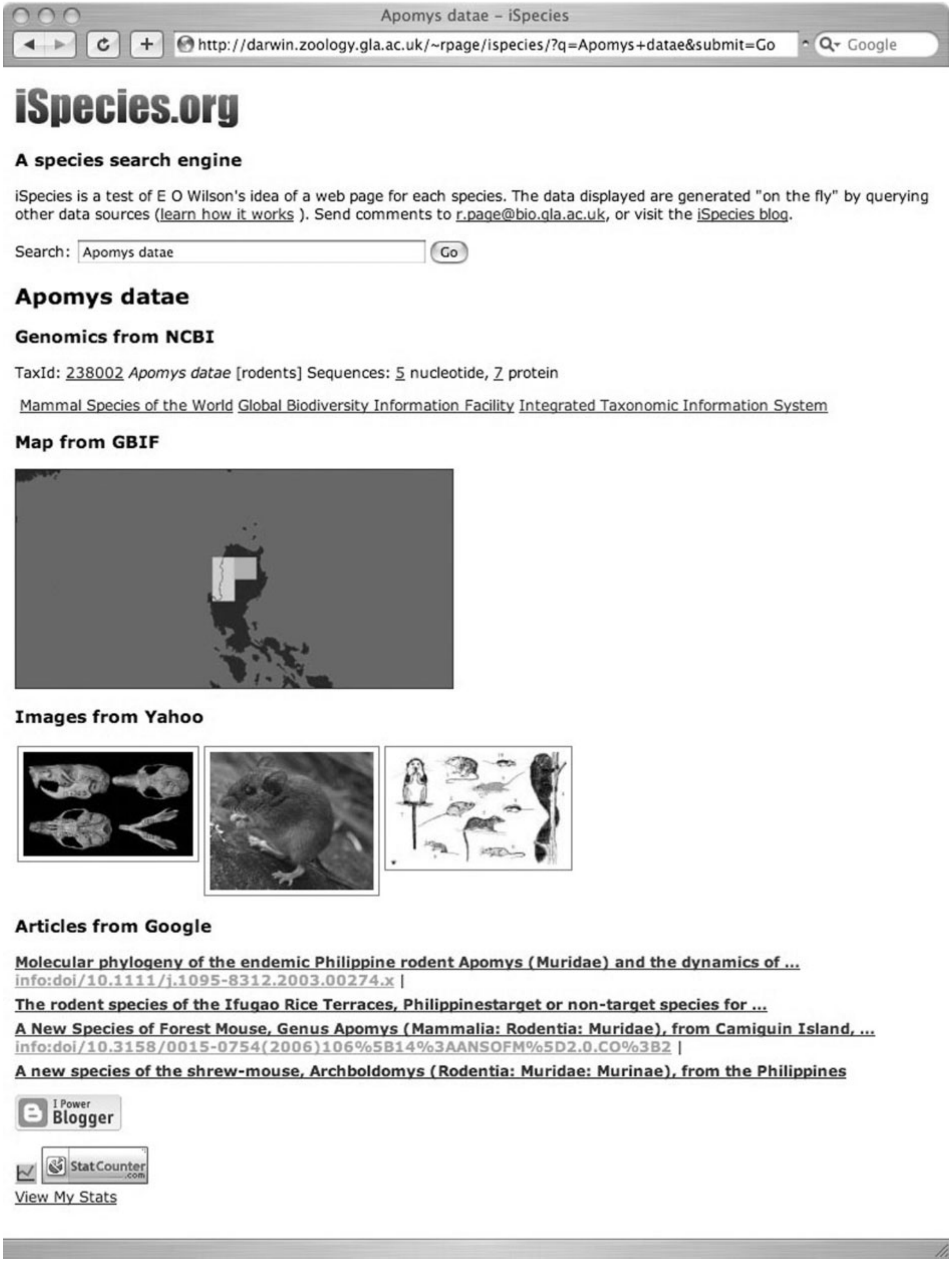
While independent confirmation of the sequence is comforting, this information is attached to different taxonomic names, and hence a user extracting data from GenBank using taxonomic names is unlikely to realize that these sequences are all from the same organism. In the same way, it will not be obvious to a user retrieving just the 28S rRNA sequence AY325951 that additional 18S rRNA and long-wavelength rhodopsin sequences are also available for this same specimen [10].

As a final example, consider a search in GenBank for sequences from the ant *Melissotarsus insularis*. At the time of writing this search finds no sequences. We are, however, not totally ignorant of the genome of this organism. If we search AntWeb (http://www.antweb.org) we discover some 288 specimens, all from Madagascar. One of these has the identifier casent0107663-d01, and the AntWeb record for this specimen informs us that it has been sent to a collaborator's lab for DNA barcoding. A literature search finds a paper on DNA barcoding Malagasy ants [12] that is accompanied by a supplementary spreadsheet of specimens sequenced, and this list includes casent0107663-d01 as the source of GenBank sequence DQ176312. If we then go back to GenBank and search on DQ176312, we discover it is from '*Melissotarsus* sp. BLF m1' (Figure 3). The obvious implication is that what GenBank refers to as '*Melissotarsus* sp. BLF m1' is *M. insularis* (Figure 4). Hence GenBank does, in fact, contain information on the genome of *M. insularis*.

These examples serve to illustrate two points. Firstly, databases and scientific publications contain a snapshot of knowledge at a given time, and are not always updated. Links between disconnected observations are made after the fact, if at all. Secondly, the links between these disconnected observations were made using shared identifiers such as sequence accession numbers and museum specimen codes. These identifiers play a crucial role in bringing together seemingly unrelated data from different sources.

IDENTIFIERS

A key prerequisite for integrating biological information from diverse sources is the use of globally


Apomys datae – iSpecies
http://darwin.zoology.gla.ac.uk/~rpage/ispecies/?q=Apomys+datae&submit=Go
Google
Go
I Power Blogger
StatCounter .com

iSpecies.org

A species search engine

iSpecies is a test of E O Wilson's idea of a web page for each species. The data displayed are generated "on the fly" by querying other data sources (learn how it works). Send comments to r.page@bio.gla.ac.uk, or visit the iSpecies blog.

Search: Apomys datae

Apomys datae

Genomics from NCBI

TaxId: 238002 *Apomys datae* [rodents] Sequences: 5 nucleotide, 7 protein

Mammal Species of the World Global Biodiversity Information Facility Integrated Taxonomic Information System

Map from GBIF

Images from Yahoo

Articles from Google

Molecular phylogeny of the endemic Philippine rodent Apomys (Muridae) and the dynamics of ...
info:doi/10.1111/j.1095-8312.2003.00274.x |

The rodent species of the Ifugao Rice Terraces, Philippinestarget or non-target species for ...

A New Species of Forest Mouse, Genus Apomys (Mammalia: Rodentia: Muridae), from Camiguin Island, ...
info:doi/10.3158/0015-0754(2006)106%5B14%3AANSOFM%5D2.0.CO%3B2 |

A new species of the shrew-mouse, Archboldomys (Rodentia: Muridae: Murinae), from the Philippines

View My Stats

Figure 1: iSpecies web site displaying information retrieved from separate queries to NCBI, GBIF, Yahoo and Google using the search term 'Apomys datae'.

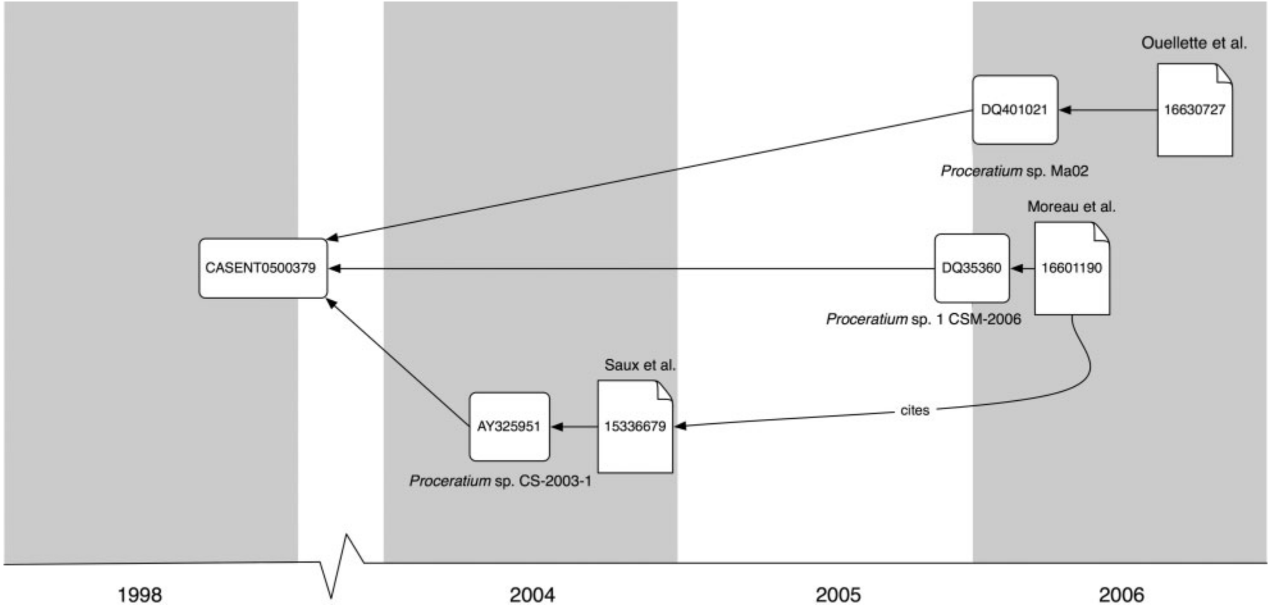


Figure 2: Sequencing history for the ant specimen casent0500379, which was collected in November 1998. Three 28S rRNA sequences have been obtained from the same specimen (casent0500379), published in three different papers [9–11], and deposited in GenBank using three different names for the ant. The sequences are placed on the timeline based on their date of submission, publications (identified by their PubMed number) by date of publication.

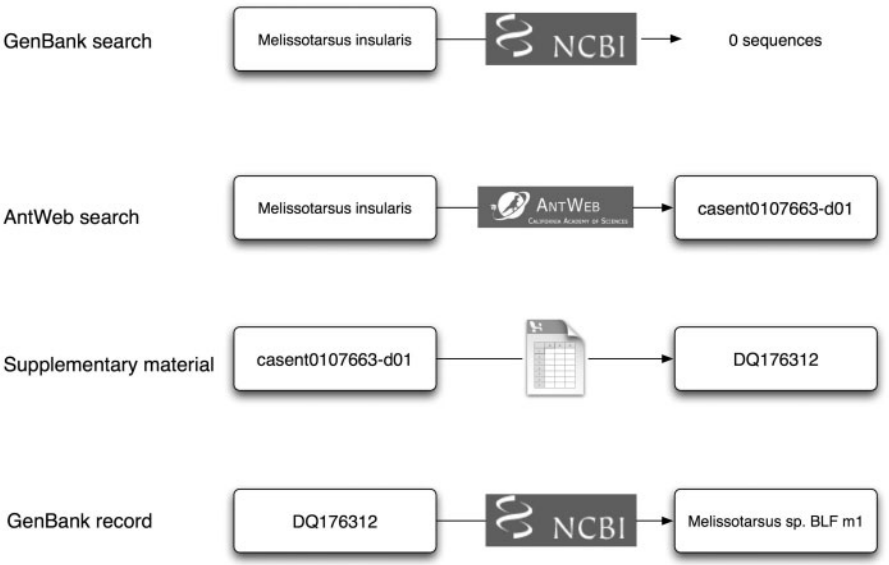


Figure 3: A search of GenBank for *Melissotarsus insularis* finds no sequences. However, a search of AntWeb finds a specimen listed as having been barcoded, and the paper publishing the barcodes has a supplementary table that lists the specimen as the source for sequence DQ176312. GenBank lists this sequence as being from taxon *Melissotarsus* sp. BLF m1.

unique identifiers (GUIDs) to consistently identify objects [2]. There are numerous schemes for generating such identifiers. Discussion within the biodiversity informatics community has focussed primarily on three alternatives, HTTP URIs (Uniform Resource Identifiers), Digital Object Identifiers (DOIs) and Life Science Identifiers (LSIDs) (http://wiki.tdwg.org/GUID).

HTTP URIs

HTTP URIs, better known as URLs, have the advantage of simplicity—all a user needs is a web

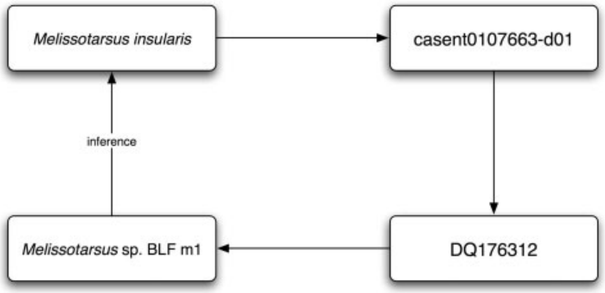


Figure 4: Linking the identifiers in Figure 3 together enables us to infer that the taxon GenBank refers to as *Melissotarsus* sp. BLF m1 is *M. insularis*.

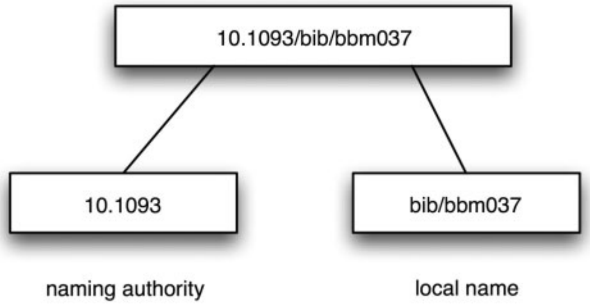


Figure 5: A doi (10.1093/bib/bbm037) and its constituent parts. This DOI identifies reference [3].

browser in which to enter the URL. They are also the identifier of choice for most Semantic Web projects [13]. Probably their greatest perceived weakness is their fragility. Parts of a URL may reflect the technology on the web server (e.g. 'may include file extensions such as '.php'), and if the web site owner changes the software used to serve web pages, the URLs may change. The enormous freedom to create, dispose or reassign URLs at will has contributed greatly to the speed of growth of the web, but as a consequence many URLs have a short life span [14].

DOI

One approach to deploying GUIDs is to provide a central authority for assigning and resolving identifiers. This is the strategy adopted by many academic publishers through CrossRef (http://www.crossref.org) which manages DOIs (http://www.doi.org) for journal articles. In some cases a field may be dominated by a single data provider that issues *defacto* GUIDs. The genomics community uses GenBank accession numbers to identify molecular sequences, and relies on the NCBI maintaining a stable API for retrieving information about those sequences.

The CrossRef model has a lot to recommend it. A DOI has two parts, the naming authority which is assigned centrally by CrossRef (and typically corresponds to a publisher), and the local identifier, assigned by the publisher (Figure 5). Because the naming authority is assigned centrally, no two publishers will have the same naming authority, and hence no two publishers will assign the same DOI to a different article. The identifiers are relatively opaque, in that it is not immediately obvious which DOI corresponds to which publisher. If one publishing company acquires another, they will also acquire the DOIs. Given that these are not explicitly 'branded' with the name of the original publisher, they can be reused as is. Indeed, this is one of the major attractions of DOIs in a fluid commercial landscape where publishing companies merge or rebrand. From the user's perspective, their ability to link to a resource is unaffected by changes in who provides that resource (note that the ability to *access* the resource may well change, but it will always retain the same identifier).

But the true value of CrossRef is not in providing persistent identifiers; rather it is the services it has built on top of those identifiers. The service most users are familiar with is resolution: by clicking on a hyperlink in a full text article, they are taken to the electronic version of that article. This service is provided by the Handle technology (http://www.handle.net) that underlies DOIs. What CrossRef adds to the Handle system are services, such as:

- given a DOI return metadata for the corresponding article (e.g., journal and article titles, volume, page numbers);
- given metadata for an article, return the corresponding DOI (if it exists).

The first service depends on publishers submitting metadata about each article to CrossRef, and forms the basis for tools such as Connotea (http://www.connotea.org), a social bookmarking service for scientists. A Connotea user need only paste in a DOI and Connotea retrieves the metadata from CrossRef, sparing the user the need to manually enter bibliographic details.

The second service enables the conversion of lists of literature cited in manuscripts to lists of links to the corresponding digital resources (identified by their DOIs). Many publishers are now submitting

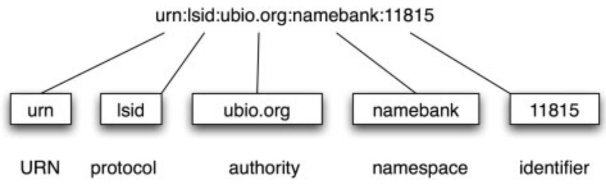


Figure 6: A LSID is prefixed with 'urn:lsid', then follows the authority, namespace and identifier components.

lists of DOIs cited in each manuscript to CrossRef. This enables 'forward linking'—the web page for an article can now display a list of articles that cite the original article. In effect, the web page is kept current and relevant well after its original publication.

LSIDs

LSIDs were developed to provide globally unique identifiers for objects in biological databases [2]. Within mainstream bioinformatics relatively few 'early adopters' have deployed LSIDs [15], possibly in part because core providers such as NCBI that provide stable identifiers and well-documented services have little incentive to add support for LSIDs. For the biodiversity informatics community the attractions of LSIDs include the distributed nature of the identifier (no central authority is required for registering or resolving identifiers), the low cost, and the convention that resolving a LSID returns metadata in Resource Description Framework (RDF). The latter facilitates integrating information from multiple sources using tools being developed for the Semantic Web [16], although, the mechanism for resolving LSIDs is not supported by existing Semantic Web tools. LSIDs are the identifier recommend by the Biodiversity Information Standards (TDWG) organization (http://www.tdwg.org).

Figure 6 shows an example LSID. Each LSID is prefixed by 'urn' indicating that the LSID is a uniform resource name (URN), 'lsid' indicates that the identifier is resolved using the LSID protocol, then follow the authority, namespace and identifier components (there may also be an optional revision component to indicate the version of the resource). The authority is a domain name that can be resolved by the Internet DNS (typically a domain name owned by the data provider), the namespace and identifier are specific to the data source that provides the resource. In this example the LSID is a taxonomic name in the uBio database (http://www.ubio.org). Note that the uniqueness of the LSID is in part guaranteed by the use of Internet domain names, which are globally unique. Providing that the data source ensures that each combination of namespace and identifier is unique within that data source, the LSID itself will be a globally unique identifier. By using the existing DNS infrastructure, LSIDs avoid the need to set up a new central naming authority.

Persistence

To be useful identifiers need to be persistent, that is, users can employ them safe in the knowledge that they won't change. Persistent identifiers are obvious candidates for inclusion in databases. For example, a developer of a database could store an identifier for the accepted name for each organism, and thus avoid the need to store all the associated data about that name. Unfortunately, not all biodiversity data sources are aware of the value of persistence. The Catalogue of Life project (http://www.catalogueoflife.org/) changes identifiers with each release—in 2006 the record number for the peacrab *Pinnotheres pisum* is 872170, whereas in the 2007 edition the record number is 3803555. Anybody populating a database with Catalogue of Life identifiers would have to completely rebuild their database with each release. Without a commitment to persistence it is unlikely that database developers will use identifiers provided by other projects.

Availability

Integration based on identifiers depends on identifiers being available for the objects of interest. At present the bulk of the records of interest to biodiversity informatics lack identifiers. To illustrate, I extracted 9152 bibliographic citations from the Hymenoptera Name Server database (which is dominated by papers about ant taxonomy), and looked up each reference in CrossRef's database using their OpenURL service. Only 251 references had DOIs, and these were concentrated in a few journals (Figure 7). Hence the bulk of the taxonomic literature on ants lies outside the mainstream digitally available literature. Unlike in most disciplines where the relevance of the bulk of the literature rapidly decays with age, in biological taxonomy publications centuries old may still be relevant, indeed vital. Hence the bias in Figure 7 towards more recently published papers is worrying.

Much of the ant literature shown in Figure 7 is accessible via URLs through antbase (http://www.antbase.org), and more recently through

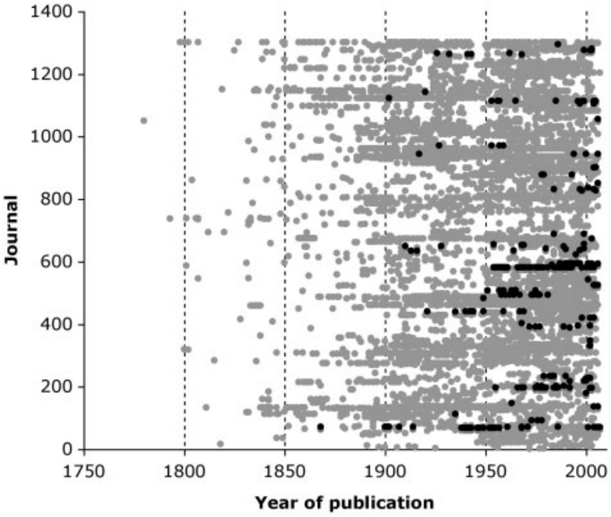


Figure 7: Digital accessibility of literature on ants and other hymenopterans. Each point represents an article in the Hymenoptera Name Server database. References that have DOIs are indicated by black dots.

a DSpace repository [17] hosted at Plazi (http://plazi.org). DSpace repositories assign Handles to objects. Although Handles are the technology that underpins DOIs, as discussed earlier, the latter have a much more developed infrastructure. Given any DOI, CrossRef's OpenURL resolver provides a single place a user can go to obtain metadata about the object, and a single place to determine whether a reference has a DOI. There are no such services for Handles, which means their utility is limited to that of being an identifier and little more. Given a Handle a user can retrieve a digital copy of an article, but without knowing the specifics of the local DSpace implementation, they cannot retrieve metadata about that article, nor can they readily discover whether a given article has a Handle.

Extracting identifiers

In many cases, a database may contain records that can be converted into identifiers. To illustrate, Figure 8 shows the metadata associated with sequence AY324463 for the Philippine rodent *Apomys datae* (the subject of the iSpecies query shown in Figure 1). In addition to the NCBI identifiers of accession number and taxon number, there are other fields that can potentially be converted into identifiers. Unlike many GenBank records, AY324463 is not linked to a corresponding entry in the PubMed literature database. However, there are sufficient bibliographic details to use an OpenURL [18] to find suitable identifiers. In this case CrossRef's OpenURL resolver (http://www.crossref.org/openurl) returns the DOI doi:10.1111/j.1095-8312.2003.00274.x, which identifies the publication by Steppan *et al.* [19]. Armed with this identifier we can find the electronic version of this publication, view it (if we have a subscription) or purchase it, as well as make use of any additional content displayed by the publisher. In this case, Blackwell list papers which cite Steppan *et al.* [19], giving us a further entry points into the scientific literature.

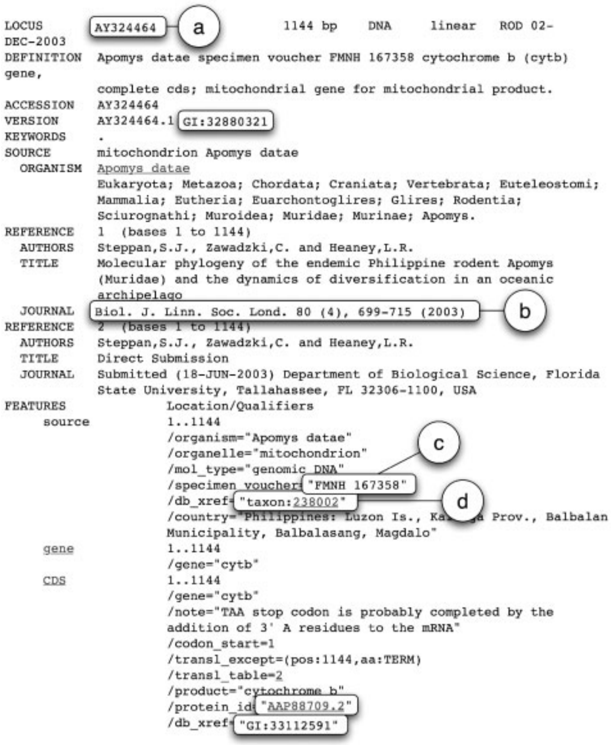

```
LOCUS       AY324464               1144 bp    DNA     linear   ROD 02-
DEC-2003
DEFINITION  Apomys datae specimen voucher FMNH 167358 cytochrome b (cytb)
gene,
            complete cds; mitochondrial gene for mitochondrial product.
ACCESSION   AY324464
VERSION     AY324464.1 GI:32880321
KEYWORDS    .
SOURCE      mitochondrion Apomys datae
  ORGANISM  Apomys datae
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Apomys.
REFERENCE   1  (bases 1 to 1144)
  AUTHORS   Steppan,S.J., Zawadzki,C. and Heaney,L.R.
  TITLE     Molecular phylogeny of the endemic Philippine rodent Apomys
            (Muridae) and the dynamics of diversification in an oceanic
            archipelago
  JOURNAL   Biol. J. Linn. Soc. Lond. 80 (4), 699-715 (2003)
REFERENCE   2  (bases 1 to 1144)
  AUTHORS   Steppan,S.J., Zawadzki,C. and Heaney,L.R.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-JUN-2003) Department of Biological Science, Florida
            State University, Tallahassee, FL 32306-1100, USA
FEATURES             Location/Qualifiers
     source          1..1144
                     /organism="Apomys datae"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /specimen_voucher="FMNH 167358"
                     /db_xref="taxon:238002"
                     /country="Philippines: Luzon Is., Kalinga Prov., Balbalan
                     Municipality, Balbalasang, Magdalo"
     gene            1..1144
                     /gene="cytb"
     CDS             1..1144
                     /gene="cytb"
                     /note="TAA stop codon is probably completed by the
                     addition of 3' A residues to the mRNA"
                     /codon_start=1
                     /transl_except=(pos:1144,aa:TERM)
                     /transl_table=2
                     /product="cytochrome b"
                     /protein_id="AAP88709.2"
                     /db_xref="GI:33112591"
```

Figure 8: Metadata in a GenBank record for the sequence AY324464 with identifiers, or potential identifiers highlighted. In addition, the sequence accession number (**a**) and taxon number (**d**), there are text strings that can be transformed into identifiers, such as a bibliographic citation (**b**) and a museum specimen code (**c**).

Similarly, the specimen voucher code has the abbreviation FMNH, which corresponds to the Field Museum of Natural History in Chicago (http://www.fmnh.org). The Field Museum's specimen records are accessible via its DiGIR provider (http://digir.sourceforge.net/), and so we can retrieve details about FMNH 167358, including the date of collection (4 April 2000) and the latitude and longitude of the collection locality (17° 27′ 30″N, 121° 4′ 6″E). This last piece of information enables us to display the specimen on a map.

Palimpsest

> ‘I likened the process of authoring a scientific paper to that of the creation of a palimpsest. Starting from original research results and working through the synthesis of a cogent explanation of the results or discovery, at each step the content becomes more abstracted from the original results, the previous work being “lost” to the reader.’ Leigh Dodds

Leigh Dodds’ provocative blog post (http://www.ldodds.com/blog/archives/000264.html) likened the process of authoring a scientific paper to creating a palimpsest. He suggested that the underlying data (which he equated to the underlying text, or *scriptio inferior*) might be more valuable than the final manuscript (the visible content). The practice of including lists of GenBank accession numbers and specimen codes in publications means that these could be extracted using text-mining tools, converted into resolvable identifiers and additional information extracted. In the *Melissotarsus* example presented above, it was some metadata attached to a specimen record, together with the supplementary material of a published paper that enabled the discovery that GenBank actually has sequences from *M. insularis*. It is disconcerting, therefore, that much of this data is consigned to the ghetto of ‘supplementary material’, where it lingers in proprietary formats such as Excel spreadsheets or Word documents, often identified by a URL, and hence vulnerable to loss [14].

LINKING

The primary motivation for linking disparate data sources is the discovery of new information. GenBank is primarily a database of molecular sequences, and supports queries that reflect its origins. Users typically search for sequences that are similar to a query sequence [20], retrieve sequences using an accession number, or browse taxonomically. By linking sequences to georeferenced specimens, for example, one could build a view of GenBank that could be queried geographically. Linking georeferenced specimens to a phylogenetic database such as TreeBASE, one could treat TreeBASE as a biogeographic resource, rather than simply a repository for evolutionary trees. The links also become a way of tracing the provenance of data.

But we can also exploit the structure of links themselves. A classic problem in information retrieval is ranking the results returned by a search engine. Algorithms such as PageRank, [21] and Hubs and Authorities [22] compute the rank of individual web pages as a function of incoming and outgoing links. The model of scientific citation directly inspired PageRank; it is not just the number of times a paper is cited but also the quality of those citations that affect the rank of a paper. The taxonomic community has long felt disadvantage by the perceived role of citation-based ‘impact factor’ in assessing the importance of taxonomic research [23], given that much of the taxonomic literature appears in relatively low impact journals. Rather than becoming embroiled in that debate, I think it more profitable to explore applying PageRank to graphs of links between biological objects. For example, specimen database queries may routinely return hundreds of matching records, with no obvious criterion with which to order those results. However, if we build a graph of all the connections between published papers, DNA sequences, images and specimens, we could compute the PageRank of each object and use that to rank the results.

To illustrate, AntWeb lists the 43 specimens of the ant species *Probolomyrmex tani* by specimen code ordered alphabetically. This gives the user no indication of which specimens might in some sense be more important than others. It would be tempting to defer to criteria such as whether a specimen was a type specimen or not, but this is a coarse criterion, and in many cases types in AntWeb are not flagged as such. Alternatively, we could regard actions such as photographing a specimen, extracting its DNA or listing it in the ‘material examined’ section of a paper as conferring value on a specimen, and base our ranking on those actions.

In the case, *P. tani*, the holotype (casent0041505) has been photographed (http://www.antweb.org/specimen.do?name=casent0041505), and three of these images have been included in the paper describing the species [24]. The holotype itself was also listed. We can depict these relationships using a graph where each object is a node, and each edge in the graph represents a link between those objects. The paper [24] is represented by a node labelled with the identifier hdl:10199/15374 (a Handle in the Plazi repository), and there are links to the images that comprise Figures 1, 3 and 5 in that paper, and these in turn are linked to the specimen they depict. There is also a direct link from the paper to the specimen, corresponding to the specimen’s presence in the list of material examined. For comparison, a specimen not figured but simply listed (casent0009766) gets a single incoming edge from the paper.

Another specimen listed in [24] has also been photographed, but these images are not included that paper. However, this specimen is the source of seven DNA sequences deposited in GenBank, and included in a phylogenetic analysis [25]. The complete graph is shown in Figure 9, along with the PageRank scores for each node. The specimen that has been both photographed and sequenced has the highest PageRank, and using this criterion would appear first in the list of specimens for *P. tani*. The graph shown in Figure 9 will grow as other data is obtained from these specimens, but also as the two papers [24, 25] are cited. Hence, as researchers confer value on the publications by citing them, this value will be transmitted down the graph to the underlying sequences, images and specimens.

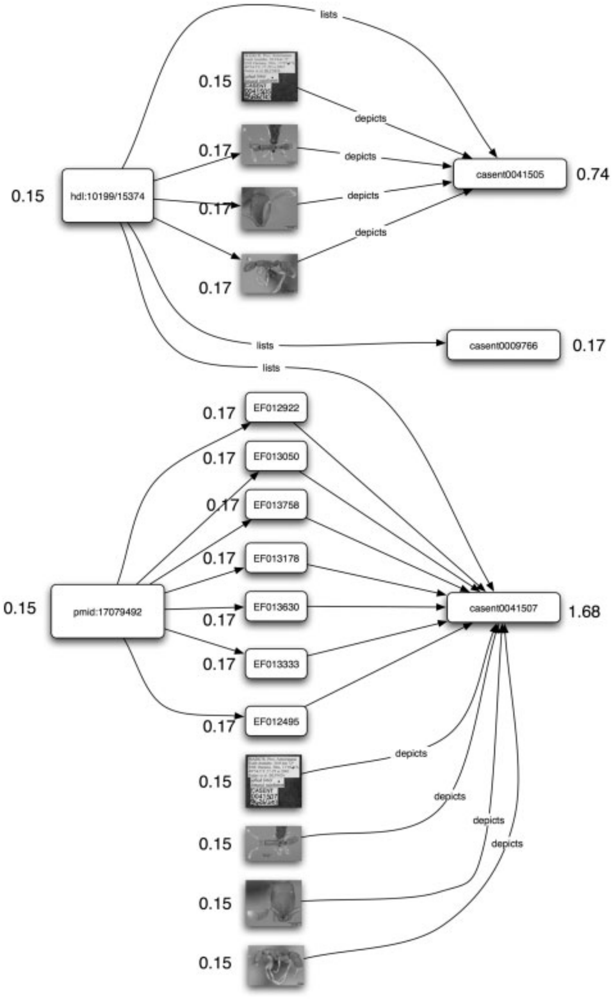


Figure 9: Graph depicting the links between three specimens of the ant *P. tani*, and the images, sequences and publications that refer to those specimens. Each node in the graph is labelled with its PageRank. The specimen that has been photographed and sequenced has the highest PageRank (1.68). The holotype of *P. tani*, casent0041505, has the next highest PageRank.

CLOSING THOUGHTS

One could argue that of all biologists, historically it is biodiversity researchers who have been the most aware of the need for globally unique identifiers. The infrastructure developed to formalize taxonomic names, conventions such as standardized abbreviations for museum collections [26] and authors [27] reflect this. What is needed in the digital age is a commitment to deploy and reuse globally unique, shared identifiers. Identifiers, by themselves, are of little benefit without services, in particular services that link identifiers. Given the diversity of data providers, such services will have to be able to consume a variety of identifiers (at a minimum HTTP URIs, Handles, DOIs, and LSIDs). A preliminary example of such a service is the bioGUID server (http://bioguid.info).

The relatively simple infrastructure CrossRef has built upon DOIs serves as a model of what can be achieved by adopting globally unique identifiers and basic services. CrossRef's task is simplified by being primarily concerned with documents of the same kind (articles, although they expand to include images and other data), and having a centralized infrastructure. In contrast, biodiversity data providers serve a broad range of data types, and are not likely to be centralized. To date most integration efforts (such as GBIF, NCBI's LinkOut and iSpecies) rely on shared taxonomic names to link data across multiple providers. As successful as this has been, a richer level of integration could be obtained through shared identifiers and services.

Key Points

- Most efforts at integrating biodiversity data integration to date have relied on taxonomic names as the shared identifier.
- Taxonomic names have limitations as identifiers, being neither stable nor globally unique.
- A wealth of information is associated with identifiers such as GenBank accession numbers, museum specimen codes and bibliographic citations.
- Integrating biodiversity resources depends on using shared global identifiers, and deploying services that link those identifiers.
- The graph of links between biodiversity data objects can be used to make new inferences between disconnected facts in different databases. This graph can also be used to compute the relative importance of these data objects using tools such as PageRank.

Acknowledgements

The author would like to thank David Patterson, David Remsen, Vince Smith and the three anonymous reviewers for their comments on the article. This work was supported by a grant from Biotechnology and Biological Sciences Research Council (BB/C004310/1).

References

1. Stein L. Integrating biological databases. *Nat Rev Genet* 2003;**4**:337–45.
2. Clark T, Martin S, Liefeld T. Globally distributed object identification for biological knowledgebases. *Brief Bioinform* 2004;**50**:59–70.
3. Sarkar IN. Biodiversity informatics: organising and linking across the spectrum of life. *Brief Bioinform* 2007;**8**:347–57.
4. Patterson DJ, Remsen D, Marino WA, *et al*. Taxonomic indexing – extending the role of taxonomy. *Syst Biol* 2006; **55**:367–73.
5. Kennedy J, Kukla R, Paterson T. Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In: Ludäscher B, Raschid L (eds). *Data Integration in the Life Sciences: Second International Workshop, DILS 2005*, San Diego, CA, USA, 20–22 July 2005 (Lecture Notes in Computer Science 3615). Heidelberg: Springer, 2005, 80–95.
6. Page RDM. TBMap: a taxonomic perspective on the phylogenetic database TreeBASE. *BMC Bioinformatics* 2007; **8**:158.
7. Crisp MD, Cook LG. Molecular evidence for definition of genera in the *Oxylobium* group (Fabaceae: Mirbelieae). *Syst Bot* 2003;**28**:705–13.
8. Chandler GT, Crisp MD, Cayzer LW, *et al*. Monograph of *Gastrolobium* (Fabaceae: Mirbelieae). *Aust Syst Bot* 2002;**15**: 619–739.
9. Saux C, Fisher BL, Spicer GS. Dracula ant phylogeny as inferred by nuclear 28S rDNA sequences and implications for ant systematics (Hymenoptera: Formicidae: Amblyoponinae). *Mol Phylogenet Evol* 2004;**33**:457–68.
10. Moreau CS, Bell CD, Vila R, *et al*. Phylogeny of the ants: diversification in the age of angiosperms. *Science* 2006;**312**: 101–4.
11. Ouellette GD, Fisher BL, Girman DJ. Molecular systematics of basal subfamilies of ants using 28S rRNA (Hymenoptera: Formicidae). *Mol Phylogenet Evol* 2006;**40**:359–69.
12. Smith MA, Fisher BL, Hebert PDN. DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philos Trans R Soc Lond* 2005; **360**:1825–34.
13. Sauermann L, Cyganiak R, Völkel M (2007), 'Cool URIs for the Semantic Web', Technical report, *DFKI Technical Memo.* http://www.dfki.uni-kl.de/~sauermann/2006/11/cooluris/.
14. Dellavalle RP, Hester EJ, Heilig LF, *et al*. Going, going, gone: lost Internet references. *Science* 2003;**302**:787–8.
15. Martin S, Hohman MM, Liefeld T. The impact of life science identifier on informatics data. *Drug Discov Today* 2005;**10**:1566–72.
16. Page RDM. Taxonomic names, metadata, and the Semantic Web. *Biodiversity Informatics* 2006;**3**. https://journals.ku.edu/index.php/jbi/article/view/25.
17. Smith M. DSpace: An open source institutional repository for digital material. *D-Lib Magazine* 2002;**8**, doi:10.1045/october2002-inbrief.
18. Apps A, MacIntyre R. Why OpenURL? *D-Lib Magazine* 2006;**12**:5.
19. Steppan SJ, Zawadzki C, Heaney LR. Molecular phylogeny of the endemic Philippine rodent Apomys (Muridae) and the dynamics of diversification in an oceanic archipelago. *Biol J Linn Soc* 2003;**80**:699–715.
20. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
21. Page L, Brin S, Motwani R, *et al*. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford University, 1998. http://dbpubs.stanford.edu/pub/1999-66.
22. Kleinberg JM. Authoritative sources in a hyperlinked environment. *J ACM* 1999;**46**:604–32.
23. Werner YL. The case of impact factor versus taxonomy: a proposal. *J Nat Hist* 2006;**40**:1285–6.
24. Fisher BL. A new species of *Probolomyrmex* from Madagascar. In: Snelling RR, Fisher BL, Ward PS, (eds). *Advances in Ant Systematics (Hymenoptera: Formicidae): Homage to E. O. Wilson – 50 Years of Contributions.* Memoirs of the American Entomological Institute, 2007;**80**:146–152.
25. Brady SG, Schultz TR, Fisher BL, *et al*. Evaluating alternative hypotheses for the early evolution and diversification of ants. *Proc Natl Acad Sci USA* 2006;**103**:18172–7.
26. Leviton AE, Gibbs RH, Heal E, *et al*. Standards in Herpetology and Ichthyology: Part 1. Standard symbolic codes for institutional resource collections in herpetology and ichthyology. *Copeia* 1985;**1985**:802–32.
27. Brummit RK, Powell CE. *Authors of Plant Names*. Royal Botanic Gardens, Kew, 1992.