Dark Research: information content in many modern research papers is not easily discoverable online

Ross Mounce[1]

[1]**Department of Biology and Biochemistry, University of Bath, UK.**
ORCID: `http://orcid.org/0000-0002-3520-2046`

ABSTRACT

Background. Research is published in indexed, online scholarly journals so that knowledge can be easily found and built upon by others. Most scholars rely on relatively few online indexing service providers to search for relevant scholarly content. It is under-appreciated that the quality of indexing can vary across different journals and that this can have an adverse effect on the quality of research.
Objective. In this short paper I compare the recall of commonly used online indexers; Google Scholar, Web of Knowledge, Scopus, Microsoft Academic Search and Mendeley Search against a selection of over 20,000 papers published in two different high-volume journals: *PLOS ONE* and *Zootaxa*.
Results. When using Google Scholar, content in *Zootaxa* has low recall for search terms that are known to occur in it, significantly lower than the near-perfect recall of the same terms in *PLOS ONE*. All other indexers tend to have lower recall than Google Scholar except Scopus which outperformed Google Scholar for recall on *Zootaxa* searches. I also elaborate *why* Dark Research is undesirable for optimal scientific progress with some recommendations for change.
Conclusion. This research is a basic proof-of-concept which demonstrates that when searching for published scholarly content, relevant studies can remain hidden as 'Dark Research' in poorly-indexed journals, even despite expertise-informed efforts to find the content. The technological capability to do full text indexing on all modern scholarly journal content certainly exists, it is perhaps just publisher-imposed access-restrictions on content that prevents this from happening.

Keywords: Information Discovery, Information Retrieval, Indexing, Taxonomy, Phylogenetics, Cladistics, Open Access, Google Scholar, Web of Knowledge, Literature Search

INTRODUCTION

Forty years ago a thorough literature search necessitated a trip to a physical library building so that researchers could systematically hand-examine relevant journals page-by-page to visually scan for the desired concepts and items of interest. More recently, the ubiquitous electronic publication of research on the Internet has enabled less-manual, more computationally-expedited methods of literature search using computers to scan articles and books for relevant terms and concepts in text-form. This paper aims to test the extent to which various academic content discovery services can actually discover search-pattern-matching journal article content in two different megajournals, using realistic search-patterns with real-use cases that are relevant to the discipline of phylogenetics and phylogenetic methods research. These are subjects which span both biomedical and non-biomedical scientific publication venues. The simple tests I have used measure recall, which in the domain of information retrieval is defined as: the fraction of the documents that match the query that are successfully retrieved by the query. I have chosen to focus on recall specifically because I have a research interest in quantifying the discoverability of *all* published studies involving some form of phylogenetic analysis. Precision is not of importance to this aim and thus has not been assessed in this study. My default assumption is that for modern, digitally-published content, full text recall should be near 100%. Older pre-2000 work can be stuck in scans/images of text ('born-analogue') but newer post-2000 scientific articles are typically 'born-digital' and thus should be easily discoverable. Even if articles are made available behind a paywall, scholarly publishers should be able to provide indexing services with special access to index the content - so whether research is open access or behind a paywall shouldn't *in theory* matter.

A concise history of online academic content discovery services

To help academics find relevant content online Thomson Reuters released the first version of Web of Knowledge (WoK) a 'research platform' for academic content discovery over a decade ago – it launched in 2002 (Anon., 2014d). Shortly afterwards, Elsevier launched a rival profit-making commercial service called Scopus (Fingerman, 2004). Both of these indexing services are now widely used by researchers in non-biomedical biological sciences. WoK only indexes the title, abstract, keywords and citations for each article or book chapter, whereas Scopus manually-adds additional metadata terms to articles from a select range of publishers (Anon., 2014c).

It is important to note here that I will not discuss PubMedCentral (PMC) - a service commonly used by most biomedical researchers because on the whole it only indexes biomedical content. My stated subject of interest is much broader than just biomedical science. Indeed many non-biomedical journals that contain a lot of phylogeny-relevant research e.g. *Zootaxa*, *Palaeontology*, and *Journal of Vertebrate Paleontology*, *et cetera*... are simply not indexed in PMC, with the exception of a few solitary articles. Thus PMC cannot be relied-upon for literature searches for non-biomedically relevant topics.
Other relevant online academic content discovery services include Google Scholar (GS; `http://scholar.google.com/`), Scirus (`http://www.scirus.com/`), Mendeley Search (MS; `http://www.mendeley.com/research-papers/search/`) and Microsoft Academic Search (MAS; `http://academic.research.microsoft.com/`).

Google Scholar (GS) first launched a decade ago as beta in November 2004 (Anon., 2015). GS can notably achieve 100% recall for some searches (Gehanno et al., 2013) and is thus often better than Scopus & WoK's recall (e.g. Beckmann and von Wehrden (2012)). But the precision of GS is often very poor (Garcia-Perez, 2012), since it searches across a much wider body of grey literature: including some blogs, newsletters and non-peer reviewed material It also offers relatively few features with which to constrain or filter searches (other than simple 'by year/journal/author'). Moreover, there is no easy mechanism provided by which hundreds of search results can be exported in a standard format (e.g. bibtex). Thus some have pointed out that GS is not useful for performing systematic literature searches (Giustini and Kamel Boulos, 2013).

Scirus (another Elsevier-provided service), when in operation allowed full text search of a limited subset of the research literature, as well as abstract-only search, and grey literature 'scientific web' searches. However it ceased to operate during the course of this research, prior to the preparation of this manuscript.

Mendeley Search (MS) is a relatively new academic search provider, also owned by Elsevier, which claims to search across a crowd-sourced database of nearly 100 million documents (Anon., 2014b).

Microsoft Academic Search (MAS) is yet another academic search provider and is still in active development, the service is described on their About page (Anon., 2014a). GS, Scirus and an early version of Microsoft Academic Search have previously been compared (Ford and O'Hara, 2008) for searches in 2006 during which GS recalled the most citations, however the aim and methodology of that study is different to the one presented herein, and I anticipate that all of the databases may have changed in performance since 2006.

METHODS

In order to rigorously examine the recall capability of academic content discovery services for finding phylogeny-related terms published in modern (post-2000 published), digitally-authored, digitally-published papers, I scored recall against sets of articles from two high-volume megajournals, to which I have legal full text local desktop access to (see Figure 1 for a visualization of this content):

- 'Zootaxa set'. The entire set of articles published in the journal *Zootaxa* from 2001 up to Issue 3690 (1) [2013-06-11] inclusive, consisting of 12490 PDF files downloaded direct from the publisher website: `http://mapress.com/zootaxa`. This set notably includes both large monographs and small erratum notices (see Figure 1). The journal only publishes articles in PDF format. No HTML, no XML, no ePub, just PDF. Zootaxa is predominantly a subscription access journal, although a minority of authors elect to pay for 'hybrid' open access to their articles.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.773v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2015, publ: 5 Jan 2015

- 'PLOS ONE set'. All articles published in *PLOS ONE* from 2006 to 2009-12-31 inclusive, consisting of 8527 research articles obtained via BioTorrents (Langille and Eisen, 2010) in PDF format. Even though full text XML is available for this journal from PubMed Central, I purposefully chose to perform analyses using the PDF articles in order to maintain a consistency of comparison with the Zootaxa set. The dump of *PLOS ONE* PDF's available via BioTorrents only provides *PLOS ONE* PDF's up to early 2010, hence the selection period of 2006 (when *PLOS ONE* first started publishing) to 2009. PLOS ONE is an open access journal that publishes articles in PDF, HTML and XML formats.

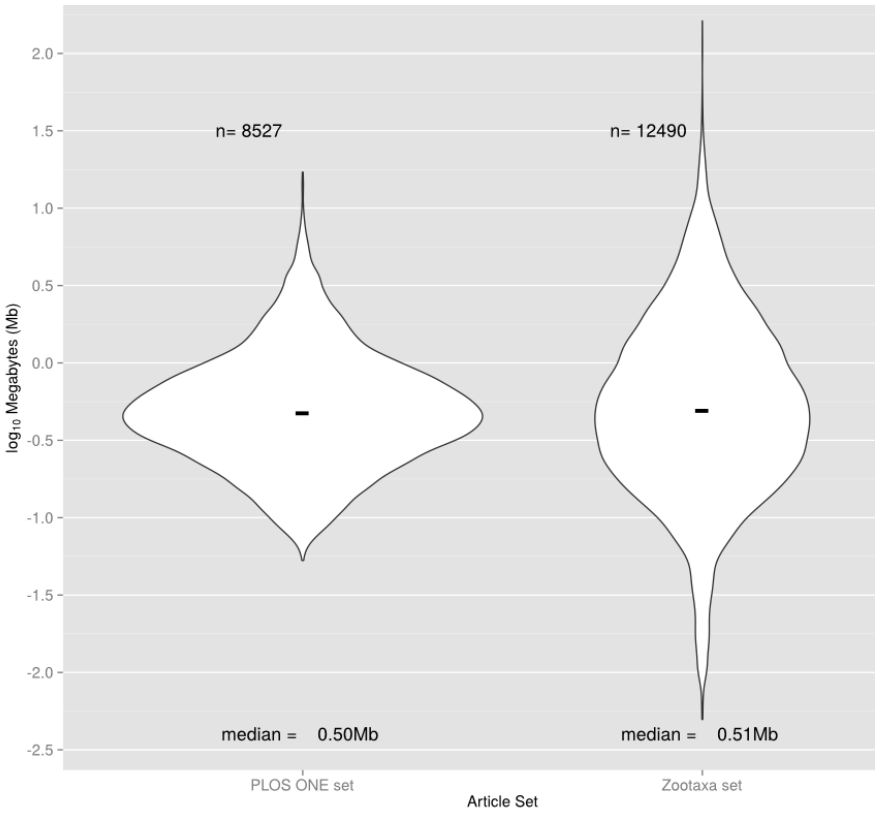

Figure 1. A comparison of the log-transformed PDF file size distribution of each article set. The distribution of the Zootaxa set varies more because this set includes both large monographs as well as small errata. The PLOS ONE set is entirely composed of research articles - no errata, essays, overviews, editorials or other articles types. Plotted with the aid of *R* (R Core Team, 2014) and *ggplot2* (Wickham, 2009)

Local command-line full text searches

The publisher-provided version of record PDF's of each set were placed in separate self-contained folders, one for the Zootaxa set, another for the PLOS ONE set. All PDFs were then converted to plain text files using *pdftotext* (`http://www.foolabs.com/xpdf/about.html`).

I then used simple command-line GNU *grep* version 2.20 (`http://git.savannah.gnu.org/cgit/grep.git/`) searches to determine which plain text documents contained phylogeny relevant word-strings that researchers may wish to search for. The *grep* searches (see Table 1) were purposefully kept very simple in order to fairly match the limited complexity of search available at the online content discovery services. If precision was desired, more complex search features such as case-sensitivity would have been used (available to grep), but as most online content discovery services do not allow case-sensitive searches, I did not employ them here. GS does not appear to support the usage of wildcards, thus I did not perform wildcard search-pattern searches with GS. The results of the grep searches were taken as the gold standard with which to measure recall against.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.773v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2015, publ: 5 Jan 2015

Table 1. The eleven different local text searches performed in this paper - the basis for the assessment of recall

Local Command-Line Search	WoK-equivalent Search
grep -iRl '\bwinclada\b'	winclada
grep -iRl '\bhennig86\b|\bhennig 86\b'	hennig86 OR 'hennig 86'
grep -iRl '\bpaup\b'	paup
grep -iRl '\bnona\b'	nona
grep -iRl '\btnt\b	tnt
grep -iRl '\bphylip\b'	phylip
grep -iRl '\bphylogeny\b'	phylogeny
grep -iRl '\bphylogen*'	phylogen*
grep -iRl '\bphylog*'	phylog*
grep -iRl '\bAedes\b'	*Aedes*
grep -iRl '\bAnopheles\b'	*Anopheles*

Many of these word-strings are the names of phylogenetic software e.g. PAUP* (Swofford, 2002), Winclada (Nixon, 2002), NONA (Goloboff, 1999), and TNT (Goloboff et al., 2008) and these are rarely mentioned in the title or abstract of papers. *Aedes* and *Anopheles* are two genera of mosquito. These are all real searches which a biologist may be interested in performing to discover academic content - these are not contrived examples. All the regular expressions searched for are in Table 1. All local *grep* searches were performed and documented in IPython notebooks (Pérez and Granger, 2007) to provide further supporting evidence for the results. This supplementary information is available on figshare (Mounce, 2015)

Searches using the online academic content discovery services

All online searches were performed on this date: 2015-01-02 (ISO 8601). Care was taken to ensure that returned 'hits' for each of the searches were constrained to the publication date ranges that I had local desktop full-text access to, which for PLOS was 2006 to 2009-12-31 (inclusive), whilst for Zootaxa this was 2001 to 2013-06-11 (inclusive). For PLOS this was easy, for Zootaxa this typically required manual removal of bibliographic records returned that were published between 2013-06-12 and 2013-12-31, outside the range of valid comparison to my local command-line searches (services such as MS could only filter by year, not exact date of publication). It was assumed that each service would not return more than one hit for the same paper - no duplicate results.

Sample search queries or URLs are given below for each service tested for the PLOS ONE 'phylogeny' query:

MS [returned 46 hits] `http://www.mendeley.com/research-papers/search/?query=phylogeny+AND+published_in%3APLOS%20ONE+AND+year+from%3A2006+year+to%3A2009`

MAS [returned 257 hits] `http://academic.research.microsoft.com/PublicationList?query=year%3E%3d2006%20year%3C%3d2009%20jour%3a%28plos%20one%29%20phylogeny&desType=4&desID=4130&start=1&end=100`

Scopus [returned 782 hits] ALL (phylogeny) AND SRCTITLE (plos one) AND PUBYEAR > 2005 AND PUBYEAR < 2010

WoK [returned 521 hits] All Databases, Advanced Search: TS=phylogeny AND SO=(PLOS ONE) AND PY=(2006-2009)

GS [returned 680 hits] `http://scholar.google.co.uk/scholar?as_q=phylogeny&as_epq=&as_oq=&as_eq=&as_occt=any&as_sauthors=&as_publication=PLOS+ONE&as_ylo=2006&as_yhi=2009&btnG=&hl=en&as_sdt=0%2C5`

PLOS ONE's own content discovery service returns 724 hits for the equivalent search: `http://www.plosone.org/search/advanced?searchName=&weekly=&monthly=&startPage=0&pageSize=15&filterKeyword=&resultView=&unformattedQuery=everything%3Aphylogeny&sort=Relevance&filterStartDate=2006-01-01&filterEndDate=2009-12-31&filterJournals=PLoSONE&filterArticleTypes=Research+Article`

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.773v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2015, publ: 5 Jan 2015

RESULTS

MAS does not appear to index *Zootaxa* articles at all, and MS appears to index a negligible few, with 0.2% recall performance across all eleven search patterns. The recall of terms in *PLOS ONE* at both MAS & MS is better than for *Zootaxa* but is still below 50% on average. The recall of WoK searches are similarly poor when searching *Zootaxa* or *PLOS ONE* averaging 17.2% and 21.9% respectively, presumably because WoK only indexes titles, abstracts and keywords. GS & Scopus, more sophisticated indexing services, are interesting to compare: Scopus has significantly better recall on *Zootaxa* articles, whilst GS has near-perfect recall on *PLOS ONE* articles. This suggests perhaps that Scopus is being given some kind of preferential access to *Zootaxa* content to which GS is not being granted (see Tables 2 & 3). The Scopus search for 'phylogeny' in the PLOS ONE set is remarkable: it somehow found 58 additional articles containing phylogeny (782 in total), that do not actually contain the word 'phylogeny' (my local searches and independent validation from PLOS ONE's own content discovery search API show there *really are* just 724 articles that contain the string 'phylogeny' in articles published between 2006 and 2009 inclusive). I exported all 782 biblographic records in bibtex format from that Scopus search for 'phylogeny' for further manual examination - the bibtex file is available in the supplementary data on figshare (Mounce, 2015). Manual examination shows that Scopus was double-counting one paper (Man et al., 2007); two separate search hits were returned for this paper in the search for 'phylogeny'. Presumably the other 57 'extra' hits were returned either through error, or because they matched some additional metadata that Scopus attaches to each bibliographic recored.

Table 2. Hits returned for all literature searches. Both local full-text and through online content discovery services against both journal article sets. n/a represents 'not applicable', GS & MAS do not appear to support wildcard searches.

	Zootaxa set						PLOS ONE set					
	grep	GS	Scopus	WoK	MS	MAS	grep	GS	Scopus	WoK	MS	MAS
winclada	151	51	105	0	0	0	2	2	0	0	0	0
hennig86	27	10	13	0	0	0	0	0	0	0	0	0
paup	688	332	444	6	0	0	131	130	33	0	0	50
nona	117	50	75	5	0	0	10	8	1	0	1	1
tnt	150	61	108	8	0	0	82	82	7	2	1	10
phylip	22	9	14	0	0	0	99	99	20	1	0	58
phylogeny	4592	2420	3903	1803	4	0	724	680	782	521	46	257
phylogen*	6849	n/a	5561	2268	6	n/a	1394	n/a	1093	619	135	n/a
phylog*	6889	n/a	5618	2300	6	n/a	1402	n/a	1111	623	141	n/a
Aedes	46	25	30	14	1	0	84	84	53	25	12	68
Anopheles	52	31	41	22	0	0	182	171	107	46	23	108

Fine-grain examination of Hennig86 searches

For one particular search I sought more fine-grain detail as to the identity of the articles found and *not found* by each discovery service. I manually examined all 27 articles in which my grep searches found 'hennig86' OR 'hennig 86' and scored what sections of the article they occurred in e.g. abstract, body-text, or references, as well as if GS or Scopus found that particular article when searching for 'Hennig86 OR Hennig 86' (data supplied at Mounce (2015)). 26 of the 27 mention Hennig86 in the body of the article but not the title, abstract or keywords. No document hits were found *solely* in the reference list. One article (Marinoni et al., 2003) clearly mentions Hennig86 in the abstract - yet only GS found this article. The set of ten articles that GS finds and twelve that Scopus finds for the same search (Hennig86) are non-overlapping, only four Hennig86-containing articles were found by both GS and Scopus. Scopus did not find the four most-recently published mentions of Hennig86 in the Zootaxa set (published in 2011, 2011, 2012 and 2013 respectively).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.773v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2015, publ: 5 Jan 2015

Table 3. Recall performance table. Recall is measured relative to the local full-text grep searches in Table 3 (service hits / grep hits) * 100, capped at the logical maximum of 100%

	Zootaxa set					PLOS ONE set				
	GS	Scopus	WoK	MS	MAS	GS	Scopus	WoK	MS	MAS
winclada	33.8	69.5	0	0	0	100	0	0	0	0
hennig86	37.0	48.1	0	0	0	n/a	n/a	n/a	n/a	n/a
paup	48.3	64.5	0.9	0	0	99.2	25.2	0	0	38.2
nona	42.7	64.1	4.3	0	0	80	10	0	10	10.0
tnt	40.7	72.0	5.3	0	0	100	8.5	2.4	1.2	12.2
phylip	40.9	63.6	0	0	0	100	20.2	1.0	0	58.6
phylogeny	52.7	85.0	39.3	0.1	0	93.9	100	72.0	6.4	35.5
phylogen*	n/a	81.2	33.1	0.1	n/a	n/a	78.4	44.4	9.7	n/a
phylog*	n/a	81.6	33.4	0.1	n/a	n/a	79.2	44.4	10.1	n/a
Aedes	54.3	65.2	30.4	2.2	0	100	63.1	29.8	14.3	81.0
Anopheles	59.6	78.8	42.3	0	0	94.0	58.8	25.3	12.6	59.3
Mean recall	**45.6**	**70.3**	**17.2**	**0.2**	**0**	**95.9**	**45.1**	**21.9**	**6.4**	**36.8**

DISCUSSION

The ability to discover previously published research is absolutely fundamental to the basic process of scientific research. If we can't discover what has been previously published, we risk overlooking valuable research and repeating experiments that have already been done. Dark research that cannot be found is a serious impediment to *systematic* literature reviews e.g. Cochrane reports, research trend analyses e.g. Von Wehrden et al. (2009) and knowledge synthesis. Discoverability of the full text content of research articles is thus crucial. This research shows that some discovery services can't even find words that occur in abstracts (e.g. Hennig86), let alone the full text content, for modern 'born-digital' research. It is worrying that content search of born-digital journals like *Zootaxa* is so poor and so variable between search providers (Table 3). If the search words aren't in the title it is very hard to accurately find *all* relevant content in *Zootaxa*. Yet the results of the *PLOS ONE* analysis offer hope. With an average document recall performance of over 95% on the eight applicable searches it demonstrates that third-party provision of near perfect recall is possible. I have not identified nor designed these experiments to find the mechanism *causing* the dark research effect. This is merely an observation study to demonstrate the effect. Separate follow-up work is needed to ascertain the causitive mechanism(s) preventing *Zootaxa* content from being more discoverable via services such as GS.

Recommendations

Recommendations for various stakeholders in research, given the results the and their implications:

- **Research funders:** Consider mechanisms with which to encourage researchers to publish their work in a fully-discoverable manner. An obvious way of achieving this would be to encourage open access publication.
- **Authors wanting to publish research:** Consider carefully where you choose to publish your work. Will the full content of your work be discoverable at the publication venue you choose? Consider the possible negative impact on citations & scientific progress if the full text of your work is not discoverable by services like Google Scholar and Scopus.
- **Researchers searching for relevant published content:** Can you find all relevant content by just using online content discovery services? This research and more (Brown et al., 2008) suggests not. If you desire rigorous *systematic* evaluation of what has been previously published in the last decade, you may need to download all relevant journals to perform full-text searches on them yourself.
- **Magnolia Press:** Consider contacting the major content discovery services to discuss with them how to improve the discoverability of work published in Magnolia Press journals.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.773v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2015, publ: 5 Jan 2015

- **Other publishers:** Check the discoverability of work published in your journals. Is content in the full text; beyond the title, abstract and keywords, discoverable?

- **Academic content discovery providers:** Make it clearer to users if you do full text searching, or just title-abstract-keyword searches. Make it clearer to publishers how their content is indexed. Consider contacting publishers to help them get their content full text indexed if it isn't already.

'Dark Research' (c.f. 'Dark Taxa' Page (2011)) - where relevant published content cannot be found online, even when specifically trying to find terms that do occur in the article - is a demonstrably real phenomenon. It can be quantified in terms of recall relative to the known text content of articles. The lower the recall, the more 'hidden-in-darkness' the research is. The near-perfect recall performance of Google Scholar on *PLOS ONE* content shows that full text discoverability with current technology is achievable - there is no valid excuse for not providing full text discoverability to modern, born-digital content. The *exact* causative mechanism impairing the discoverability of full-text content in *Zootaxa* is not identified by this research. However, it would seem reasonable to speculate that the cause could be the access-restriction mechanism that Magnolia Press use. If Google Scholar's crawler/indexer bots are not being allowed past the paywalls then they can only index the title, abstract, keywords and references, at best.

Future Research

I have examined just two journals here to provide a first-pass proof-of-concept that 21st-century published 'Dark Research' is a real phenomenon, even despite the impressive capability of modern web technology e.g. Google Scholar. It is obvious that more work urgently needs to be done to explore the discoverability of born-digital, 21st-century published content in a wider range of journals at a wider range of publishers to get a fuller picture on the extent of Dark Research. Is all open access journal content near 100% full text discoverable like *PLOS ONE*? Is all research published behind a paywall less discoverable than open access research when using Google Scholar, or is it just Magnolia Press journals? It is already known that open access typically confers more downloads, views and citations (Lawrence, 2001; Hajjem et al., 2005; Eysenbach, 2006; Gargouri et al., 2010; Davis, 2011), perhaps discoverability might be formally added to the list of advantages of open access?

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

The author declares there are no competing interests.

Author Contributions

Ross Mounce conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and tables, reviewed drafts of the paper.

Funding

The author declares there was no funding for this work.

ACKNOWLEDGMENTS

An earlier version of this manuscript formed a chapter of my thesis: 'Mounce, R. 2014. Comparative Cladistics: Fossils, Morphological Data Partitions and Lost Branches in the Fossil Tree of Life. PhD thesis, University of Bath'. It has been rewritten here and considerably shortened to fine-tune clarity and focus, but the data and results are largely the same. I would like to thank PLOS for providing open access to most PLOS ONE content under the Creative Commons Attribution Licence, which allowed me to do this research in a scalable manner without asking permission. I should also thank the UK government for their recent copyright reform. The specific copyright exception for 'copying of works for use by text and data analytics' enabled me to legally analyze the discoverability of Zootaxa content without fear of being sued for copyright infringement. Finally, I would also like to thank the Natural History Museum, London for providing me legal, paid-for, electronic access to Zootaxa content.

REFERENCES

Anon. (2014a). About microsoft academic search. http://academic.research.microsoft.com/About/help.htm#5.

Anon. (2014b). Compare mendeley. http://www.mendeley.com/compare-mendeley/.

Anon. (2014c). Scopus content fact sheet. http://www.elsevier.com/__data/assets/pdf_file/0006/155427/Scopus-Content.pdf.

Anon. (2014d). Web of knowledge: Who we are. http://wokinfo.com/about/whoweare/.

Anon. (2015). Our history in depth. http://www.google.com/about/company/history/#2004.

Beckmann, M. and von Wehrden, H. (2012). Where you search is what you get: literature mining – google scholar versus web of science using a data set from a literature search in vegetation science. *J Veg Sci*, 23(6):1197–1199.

Brown, L. E., Dubois, A., and Shepard, D. B. (2008). Inefficiency and bias of search engines in retrieving references containing scientific names of fossil amphibians. *Bulletin of Science, Technology & Society*, 28(4):279–288.

Davis, P. M. (2011). Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *The FASEB Journal*, 25(7):2129–2134.

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biol*, 4(5):e157.

Fingerman, S. (2004). Elsevier holds official launch events for scopus.

Ford, L. and O'Hara, L. H. (2008). It's All Academic: Google Scholar, Scirus, and Windows Live Academic Search. *Journal of Library Administration*, 46(3-4):43–52.

Garcia-Perez, M. (2012). Reviewer's report on: Is google scholar sensitive enough to be used alone for systematic reviews? *BMC Medical Informatics and Decision Making*.

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., and Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PloS one*, 5(10):e13636.

Gehanno, J.-F. F., Rollin, L., and Darmoni, S. (2013). Is the coverage of google scholar enough to be used alone for systematic reviews. *BMC medical informatics and decision making*, 13(1):7+.

Giustini, D. and Kamel Boulos, M. N. (2013). Google scholar is not enough to be used alone for systematic reviews. *Online Journal of Public Health Informatics*, 5(2).

Goloboff, P. (1999). Pee-wee and nona, versions 3.0, programs and documentation. *Published by the author*.

Goloboff, P. A., Farris, J. S., and Nixon, K. C. (2008). TNT, a free program for phylogenetic analysis. *Cladistics*, 24(5):774–786.

Hajjem, C., Harnad, S., and Gingras, Y. (2005). Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. *IEEE Data Engineering Bulletin*, 28(4):39–47.

Langille, M. G. I. and Eisen, J. A. (2010). Biotorrents: A file sharing service for scientific data. *PLoS ONE*, 5(4):e10071+.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature*, 411(6837):521.

Man, O., Willhite, D., Crasto, C., Shepherd, G., and Gilad, Y. (2007). A framework for exploring functional variability in olfactory receptor genes. *PLoS ONE*, 2(1). cited By 4.

Marinoni, L., Steyskal, G. C., and Knutson, L. V. (2003). Revision and cladistic analysis of the neotropical genus thecomyia perty (diptera: Sciomyzidae). revisión y análisis cladístico del género neotropical thecomyia perty (diptera: Sciomyzidae). *Zootaxa.*, (191):1–36.

Mounce, R. (2015). Dark research - supplementary files. ***figshare***. http://dx.doi.org/10.6084/m9.figshare.1284356.

Nixon, K. (2002). Winclada ver. 1.00. 08. *Published by the author, Ithaca, NY*.

Page, R. (2011). Dark taxa: Genbank in a post-taxonomic world. http://iphylo.blogspot.co.uk/2011/04/dark-taxa-genbank-in-post-taxonomic.html.

Pérez, F. and Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.773v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2015, publ: 5 Jan 2015

Swofford, D. L. (2002). Paup* phylogenetic analysis using parsimony (*and other methods).
Von Wehrden, H., Hanspach, J., Bruelheide, H., and Wesche, K. (2009). Pluralism and diversity: trends in the use and application of ordination methods 1990-2007. *Journal of Vegetation Science*, 20(4):695–705.
Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.773v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2015, publ: 5 Jan 2015