

Wikipedia as an encyclopaedia of life

Roderic D. M. Page

Division of Ecology and Evolutionary Biology, Faculty of Biomedical Life Sciences,
University of Glasgow, Glasgow 12 8QQ, UK

Introduction

In his 2003 essay E O Wilson (Wilson 2003) outlined his vision for an “encyclopaedia of life” comprising “an electronic page for each species of organism on Earth”, each page containing “the scientific name of the species, a pictorial or genomic presentation of the primary type specimen on which its name is based, and a summary of its diagnostic traits.” Although the “quiet revolution” in biodiversity informatics (Bisby 2000) has generated numerous online resources, including some directly inspired by Wilson's essay (e.g., <http://ispecies.org>, <http://www.eol.org>), we are still some way from the goal of having available online all relevant information about a species, such as its taxonomy, evolutionary history, genomics, morphology, ecology, and behaviour. While the biodiversity community has been developing a plethora of databases, some with overlapping goals and duplicated content (Thomas 2009), Wikipedia has been slowly growing to the point where it now has over 100,000 pages on biological taxa. My goal in this essay is to explore the idea that, largely independent of the efforts of biodiversity informatics and well-funded international efforts, Wikipedia (http://en.wikipedia.org/wiki/Main_Page) has emerged as potentially the best platform for fulfilling E O Wilson's vision.

Wikis

Wilson (2003) envisaged a system where each “page is indefinitely expandable. Its contents are continuously peer reviewed and updated with new information. All the pages together form an encyclopaedia, the content of which is the totality of comparative biology.” Although Wilson did not mention wikis (which were in their infancy when he wrote his article), to today's reader his vision has echoes of several key features of a wiki. Wikis are expandable, and can be continuously edited and updated. Perhaps just as significant as the pages themselves is the emergence of the large community of contributors to sites such as Wikipedia. The potential of this community to make significant contributions to the task of biological annotation is already being explored by other biologists (Waldrop 2008), notably in the Gene Wiki project which has created numerous Wikipedia pages for human genes (J. W. Huss et al. 2009; Jon W. Huss et al. 2008). This project was motivated by the realisation that centralised annotation by a small pool of experts simply couldn't keep pace with the rapid growth of biomedical literature. Increasing concerns about the accuracy of DNA sequences and their annotations in the GenBank sequence repository (Bridge et al. 2003), and the difficulty of correcting these has led to calls to “wikify” GenBank (Bidartondo 2008), so that errors can be rapidly corrected by the biological community, although it has to be said that GenBank curators have not greeted this proposal with enthusiasm (Pennisi 2008).

A further reason for taking Wikipedia seriously is its dominance of internet search results. The first thing many people will do when encountering an unfamiliar taxonomic name is search for information about that name using Google's search

engine. In many cases Wikipedia will provide one of the top “hits”. To illustrate this, I took the 5416 mammal species names from the Mammals Species of the World database (<http://www.bucknell.edu/MSW3/>) (Wilson & Reeder 2005) and looked up each name in Google, recording the URLs of the first ten web sites that Google found. Wikipedia dominated the search rankings, with 97% of mammal species having a Wikipedia page in the top ten search hits. If we look at the rankings of each hit within the search results then Wikipedia’s dominance becomes even more striking. For almost half of the mammals Wikipedia is the first hit found by Google, and for just under three quarters of the species Wikipedia is either the first or second hit (details available at <http://iphylo.blogspot.com/2009/09/google-wikipedia-and-eol.html>). It might be tempting to think that Wikipedia’s dominance is taxon-specific – mammals are charismatic and hence are likely to have a strong presence in Wikipedia compared to other, less popular taxa. However, many mammal pages in Wikipedia are small “stubs,” giving little more than basic taxonomic information, yet even these stubs appear near the top of Google’s search results. Furthermore, if we extend the analysis to all species names in Wikipedia, the encyclopaedia’s dominance becomes even more apparent. For the 72,000 species names it contains, Wikipedia is an order of magnitude more likely than other web sites to be the first search result returned by Google (Fig. 1).

If web visibility and the number of potential contributors were the sole considerations, creating the encyclopaedia of life within Wikipedia would seem the obvious solution do. But given the considerable resources that have been invested in existing biodiversity informatics initiatives (Thomas 2009), perhaps we should first evaluate the current state of taxonomic information in Wikipedia.

Taxonomic information in Wikipedia

The taxonomic classification in the English language Wikipedia follows the Linnaean system where most taxa have ranks, such as Kingdom, Order, Class, Genus, and Species. Each taxon in Wikipedia has a corresponding page. A typical taxon page in Wikipedia uses a template called a “Taxobox,” which lists a set of attributes for that taxon, such as a simple classification, the scientific name for the organism, and its conservation status (Fig. 2).

Using taxon names as Wikipedia page names poses several problems: a taxon may have more than one name, and the same name may correspond to more than one taxon. These problems of synonymy and homonymy, respectively, are not unique to taxonomy. Many people, places, and concepts may have more than one name, and the same term can mean many different things (for example, “bank” can mean a financial institution or a river bank, to name just a few meanings). Wikipedia handles synonyms (and common names) using “redirection”. For example, the page for “*Morus bassanus*” automatically sends the user to the page “Northern Gannet”, which contains the Wikipedia entry for *Morus bassanus*.

Wikipedia’s mechanism for dealing with homonyms is to have a disambiguation page, which lists the possible meanings of the word from which the user can choose the one they intended. For example, *Morus* is a genus of birds (gannets) and a genus of plants (mulberry trees). The Wikipedia page for “Morus” (<http://en.wikipedia.org/wiki/Morus>) lists both genera, and the user can decide whether they meant the bird or the plant.

Gregg's paradox

Wikipedia does run into some problems that are specific to taxonomy, notably Gregg's paradox (Buck & Hull 1966). Gregg (1954) argued that if we (a) treat taxa as sets defined by extension (i.e., by listing all the members of that set), and (b) accept that two sets with exactly the same content must be the same set, then many traditional biological classifications contain redundancy because the same taxon may be assigned to multiple levels in the Linnaean hierarchy. For example, the aardvark, *Orycteropus afer*, is the only extant species of the genus *Orycteropus*, which is the only extant member of the family Orycteropodidae, which in turn is the sole extant representative of the order Tubulidentata. Under Gregg's model, Tubulidentata, Orycteropodidae, *Orycteropus* are all redundant as they have exactly the same content (namely the aardvark *Orycteropus afer*).

The Wikipedia page for the aardvark (<http://en.wikipedia.org/wiki/Aardvark>) exemplifies Gregg's paradox (Fig. 2). If the aardvark is the sole representative of the Tubulidentata, then there is no relevant information that could be put on a Tubulidentata page (or pages for Orycteropodidae and *Orycteropus*) that doesn't also belong on the page for the aardvark. Hence these taxa are listed in the Taxobox without links. It would only make sense to create pages for them if additional taxa existed that could be assigned to other species of *Orycteropus*, other genera of Orycteropodidae, or other families of Tubulidentata. Such taxa do exist (Pickford 1974; Thomas Lehmann 2009; T Lehmann et al. 2006), and so if and when they are added to Wikipedia these higher taxa would merit their own pages.

Gregg's paradox is a consequence of ranks and requiring each rank (or at least a reasonable subset of them) to exist in a classification. If we ignore ranks, then there's no reason to put any taxa between Afrotheria and *Orycteropus afer*. However, by itself this won't obviate the paradox as many species, notably fossils, belong in monotypic genera. Wikipedia will contain a page for the genus, or the species, but not both.

Classification

A biological classification can be represented a rooted tree in which each node has a single parent node (it's "ancestor", the root is the one exception as it has no ancestor) and one or more child nodes (descendants) (Fig. 3). The simplest way for Wikipedia to depict a classification of a given taxon would be to simply list the page corresponding to the parent node of that taxon (e.g., for Afrotheria this would be Mammalia). This would ensure the classification would be a tree, at the cost of being rather minimal. Instead Wikipedia pages list the complete lineage of a taxon, and frequently its child taxa as well. Hence, a Wikipedia page both "has parent: and "has child" relationships. Because one necessarily entails the other, having both two relationships is redundant (in Fig. 3 node B is a child of node A then by definition node A is the parent of node B). Because these relationships are entered manually into different Wikipedia pages (often by different contributors at different times) they can become inconsistent. For example, the Wikipedia page for Amphibia lists the children of Amphibia as the order Temnospondyli, and the subclasses Lepospondyli and Lissamphibia (Anura, Caudata, and Gymnophiona) (Fig. 4a) Hence we would expect the Lissamphibia to be the parent of the Anura, Caudata, and Gymnophiona, whereas this is only the case for the page for the Caudata (salamanders) — the Anura and Gymnophiona both link directly back to the Amphibia page. While this may seem a

mild inconsistency, there are over 200 Wikipedia pages for amphibian genera that list Amphibia as the parent page, despite the Amphibia page itself listing only four immediate children (Fig. 4b). These additional genera represent fossils of uncertain affinity, so in effect the implied Wikipedia tree for Amphibia (at least, the tree defined by the “has parent” relationship) has a large basal polytomy reflecting this ignorance (Nelson & Platnick 1980).

Quality of Wikipedia pages

Wikipedia is community-edited — literally anyone can edit almost any article. This can be seen as both a strength (potential contributors aren't excluded) and as a weakness (in the most extreme case, the page may be vandalised), and has raised questions concerning the reliability of the content of Wikipedia articles (J. Giles 2005).

Citations

One approach to evaluating the quality of a Wikipedia page is to evaluate the quality of the sources the page cites. Nielsen (2007) found that many Wikipedia pages cited the primary scientific literature, and that journals most highly cited by Wikipedia were high impact factor journals such as *Nature* and *Science*. In other words, Wikipedia citation patterns reflect citation patterns typical of the scientific literature. From a taxonomic perspective it is interesting that journals such as *Australian Systematic Botany* and *Nuytsia* have higher citation rates than predicted by their impact factor. A subsequent analysis (Nielsen 2008) ranked the high-volume taxonomic journal *Zootaxa* eleventh among all journals cited by Wikipedia (http://neuro.imm.dtu.dk/services/wikipedia/enwiki-20080312-ref-articlejournal_highlycited.html). Given the ongoing controversy about whether impact factor adequately measures the worth of taxonomic publications (Krell 2002; Garfield 2001; Werner 2006), creating well-referenced articles in Wikipedia could be one way for taxonomists to increase the visibility of their work, which in turn may lead to increased citations (Lawrence 2001).

It is worth noting that whereas many Wikipedia articles contain direct links to the primary literature via DOIs or PubMed numbers, this cannot be said of any of the flagship biodiversity databases such as EOL or the Catalogue of Life. These databases treat literature as if the web did not exist, simply displaying citations as text strings without links to the actual publications. Whereas a reader of a Wikipedia article is provided with numerous points of departure for further browsing, a user of a typical biodiversity database is faced with the prospect of cutting and pasting text into Google to try and locate what references the database may provide.

Controversy

Wikipedia pages are open to anybody to edit, and some controversial topics have been the subject of “edit wars” where contributors with one viewpoint repeatedly delete content added by contributors holding a rival view. The disciplines of taxonomy and systematics are not without their own controversies (Hull 1990), hence it will come as no surprise that there are taxon pages in Wikipedia that have been the subject of edit wars. Some of the bitterest disputes are over relatively trivial taxonomic details, such as the correct name for the sperm whale, a debate conducted somewhat more civilly in scientific literature (Schevill 1986; Holthuis 1987; Schevill 1987). Because Wikipedia

retains a complete history of every edit made, a user can browse this history to see whether the current wiki page (which displays only the most recent version) is a fair reflection of the debate between the rival factions. This transparency has inspired the development of tools to quantify and visualise the edits made to a Wikipedia page (Vuong et al. 2008; Viégas et al. 2004). One of the most attractive visualisations is “history flow” (Viégas et al. 2004), which displays a timeline of successive edits to a Wikipedia page, colour-coded by contributor, in which one can see the fate of individual contributions (Fig. 6).

Internal consistency

Wikipedia pages are essentially text documents, which can be edited independently by different users at different times. In this sense Wikipedia is rather different from a database, where records can be interlinked so that a change in one record (say a customer’s address) can be propagated throughout the rest of the data (so that every order a customer makes is linked to their new address). Because it lacks these checks on consistency it is possible for Wikipedia pages to become mutually inconsistent, as we saw with the Amphibian example (Fig. 5). The mammal pages in Wikipedia generally follow the Mammals Species of the World classification (Wilson & Reeder 2005) (except for fossil taxa which aren’t included in Mammals Species of the World). When I extracted the mammal pages from Wikipedia and attempted to build a tree for mammals using the “has parent” link, instead of a single tree (as one might expect) I obtained a graph with over 800 distinct components (sets of nodes connect to each other but not to other nodes in the graph). The largest component in the graph corresponds to a tree (Fig. 5) closely resembled the Mammals Species of the World classification, which is reassuring, but the remaining components represent orphaned” pages, that is, pages that are not linked to the page for the relevant higher taxon.

To wiki or not to wiki

If a primary goal of biodiversity informatics is to make basic information about taxa widely available (and, more to the point, *findable*) then Wikipedia’s dominance of Google’s search results suggests that this is where we should be focussing our efforts. No other source of information on the web comes close to Wikipedia in terms of web presence, and potential size of contributors. The relative prominence of citations to the primary taxonomic literature is another incentive, given the widespread feeling that measures such as impact factor are poor metrics for this field (Krell 2002).

Looking ahead, linked data provides another reason for engaging with Wikipedia. The web most of us are familiar with is a web of documents, such as web pages, images, movies, and other media (including PDF files), designed to be viewed (or watched and listened to) by people. But the web is also full of data, and linked data (<http://linkeddata.org/>) is an approach to connecting this data across the web, making the web in effect a single enormous database. The “links” in “linked data” depend on shared identifiers, so that different data sets use the same identifier when referring to the same entity. Because of its size and scope of coverage, Wikipedia (through the DBpedia project) has emerged as the natural source of many of these identifiers (Bizer et al. 2009). Organisations such as the BBC that are seeking to organise their own extensive media collections and integrate these with other databases (Kobilarov et al. 2009) are reusing Wikipedia-derived identifiers for taxa, adaptations, and ecosystems. It is likely that Wikipedia-derived identifiers will be central to any efforts

to integrate information from taxonomy and systematics with information derived from other disciplines (e.g., geography, climate, economics, history).

Dominance of search result rankings, contributor size, and potential linkage to other data are all strengths of Wikipedia, but as we have seen above, it is not without its flaws. One approach to addressing the limitations of Wikipedia's taxonomic content is that adopted earlier this year by EOL (<http://www.eol.org>), which has started incorporating content from Wikipedia into its own pages, albeit the Wikipedia-derived content is held in quarantine and flagged as "unreviewed". This approach contrasts with that adopted, say, by the BBC, which reuses Wikipedia content, editing Wikipedia pages directly if it is felt that an article is not of sufficient quality.

If adopting Wikipedia as the platform for the encyclopaedia of life seems a step too far, I would argue that wikis in general will still have a major role to play in mobilising biodiversity data. Even if we restrict ourselves to biodiversity, we face a major challenge trying to link disparate databases together (Thomas 2009). These databases are replete with identifiers such as taxonomic names, bibliographic identifiers and citations, museum specimen codes, and GenBank accession numbers (R. D. M. Page 2008). Inconsistency in the use of taxonomic names and the way bibliographic records are treated (Roderic DM Page 2007), coupled with database errors can make data integration a time consuming task. Furthermore, ambitious digitisation efforts such as the Biodiversity Heritage Library (<http://www.biodiversitylibrary.org>) (Rinaldo 2009) are generating huge volumes of text extracted by optical character recognition (OCR) from images. This text is of variable accuracy, but contains a wealth of information about the biology and taxonomy of the Earth's biota. Automated efforts to extract information from this OCR text have met with variable success (Lu et al. 2008). By opening the process of annotation, correction, and linking to more participants, wikis may hasten the time when biodiversity data becomes truly integrated, and we become a step closer to realising the dream of an encyclopaedia of life.

Acknowledgements

All analyses reported here were performed on the June 18, 2009 dump of Wikipedia. Many of the ideas outlined here were first explored on my blog <http://iphylo.blogspot.com>. I thank the numerous people who provided feedback on those blog posts, and to audiences at the Sloan Foundation and Sheffield University who have heard me talk about this topic. Alex Wild's blog post <http://myrmecos.wordpress.com/2009/06/06/pyramica-vs-strumigenys-why-does-it-matter/> brought the edit war over *Pyramica* to my attention, and Tony Rees alerted me to the rather intemperate language being used in the debate over whether the proper name for the sperm whale is *Physeter catodon* or *Physeter macrocephalus*. I thank Rudolf Meier for inviting me to write this essay, and for his patience as deadlines began to slip.

Literature cited

- Bidartondo, M.I., 2008. Preserving Accuracy in GenBank. *Science*, 319(5870), 1616a-1616a.
- Bisby, F.A., 2000. The Quiet Revolution: Biodiversity Informatics and the Internet. *Science*, 289(5488), 2309-2312.
- Bizer, C. et al., 2009. DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), 154-165.
- Bridge, P.D. et al., 2003. On the unreliability of published DNA sequences. *New Phytologist*, 160(1), 43-48.
- Buck, R.C. & Hull, D.L., 1966. The Logical Structure of the Linnaean Hierarchy. *Systematic Zoology*, 15(2), 97.
- Garfield, E., 2001. Taxonomy is small, but it has its citation classics. *Nature*, 413(6852), 107.
- Giles, J., 2005. Internet encyclopaedias go head to head. *Nature*, 438(7070), 900-901.
- Gregg, J.R., 1954. *The language of taxonomy: an application of symbolic logic to the study of classificatory systems*, New York: Columbia University Press.
Available at: [Accessed February 18, 2010].
- Holthuis, L.B., 1987. The Scientific Name of the Sperm Whale. *Marine Mammal Science*, 3(1), 87-89.
- Hull, D., 1990. *Science as a process : an evolutionary account of the social and conceptual development of science* Paperback ed., Chicago: University of Chicago Press.
- Huss, J.W. et al., 2009. The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Research*, 38(Database), D633-D639.
- Huss, J.W. et al., 2008. A Gene Wiki for Community Annotation of Gene Function. *PLoS Biology*, 6(7), e175.
- Kobilarov, G. et al., 2009. Media Meets Semantic Web --- How the BBC Uses DBpedia and Linked Data to Make Connections. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*. Heraklion, Crete, Greece: Springer-Verlag, pp. 723-737.
- Krell, F., 2002. Why impact factors don't work for taxonomy. *Nature*, 415(6875), 957.

- Lawrence, S., 2001. Free online availability substantially increases a paper's impact. *Nature*, 411(6837), 521.
- Lehmann, T. et al., 2006. A sub-complete fossil aardvark (Mammalia, Tubulidentata) from the Upper Miocene of Chad. *Comptes Rendus Palevol*, 5(5), 693-703.
- Lehmann, T., 2009. Phylogeny and systematics of the Orycteropodidae (Mammalia, Tubulidentata). *Zoological Journal of the Linnean Society*, 155(3), 649-702.
- Lu, X. et al., 2008. A metadata generation system for scanned scientific volumes. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. Pittsburgh PA, PA, USA: ACM, pp. 167-176. Available at: <http://portal.acm.org/citation.cfm?id=1378918> [Accessed February 24, 2010].
- Nelson, G. & Platnick, N.I., 1980. Multiple Branching in Cladograms: Two Interpretations. *Systematic Zoology*, 29(1), 86.
- Nielsen, F.A., 2008. Clustering of scientific citations in Wikipedia. 0805.1154. Available at: <http://arxiv.org/abs/0805.1154> [Accessed February 18, 2010].
- Nielsen, F.A., 2007. Scientific citations in Wikipedia. 0705.2106. Available at: <http://arxiv.org/abs/0705.2106> [Accessed February 18, 2010].
- Page, R.D.M., 2008. Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics*, 9(5), 345-354.
- Page, R.D., 2007. TBMap: a taxonomic perspective on the phylogenetic database TreeBASE. *BMC Bioinformatics*, 8(1), 158.
- Pennisi, E., 2008. DNA DATA: Proposal to 'Wikify' GenBank Meets Stiff Resistance. *Science*, 319(5870), 1598-1599.
- Pickford, M., 1974. New fossil Orycteropodidae (Mammalia, Tubulidentata) from East Africa. *Orycteropus minutus* sp. nov. and *Orycteropus chemeldoi* sp. nov. *Netherlands Journal of Zoology*, 25(1), 57-88.
- Rinaldo, C., 2009. The Biodiversity Heritage Library: Exposing the Taxonomic Literature. *Journal of Agricultural & Food Information*, 10(3), 259-265.
- Schevill, W.E., 1987. Mr. Schevill replies: *Marine Mammal Science*, 3(1), 89-90.
- Schevill, W.E., 1986. The International Code of Zoological Nomenclature and a paradigm: the name *Physeter catodon* Linnaeus 1758. *Marine Mammal Science*, 2(2), 153-157.
- Thomas, C., 2009. Biodiversity Databases Spread, Prompting Unification Call. *Science*, 324(5935), 1632-1633.
- Viégas, F.B., Wattenberg, M. & Dave, K., 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the 2004*

conference on Human factors in computing systems - CHI '04. the 2004 conference. Vienna, Austria, pp. 575-582. Available at: <http://portal.acm.org/citation.cfm?doid=985692.985765>.

Vuong, B. et al., 2008. On ranking controversies in wikipedia. In *Proceedings of the international conference on Web search and web data mining - WSDM '08*. the international conference. Palo Alto, California, USA, p. 171. Available at: <http://portal.acm.org/citation.cfm?doid=1341531.1341556>.

Waldrop, M., 2008. Big data: Wikiomics. *Nature*, 455(7209), 22-25.

Werner, Y.L., 2006. The case of impact factor versus taxonomy: a proposal. *Journal of Natural History*, 40(21), 1285.

Wilson, D. & Reeder, D.M. eds., 2005. *Mammal species of the world : a taxonomic and geographic reference* 3rd ed., Baltimore: Johns Hopkins University Press.

Wilson, E., 2003. The encyclopedia of life. *Trends in Ecology & Evolution*, 18(2), 77-80.

Figures

Fig. 1 Number of times a web site is the first hit in a Google search for a binomial name that has a page in Wikipedia.

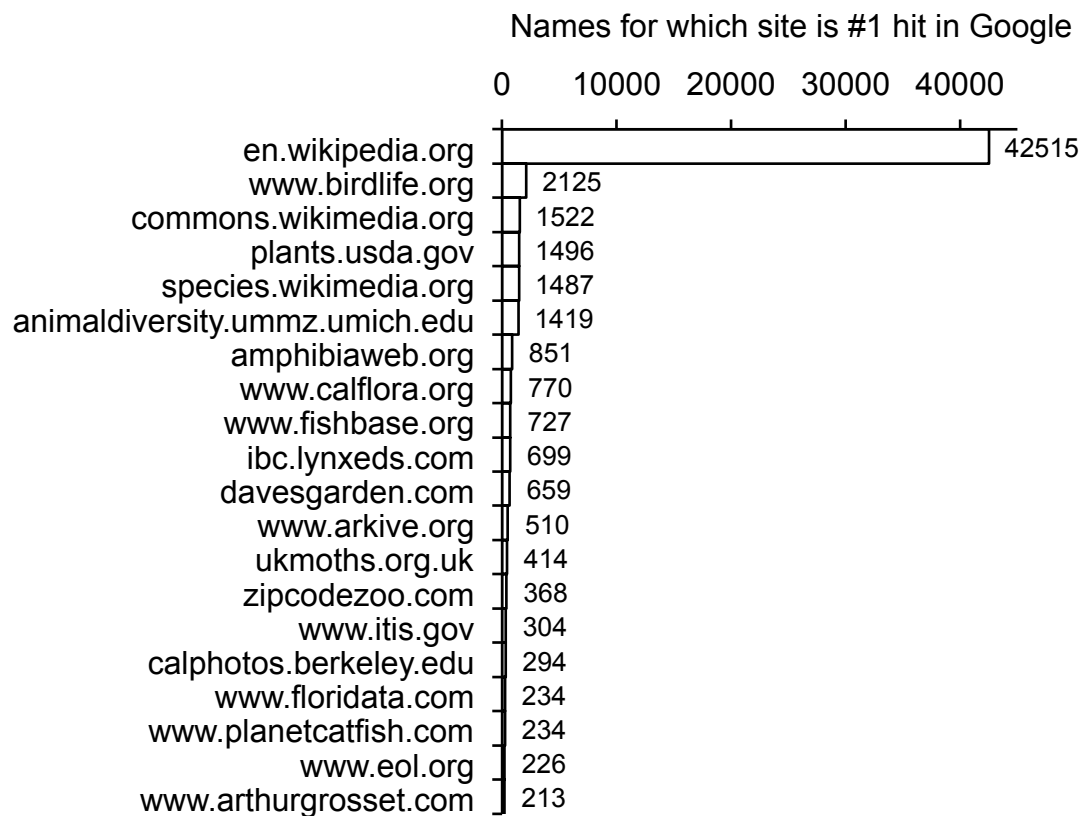


Fig. 2 Wikipedia Taxobox for the aardvark (*Orycteropus afer*), accessed October 6, 2009. Note that the taxon names Tubulidentata, Orycteropodidae, and *Orycteropus* are not clickable links (see text).

Aardvark

Fossil range: Early Miocene–Recent

PreЄ

Є

O

S

D

C


P

T

J

K

PgN



Conservation status

Extinct

Threatened

Least Concern

EX

EW

CR

EN

VU

NT

LC

Least Concern (IUCN 3.1)^[1]

Scientific classification

Kingdom:

Animalia

Phylum:

Chordata

Class:

Mammalia

Superorder:

Afrotheria

Order:

Tubulidentata

Huxley, 1872

Family:

Orycteropodidae

Gray, 1821

Genus:

Orycteropus

G. Cuvier, 1798

Species:

O. afer

Fig. 3 A tree showing the two types of links (“has child” and “has parent”) that can be used to specify relationships between the nodes. Only one kind of relationship is necessary to define the tree.

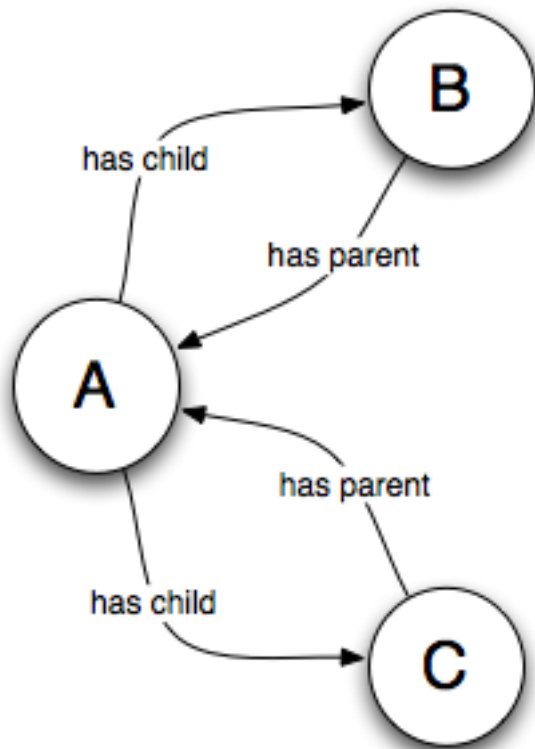


Fig. 4 Trees for Amphibia constructed from Wikipedia, based on (a) child links and (b) parent links (see Fig. 3) inconsistent. The tree based on parent links (b) has many more taxa descending from the Amphibia, including 274 genera of fossil amphibians.

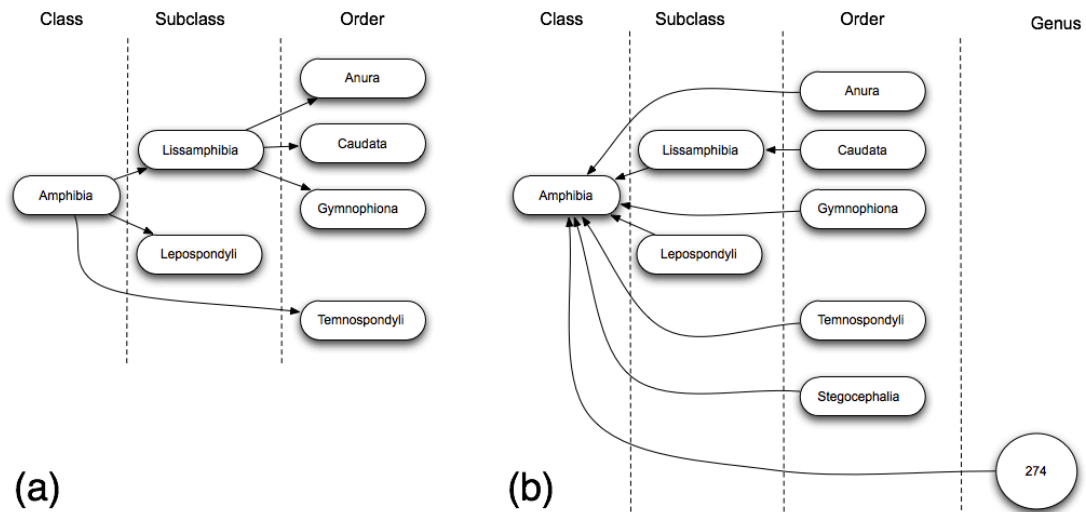


Fig. 5 Tree for the majority of mammal species in Wikipedia (see text).

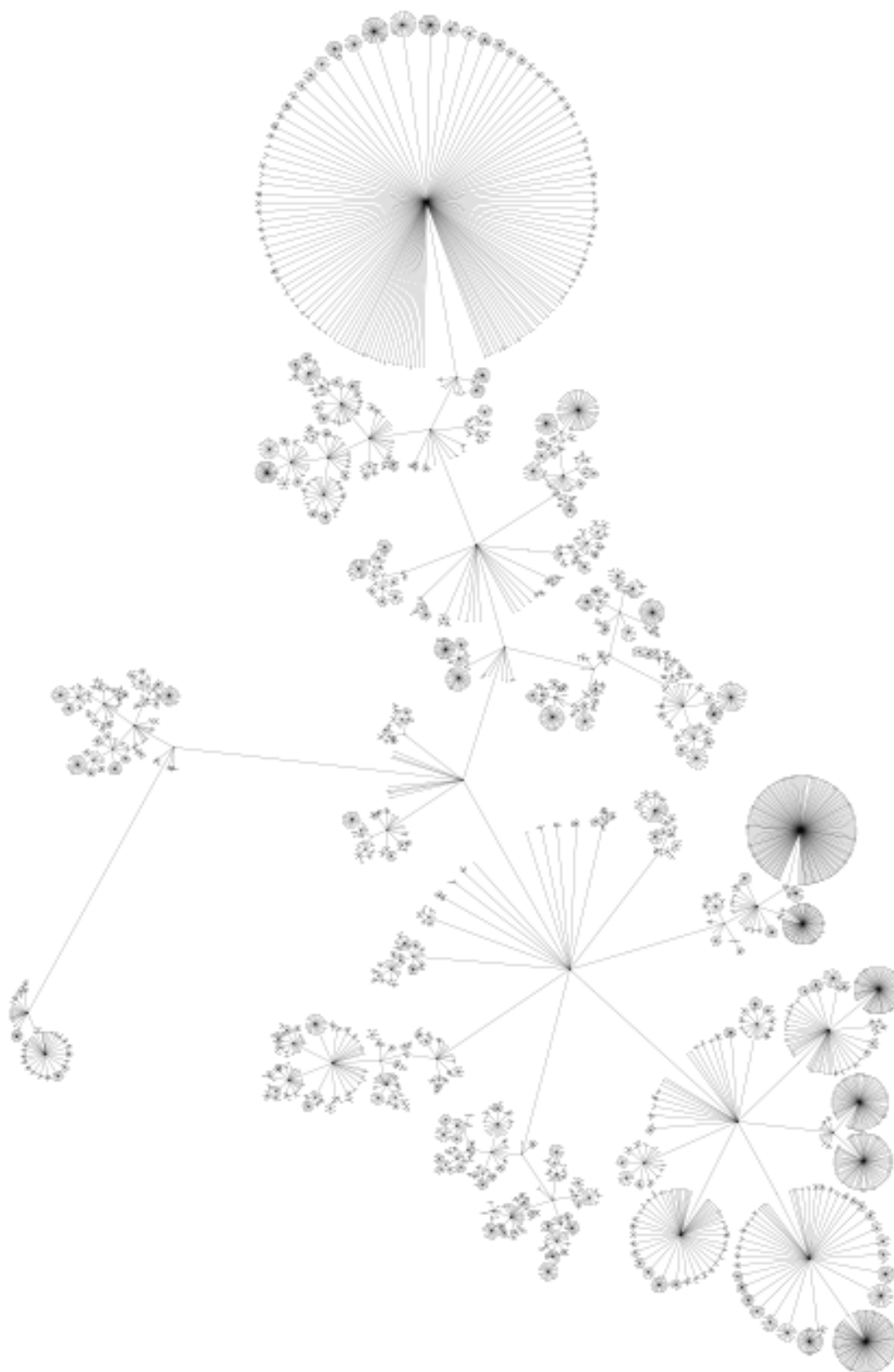


Fig. 6 A small “edit war” on Wikipedia for the page on the ant genus *Pyramica* visualised using history flow (Viégas et al. 2004). Vertical bars represent revisions of the page over time, with each block of text coloured by the user that contributed that text. The height of the vertical bar is proportional to the total size of the page, and horizontal bars connect blocks of text that remain unchanged between each edit. Significant events in the history of the page are highlighted, together with the editor’s user name and their stated reason for the edit.

