Towards realising Darwin’s dream: setting the trees free

Roderic D. M. Page (r.page@bio.gla.ac.uk)

DEEB, FBLS, University of Glasgow, Glasgow G12 8QQ, UK

Summary

The fact that all living organisms are related by common descent is one of the central principles of modern biology. Since the early 1990’s the amount of data available to evolutionary biologists has exploded, and Elsevier’s journal *Molecular Phylogenetics and Evolution*, has become the largest single publisher of evolutionary trees (phylogenies). These trees and their supporting data potentially form a tremendous resource for biologists, with applications in genomics, evolutionary biology, biodiversity, and public health. However, most published trees are not available in any public database, but instead languish as images, “locked up” in the pages of journals. A long term solution to this problem is to invert the relationship between journal and database, such that the database is the primary repository, and the journal article becomes effectively a “report” on that data, albeit a report that is citable, and thus has the same status as a scientific article. This vision is some way off being achieved. However, publishers could greatly enhance the scientific value of their digital content by expanding article metadata to include taxonomic names, shared digital identifiers, geographical coordinates, and data structures such as evolutionary trees. This metadata doesn’t require making the full text available, but could substantially improve the findability of that text.

Project Goals, Purpose, and Outcomes

From publishing articles to publishing data

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. But at present the machine-readable content is still a second-class citizen, often limited to bibliographic metadata about the article. Much of the key scientific content in a paper resides as bitmap images embedded in PDF or HTML documents, or in the body of the text. Efforts to increase the usability of these data incur a considerable cost in extracting data from articles and converting them into a format suitable for databasing.

A long term solution to this problem is to invert the relationship between journal and database, such that the database is the primary repository, and the journal article becomes effectively a “report” on that data, albeit a report that is citable, and thus has the same status as a scientific article.

The goal of this proposal is to make the case for pursing this vision by showing how a particular class of scientific study (evolutionary trees) can be made more accessible by users, and more connected to other digital resources. By expanding article metadata to include taxonomic names, shared digital identifiers, geographical coordinates, and data structures such as evolutionary trees, publishers can greatly increase the scientific value of their digital content. At the same time this will provide novel ways for users to discover this content. By developing simple tools to extract extended metadata, and demonstrating the value of this information, the project endeavours to make the case for making generating this metadata an intrinsic step in the publishing process.

Making documents findable

Electronic publications are essentially opaque "black boxes", their digital content locked inside PDFs or HTML pages (Figure 1(a)). In order to make these documents findable, indexing and abstracting services such as Google Scholar and PubMed make use of fragments of text (e.g., the abstract) or full-text indexing to help users find relevant content (Figure 1(b)).

Document findability would be significantly improved if additional information was extracted, 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) (Figure 1 (c)). For example, querying on the taxonomic name "Aves" would find all papers on birds, even if those documents didn't all include the term "Aves". Extracting identifiers enables the document to be linked to external databases, leading to the development of metrics of data citation, and making provenance traceable (described further below). 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 query that finds publications whose content is spatially close to that article (e.g., "find other articles on species within 100 km of the centre-point of this study").

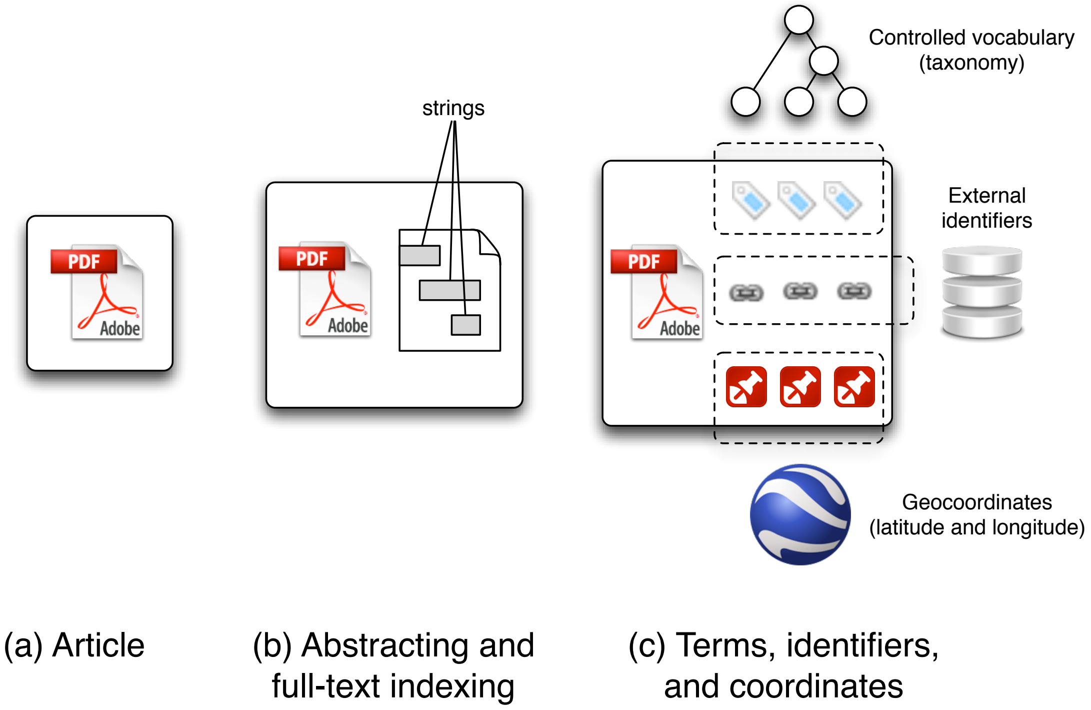


Figure 1 Increasing the findability of documents by increasingly sophisticated indexing. (a) A document, such as a PDF file. (b) Indexing strings, such as keywords, facilitates basic text searching. (c) Indexing terms from controlled vocabularies, database identifiers, and geographical coordinates.

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). Basic bibliographic metadata is already made available to abstracting services. For relatively little effort, additional metadata could be exposed, greatly increasing the document's findability and usability. A model of this approach is Nature's Open Text Mining Initiative (http://opentextmining.org/wiki/Main_Page).

Beyond linking to just the literature

Through CrossRef the publishing industry has made great strides in linking articles together in citation networks. A reader browsing a paper online can typically click on links to papers cited by (or citing) the current paper. However, this still treats the scientific publication as a single atomic entity, rather than as an aggregation of information (Figure 1 (c)). For example, modern phylogenetic studies typically refer to macromolecular sequences in the Genbank database, and voucher specimens in museum collections. A given DNA sequence may be used in multiple studies, and a given specimen may be the source of multiple sequences. A scientific paper cites these objects, just as it cites previous literature. By extracting the identifiers corresponding to these objects, we can expand citation networks to include these objects. Just as we can compute a journal's impact factor, we could compute the impact factor of, say, a DNA sequence, based on the number of times that sequence is cited. I have explored the use of Google's PageRank algorithm in this context (Page, 2008). Linking to identifiers also provides a mechanism for readers to discover related content (e.g., the publisher's web site could list other papers that use the same data), and enables authors to see who is using their data. An additional benefit from linking using identifiers is the acquisition of further metadata by "following the links" (Figure 2).

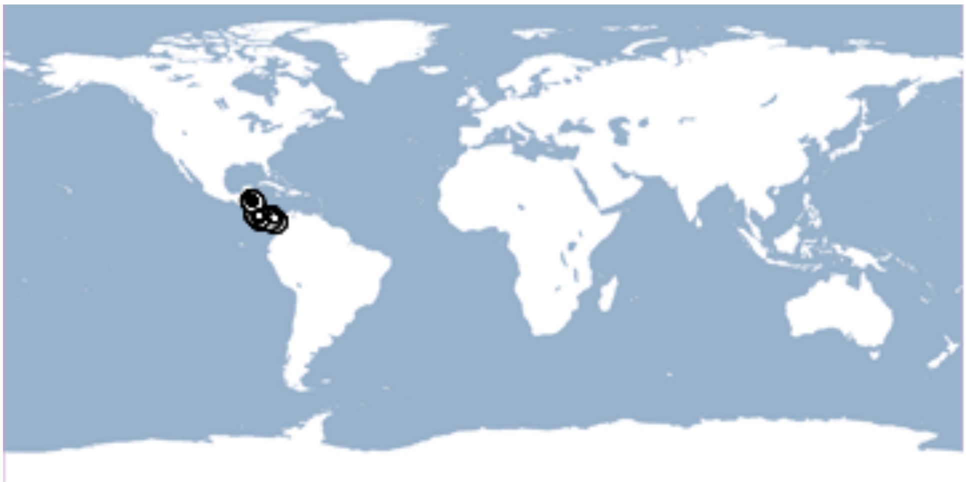


Figure 2 A map generated for Wang et al. (2008). The localities were obtained by (1) extracting the GenBank DNA sequences linked to the PubMed entry for this paper, then (2) extracting museum specimen codes from those sequences, and finally (3) querying the museum specimen databases for information about the corresponding specimens. This harvesting of linked metadata was done automatically using tools that are part of the iPhylo demo.

Data loss

Elsevier's journal *Molecular Phylogenetics and Evolution* (MPE; ISSN 1055-7903) is "dedicated to bringing Darwin's dream - to 'have fairly true genealogical trees of each great kingdom of Nature' - within grasp", and has become the largest single source of published evolutionary trees. However, few of these trees are available for analysis in public databases such as TreeBASE (http://www.treebase.org) (Figure 3).

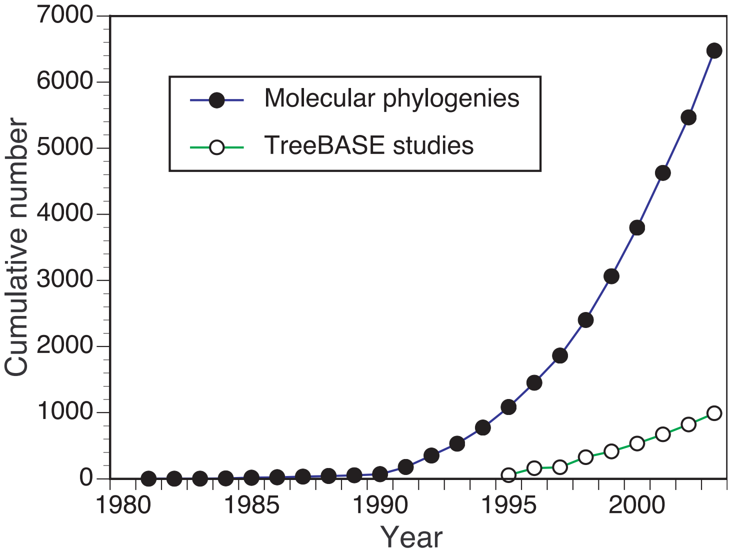
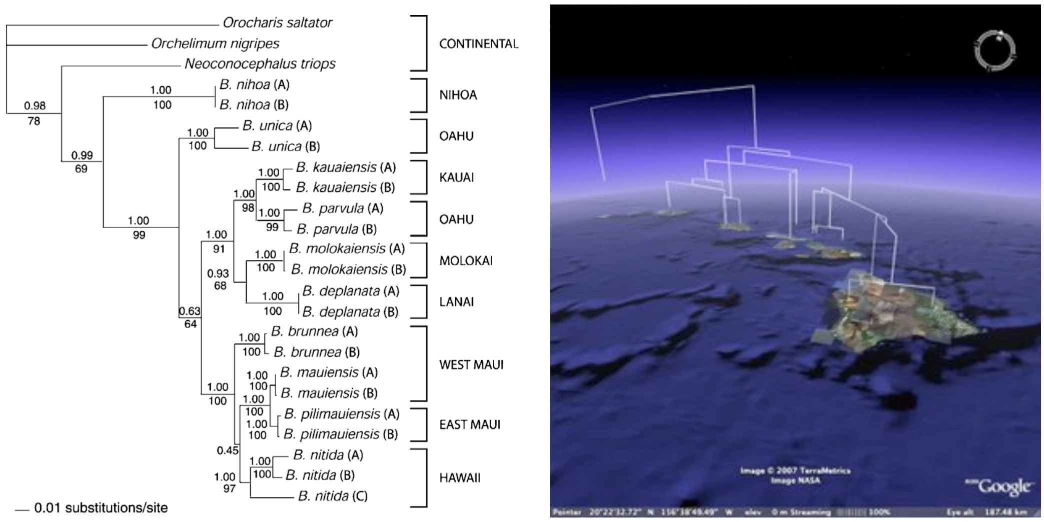


Figure 3 Cumulative growth of publications on phylogenetics, based on the number of papers found in the Web of Science by searching on the key words ``molecular" and ``phylogenetic" since 1981 compared with the growth of the phylogenetic repository TreeBASE, which launched in 1996 (a study in TreeBASE is equivalent to a single paper). From Page (2007).

One reason for the growing gap between publications and database is what authors perceive as the tedium of entering data and trees into yet another place. In other words, databasing is not an intrinsic part of the publishing process. As an intermediate step towards journals evolving into databases, publishers could increase the quantity of domain-specific metadata they provide for each article, and this metadata could be harvested automatically by external databases. Without such a step, the gap between the two lines in Figure 3 is likely to grow.

Novel visualisations

In addition to the data "loss" incurred by the lack of adequate databasing, the constraints of a journal page (whether printed or online) places severe limits on displaying large evolutionary trees. Many of these limitations disappear if the user has access to the trees themselves. Tools such as Google Earth open up possibilities for novel visualisations that are easy to construct, and visually compelling. For example, Figure 4(a) shows a phylogeny for Hawaiian for *Banza* katydids (small cricket-like insects), taken from Fig. 3 in Shapiro et al. (2006). By adding the geographical coordinates published in Appendix A in the same paper, we can display this tree in geographical space (Figure 4 (b)). We can further refine this visualization by making the altitude of the nodes in the tree proportional to genetic divergence, or geological time, yielding insights into the biogeographic history of this group of organisms. Constructing these visualisations requires access to the phylogenetic trees and associated metadata.



(a) tree

(b) Google Earth tree

Figure 4 Novel visualization of a phylogenetic tree. (a) Phylogenetic tree as published in *Molecular Phylogenetics and Evolution* (Shapiro et al., 2006). (b) Same tree displayed in Google Earth (see http://iphylo.blogspot.com/2007/06/google-earth-phylogenies.html).

Specific goals

For the entire corpus of MPE, the goals of this project are to extract

1. digital identifiers, such as GenBank sequence numbers and museum specimen codes from text
2. latitude and longitude coordinates from text
3. phylogenetic trees from figures

In addition, all PubMed records and linked GenBank sequences for MPE articles will be retrieved from NCBI. Where possible, all identifiers will be resolved and the metadata retrieved will be stored in a database. Any geographical coordinates retrieved from sequence or specimen records will be used to supplement latitude and longitudes extracted directly from the text. Task 3 is more challenging and depends on the success of tools to automatically extract phylogenetic tree descriptions from images. Trees so obtained will be stored in the database, and linked to the corresponding publication.

Detailed Project Description: Content and Functionalities

This project would make use of the journal *Molecular Phylogenetics and Evolution* (other Elsevier journals also publish phylogenies, but MPE is the largest journal devoted solely to this topic). Text mining would use both the XML and PDF versions of the articles, phylogenetic tree extraction would use the PDFs. Raw text and image extraction would use the Open Source xpdf package (programs `pdftotext` and `pdfimages`). I have written Perl scripts to extract identifiers and geographical coordinates from text and PDF files. The uBio findIT web service would be used to extract taxonomic names. Dr Joseph Hughes (a researcher in my lab) has developed a tool (Tree ripper) to automatically extract a phylogenetic tree from a bitmap using image cleaning and OCR techniques. Identifiers, such as GenBank accession numbers and specimen codes would be resolved using tools the author as developed. Metadata extracted will be stored in a MySQL database using an entity-attribute-value model. Google Earth KML files will be generated for each article that contains geographical coordinates. The output will be a web interface to all of MPE, modelled on the iPhylo demo (http://iphylo.org/~rpage/demo1/), to enable users to explore the links between publication and data. I will also construct a mock-up of a MPE web page showing how the additional links might be represented on the publisher's page.

Project Background

This proposal builds on work that I have documented on my blog "iPhylo" (http://iphylo.blogspot.com), and in various publications (Page, 2007, 2008). I have also created a web site (http://darwin.zoology.gla.ac.uk/~rpage/elsevier/) that provides more details on the examples discussed above. Much of the database and extraction software has already been developed. For example, my service for extracting latitude and longitudes from PDFs is available at http://bioguid.info.services. A demonstration site showing papers linked to sequences, specimens, and geography is online at http://iphylo.org/~rpage/demo1. The intention for this competition would be to refine this interface, and populate it solely with MPE-derived content.

Methods

The timeline for this project is

Month	Task
August	Develop scripts to parse MPE content, populate database with basic bibliographic metadata.
September	Refine and apply my identifier and locality extraction algorithms.
October	Refine automatic tree extraction algorithm. This is a challenging problem that we have made some progress on. Although it would be a great asset to the project to extract the trees, if we do not have sufficient success in the allotted time we will still have generated considerable body of additional metadata.
November	Integrate additional metadata into database, create web site based on my work on iPhylo, and generate a mock-up of Elsevier web page with additional content.

References

Page, RDM. 2007. Towards a Taxonomically Intelligent Phylogenetic Database. Available from Nature Precedings doi:10.1038/npre.2007.1028.1

Page, RDM. 2008. Biodiversity informatics: the challenge of linking data and the role of shared identifiers. Briefings in Bioinformatics. doi:10.1093/bib/bbn02

Shapiro, LH, Strazanac, JS, Roderick, GK. 2006. Molecular phylogeny of *Banza* (Orthoptera: Tettigoniidae), the endemic katydids of the Hawaiian Archipelago. Molecular Phylogenetics and Evolution 41 (1) 53. doi:10.1016/j.ympev.2006.04.006

Wang, IJ, Crawford, AJ, Bermingham, E. 2008. Phylogeography of the Pygmy Rain Frog (*Pristimantis ridens*) across the lowland wet forests of isthmian Central America. Molecular phylogenetics and evolution 47 (3) 992. doi:10.1016/j.ympev.2008.02.021