Enhanced display of scientific articles using extended metadata

Roderic D. M. Page [a,]
[a] University of Glasgow, Glasgow G12 8QQ, UK

Abstract

Although the Web has transformed science publishing, scientific papers themselves are still essentially "black boxes", with much of their content intended for human readers only. Typically, computer-readable metadata associated with an article is limited to bibliographic details. By expanding article metadata to include taxonomic names, identifiers for cited material (e.g., publications, sequences, specimens, and other data), and geographical coordinates, publishers could greatly increase the scientific value of their digital content. At the same time this will provide novel ways for users to discover and navigate through this content, beyond the relatively limited linkage provided by bibliographic citation.

As a proof of concept, my entry in the Elsevier Grand Challenge extracted extended metadata from a set of articles from the journal Molecular phylogeny and evolution and used it to populate a entity-attribute-value database. A simple web interface to this database enables an enhanced display of the content of an article, including a map ofl ocalities mentioned either explicitly or implicitly (through links to geotagged data), taxonomic coverage, and both data and citation links. Metadata extraction was limited to information listed in tables in the articles, such as GenBank sequences and specimen codes. The body of the article wasn't used, a restriction that was deliberate to demonstrate that making extended metadata available doesn't require a journal's publisher to make the full-text freely available (although this is desirable for other reasons).

Key words: data citation, geotagging, identifiers, scientific publication, biodiversity informatics, Elsevier Grand Challenge

1. Introduction

Scientific publishing is in transition. The classic model of publishers delivering human-only readable content, whether in print or electronically, is being supplemented by machine-readable content [29]. But machine-readable content is still a second-class citizen, often limited to bibliographic metadata about the article. Hence, most electronic publications are essentially opaque "black boxes", their digital content locked inside PDFs or HTML pages (Fig. 1a). In order to make these documents findable, indexing and abstracting services such as Google Scholar [1] and PubMed [2] make use of fragments of text (e.g., the abstract) or full-text indexing to help users find relevant content (Fig. 1b).

Document findability would be significantly improved ifi ndexing services could extract additional information, such as terms for which we have a controlled vocabulary (e.g., taxonomic names of organisms), database identifiers (such as macromolecular sequence accession numbers, and museum specimen codes), and geographic coordinates (latitude and longitude pairs) (Fig. 1c). For example, querying on the taxonomic name "Aves" (birds) should find all papers on birds, even if those documents don't all include the term "Aves". Extracting iden-

Corresponding author. Tel: +44 141 330 4778
Email address: r.page@bio.gla.ac.uk (Roderic D. M. Page).

[1] http://scholar.google.com
[2] http://www.ncbi.nlm.nih.gov/pubmed/

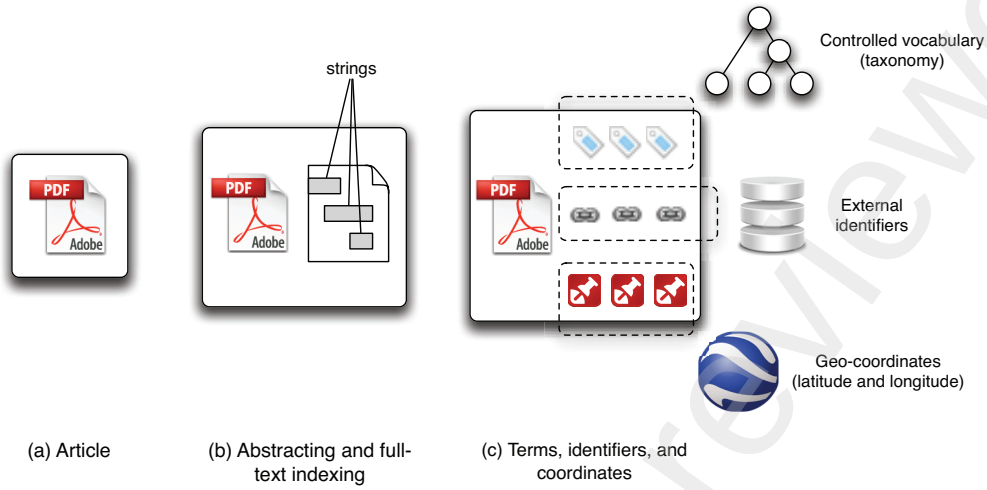

Fig. 1. Increasing the findability of documents with more sophisticated indexing. (a) A raw document, such as a PDF file. (b) Indexing strings, such as keywords, facilitates basic text searching. (c) Indexing terms from controlled vocabularies, database identifiers, and geographical coordinates enables more sophisticated queries.

tifiers enables the document to be linked to external databases, leading to the development of metrics of data citation, and making provenance traceable. Extracting geographical coordinates would enable publishers to provide interfaces that query their journal content by geographic area (e.g., "find all articles that include species living in Madagascar"), or for a given article provide a list of geographically proximate publications (e.g., "find other articles on species within 100 km of the centre-point of this study").

Note that exposing these elements (terms, identifiers, and coordinates) need not require the publisher to make available their primary digital asset (the full text of the article). Additional metadata – beyond standard bibliographic details – could be readily exposed, for exampling using Open Text Mining Interface [3], greatly increasing a document's findability and usability.

1.1. *Citation linking*

Article interrelationships can include "citation", "bibliographic coupling", and "co-citation" (Fig. 2). Establishing these links requires a mechanism to uniquely identify a publication, and tools for finding appropriate identifiers from article metadata. These services are provided by CrossRef [4], 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 "forward linking", so that a publisher displaying paper B can tell the user that B is cited by the more recent paper, A (Fig. 2). Metrics based on the links between publications 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.

1.2. *Data citation*

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, implicit links. For example, a paper in systematics may list DNA sequences from GenBank, but not the voucher specimens from which those sequences were obtained. This potentially deprives natural his-

[3] http://opentextmining.org/wiki/Main_Page

[4] http://www.crossref.org

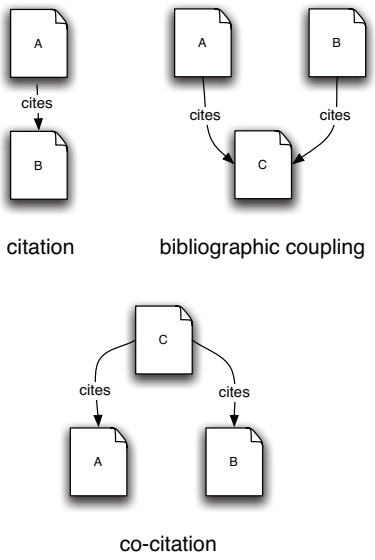

Fig. 2. Relationships between publications. One document may cite another. Two documents (e.g., A and B) are bibliographically coupled if they both cite a third (C). Similarly, co-citation occurs when two documents (e.g., A and B) are cited by a third (C).

tory museums, for example, 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. Similarly, some have argued that failure to cite the authorities of taxonomic names contributes to the relatively low impact factor of taxonomic publications ([31], but see [9]).

An obvious extension to current publishing practice is to extend linking beyond publications alone, thus embedding those publications in a broader web of biodiversity data [23,25]. Some publishers have already made efforts in this direction. BioOne [5] converts putative taxonomic names to links to ITIS [6], others provide a discount on the cost of publication if authors use bibliographic software to construct their list of references, and mandate that some data links (e.g., GenBank accession numbers) are marked-up.

Extending this linking of resources to include museum specimens, for example, would add an important extra dimension to biodiversity publications, beyond providing provenance for data [27]. In particular, many museum specimens have been georeferenced [10], providing a wealth of spatial data that has been harvested globally by GBIF [7]. 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 [1]. Many records in GenBank are themselves geo-referenced, adding a further source of geographic information.

However, there are significant obstacles to extending linking beyond bibliographic citation. The Digital Object Identifiers (DOIs) used by publishers have an underlying technical and social 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 [6]. 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). While publishers have a stake in maintaining the availability of the articles that they publish, they currently have less incentive to maintain access to underlying data, as evidenced by the frequent failure to adequately archive supplementary data [8].

Museum specimen records are more isolated digitally than records in publication or genomic databases, 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 with any widely used identifiers associated with those names (such as NCBI taxonomy ids).

1.3. *Modeling the links*

Combining information on publications, sequences, and specimens, we arrive at a simple model that includes the key entities of interest in biodiversity informatics (Fig. 3). 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 either of these records may be geo-referenced. The record for a specimen lists the

[5] http://www.bioone.org
[6] http://www.itis.gov
[7] http://www.gbif.org

name of the corresponding organism, and GenBank records for parasites may list their host organism. Both of these names can be assigned an identifier using webservices provided by uBio [8].

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

2. Challenge entry

My entry in the Elsevier Grand Challenge "Knowledge Enhancement in the Life Sciences" contest [9] used the model shown in Fig. 3 as the framework for a database that was populated with content from the journal *Molecular Phylogenetics and Evolution*. I extracted citation links to both papers and data, such as Genbank sequences, taxonomic names, and museum specimens, together with geotagged localities, and built a "web" of entities linked by typed relationships [10]. Assembling the entry was a three-step process, involving harvesting, assembling, and displaying data.

2.1. *Harvesting*

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 using a XSLT style sheet. The summary document listed basic bibliographic metadata, the references cited, as well as the content of any tables. Scripts written in PHP were used to extract identifiers and localities from the tables, such as specimen codes, GenBank accession numbers, and latitude and longitude pairs. The body of the article was not searched for these identifiers, partly to see how much could be gleaned without using that text, and also because discovering identifiers within the text itself can be problematic. For example, in experiments with other documents, I discovered that a simple regular expression to extract GenBank identifiers also matched UTM grid references (for example UTM grid reference DQ402119 in [19] is also a GenBank sequence for human herpesvirus). For each reference cited in a paper's bibliography, the bioGUID OpenURL resolver [11] [22] was used to search for available identifiers (such as DOIs, PubMed numbers, Handles, or URLs). This service also returned metadata for GenBank accession numbers and specimen codes.

2.2. *Assembling*

Harvested metadata was stored in an Entity-Attribute-Value (EAV) database [18,20] implemented in MySQL. Entity attributes (metadata) were stored in typed attribute tables (i.e., there are separate tables for dates, integers, real numbers, strings, and large chunks of text). A separate table listed all available globally unique identifiers (GUID) for each entity (e.g., DOIs, PMID, Handles, GenBank accession numbers, etc.). For each entity a 32-digit hexadecimal MD5 hash of the default GUID was used as the unique internal identifier. To avoid duplication due to the same publication being added from different sources (and having different GUIDs), a Journal Article Citation Convention (JACC) identifier [5] was created using the journal ISSN, volume, and starting page – two articles with the same JACC were regarded as the same. Typed links between entities were stored in a separate table. A table with a MySQL SPATIAL index was used to store localities, permitting basic spatial queries. To support taxonomic queries a nested-sets representation of the NCBI Taxonomy was created, following [24].

2.3. *Display*

Figure 4 shows a screen shot of my entry [12] displaying a typical article [30]. Each object in the database, whether article, sequence, specimen, taxon, or author has its own webpage that displays metadata about that object, as well as any links it may have to other objects (Fig. 4). In addition to classical links between papers, such as citation

[8] http://www.ubio.org
[9] http://www.elseviergrandchallenge.com
[10] http://iphylo.org/∼rpage/challenge/www
[11] http://bioguid.info/openurl
[12] http://iphylo.org/∼rpage/challenge/www/

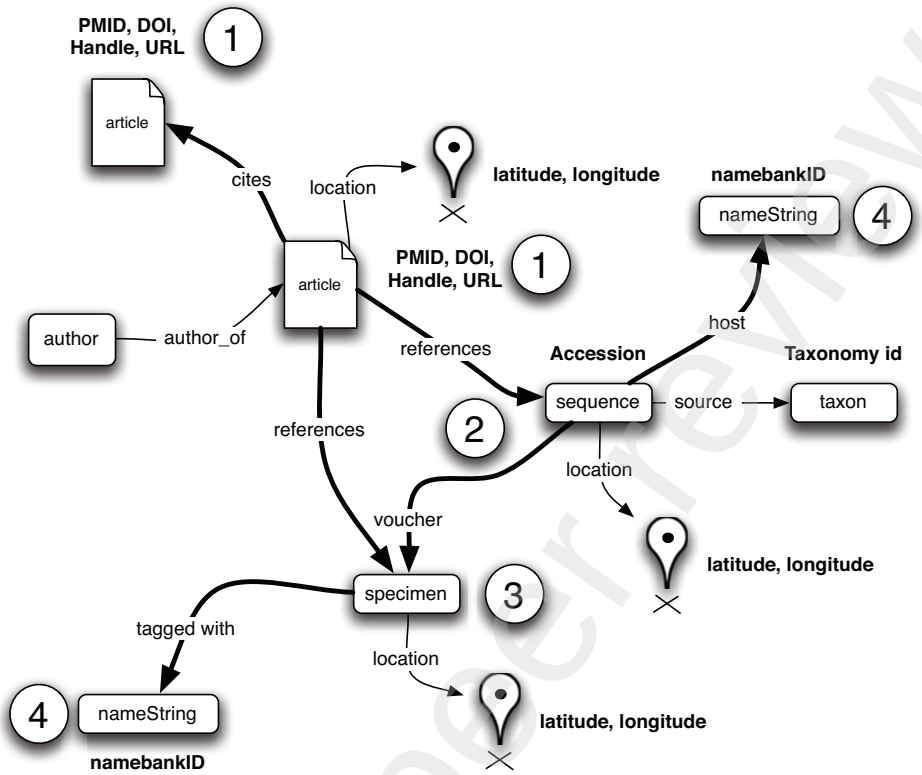

Fig. 3. Data model linking publications, authors, sequences, specimens, taxonomic names, and locations. Thick lines indicate relationships that have to be extracted from the text of an article, or from PubMed, GenBank, or specimen records. Circled numbers indicate identifiers that can be obtained for the corresponding entity. 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.

(Fig. 2), papers may be linked by shared data, that is, they have links to one or more GenBank sequences or specimens in common (analogous to bibliographic coupling, Fig. 2) [15]. These shared links (if present) are listed in a section entitled "Shares data with" (note that these papers need not directly cite, nor be cited by, the current article).

Taxonomic information is displayed in a variety of ways. The taxa in a study are listed on the article web page, and summarised in a split layout treemap [7] populated with images of the taxa [13] . The size of each cell in the treemap is proportional to the number of taxa in that group (typically a genus). Each taxon has its own separate page, which lists papers referring to that taxon, and displays the NCBI classification for that taxon using a PygmyBrowser [2].

Localities extracted from tables in the articles, or via objects that are georeferenced (e.g., GenBank sequences, or voucher specimens for those sequences) are used to generate a map of all localities linked to the article. The web page displays up to five studies whose geographic coverage overlaps that of the article being displayed.

2.4. *Obstacles and limitations*

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 be enormous variation in how latitude and

[13] The images were obtained using Yahoo's image search APIs and Flickr.

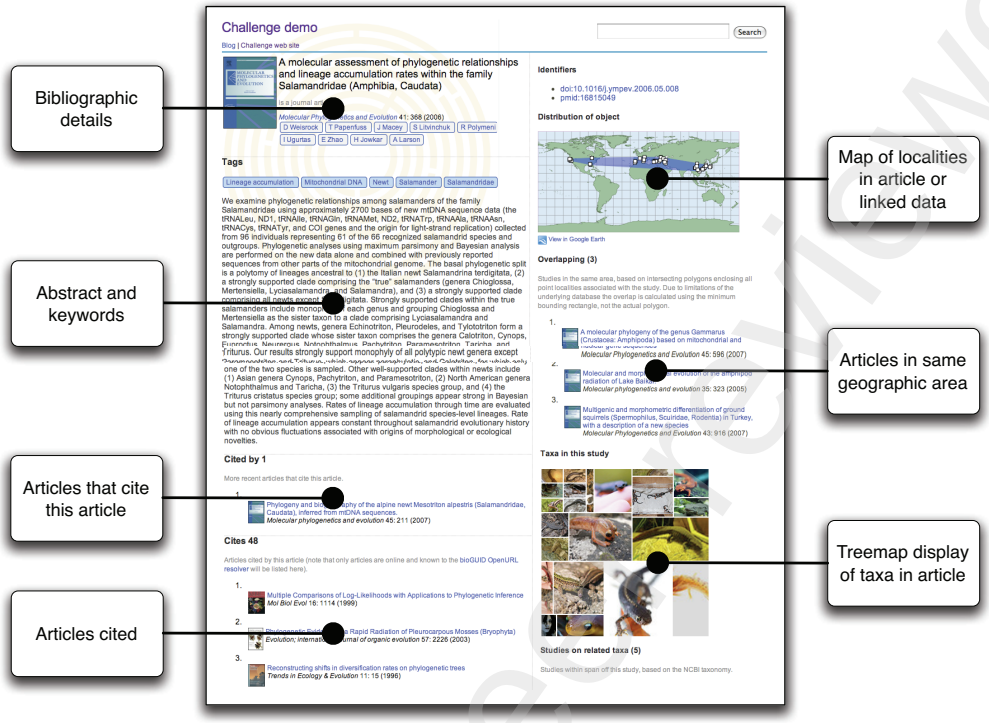

Fig. 4. Screen shot showing basic bibliographic metadata being displayed for an article [30], enhanced with a map of localities linked to the article, a taxonomic summary as a treemap, and lists of other articles that are related either through citation, geographic or taxonomic overlap. These related studies are discovered by harvesting identifiers and geographical co-ordinates from the article.

GPS position	Latitude	Longitude
27:30:08N; 80:18:48W	23°03′44′′N	161°55′34′′W
36:23:49N; 6:12:16W	23°03′44′′N	161°55′34′′W

Lat/Long	Latitude (N)	Longitude (W)
	Tokyo, Japan	
N 43.97493° W 122.16516°	34.05	135.35
N 35.72209° W 94.40863°		

Fig. 5. Examples of different ways latitude and longitude are reported in four articles [13,11,16,28] from the journal *Molecular Phylogenetics and Evolution*.

longitudes are written, both individually, and when written as a pair of co-ordinates. In addition, there is considerable variation in how latitude and longitude pairs are reported in tables in *Molecular Phylogenetics and Evolution*. 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 (Fig. 5). 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 rather crudely implemented. Very simple full-text searching is available for text, implemented using MySQL's FULLTEXT index. Spatial searching relies on MySQL's spatial extensions, which have limited functionality in the current version of that software. For example, the search for overlapping polygons is actually implemented as a search for overlapping minimum bounding rectangles. Even if this issue was addressed, the use of bounding polygons is in itself too crude to adequately represent a geographic distribution. The taxonomic search returns taxa within the taxonomic span of the article.

However (and analogously to the bounding polygon), if a study includes a few disparate taxa, then the list of taxa returned may be more diverse than the user anticipates.

An alternative approach to implementing the challenge entry would have been using RDF and a triple store, with SPARQL as the query language. RDF requires vocabularies to describe entities and their relationships. For some classes of entity (such as publications) there are several competing vocabularies in existence, whereas for others there are none. SPARQL has considerable potential, but in the context of the Challenge entry its use was not feasible as it does not natively support the geospatial, full-text searching, and nested sets queries needed to create the visualisations (e.g., Fig. 4).

The entry has merely scratched the surface of possible visualisations. The original proposal [26] included the use of Google Earth to display phylogenies. This is technically achievable [14], but preliminary work found that automated extracting trees from bitmap images in journal pages was more difficult than anticipated, and I abandoned this task. However, the maps displayed on each web page (Fig. 4) can be exported in KML and viewed in Google Earth.

3. Conclusion and outlook

My challenge entry made very limited use of the full-text documents. This was partly to keep the task manageable, but also to see what could be achieved without the publisher handing over its "crown jewels" (i.e., the full text) to the reader. Part 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 [3]) incomplete citation data can still be useful [12]).

The key point is that much of the information displayed in the demonstration comes from a small fraction of the information in the journal articles. If for each article publishers were to output metadata listing the papers cited, GenBank accession numbers, specimen codes, and any latitude-longitude pairs, then a system like this entry could be created without requiring 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 when this process is automated, it would be desirable to provide authors with simple tools to mark up their manuscripts, for example, by flagging GenBank accessions, museum codes, and latitudes and longitudes (the later being written in standard format).

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 (in the strict sense of enabling copying and repurposing of the published content [17]). Purely from the narrow point of this Challenge entry, extracting identifiers and geotags is error-prone, and having the full text available would facilitate detecting and correcting errors. Furthermore, papers themselves can contain errors. On several occasions in this study I found cases where GenBank accession numbers were clearly incorrect. For example, [21] 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, [14] 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. One way forward is to treat a scientific article not as an immutable text, but rather as version 1.0 of a series of annotated and edited versions. Supporting community revision suggests a wiki-style interface, and may be yet another step in the convergence of journals and databases [4].

4. Acknowledgements

I thank Anita de Ward and Noelle Gracy of Elsevier for facilitating access to full-text XML of *Molecular Phylogenetics and Evolution*, Kehan Harman and Benjamin Good for comments on the entry, and the three anonymous reviewers who provided helpful comments on the manuscript.

[14] http://iphylo.blogspot.com/2007/06/google-earth-phylogenies.html

References

[1] Anonymous, A place for everything, Nature 453 (2008) 2.

[2] Z. Band, R. W. White, Pygmybrowse: a small screen tree browser, in: CHI '06: CHI '06 extended abstracts on Human factors in computing systems, ACM, New York, NY, USA, 2006.

[3] E. V. Bernstam, J. R. Herskovic, Y. Aphinyanaphongs, C. F. Aliferis, M. G. Sriram, W. R. Hersh, Using citation data to improve retrieval from medline, Journal of the American Medical Informatics Association : JAMIA 13 (2006) 96–105.

[4] P. Bourne, Will a biological database be different from a biological journal?, PLoS Comput Biol 1 (3) (2005) e34.

[5] R. D. Cameron, Scholar-friendly DOI suffixes with JACC: Journal Article Citation Convention, Tech. Rep. CMPT TR 1998-08, School of Computing Science, Simon Fraser University, avalable from http://elib.cs.sfu.ca/USIN/JACC.html (1998).

[6] R. P. Dellavalle, E. J. Hester, L. F. Heilig, A. L. Drake, J. W. Kuntzman, M. G. L. M. Schillin, Going, going, gone: Lost internet references, Science 302 (2003) 787.

[7] B. Engdahl, Ordered and unordered treemap algorithms and their applications on handheld devices, Master's degree project, Department of Numerical Analysis and Computer Science, Stockholm Royal Institute of Technology, SE-100 44 Stockholm, Sweden (2005).

[8] E. Evangelou, T. Trikalinos, J. Ioannidis, Unavailability of online supplementary scientific information from articles published in major journals, The FASEB Journal 19 (2005) 1943–1944.

[9] E. Garfield, Taxonomy is small, but it has its citation classics, Nature 413 (2002) 107.

[10] R. P. Guralnick, J. Wieczorek, R. Beaman, R. J. Hijmans, BioGeomancer: automated georeferencing to map the world's biodiversity data, PLoS Biology 4 (2006) e381.

[11] B. E. Hendrixson, J. E. Bond, Molecular phylogeny and biogeography of an ancient Holarctic lineage of mygalomorph spiders (Araneae: Antrodiaetidae: Antrodiaetus), Molecular Phylogenetics and Evolution 42 (2007) 738–755.

[12] J. R. Herskovic, E. V. Bernstam, Using incomplete citation data for medline results ranking., AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium (2005) 316–320.

[13] J. R. Hyde, R. D. Vetter, The origin, evolution, and diversification of rockfishes of the genus *Sebastes* (Cuvier), Molecular phylogenetics and evolution 44 (2007) 790–811.

[14] L. Joseph, T. Wilke, E. Bermingham, D. Alpers, R. Ricklefs, Towards a phylogenetic framework for the evolution of shakes, rattles, and rolls in *Myiarchus* tyrant-flycatchers (Aves: Passeriformes: Tyrannidae), Molecular Phylogenetics and Evolution 31 (2004) 139–152.

[15] M. M. Kessler, Bibliographic coupling between scientific papers, American Documentation 14 (1963) 10–25.

[16] S. López-Legentil, X. Turon, Lack of genetic variation in mtDNA sequences over the amphiatlantic distribution range of the ascidian *Ecteinascidia turbinata*, Molecular Phylogenetics and Evolution 45 (2007) 405–408.

[17] C. J. MacCallum, When is open access not open access?, PLoS Biology 5 (2007) e285.

[18] L. Marenco, N. Tosches, C. C. G. Shepherd, P. L. Miller, P. M. Nadkarni, Achieving evolvable web-database bioscience applications using the EAV/CR framework: recent advances, Journal of the American Medical Informatics Association 10 (2003) 444–453.

[19] R. Mesibov, 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 62 (2005) 103–146.

[20] P. Nadkarni, L. Marenco, R. Chen, E. Skoufos, G. Shepherd, P. Miller, Organization of heterogeneous scientific data using the EAV/CR representation, Journal of the American Medical Informatics Association : JAMIA 6 (1999) 478–493.

[21] E. Nikulinaa, R. Hanelb, P. Schafera, Cryptic speciation and paraphyly in the cosmopolitan bryozoan *Electra pilosa*: impact of the Tethys closing on species evolution, Molecular Phylogenetics and Evolution 45 (2007) 765–776.

[22] R. Page, bioGUID: resolving, discovering, and minting identifiers for biodiversity informatics, BMC Bioinformatics 10 (Suppl 14) (2009) S5.

[23] R. D. M. Page, Taxonomic names, metadata, and the Semantic Web, Biodiversity Informatics 3 (2006) 1–15.

[24] R. D. M. Page, TBMap: a taxonomic perspective on the phylogenetic database TreeBASE, BMC Bioinformatics 8 (1) (2007) 158.

[25] R. D. M. Page, Biodiversity informatics: the challenge of linking data and the role of shared identifiers, Brief Bioinform 9 (5) (2008) 345–354.

[26] R. D. M. Page, Towards realising Darwin's dream: setting the trees free. Available from Nature Precedings http://dx.doi.org/10.1038/npre.2008.2217.1.

[27] A. T. Peterson, R. G. Moylea, A. S. Nyária, M. B. Robbins, R. T. Brumfield, J. Remsen, The need for proper vouchering in phylogenetic studies of birds, Molecular Phylogenetics and Evolution 45 (3) (2007) 1042–1044.

[28] L. H. Shapiro, J. S. Strazanac, G. K. Roderick, Molecular phylogeny of *Banza* (Orthoptera: Tettigoniidae), the endemic katydids of the Hawaiian Archipelago, Molecular Phylogenetics and Evolution 41 (2006) 53–63.

[29] D. Shotton, Semantic publishing: the coming revolution in scientific journal publishing, Learned Publishing 22 (2009) 85–94.

[30] D. W. Weisrock, T. J. Papenfuss, J. R. Macey, S. N. Litvinchuk, R. Polymeni, I. H. Ugurtas, E. Zhao, H. Jowkar, A. Larson, A molecular assessment of phylogenetic relationships and lineage accumulation rates within the family Salamandridae (Amphibia, Caudata), Molecular Phylogenetics and Evolution 41 (2) (2006) 368 – 383.

[31] Y. L. Werner, The case of impact factor versus taxonomy: a proposal, Journal of Natural History 40 (2006) 1285–1286.

Enhanced display of scientific articles using extended metadata

Roderic D. M. Page [a,*]
[a] *University of Glasgow, Glasgow G12 8QQ, UK*

Abstract

Although the Web has transformed science publishing, scientific papers themselves are still essentially "black boxes", with much of their content intended for human readers only. Typically, computer-readable metadata associated with an article is limited to bibliographic details. By expanding article metadata to include taxonomic names, identifiers for cited material (e.g., publications, sequences, specimens, and other data), and geographical coordinates, publishers could greatly increase the scientific value of their digital content. At the same time this will provide novel ways for users to discover and navigate through this content, beyond the relatively limited linkage provided by bibliographic citation.

As a proof of concept, my entry in the Elsevier Grand Challenge extracted extended metadata from a set of articles from the journal *Molecular phylogeny and evolution* and used it to populate a entity-attribute-value database. A simple web interface to this database enables an enhanced display of the content of an article, including a map of localities mentioned either explicitly or implicitly (through links to geotagged data), taxonomic coverage, and both data and citation links. Metadata extraction was limited to information listed in tables in the articles, such as GenBank sequences and specimen codes. The body of the article wasn't used, a restriction that was deliberate to demonstrate that making extended metadata available doesn't require a journal's publisher to make the full-text freely available (although this is desirable for other reasons).

Key words: data citation, geotagging, identifiers, scientific publication, biodiversity informatics, Elsevier Grand Challenge

1. Introduction

Scientific publishing is in transition. The classic model of publishers delivering human-only readable content, whether in print or electronically, is being supplemented by machine-readable content [29]. But machine-readable content is still a second-class citizen, often limited to bibliographic metadata about the article. Hence, most electronic publications are essentially opaque "black boxes", their digital content locked inside PDFs or HTML pages (Fig. 1a). In order to make these documents findable, indexing and abstracting services such as Google Scholar [1] and PubMed [2] make use of fragments of text (e.g., the abstract) or full-text indexing to help users find relevant content (Fig. 1b).

Document findability would be significantly improved if indexing services could extract additional information, such as terms for which we have a controlled vocabulary (e.g., taxonomic names of organisms), database identifiers (such as macromolecular sequence accession numbers, and museum specimen codes), and geographic coordinates (latitude and longitude pairs) (Fig. 1c). For example, querying on the taxonomic name "Aves" (birds) should find all papers on birds, even if those documents don't all include the term "Aves". Extracting iden-

* Corresponding author. Tel: +44 141 330 4778
Email address: r.page@bio.gla.ac.uk (Roderic D. M. Page).

[1] http://scholar.google.com
[2] http://www.ncbi.nlm.nih.gov/pubmed/

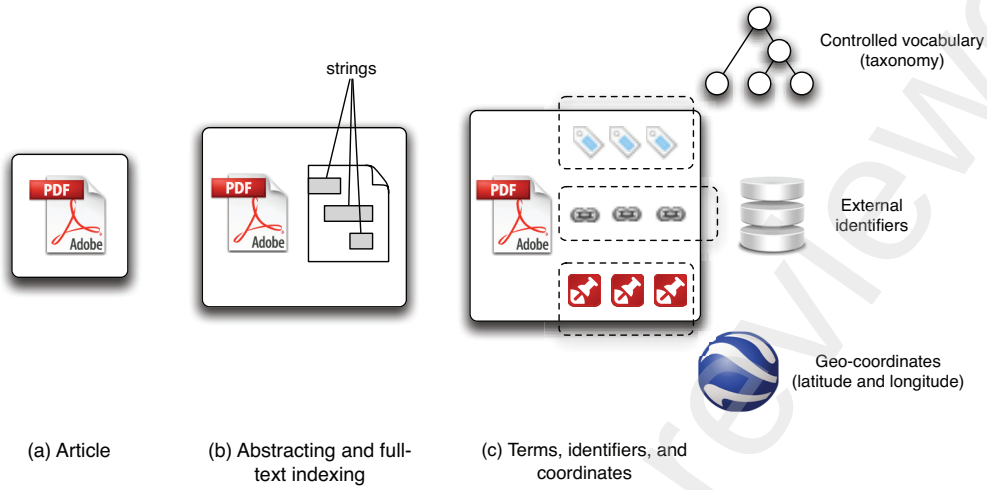

Fig. 1. Increasing the findability of documents with more sophisticated indexing. (a) A raw document, such as a PDF file. (b) Indexing strings, such as keywords, facilitates basic text searching. (c) Indexing terms from controlled vocabularies, database identifiers, and geographical coordinates enables more sophisticated queries.

tifiers enables the document to be linked to external databases, leading to the development of metrics of data citation, and making provenance traceable. Extracting geographical coordinates would enable publishers to provide interfaces that query their journal content by geographic area (e.g., "find all articles that include species living in Madagascar"), or for a given article provide a list of geographically proximate publications (e.g., "find other articles on species within 100 km of the centre-point of this study").

Note that exposing these elements (terms, identifiers, and coordinates) need not require the publisher to make available their primary digital asset (the full text of the article). Additional metadata – beyond standard bibliographic details – could be readily exposed, for exampling using Open Text Mining Interface [3], greatly increasing a document's findability and usability.

1.1. *Citation linking*

Article interrelationships can include "citation", "bibliographic coupling", and "co-citation" (Fig. 2). Establishing these links requires a mechanism to uniquely identify a publication, and tools for finding appropriate identifiers from article metadata. These services are provided by CrossRef [4], 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 "forward linking", so that a publisher displaying paper B can tell the user that B is cited by the more recent paper, A (Fig. 2). Metrics based on the links between publications 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.

1.2. *Data citation*

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, implicit links. For example, a paper in systematics may list DNA sequences from GenBank, but not the voucher specimens from which those sequences were obtained. This potentially deprives natural his-

[3] http://opentextmining.org/wiki/Main_Page

[4] http://www.crossref.org

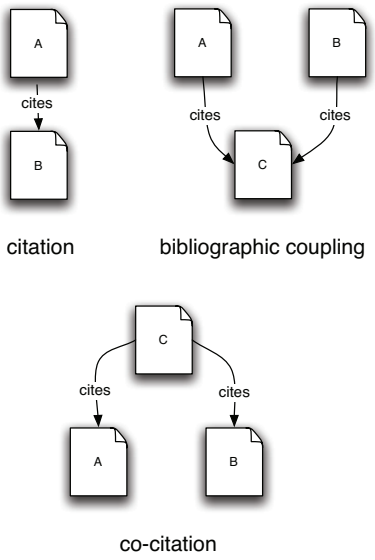

Fig. 2. Relationships between publications. One document may cite another. Two documents (e.g., A and B) are bibliographically coupled if they both cite a third (C). Similarly, co-citation occurs when two documents (e.g., A and B) are cited by a third (C).

tory museums, for example, 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. Similarly, some have argued that failure to cite the authorities of taxonomic names contributes to the relatively low impact factor of taxonomic publications ([31], but see [9]).

An obvious extension to current publishing practice is to extend linking beyond publications alone, thus embedding those publications in a broader web of biodiversity data [22,24]. Some publishers have already made efforts in this direction. BioOne [5] converts putative taxonomic names to links to ITIS [6], others provide a discount on the cost of publication if authors use bibliographic software to construct their list of references, and mandate that some data links (e.g., GenBank accession numbers) are marked-up.

Extending this linking of resources to include museum specimens, for example, would add an important extra dimension to biodiversity publications, beyond providing provenance for data [27]. In particular, many museum specimens have been georeferenced [10], providing a wealth of spatial data that has been harvested globally by GBIF [7]. 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 [1]. Many records in GenBank are themselves geo-referenced, adding a further source of geographic information.

However, there are significant obstacles to extending linking beyond bibliographic citation. The Digital Object Identifiers (DOIs) used by publishers have an underlying technical and social 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 [6]. 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). While publishers have a stake in maintaining the availability of the articles that they publish, they currently have less incentive to maintain access to underlying data, as evidenced by the frequent failure to adequately archive supplementary data [8].

Museum specimen records are more isolated digitally than records in publication or genomic databases, 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 with any widely used identifiers associated with those names (such as NCBI taxonomy ids).

1.3. *Modeling the links*

Combining information on publications, sequences, and specimens, we arrive at a simple model that includes the key entities of interest in biodiversity informatics (Fig. 3). 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 either of these records may be geo-referenced. The record for a specimen lists the

[5] http://www.bioone.org
[6] http://www.itis.gov
[7] http://www.gbif.org

name of the corresponding organism, and GenBank records for parasites may list their host organism. Both of these names can be assigned an identifier using webservices provided by uBio [8].

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

2. Challenge entry

My entry in the Elsevier Grand Challenge "Knowledge Enhancement in the Life Sciences" contest [9] used the model shown in Fig. 3 as the framework for a database that was populated with content from the journal *Molecular Phylogenetics and Evolution*. I extracted citation links to both papers and data, such as Genbank sequences, taxonomic names, and museum specimens, together with geotagged localities, and built a "web" of entities linked by typed relationships [10]. Assembling the entry was a three-step process, involving harvesting, assembling, and displaying data.

2.1. *Harvesting*

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 using a XSLT style sheet. The summary document listed basic bibliographic metadata, the references cited, as well as the content of any tables. Scripts written in PHP were used to extract identifiers and localities from the tables, such as specimen codes, GenBank accession numbers, and latitude and longitude pairs. The body of the article was not searched for these identifiers, partly to see how much could be gleaned without using that text, and also because discovering identifiers within the text itself can be problematic. For example, in experiments with other documents, I discovered that a simple regular expression to extract GenBank identifiers also matched UTM grid references (for example UTM grid reference DQ402119 in [19] is also a GenBank sequence for human herpesvirus). For each reference cited in a paper's bibliography, the bioGUID OpenURL resolver [11] [26] was used to search for available identifiers (such as DOIs, PubMed numbers, Handles, or URLs). This service also returned metadata for GenBank accession numbers and specimen codes.

2.2. *Assembling*

Harvested metadata was stored in an Entity-Attribute-Value (EAV) database [18,20] implemented in MySQL. Entity attributes (metadata) were stored in typed attribute tables (i.e., there are separate tables for dates, integers, real numbers, strings, and large chunks of text). A separate table listed all available globally unique identifiers (GUID) for each entity (e.g., DOIs, PMID, Handles, GenBank accession numbers, etc.). For each entity a 32-digit hexadecimal MD5 hash of the default GUID was used as the unique internal identifier. To avoid duplication due to the same publication being added from different sources (and having different GUIDs), a Journal Article Citation Convention (JACC) identifier [5] was created using the journal ISSN, volume, and starting page – two articles with the same JACC were regarded as the same. Typed links between entities were stored in a separate table. A table with a MySQL SPATIAL index was used to store localities, permitting basic spatial queries. To support taxonomic queries a nested-sets representation of the NCBI Taxonomy was created, following [23].

2.3. *Display*

Figure 4 shows a screen shot of my entry [12] displaying a typical article [30]. Each object in the database, whether article, sequence, specimen, taxon, or author has its own webpage that displays metadata about that object, as well as any links it may have to other objects (Fig. 4). In addition to classical links between papers, such as citation

[8] http://www.ubio.org
[9] http://www.elseviergrandchallenge.com
[10] http://iphylo.org/∼rpage/challenge/www
[11] http://bioguid.info/openurl
[12] http://iphylo.org/∼rpage/challenge/www/

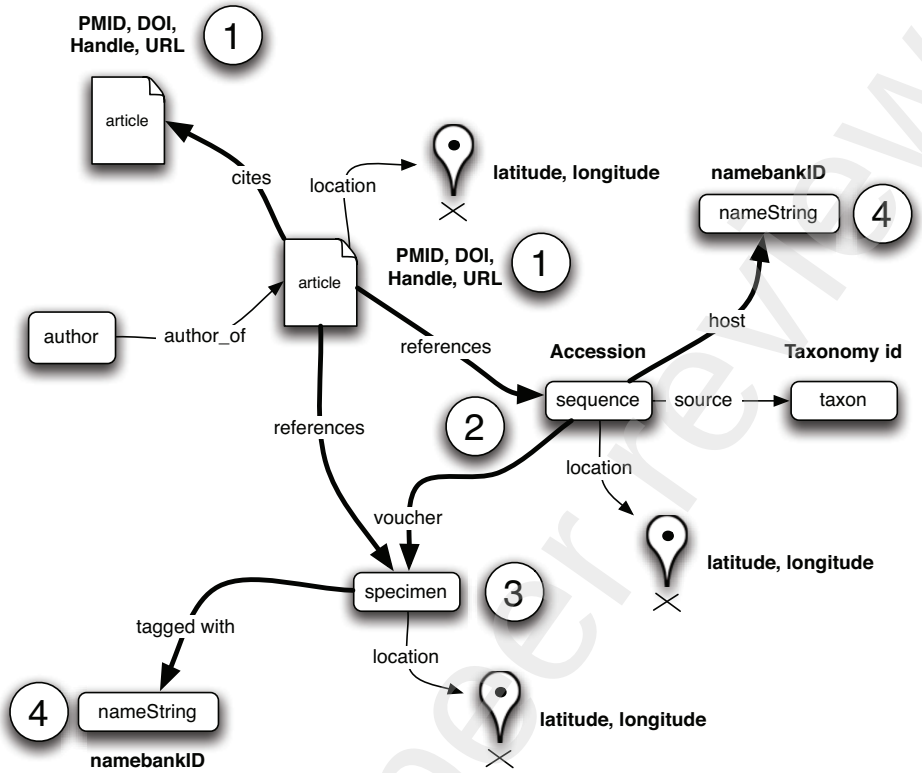

Fig. 3. Data model linking publications, authors, sequences, specimens, taxonomic names, and locations. Thick lines indicate relationships that have to be extracted from the text of an article, or from PubMed, GenBank, or specimen records. Circled numbers indicate identifiers that can be obtained for the corresponding entity. 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.

(Fig. 2), papers may be linked by shared data, that is, they have links to one or more GenBank sequences or specimens in common (analogous to bibliographic coupling, Fig. 2) [15]. These shared links (if present) are listed in a section entitled "Shares data with" (note that these papers need not directly cite, nor be cited by, the current article).

Taxonomic information is displayed in a variety of ways. The taxa in a study are listed on the article web page, and summarised in a split layout treemap [7] populated with images of the taxa [13] . The size of each cell in the treemap is proportional to the number of taxa in that group (typically a genus). Each taxon has its own separate page, which lists papers referring to that taxon, and displays the NCBI classification for that taxon using a PygmyBrowser [2].

Localities extracted from tables in the articles, or via objects that are georeferenced (e.g., GenBank sequences, or voucher specimens for those sequences) are used to generate a map of all localities linked to the article. The web page displays up to five studies whose geographic coverage overlaps that of the article being displayed.

2.4. *Obstacles and limitations*

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 be enormous variation in how latitude and

[13] The images were obtained using Yahoo's image search APIs and Flickr.

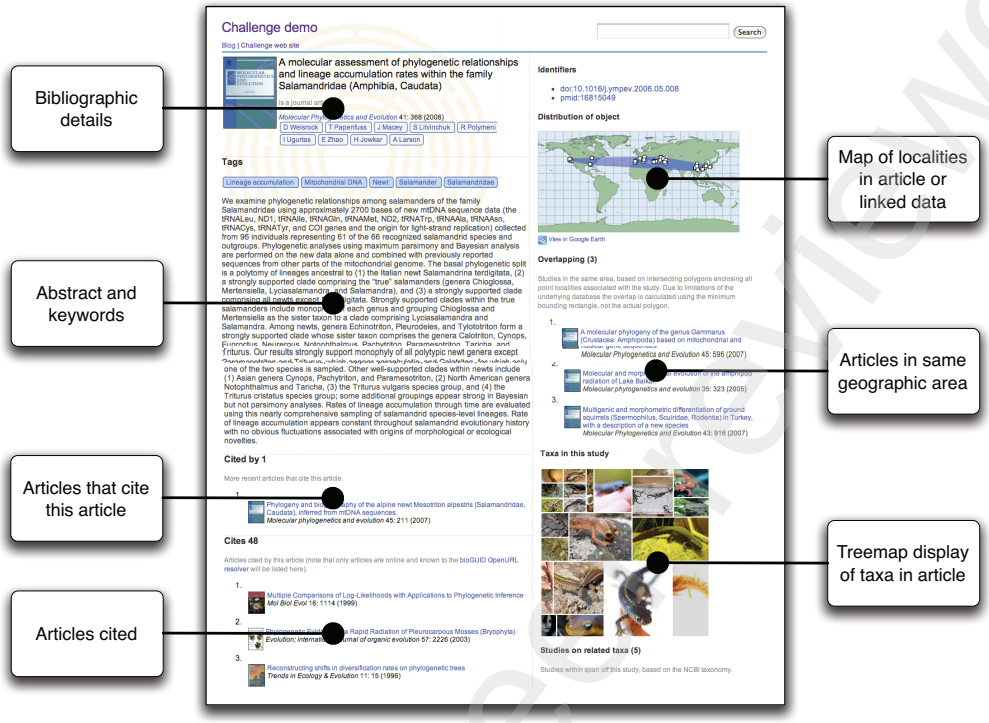

Fig. 4. Screen shot showing basic bibliographic metadata being displayed for an article [30], enhanced with a map of localities linked to the article, a taxonomic summary as a treemap, and lists of other articles that are related either through citation, geographic or taxonomic overlap. These related studies are discovered by harvesting identifiers and geographical co-ordinates from the article.

GPS position	Latitude	Longitude
27:30:08N; 80:18:48W	23°03′44′′N	161°55′34′′W
36:23:49N; 6:12:16W	23°03′44′′N	161°55′34′′W

Lat/Long	Latitude (N)	Longitude (W)
	Tokyo, Japan	
N 43.97493° W 122.16516°	34.05	135.35
N 35.72209° W 94.40863°		

Fig. 5. Examples of different ways latitude and longitude are reported in four articles [13,11,16,28] from the journal *Molecular Phylogenetics and Evolution*.

longitudes are written, both individually, and when written as a pair of co-ordinates. In addition, there is considerable variation in how latitude and longitude pairs are reported in tables in *Molecular Phylogenetics and Evolution*. 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 (Fig. 5). 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 rather crudely implemented. Very simple full-text searching is available for text, implemented using MySQL's FULLTEXT index. Spatial searching relies on MySQL's spatial extensions, which have limited functionality in the current version of that software. For example, the search for overlapping polygons is actually implemented as a search for overlapping minimum bounding rectangles. Even if this issue was addressed, the use of bounding polygons is in itself too crude to adequately represent a geographic distribution. The taxonomic search returns taxa within the taxonomic span of the article.

However (and analogously to the bounding polygon), if a study includes a few disparate taxa, then the list of taxa returned may be more diverse than the user anticipates.

An alternative approach to implementing the challenge entry would have been using RDF and a triple store, with SPARQL as the query language. RDF requires vocabularies to describe entities and their relationships. For some classes of entity (such as publications) there are several competing vocabularies in existence, whereas for others there are none. SPARQL has considerable potential, but in the context of the Challenge entry its use was not feasible as it does not natively support the geospatial, full-text searching, and nested sets queries needed to create the visualisations (e.g., Fig. 4).

The entry has merely scratched the surface of possible visualisations. The original proposal [25] included the use of Google Earth to display phylogenies. This is technically achievable [14], but preliminary work found that automated extracting trees from bitmap images in journal pages was more difficult than anticipated, and I abandoned this task. However, the maps displayed on each web page (Fig. 4) can be exported in KML and viewed in Google Earth.

3. Conclusion and outlook

My challenge entry made very limited use of the full-text documents. This was partly to keep the task manageable, but also to see what could be achieved without the publisher handing over its "crown jewels" (i.e., the full text) to the reader. Part 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 [3]) incomplete citation data can still be useful [12]).

The key point is that much of the information displayed in the demonstration comes from a small fraction of the information in the journal articles. If for each article publishers were to output metadata listing the papers cited, GenBank accession numbers, specimen codes, and any latitude-longitude pairs, then a system like this entry could be created without requiring 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 when this process is automated, it would be desirable to provide authors with simple tools to mark up their manuscripts, for example, by flagging GenBank accessions, museum codes, and latitudes and longitudes (the later being written in standard format).

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 (in the strict sense of enabling copying and repurposing of the published content [17]). Purely from the narrow point of this Challenge entry, extracting identifiers and geotags is error-prone, and having the full text available would facilitate detecting and correcting errors. Furthermore, papers themselves can contain errors. On several occasions in this study I found cases where GenBank accession numbers were clearly incorrect. For example, [21] 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, [14] 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. One way forward is to treat a scientific article not as an immutable text, but rather as version 1.0 of a series of annotated and edited versions. Supporting community revision suggests a wiki-style interface, and may be yet another step in the convergence of journals and databases [4].

4. Acknowledgements

I thank Anita de Ward and Noelle Gracy of Elsevier for facilitating access to full-text XML of *Molecular Phylogenetics and Evolution*, Kehan Harman and Benjamin Good for comments on the entry, and the three anonymous reviewers who provided helpful comments on the manuscript.

[14] http://iphylo.blogspot.com/2007/06/google-earth-phylogenies.html

References

[1] Anonymous, A place for everything, Nature 453 (2008) 2.

[2] Z. Band, R. W. White, Pygmybrowse: a small screen tree browser, in: CHI '06: CHI '06 extended abstracts on Human factors in computing systems, ACM, New York, NY, USA, 2006.

[3] E. V. Bernstam, J. R. Herskovic, Y. Aphinyanaphongs, C. F. Aliferis, M. G. Sriram, W. R. Hersh, Using citation data to improve retrieval from medline, Journal of the American Medical Informatics Association : JAMIA 13 (2006) 96–105.

[4] P. Bourne, Will a biological database be different from a biological journal?, PLoS Comput Biol 1 (3) (2005) e34.

[5] R. D. Cameron, Scholar-friendly DOI suffixes with JACC: Journal Article Citation Convention, Tech. Rep. CMPT TR 1998-08, School of Computing Science, Simon Fraser University, avalable from http://elib.cs.sfu.ca/USIN/JACC.html (1998).

[6] R. P. Dellavalle, E. J. Hester, L. F. Heilig, A. L. Drake, J. W. Kuntzman, M. G. L. M. Schillin, Going, going, gone: Lost internet references, Science 302 (2003) 787.

[7] B. Engdahl, Ordered and unordered treemap algorithms and their applications on handheld devices, Master's degree project, Department of Numerical Analysis and Computer Science, Stockholm Royal Institute of Technology, SE-100 44 Stockholm, Sweden (2005).

[8] E. Evangelou, T. Trikalinos, J. Ioannidis, Unavailability of online supplementary scientific information from articles published in major journals, The FASEB Journal 19 (2005) 1943–1944.

[9] E. Garfield, Taxonomy is small, but it has its citation classics, Nature 413 (2002) 107.

[10] R. P. Guralnick, J. Wieczorek, R. Beaman, R. J. Hijmans, BioGeomancer: automated georeferencing to map the world's biodiversity data, PLoS Biology 4 (2006) e381.

[11] B. E. Hendrixson, J. E. Bond, Molecular phylogeny and biogeography of an ancient Holarctic lineage of mygalomorph spiders (Araneae: Antrodiaetidae: Antrodiaetus), Molecular Phylogenetics and Evolution 42 (2007) 738–755.

[12] J. R. Herskovic, E. V. Bernstam, Using incomplete citation data for medline results ranking., AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium (2005) 316–320.

[13] J. R. Hyde, R. D. Vetter, The origin, evolution, and diversification of rockfishes of the genus *Sebastes* (Cuvier), Molecular phylogenetics and evolution 44 (2007) 790–811.

[14] L. Joseph, T. Wilke, E. Bermingham, D. Alpers, R. Ricklefs, Towards a phylogenetic framework for the evolution of shakes, rattles, and rolls in *Myiarchus* tyrant-flycatchers (Aves: Passeriformes: Tyrannidae), Molecular Phylogenetics and Evolution 31 (2004) 139–152.

[15] M. M. Kessler, Bibliographic coupling between scientific papers, American Documentation 14 (1963) 10–25.

[16] S. López-Legentil, X. Turon, Lack of genetic variation in mtDNA sequences over the amphiatlantic distribution range of the ascidian *Ecteinascidia turbinata*, Molecular Phylogenetics and Evolution 45 (2007) 405–408.

[17] C. J. MacCallum, When is open access not open access?, PLoS Biology 5 (2007) e285.

[18] L. Marenco, N. Tosches, C. C. G. Shepherd, P. L. Miller, P. M. Nadkarni, Achieving evolvable web-database bioscience applications using the EAV/CR framework: recent advances, Journal of the American Medical Informatics Association 10 (2003) 444–453.

[19] R. Mesibov, 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 62 (2005) 103–146.

[20] P. Nadkarni, L. Marenco, R. Chen, E. Skoufos, G. Shepherd, P. Miller, Organization of heterogeneous scientific data using the EAV/CR representation, Journal of the American Medical Informatics Association : JAMIA 6 (1999) 478–493.

[21] E. Nikulinaa, R. Hanelb, P. Schafera, Cryptic speciation and paraphyly in the cosmopolitan bryozoan *Electra pilosa*: impact of the Tethys closing on species evolution, Molecular Phylogenetics and Evolution 45 (2007) 765–776.

[22] R. D. M. Page, Taxonomic names, metadata, and the Semantic Web, Biodiversity Informatics 3 (2006) 1–15.

[23] R. D. M. Page, TBMap: a taxonomic perspective on the phylogenetic database TreeBASE, BMC Bioinformatics 8 (1) (2007) 158.

[24] R. D. M. Page, Biodiversity informatics: the challenge of linking data and the role of shared identifiers, Brief Bioinform 9 (5) (2008) 345–354.

[25] R. D. M. Page, Towards realising Darwin's dream: setting the trees free. Available from Nature Precedings http://dx.doi.org/10.1038/npre.2008.2217.1.

[26] R. Page, bioGUID: resolving, discovering, and minting identifiers for biodiversity informatics, BMC Bioinformatics 10 (Suppl 14) (2009) S5.

[27] A. T. Peterson, R. G. Moylea, A. S. Nyária, M. B. Robbins, R. T. Brumfield, J. Remsen, The need for proper vouchering in phylogenetic studies of birds, Molecular Phylogenetics and Evolution 45 (3) (2007) 1042–1044.

[28] L. H. Shapiro, J. S. Strazanac, G. K. Roderick, Molecular phylogeny of *Banza* (Orthoptera: Tettigoniidae), the endemic katydids of the Hawaiian Archipelago, Molecular Phylogenetics and Evolution 41 (2006) 53–63.

[29] D. Shotton, Semantic publishing: the coming revolution in scientific journal publishing, Learned Publishing 22 (2009) 85–94.

[30] D. W. Weisrock, T. J. Papenfuss, J. R. Macey, S. N. Litvinchuk, R. Polymeni, I. H. Ugurtas, E. Zhao, H. Jowkar, A. Larson, A molecular assessment of phylogenetic relationships and lineage accumulation rates within the family Salamandridae (Amphibia, Caudata), Molecular Phylogenetics and Evolution 41 (2) (2006) 368 – 383.

[31] Y. L. Werner, The case of impact factor versus taxonomy: a proposal, Journal of Natural History 40 (2006) 1285–1286.