Visualising a scientific article

Roderic D M Page*[1]

[1]Division of Ecology and Evolutionary Biology, Faculty of Biomedical and Life Sciences, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, UK

Email: Roderic D M Page*- r.page@bio.gla.ac.uk;

*Corresponding author

Abstract

This paper describes my entry in the Elsevier Grand Challenge "Knowledge Enhancement in the Life Sciences" contest. The entry takes a collection of fulltext issues of *Molecular Phylogenetics and Evolution* as the starting point, then extracts citation links to both papers and data, such as Genbank sequences and specimens, together with geotagged localities, and builds a "web" of objects linked by typed relationships. Each object (such as a publication, a sequence, a specimen, a taxon name, etc.) is treated equally, so that you can take a publication and see what taxa it refers to, or take the taxon and find all the publications that refer to the taxon. Although the database has been seeded with some articles from *Molecular Phylogenetics and Evolution*, much of the data comes from GenBank, PubMed, and specimen databases. These are accessed through http://bioguid.info, a tool I constructed to resolve globally unique identifiers and return associated metadata.

Background

Modern scientific publication

Scientific publishers have created a simple, robust system for identifying publications, and for linking publications together. Articles relationships include "citation", "bibliographic coupling", and "co-citation" (Fig. 1). Establishing these links requires a mechanism to uniquely identify a publication, and tools for finding appropriate identifiers from article metadata (for example, to convert the list of references in a

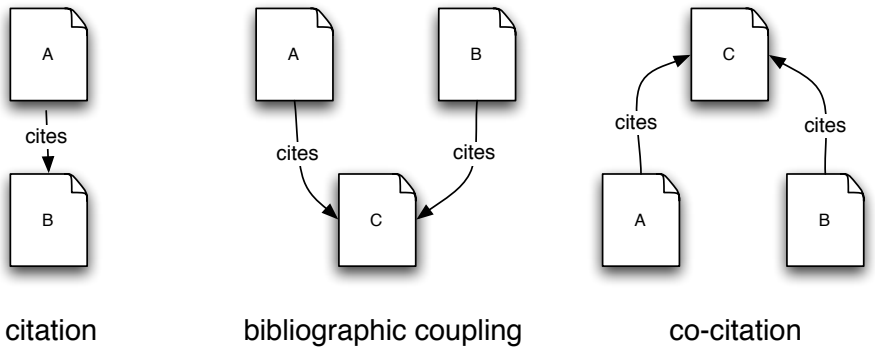

Figure 1: Relationships between publications. A document may cite another. Two documents (e.g., A and B) are bibliographically coupled if they cite a third (C). Similarly, co-citation occurs when two documents (e.g., A and B) are cited by a third (C).

manuscript into a list of identifiers). These services are provided by CrossRef [1], which stores basic metadata about a publication (such as journal title, volume, starting page number, first author), but can also store more comprehensive metadata, including lists of cited references. This enables CrossRef to provide "cited-by linking", so that a publisher displaying paper B can tell the user that B is cited by the more recent paper A.

In this system scientific publications are essentially opaque "black boxes", their digital content locked inside PDFs or HTML pages, reserved for those with access rights. All that is required is that articles be uniquely identified, and that lists of cited references can be extracted. Metrics based on the links between references (Fig. 1) can be used to quantify the value of an article (e.g., how many times it has been cited), which can be used to measure the value of the journal publishing those articles (e.g., "impact factor"), or the impact that an author is having on their field of research.

Data

In data-rich subjects such as molecular biology, taxonomy, systematics, and ecology, a paper may contain a wealth of potential links to data (such as DNA sequences, museum specimens, taxonomic names, ecological observations), many of which have a digital presence. There may also be additional, indirect links. For example, a paper may list DNA sequences from GenBank, but not the voucher specimens those sequences were obtained from. This potentially deprives museums from an opportunity to demonstrate the value of their collections by being able to list publications that use (or cite) data derived from material in their care. Some taxonomists have even argued that referring to a taxonomic name in a publication should be

accompanied by a literature citation [2], to counter balance the relatively low impact factor of taxonomic publications (for debate about the impact factor of taxonomic publ;ications see [3–5]).

An obvious extension to current publishing practice is to extend linking to these additional objects, thus building a web of data biodiversity data [6, 7]. Some publishers have made efforts in this direction. For example, BioOne [8] converts putative taxonomic names to links to ITIS [9] (although there is no guarantee that the name actually exists in ITIS, leading to some blind links). Some publishers encourage authors to mark-up their manuscripts in ways that facilitate link extraction. BioMed Central [10] requires GenBank accessions to be indicated using square brackets, e.g. [GenBank:U49845], and provides a discount on the cost of publication if authors use bibliographic software to construct their list of references.

However, there are significant obstacles to doing this extending linking beyond bibliographic citation, such as digital persistence. The Digital Object Identifiers (DOIs) used by publishers have an underlying infrastructure that supports their persistence. In contrast, many resources that are identified by URLs in scientific papers lack this infrastructure, and may disappear at a moment's notice [11]. Hence, publishers may be reluctant to populate their documents with links that may break.

There may also be significant amounts of data (and/or links to data) in supplementary appendices (typically stored in Word, Excel, or PDF files). Whereas publishers have a stake in maintaining the availability of the articles that they publish, they have less incentive to maintain access to underlying data, as evidenced by the frequent failure to adequately archive supplementary data [12].

Links

Many articles are published online with a rather limited set of connections to other digital entities, such as links to cited articles (identified by DOIs), and a link to the parent journal (identified by an International Standard Serial Numbers, ISSN). Authors themselves typically don't have identifiers.

Some indexing services have a richer set of links than those provided by publishers. NCBI PubMed, for example, links publications to molecular sequences (identified by accession numbers), which are in turn linked to taxon names in the NCBI taxonomy database (identified by NCBI taxonomy ids). In addition, some GenBank sequences are geotagged. PubMed has limited information on citations, derived from Open Access content in PubMed Central [13, 14]. Note that PubMed does not index all the papers that publish sequences, including a significant fraction of the phylogenetic literature. Consequently many sequences in GenBank are not associated with a PubMed identifier.

Missing from either Fig. 2 or Fig. 3 are connections to other major sources of biodiversity data, such as

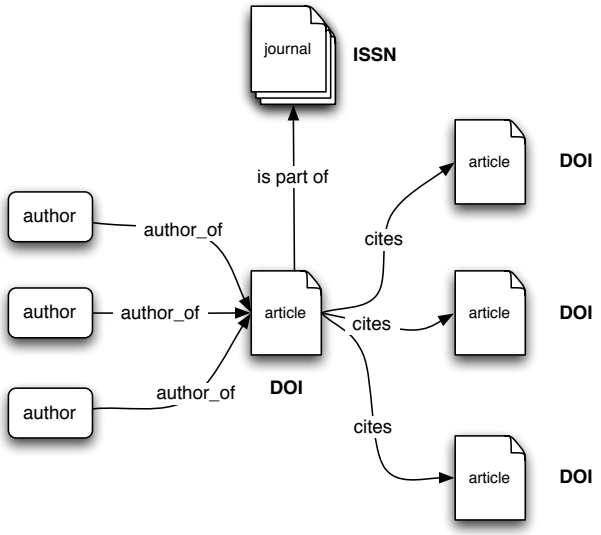

Figure 2: Model of an article published in a journal. The publication has one or more globally unique identifiers (such as a DOI), and is linked to other publications that it cites. The article is part of a journal, which is identified by an ISSN.

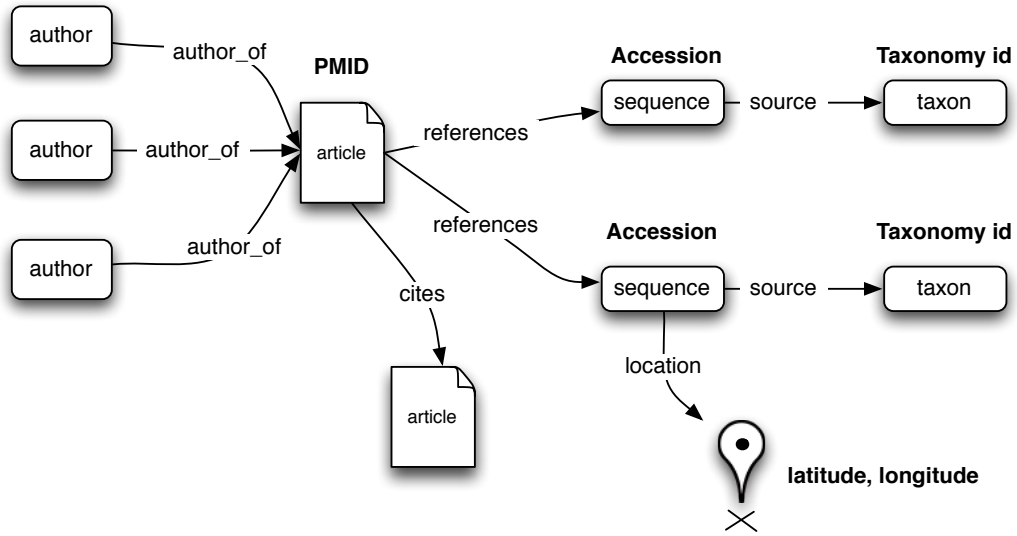

Figure 3: Model of a publication record in PubMed.

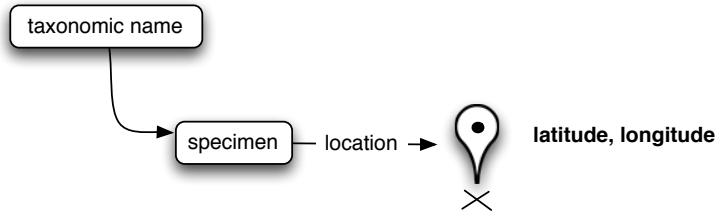

Figure 4: Model of a museum specimen, which may be geo-referenced, and which is typically associated with a taxonomic name.

digitised museum collections. Museum records are much more digitally isolated than records in publication or genomic databases (Fig. 4), in part because of the lack of a simple, resolvable identifiers for museum specimens. This is an obstacle to linking museum records to publications and sequences. Furthermore, specimens are tagged with taxonomic names, but not any widely used identifiers associated with those names (such as NCBI taxonomy ids).

Linking to museum specimens adds an important extra dimension to biological publication, beyond providing provenance for data [15]. In particular, many museum specimens have been georeferenced [16], providing a wealth of spatial data [17]. Linking sequences and publications to geo-referenced specimens will enable spatial queries to be performed, bringing an additional dimension to information retrieval from scientific literature [18].

Model

Combining information on publications (Fig. 2), sequences (Fig. 3), and specimens (Fig. 4) we arrive at a simple model that includes the key objects of interest in biodiversity informatics. Authors are linked to publications, which may be linked to other publications via citation links. Publications may cite nucleotide sequences and specimens (typically listed in tables in the body of the text), and may also list localities. Specimens may be listed in GenBank records, and both of these records may be geo-referenced. The record for a specimen lists the name of the corresponding organism, and GenBank records for parasites may list the host organism. Both of these names can be assigned an identifier using webservices provided by uBio [19][1].

Note that any one source may provide only a partial sets of links. PubMed records don't always list the associated GenBank sequences, and when they do they are usually just the sequences that are newly

[1]This model is incomplete. For example, authors may also be linked to taxonomic names, and names from uBio could be linked to NCBI taxon names. Some of these mappings are straightforward, some are not.

published by that paper, which may be a small fraction of the sequences actually analysed (phylogenetic studies typically build on previous work). GenBank records may omit the names of voucher specimens, which may instead be found in the publication. GenBank locations may be geo-referenced, but the museum records for a specimen may lack this information, and *visa versa*.

Methods

Assembling the demo is a three-step process, involving harvesting, assembling, and displaying data.

Harvesting

The core tool for populating the demo database second source of data comes from my bioGUID project [20], which provides an OpenURL [21] wrapper around data sources such as CrossRef, NCBI (PubMed and GenBank), uBio, and various museum DiGIR providers. As well as reformatting metadata provider by the original providers into JSON, the bioGUID OpenURL service attempts to populate the metadata with any additional identifiers and metadata. For example, a DOI will be linked to the corresponding PMID, publications in a GenBank record will be associated with their DOIs (if they exist), and taxonomic names in specimen records are linked to uBio namebankIDs.

The database was seeded with a full text collection of articles for *Molecular Phylogenetics and Evolution*. The XML documents were converted into a summary document in JSON format that listed the references cited, and any specimen codes, GenBank accession numbers, and latitude and longitude pairs that were listed in tables. The body of the article was not searched for these identifiers, as discovering identifiers within the text itself can be problematic[2]. All references found parsed to the bioGUID OpenURL resolver, which returned one or more identifiers (such as a DOI, Handle, PMID, or URL[3]). GenBank accession numbers and specimen codes were also resolved (only a subset of specimen codes could be interpreted as museum catalogue numbers).

Assembling

Harvested metadata are stored in an Entity-Attribute-Value (EAV) database [23, 24] implemented in MySQL. Object attributes (metadata) are stored in typed attribute tables (i.e., there are separate tables for

[2]In experiments with other documents, I discovered that a simple regular expression to extract GenBank identifiers found numerous accession numbers in a paper on millipedes [22], although that paper doesn't list any sequences. The putative accession numbers were UTM grid references for localities in Tasmania. In this case, the grid reference DQ402119 corresponds to the coordinates 41° 26' 31" S 146° 17' 02" E, but is also a GenBank sequence for human herpesvirus.

[3]The bioGUID OpenURL resolver has been populated with several thousand references that are not in CrossRef, and hence can return identifiers even if CrossRef's OpenURL resolver doesn't.

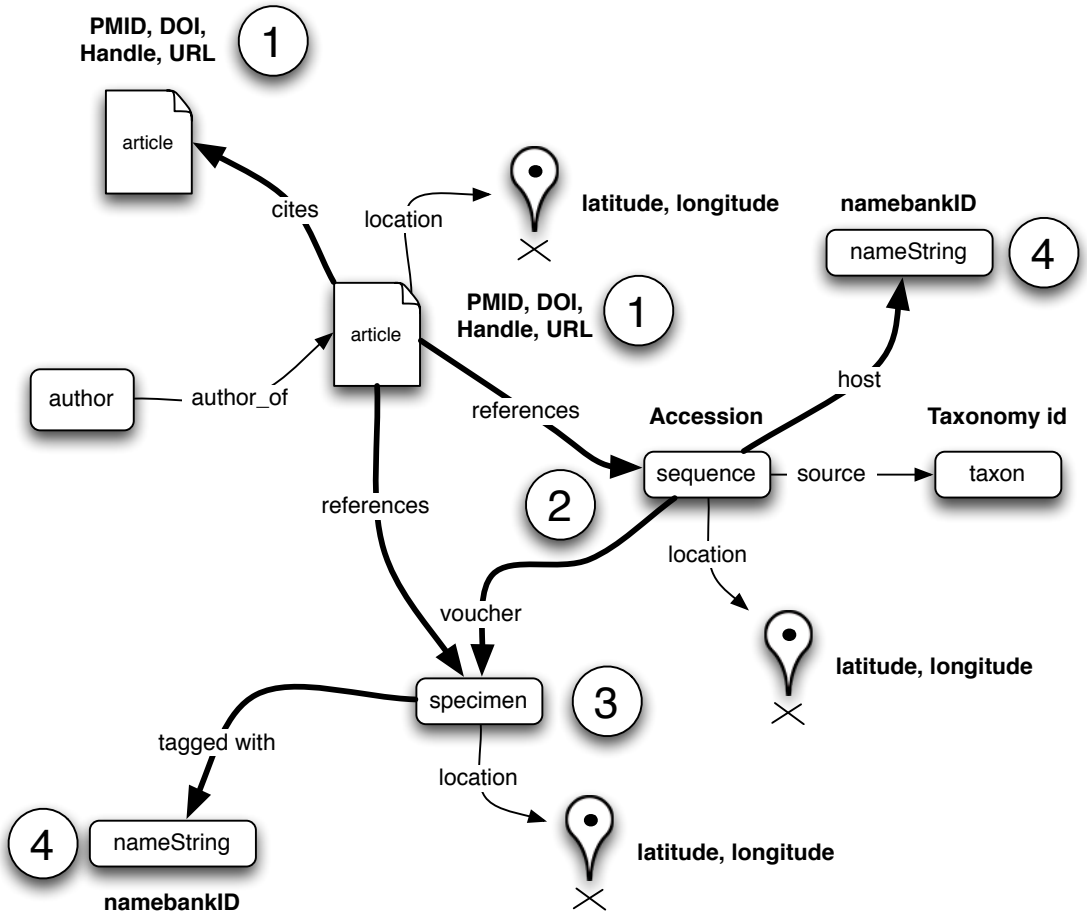

Figure 5: Data model linking publications, authors, sequences, specimens, taxonomic names, and locations. Thick lines indicate relationships that have to be extracted from the text of an article, or from PubMed, GenBank, or specimen records. Circled numbers indicate identifiers that can be obtained for the corresponding object. These include (1) DOIs, PubMed identifiers (PMIDs), Handles, or URLs for publications; (2) GenBank accession numbers (either listed in the publication, or obtained from NCBI); (3) specimen codes (either listed in the publication, or in the GenBank record); and, (4) uBio namebankIDs for taxonomic names.

dates, integers, real numbers, strings, and large objects of text). A separate table lists all available globally unique identifiers (GUID) for each object (e.g., DOIs, PMID, Handles, GenBank accession numbers, etc.). Each object is assigned a unique internal identifier, which is a 32 digit hexidecimal MD5 hash of the default GUID for the object[4]. To avoid duplication due to the same publication being added from different sources (for example, from a publisher's web site where it is identified by a DOI, and from PubMed where it is identified by a PMID) a Journal Article Citation Convention (JACC) identifier [25] is created using the journal ISSN, volume, and starting page. Two articles with the same JACC are regarded as the same. Typed links between objects are stored in a separate table. Locality information is stored in a separate table with a MySQL SPATIAL index, permitting basic spatial queries. To support taxonomic queries a nested-sets representation of the NCBI Taxonomy was created [26].

Display

Figure 6 shows a screen shot of a typical article display. In addition to the typical information one might expect about an article (such as bibliographic metadata and lists of references cited), the page lists articles that are related by geography, taxonomy, and data.

Data coupling

In addition to classical links between papers, such as citation (Fig. 1), papers may be linked by shared data, that is they refer to one or more GenBank sequences or specimens in common (analogous to bibliographic coupling, Fig. 1 [27]). To explore this the web page has a section entitled "Shares data with" lists papers that refer to sequences referenced by the article being displayed (note that these may or be papers cited by, or that cite the current article).

Taxonomy

Taxonomic information is displayed in a variety of ways. The taxa in a study are listed on the article web page. In order to convey a gestalt sense of what an article is about, I use a treemap [28] to display a set of images of the taxa in the article. The size of each cell in the treemap is proportional to the number of taxa in that group (typically a genus).

Each NCBI taxon has its own separate page, which The NCBI classification is displayed using a PygmyBrowser [29]. In the current version of the demo this browser is interactive, but doesn't enable the

[4]Web sites such as Connoteas and delicious use this approach

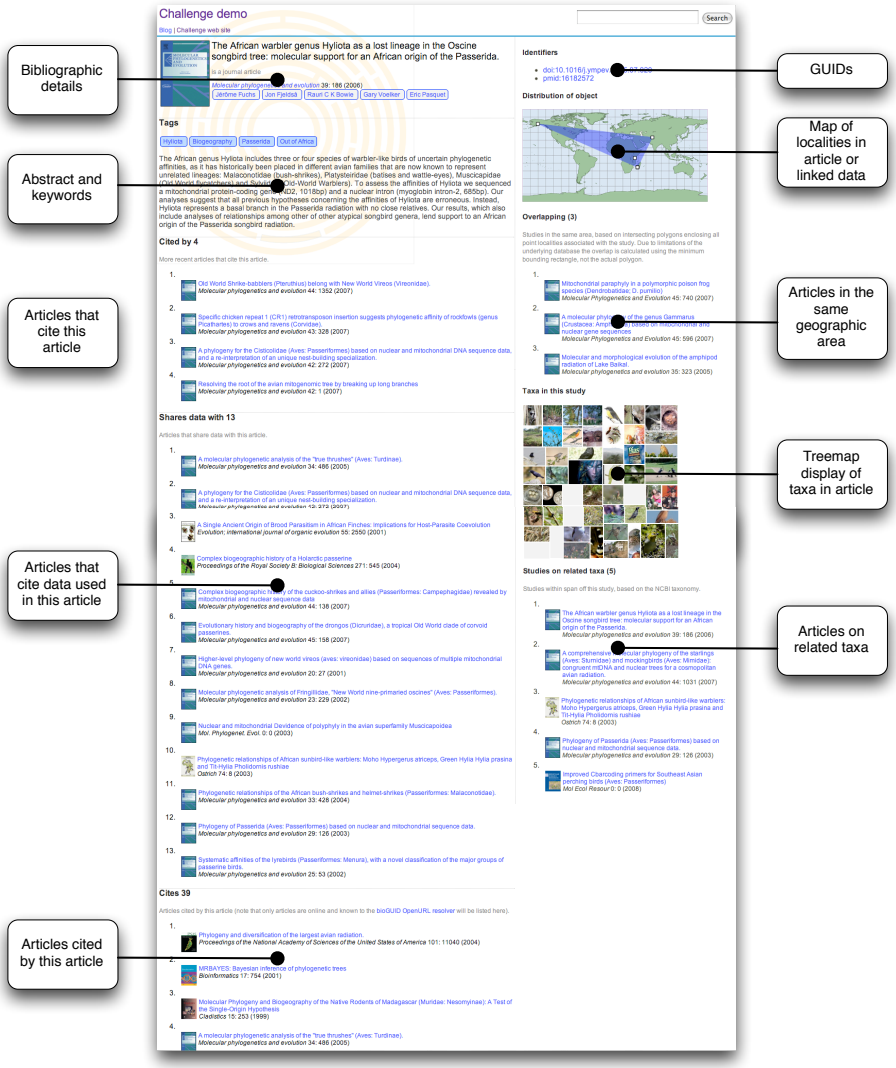

Figure 6: Screen shot showing one article being displayed, together with a map, taxonomic summary, and links to other studies.

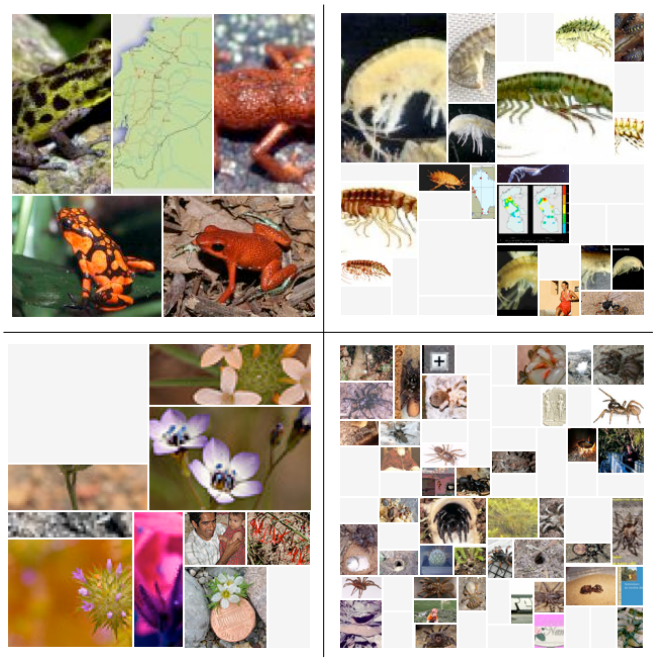

Figure 7: Four examples of treemaps summarising the taxa in an article.

user to do anything with it.

Geotagging

Localities extracted from tables in the articles, or via objects that are georeferenced (Fig. 8) are used to generate a map of all localities relevant to the article. The web page dispalys up to five studies that occur in the same geographic area as the article being displayed.

Results and Discussion

The demo is available at http://iphylo.org/~rpage/challenge/www/. The front page lists some starting points for browsing. Individual articles can also be viewed by appending the DOI to http://iphylo.org/~rpage/challenge/www/doi/, for example http://iphylo.org/~rpage/challenge/www/doi/10.1016/j.ympev.2007.06.010.

Outstanding issues

Automatic extraction of identifiers can run into problems. As noted above, GenBank identifiers can be identical to UTM grid references. Museum specimen codes are written in a variety of styles, and there can

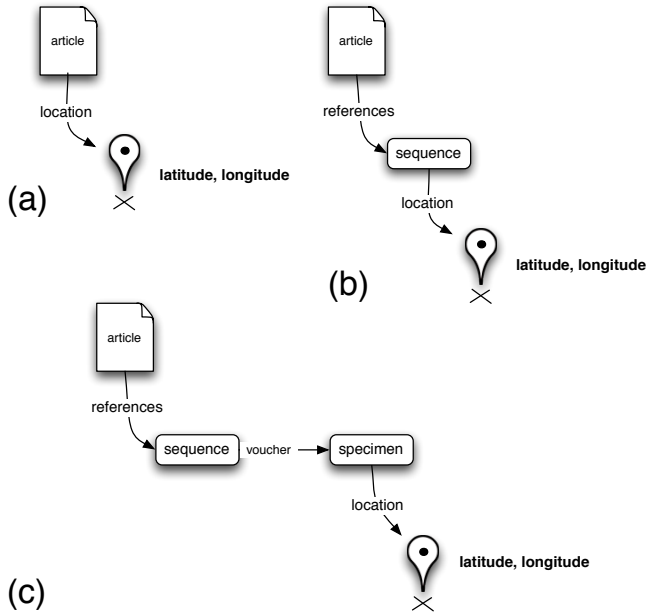

Figure 8: Adding geotags to an article via (a) localities listed in the article itself, from (b) georeferenced sequences in GenBank, and (c) georeferenced specimens linked to sequences linked to the article.

be enormous variation in how latitude and longitudes are written, both individually, and when written as a pair of co-ordinates:

- 23°03'44"N
- N10° 54.448'
- 35° 56.218′ N
- 5° 67'
- S 9°3'; W 72°44'
- 36°57'N, 10°37'E
- 115.59E/37.64N
- 39:49:35N; 3:08:50E
- S 4.45' W 73.57'

In addition, there is considerable variation in how latitude and longitude pairs are reported in tables in *Molecular Phylogenetics and Evolution* XML documents. Authors may include the (latitude, longitude) pair in a single column, split it between two columns, or put one value under another. In some cases the individual values include information on which hemisphere they refer to, in others this information is in the table header. Taken together this makes parsing georeferences somewhat challenging, and is a good argument for authors (and copy editors) adopting a standard way to include this information in manuscripts.

Search is very crudely implemented in the demo. Very simple full text searching is available for text, implemented using MySQL's build in FULLTEXT index. Spatial searching relies on MySQL's spatial extensions, which are very limited. For example, the search for overlapping polygons is actually implemented as a search for overlapping minimum bounding rectangles. Even if this issue is addressed, the use of bounding polygons[5] is in itself too crude to adequately represent a distribution. The taxonomic search returns taxa within the taxonomic span of the article. However (and somewhat analogously to the bounding polygon), if a study includes a few disparate taxa (say, as outgroups), then the list of taxa returned may be more diverse than the user anticipates. All these search and indexing problems could be resolved, in time.

Future

In some ways the demo makes very limited use of the full text documents. This was both to keep the task manageable, but also to see what could be achieved without, in a sense, the publisher handing over it's crown jewels (i.e., the full text) to the reader. Much of the database created here could have been populated from sources such as PubMed, without access to *Molecular Phylogenetics and Evolution* full text at all, although it would lack the citation links, and many GenBank and museum links, as well as some locality information. Some citation links could be recovered from PubMed, and for some purposes (such as using citation links to improve information retrieval [30]) incomplete citation data can still be useful [31]). The key point is that much of the value of the demo comes from a small fraction of the information in the journal articles. If for each article, publishers output metadata listing the papers cited, GenBank accession numbers, specimen codes, and any latitude-longitude pairs, then the demo could be created without requiring free access to the underlying text. The task facing the publisher, then, is to extract the metadata and make it available. Given the potential for errors in this process, it would be desirable to provide

[5]The bounding polygon is a computed as a convex hull enclosing all localities linked to an article.

authors with simple tools to mark up their manuscripts, for example by flagging GenBank accessions, museum codes, and latitude and longitudes (the later being written in a single format).

The demo has scratched the surface of possible visualisations. The original proposal [32] mentioned the use of Google Earth to display phylogenies. This is technically achievable, but preliminary work suggests that extracting trees from bitmap images in journal pages is difficult, and I abandoned this task. Google Earth itself could easily be used to create a visualisation where localities mentioned in *Molecular Phylogenetics and Evolution* papers are shown, and when the user clicks on a placemark, a link to the relevant articles, sequences, and specimens is dislayed.

Why Open Access?

Given that making just a little more metadata available can significantly enhance what publishers (and their readers) could do with an article, even without making the full text freely available, one might wonder what is gained from moving to Open Access[6]. From the purely narrow point of this demo, the major advantage would be ease of detecting and correcting errors. Extracting identifiers and geotags is error prone, and having the full text available for annotating would mean that readers could correct these (through, for example, a wiki-style interface [34]). Furthermore, papers contain errors. On several occasions the demo found cases where GenBank accession numbers were clearly incorrect. For example, [35] is a paper on bryozoans, yet the demo linked this study to *Homo sapiens*. This is a result of a table in the paper listing the incorrect GenBank accession numbers (AJ711044-50 should have been AJ971044-50). In the same way, [36] is a study on birds, yet contains a stray fish sequence due to an error in one of the tables. The existing model of relying on authors detecting these errors and arranging for errata to be published is inefficient, and means that many errors in the scientific literature are likely to go undetected and uncorrected.

Abbreviations

- DOI, Digital Object Identifier
- GBIF, Global Biodiversity Informatics Facility
- GUID, Globally Unique Identifier

[6]There seems to be some confusion among publishers as to what "Open Access" means. See [33] for an explicit statement that licenses such as the Creative Commons Attribution License and the Creative Commons Attribution Non-Commercial License are consistent with Open Access. Other licenses permit "Free Access," which merely provides access to the text at no cost. This is not Open Access.

- ISSN, International Standard Serial Number
- JACC,Journal Article Citation Convention
- JSON, JavaScript Object Notation
- PMID, PubMed Identifier
- URL, Uniform Resource Locator

Acknowledgements

This paper is part of an entry into the Elsevier Grand Challenge "Knowledge Enhancement in the Life Sciences". I thank Anita de Ward and Noelle Gracy for facilitating access to the digital copies of *Molecular Phylogenetics and Evolution*.

References

1. **CrossRef** [http://www.crossref.org/].
2. Werner YL: **The case of impact factor versus taxonomy: a proposal**. *Journal of Natural History* 2006, **40**:1285–1286, [http://dx.doi.org/10.1080/00222930600903660].
3. Krell FT: **Impact factors aren't relevant to taxonomy**. *Nature* 2002, **405**:507–508, [http://dx.doi.org/10.1038/35014664].
4. Krell FT: **Why impact factors don't work for taxonomy**. *Nature* 2002, **415**:957, [http://dx.doi.org/10.1038/415957a].
5. Garfield E: **Taxonomy is small, but it has its citation classics**. *Nature* 2002, **413**:107, [http://dx.doi.org/10.1038/35093267].
6. Page RDM: **Taxonomic names, metadata, and the Semantic Web**. *Biodiversity Informatics* 2006, **3**, [http://jbi.nhm.ku.edu/index.php/jbi/article/view/25].
7. Page RDM: **Biodiversity informatics: the challenge of linking data and the role of shared identifiers**. *Brief Bioinform* 2008, **9**(5):345–354.
8. **BioOne** [http://www.bioone.org/].
9. **Integrated Taxonomic Information System** [http://www.itis.gov/].
10. **BioMed Central** [http://www.biomedcentral.com/].
11. Dellavalle RP, Hester EJ, Heilig LF, Drake AL, Kuntzman JW, Schillin MGLM: **Going, Going, Gone: Lost Internet References**. *Science* 2003, **302**:787, [http://dx.doi.org/10.1126/science.1088234].
12. Evangelou E, Trikalinos T, Ioannidis J: **Unavailability of online supplementary scientific information from articles published in major journals**. *The FASEB Journal* 2005, **19**:1943–1944, [http://dx.doi.org/10.1096/fj.05-4784lsf].
13. Roberts RJ: **PubMed Central: The GenBank of the published literature**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(2):381–382.
14. **PubMed Central** [http://www.pubmedcentral.nih.gov/].
15. Peterson AT, Moylea RG, Nyária AS, Robbins MB, Brumfield RT, Remsen J: **The need for proper vouchering in phylogenetic studies of birds**. *Molecular Phylogenetics and Evolution* 2007, [http://dx.doi.org/10.1016/j.ympev.2007.08.019].

16. Guralnick RP, Wieczorek J, Beaman R, Hijmans RJ: **BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data**. *PLoS Biology* 2006, **4**:e381, [http://dx.doi.org/10.1371/journal.pbio.0040381].

17. **Global Biodiversity Informatics Facility (GBIF)** [http://www.gbif.org/].

18. **A place for everything**. *Nature* 2008, **453**:2, [http://dx.doi.org/10.1038/453002a].

19. **Universal Biological Indexer and Organizer (uBio)** [http://www.ubio.org/].

20. **bioGUID** [http://bioguid.info/].

21. de Sompel HV, Beit-Arie O: **Open Linking in the Scholarly Information Environment Using the OpenURL Framework**. *D-Lib Magazine* 2001, **7**(3), [http://dx.doi.org/10.1045/march2001-vandesompel].

22. Mesibov R: **The millipede genus *Lissodesmus* Chamberlin, 1920 (Diplopoda: Polydesmida: Dalodesmidae) from Tasmania and Victoria, with descriptions of a new genus and 24 new species**. *Memoirs of Museum Victoria* 2005, **62**:103–146.

23. Nadkarni P, Marenco L, Chen R, Skoufos E, Shepherd G, Miller P: **Organization of heterogeneous scientific data using the EAV/CR representation.** *Journal of the American Medical Informatics Association : JAMIA* 1999, **6**:478–493, [http://view.ncbi.nlm.nih.gov/pubmed/10579606].

24. Marenco L, Tosches N, Shepherd CCG, Miller PL, Nadkarni PM: **Achieving Evolvable Web-Database Bioscience Applications Using the EAV/CR Framework: Recent Advances**. *Journal of the American Medical Informatics Association* 2003, **10**:444–453, [http://dx.doi.org/10.1197/jamia.m1303].

25. Cameron RD: **Scholar-Friendly DOI Suffixes with JACC: Journal Article Citation Convention**. Tech. Rep. CMPT TR 1998-08, School of Computing Science, Simon Fraser University 1998, [http://elib.cs.sfu.ca/USIN/JACC.html].

26. Page RDM: **TBMap: a taxonomic perspective on the phylogenetic database TreeBASE**. *BMC Bioinformatics* 2007, **8**:158, [http://dx.doi.org/10.1186/1471-2105-8-158].

27. Kessler MM: **Bibliographic coupling between scientific papers**. *American Documentation* 1963, **14**:10–25.

28. Shneiderman B: **Tree visualization with tree-maps: 2-d space-filling approach**. *ACM Trans. Graph.* 1992, **11**:92–99.

29. Band Z, White RW: **PygmyBrowse: a small screen tree browser**. In *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, New York, NY, USA: ACM 2006:514–519.

30. Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Sriram MG, Hersh WR: **Using citation data to improve retrieval from MEDLINE.** *Journal of the American Medical Informatics Association : JAMIA* 2006, **13**:96–105, [http://dx.doi.org/10.1197/jamia.M1909].

31. Herskovic JR, Bernstam EV: **Using incomplete citation data for MEDLINE results ranking.** *AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2005, :316–320, [http://www.ncbi.nlm.nih.gov/pubmed/16779053].

32. Page RDM: **Towards realising Darwin's dream: setting the trees free** 2008, [http://dx.doi.org/10.1038/npre.2008.2217.1].

33. MacCallum CJ: **When Is Open Access Not Open Access?** *PLoS Biology* 2007, **5**:e285, [http://dx.doi.org/10.1371/journal.pbio.0050285].

34. Waldrop M: **Big data: Wikiomics**. *Nature* 2008, **455**:22–25.

35. Nikulinaa E, Hanelb R, Schafera P: **Cryptic speciation and paraphyly in the cosmopolitan bryozoan Electra pilosa Impact of the Tethys closing on species evolution**. *Molecular Phylogenetics and Evolution* 2007, **45**:765–776, [http://dx.doi.org/10.1016/j.ympev.2007.07.016].

36. Joseph L, Wilke T, Bermingham E, Alpers D, Ricklefs R: **Towards a phylogenetic framework for the evolution of shakes, rattles, and rolls in Myiarchus tyrant-flycatchers (Aves: Passeriformes: Tyrannidae)**. *Molecular phylogenetics and evolution* 2004, **31**:139–152, [http://dx.doi.org/10.1016/S1055-7903(03)00259-8].