Cite this article as: BMJ, doi:10.1136/bmj.38422.611736.E0 (published 12 April 2005)

Information in practice

BMJ: first published as 10.1136/bmj.38422.611736.E0 on 12 April 2005. Downloaded from http://www.bmj.com/ on 20 April 2024 by guest. Protected by copyright.

Open access and openly accessible: a study of scientific publications shared via the internet

Jonathan D Wren

Abstract

Objectives To determine how often reprints of scientific publications are shared online, whether journal readership level is a predictor, how the amount of file sharing changes with the age of the article, and to what degree open access publications are shared on non-journal websites.

Design The internet was searched using an application programming interface to Google, a popular and freely available search engine.

Main outcome measures The proportion of reprints of journal articles published between 1994 and 2004 from within 13 subscription based and four open access journals that could be located online at non-journal websites.

Results The probability that an article could be found online at a non-journal website correlated with the journal impact factor and the time since initial publication. Papers from higher impact journals and more recent articles were more likely to be located. On average, for the high impact journal articles published in 2003, over a third could be located at non-journal websites. Similar trends were observed for the delayed or full open access publications.

Conclusions Decentralised sharing of scientific reprints through the internet creates a degree of de facto open access that, though highly incomplete in its coverage, is none the less biased towards publications of higher popular demand.

Introduction

Widespread access to the internet has changed the paradigm of scientific publishing and research. For publishers, the internet is not just an alternative to the traditional print medium but a superior one that enables a broader range of content to be provided, such as animations, searchable databases, and large datasets. It also lowers the barriers to publishing, providing immediate worldwide electronic dissemination at a lower cost.[1] Consequently, an increasing number of journals are being published in many different specialties, leading to greater potential subscription costs in a time of shrinking library budgets. For researchers, however, increased costs may impede their ability to access scientific research. The lower cost of electronic publication and dissemination in combination with increases in total or potential subscription costs has given rise to the recent debate about "open access"—moving from a publishing model where readers pay for access to one where authors pay for publication.[2–5] Recently, the National Institutes of Health announced its intention to require open access publication of all its funded research,[6 7] which could profoundly affect subscription based journal publication.

The crux of the open access debate primarily argues that costs will prevent some people from accessing published scientific research. As the Wellcome Trust accurately notes, "the benefits of research are derived principally from access to research results."[8] Certainly, this access is important to further both future research efforts and public education, but there is a factor in this debate that has been neglected because the extent to which it occurs is unknown. While some scientific publications may not have been published in an open access journal, they are, none the less, openly accessible to the public on non-journal websites[9] and can be located by using freely available internet search engines. This phenomenon is largely unstudied and pertinent to the open access debate, as arguments about barriers to access should take into account free alternative means of access—to the extent that they exist. Without a better understanding of how commonly scientific publications are shared online, what types of publications are shared, and whether or not this is changing as we progress further into the internet age, it is difficult, if not impossible, to factor this in at all.

I examined the extent of scientific file sharing, including how commonly scientific publications are shared online, whether journal readership level is a predictor, how the amount of file sharing changes with the age of the article, and to what degree open access publications are shared on non-journal websites.

Methods

I wrote a program in Visual Basic .NET to read Medline records and interface with the Google application programming interface (API), which is also available in Visual Basic .NET, enabling queries to be sent to Google in an automated manner (that is, without a user having to type in each one manually).

Selection of journals

I chose 13 subscription based journals for analysis on the basis of their 2002 journal impact factor, which correlates with the level of readership (box). All journals had articles indexed in Medline dating back at least to 1994 and were subscription based. Four journals had high impact factors, five had relatively low impact factors, and four had impact factors around 10 (medium range). Impact factors were obtained from ISI's Journal Citation Reports.[10]

The query target

As my query target I chose PDF files rather than HTML files for several reasons. Firstly, because all necessary information (such as figures and tables) is in one file, it is easier to post a PDF than recreate a HTML file with all associated images. Secondly, journal reprints are typically distributed as PDF files and readers

Copyright 2005 BMJ Publishing Group Ltd

- **Google** (www.google.com) – a popular, freely available internet search engine that can be accessed through any web browser
- **Google API** (application programming interface) – a means of interacting with Google from within a computer program, rather than through a web browser, enabling queries and results to be handled programmatically
- **HTML** (hypertext markup language) – HTML is a means of formatting the display of a document for a web browser through the use of delimiting tags that are not themselves displayed
- **PDF** (portable document format) – PDF enables documents to be reproduced almost precisely as they originally appeared in print. The format was created by Adobe (www.adobe.com), which provides freely available software for viewing PDFs on virtually any operating system

prefer them because they can be printed out without loss of formatting (in HTML, tables and figures are often separate from the main document). Other formats that permit this (such as Microsoft Word or Postscript) were relatively uncommon among articles indexed in Medline (data not shown). Thirdly, PDFs enable specific page numbers to be used as part of the query.

Constructing Google queries to locate Medline articles online

Constructing queries with the digital object identifier (DOI) corresponding to each published article would be an ideal means of retrieving articles as DOIs are unique. Unfortunately, though DOIs are recorded by PubMed, they are not provided in the distributed version used to obtain article information. Even if they were, there is still variance among and within journals regarding the inclusion of DOIs within reprints (PDFs), and most journals adopted their DOI inclusion policy at varying points within the period being studied (1994-2004).

I therefore had to design highly restrictive queries to send to Google, the goal being to return only unique matches in response to the query. To narrow the search results to include only the PDF of the full text Medline article being queried rather than other articles that cite it, I had to include information other than words or phrases normally found within citations, such as title and author names. The first query term was the rarest of the authors' last names, which I determined by calculating frequencies of all last names of authors within Medline. The second query term was the rarest word found within the authors' affiliation field to help screen out Google matches to citations (which include authors' names but not affiliation information). Thirdly, I used the title in quotes so that only exact matches would be returned. I excluded Greek letters commonly used in gene names from queries because of the frequent disparity between Medline entries and actual article titles. For example, a Medline title might contain the string "TGF-beta 1 protein" but the corresponding journal article might read "TGF-β 1 protein," which would not be located by the Google queries used. Thus, the query string used for this paper would be split around this character ("TGF-" "1 protein").

One of the most important narrowing criteria for keyword queries was the use of implicit page numbers, which are normally not present within HTML files but are present within reprints. For example, when the information "p. 341-5" is part of a reference, this explicitly states that page 341 is part of the citation and implies that so are pages 342, 343, 344, and 345. These implicit page numbers screen out false positive results that arise from self citation, which is relatively common. When an author cites his or her own work, affiliations associated with the cited work will probably be present in the citing paper and there is a higher probability that, due to continued collaborations, the same rare author name might also be present in the citing paper. Plus, implicit page numbers reduce the probability that the article being returned is from a self archived document before peer review, a practice that has grown in some specialties[11] with acceptance varying among journals. This search process will miss versions of articles that differ from journal reprints. *BMJ*, for example, has an ELPS (electronic long, paper short) policy that creates two PDF versions—one as it appears in the journal and the other with additional or supplementary information that, out of necessity, has a different page numbering scheme. In these cases, the queries will only find the (shorter) journal reprint version.

Queries submitted to Google were thus of the form: "<rarest author last name> <rarest affiliation word> <first implicit page number (if one exists)> <second implicit page number (if one exists)> <exact title (in quotes) of article being queried>." Once query results had been compiled, I sorted them by URL and excluded websites associated with the journals from analysis. I restricted analysis to papers classified by PubMed as "journal articles" because non-research articles such as commentaries and editorials tend to have fewer narrowing criteria (that is, they have fewer author names, affiliation information is not always present, and they are often one page long and thus have no implicit page numbers). The end result was a list of journal articles indexed by Google and freely available online at non-journal websites.

Benchmarking query recall

I chose the *Journal of Biological Chemistry* (*J Biol Chem*) to benchmark how well the constructed Google queries located Medline articles online because *J Biol Chem* makes its articles open access at the end of the calendar year.[12] Although search engines also index subscription only PDFs, which could also be used for testing, journal policy with regards to permitting search engines ("webbots") access to index their website content is often unstated and varies with respect to which search engines are

Journals analysed (impact factor)

Subscription based journals
New England Journal of Medicine (*N Engl J Med*) (32)
Nature (30)
Science (29)
Cell (27)
Current Opinion in Neurobiology (*Curr Opin Neurobiol*) (11)
American Journal of Human Genetics (*Am J Hum Genet*) (11)
EMBO Journal (*EMBO J*) (11)
Circulation (10)
Glia (5)
Prostate (3)
Nutrition Reviews (*Nutr Rev*) (2)
Chemotherapy (1)
Journal of Spinal Disorders (*J Spin Disord*) (0.7)

Open access journals
Proceedings of the National Academy of Sciences (*Proc Natl Acad Sci U S A*) (11)
Molecular and Cell Biology (*Mol Cell Biol*) (9)
British Medical Journal (*BMJ*) (8)
Journal of Biological Chemistry (*J Biol Chem*) (7)

BMJ: first published as 10.1136/bmj.38422.611736.E0 on 12 April 2005. Downloaded from http://www.bmj.com/ on 20 April 2024 by guest. Protected by copyright.

Benchmarking query efficiency with open access articles published in *J Biol Chem**

Year	Total articles published	No (%) indexed	No found offsite (% of total/% of indexed)	No found onsite (% of total/% of indexed)
1996	4947	1819 (37)	181 (4/10)	1567 (32/86)
1997	4819	1760 (37)	174 (4/10)	1522 (32/86)
1998	4953	1777 (36)	217 (4/12)	1514 (31/85)
1999	5292	1829 (35)	297 (6/16)	1602 (30/88)
2000	5633	1987 (35)	394 (7/19)	1747 (31/88)
2001	6519	2219 (34)	435 (7/19)	2039 (31/92)
2002	6531	2294 (35)	451 (7/19)	2142 (33/93)
2003	6588	2757 (42)	417 (6/15)	2490 (38/90)
Total	45 282	16 442 (36.3)	2566 (6/16)	14 623 (32/89)

*Number of PDF based *J Biol Chem* articles indexed by Google varied by year, averaging about 36% of article URLs listed on *J Biol Chem* website. Averaging indexed PDFs by year, query routine as described found average of 89% (SD 3%).

allowed (see standards for webbot exclusion[13]). Therefore, I thought it preferable to benchmark using journals that have declared certain content freely available to the public. Additionally, *J Biol Chem* publishes more journal articles per year than most other journals (roughly twice that in the next highest journal in the 17 examined), offering a greater sample size.

Online documents can be found by querying a search engine only if the search engine itself has located and indexed them. Thus, if queries locate only half of the journal articles known to exist within Medline on a website that should contain all of them, one might initially conclude that recall is low. However, search engines are not comprehensive in their indexing of web accessible documents.[14] Thus, before I could estimate query recall using *J Biol Chem*, I had to measure the number of *J Biol Chem* journal article PDFs indexed by Google. I downloaded the URLs corresponding to the location of full text PDF articles published between 1996 and 2003 from the *J Biol Chem* website and used them as the query string submitted to the Google API. For example, on the *J Biol Chem* website, the URL www.jbc.org/cgi/reprint/275/2/1007.pdf corresponds to the PDF of a specific Medline article published in *J Biol Chem*.[15] When this URL is used in a Google query, the link will be displayed if the URL is indexed. I queried 45 282 PDF URLs from *J Biol Chem* on three separate occasions in 2004: 1 July, 2 August, and 13 September. The total number of article PDFs indexed by Google varied from 19 194 (42.4% of the total) on the July run to 25 084 (55.4%) on the August run to 16 442 (36.3%) in September. This suggested that overall statistics on query performance need to be gathered as close as possible to the time the index benchmarking took place.

To see if it was reasonable to use the rate of *J Biol Chem* article indexing by Google as a measure of overall recall, I ran a similar batch of queries on 9 August using 22 819 journal article PDF URLs corresponding to articles published during the same period (1996-2003) extracted directly from the Proceedings of the National Academy of Sciences (*Proc Natl Acad Sci U S A*) website, finding a total of 4022 (18%). Thus, while query performance versus indexed documents can be estimated, it is difficult to extrapolate these numbers to estimate the true recall of the queries (that is, what percentage of all web accessible journal articles are found). It should also be noted that the fraction of PDFs locatable on one website does not necessarily reflect the fraction that could be locatable on the internet in general. For example, if these 22 819 *Proc Natl Acad Sci U S A* articles were scattered across many websites instead of one, we cannot estimate a priori how many would be found by Google (without knowing the specifics of Google's indexing software). To obtain the answer, they would have to be queried individually as was done here.

Results

Evaluating query performance (precision and recall)

I tested the ability of the constructed queries to find known journal article PDFs on the *J Biol Chem* website on 11-12 September, and the table summarises the results. With the September run as a benchmark, the approximate recall of the constructed queries on indexed documents was 89% (SD 3%). Thus, the queries should locate about nine out of 10 Medline documents that are both located on the internet and indexed by Google.

I estimated precision by manually examining three sets of 50 PDFs identified as potential reprints of journal articles by the Google queries. Each set of PDFs was chosen randomly from within the entire list of queried article reprints and only PDFs found at non-journal websites were examined. A query was considered successful only if the first PDF it returned corresponded to the journal article being queried. I did not examine any other entries after the first one. Six documents returned either a blank page or a "404 not found" error, and these were excluded from further analysis (URLs corresponding to the location of these PDFs were examined on 16 September 2004). A total of 38/48 (79%), 37/48 (77%), and 34/48 (71%) top query results corresponded to the article being sought. The mean precision was 76% (SD 4%).

Journal queries

I queried 48 516 journal articles indexed by Medline within the 13 subscription based journals with a publication date between January 1994 and July 2004. Figure 1 shows the results. Several trends are apparent. Firstly, journals with higher impact have a larger fraction of papers that can be found online at non-journal sites. A two tailed t test comparing the areas under the curve for high, medium, and low impact journals yielded: high v medium ($P<0.02$) and medium v low ($P<0.07$) and high v low ($P<0.0002$). Secondly, for these journals, the probability a paper could be found correlates with how recently it was published. Thirdly, many of these journals showed a recent drop in online availability. This is probably artificial, however, as journal citations often appear in Medline after a paper is accepted for publication but before it appears in print (or PDF), sometimes several months before. It is also possible that online posting tends to lag publication date.

For all the PDFs found online at non-journal websites, the total number of unique root domains (for example, www.ou.edu is the root domain for the website URL www.ou.edu/web/academics) was 5086, and the most PDFs found at one root domain was 138. This suggests that file sharing is highly distributed and that no central repository is contributing significantly to this phenomenon. Figure 2 summarises the distribution of file

BMJ: first published as 10.1136/bmj.38422.611736.E0 on 12 April 2005. Downloaded from http://www.bmj.com/ on 20 April 2024 by guest. Protected by copyright.

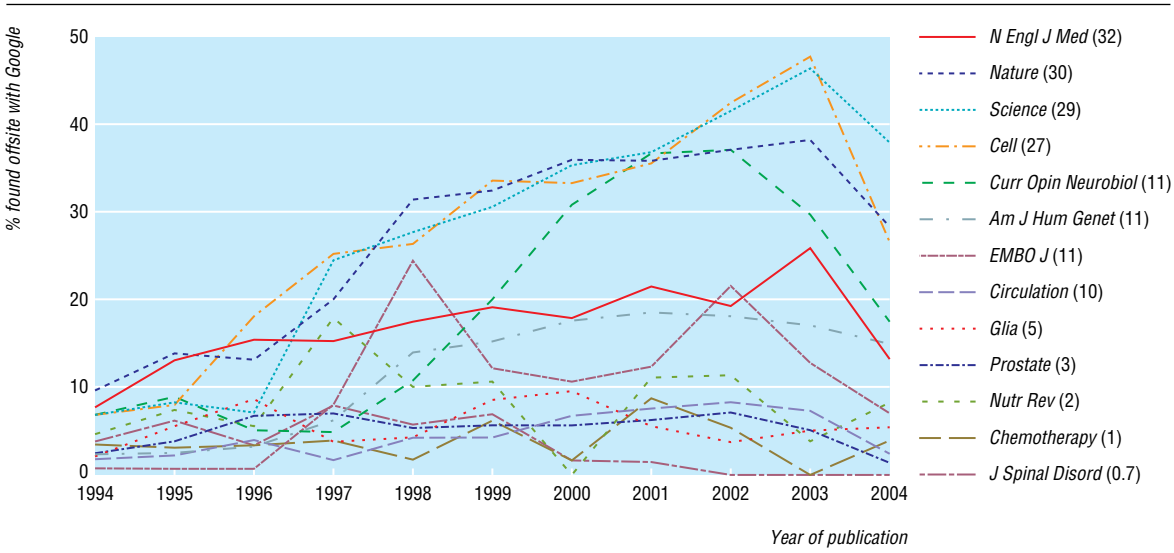

Fig 1 Subscription based journal articles locatable with Google at non-journal websites, with approximate impact factors for 2002 in parentheses. No articles were found in Medline for *J Spinal Disord* from 2002-3

sharing by top level domain (for example, ".edu" is the top level domain for the URL above).

I also examined file sharing for four open access or delayed open access journals. The free availability of these articles could obviate the need to share them on non-journal websites—website authors could just as easily provide a link to the journal's PDF rather than download and provide their own. On the other hand, the free availability of articles might encourage them to be copied and shared.[16] Over the same period (1994-2004) 102 404 articles were queried for *Proc Natl Acad Sci U S A*, *J Biol Chem*, the *BMJ*, and *Mol Cell Biol* (fig 3). Many open access articles were also found at non-journal websites, with the same time dependent trend. I used a *t* test used to compare the area under the curve of these four open access journals with their subscription based counterparts and found that their online availability trends were more similar to the mid-range impact factor group ($P<0.46$) than the high ($P<0.003$) or low ($P<0.24$). As the impact factors of these open access journals are in this mid-range, this suggests that the probability that a journal article can be found on a non-journal website is less a function of copyright or ownership than it is of impact factor or journal readership levels.

Discussion

The number of full text scientific research articles openly accessible online at non-journal websites correlates most strongly with the publishing journal's impact factor and inversely with time since original publication. Cost barriers to access alone, however, do not explain the prevalence of file sharing because a relatively large fraction of open access and delayed open access publications were also found on non-journal websites. Perhaps some of this could be attributed to a "supply and demand" model—a high demand from readers to view current important papers is met by some party supplying the paper. Also, because the online visibility and accessibility of an article or articles correlates with readership and citation level,[17 18] some authors may simply be trying to increase awareness of their work. Or, perhaps, somewhat cynically, file sharing may arise from a "trophy effect"—the desire for researchers to display their accomplishments—which would explain why high impact publications are more common online. Examination of some of the URL names in the random samples taken, however, suggests that several of them were probably intended to be there only temporarily (for example, URLs containing the word "journal_club") for the purpose of sharing important information. This would also explain the observed trend as journal clubs tend to focus more on recent and high impact developments. More studies will need to be done, however, before motivations for sharing scientific publications are better understood. The number of different times that an article appears online and the number of websites

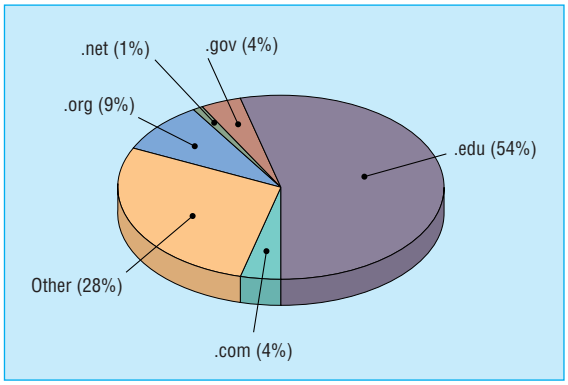

Fig 2 Distribution of journal article file sharing by top level domain

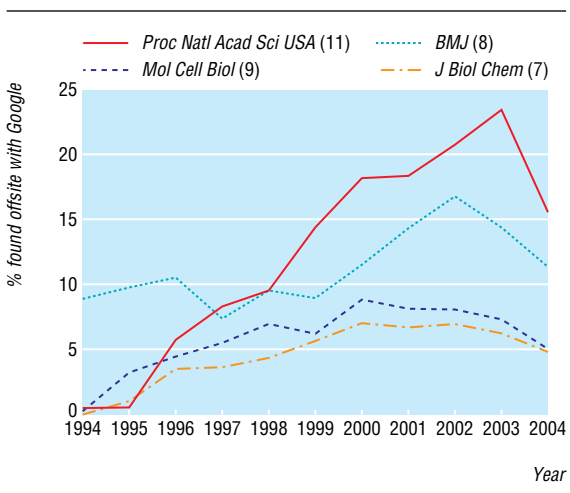

Fig 3 Open access journal articles locatable with Google at non-journal websites

BMJ: first published as 10.1136/bmj.38422.611736.E0 on 12 April 2005. Downloaded from http://www.bmj.com/ on 20 April 2024 by guest. Protected by copyright.

it appears on could reasonably be considered an alternative means of measuring the scientific impact of individual articles, as could the number of citations identifiable within online publications.

One weakness of my study is that it is difficult to assess the true fraction of journal articles accessible at non-journal websites because of incomplete search engine indexing. Consequently, the reported numbers almost certainly underestimate the real numbers. This incomplete indexing is not specific to Google. I also checked Yahoo and MetaCrawler by submitting a sample of 30 identical queries to each search engine requesting an exact match on randomly chosen unique strings found within 30 PDFs. Yahoo and Metacrawler displayed similar performance, although different engines failed to index different PDFs (data not shown). This differential and incomplete coverage of search engine indexes was previously noted by Lawrence,[14] although his 1998 study did not include Google. The relatively low proportion of indexed articles may be due partly to difficulties searching PDF content. Early in the study, before I chose PDFs as the query targets because of their implicit page numbers, I found that HTML based articles were indexed at higher rates (data not shown). The appearance of new search engines specifically for academics, such as Google Scholar (http://scholar.google.com), should help researchers to locate these full text articles with greater precision, although incomplete web page indexing will probably remain an issue.

Finally, a straightforward interpretation of figure 1 suggests that publications are becoming increasingly available online as time goes by. It could be equally hypothesised, however, that most of the observed trend is due to a relatively constant rate of article posting in combination with a time dependent decay in URL availability, which has been well established not only as a general phenomenon but also in scientific publishing.[19–21]

The National Library of Medicine provided electronic Medline records in XML format. I thank the API development team at Google for permitting use of their web search engine interface as well as Robert Dellavalle, Lisa Schilling, Peter Suber, and Tim Cole for helpful manuscript reviews.

Contributors: JDW is the sole author.

Funding: This work was funded in part by a grant from NSF-EPSCoR (EPS-0132534).

Competing interests: None declared.

Ethical approval: Not required.

1 Mayor S. Open access could reduce cost of scientific publishing. *BMJ* 2004;328:1094.
2 Shattil SJ. Open access, yes! Open excess, no! *Blood* 2004;103:3257.
3 Plutchak TS. Embracing open access. *J Med Libr Assoc* 2004;92:1-3.
4 Graczynski MR, Moses L. Open access publishing—panacea or Trojan horse? *Med Sci Monit* 2004;10:ED1-3.

What is already known on this topic

The internet is unregulated and allows people to share files of any type online, which sometimes includes copyrighted works

Articles from subscription only journals may appear on non-journal websites, sometimes with permission and sometimes without

What this study adds

This study examined the posting of journal reprints on non-journal websites and compared posting trends between open access and subscription based journal articles

The higher the impact of the publishing journal and the more recent the article, the more likely it is that the article can be found online at a non-journal website

5 Kaiser J. Scientific publishing. Seeking advice on 'open access,' NIH gets an earful. *Science* 2004;305:764.
6 Roehr B. NIH moves towards open access. *BMJ* 2004;329:590.
7 Enhanced public access to NIH research information. http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-04-064.html (accessed 4 April 2005).
8 Wellcome Trust. *Costs and business models in scientific research publishing.* www.wellcome.ac.uk/doc_wtd003185.html 2004 (accessed 4 April 2005).
9 Dufva M. Open access will deter illegal file-sharing. *Nature* 2003;426:15.
10 ISI journal citation reports. http://www.isinet.com (accessed 4 April 2005).
11 Harnad S. The self-archiving initiative. *Nature* 2001;410:1024-5.
12 ASBMB: the open access publisher. www.jbc.org/misc/JBC_Open_Access.shtml (accessed 4 April 2005).
13 A standard for robot exclusion. www.robotstxt.org/wc/norobots.html (accessed 4 April 2005).
14 Lawrence S, Giles CL. Searching the world wide web. *Science* 1998;280:98-100.
15 Goto JJ, Zhu H, Sanchez RJ, Nersissian A, Gralla EB, Valentine JS, et al. Loss of in vitro metal ion binding specificity in mutant copper-zinc superoxide dismutases associated with familial amyotrophic lateral sclerosis. *J Biol Chem* 2000;275:1007-14.
16 Suber P. *SPARC Open access newsletter.* 2004. www.earlham.edu/~peters/fos/newsletter/01-02-04.htm≠manycopy (accessed 4 April 2005).
17 Lawrence S. Free online availability substantially increases a paper's impact. *Nature* 2001;411:521.
18 Perneger TV. Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ. *BMJ* 2004;329:546-7.
19 Spinellis D. The decay and failures of web references. *Commun ACM* 2003;46:71-7.
20 Dellavalle RP, Hester EJ, Heilig LF, Drake AL, Kuntzman JW, Graber M, et al. Information science. Going, going, gone: lost internet references. *Science* 2003;302:787-8.
21 Wren JD. 404 not found: the stability and persistence of URLs published in Medline. *Bioinformatics* 2004;20:668-72.

(Accepted 11 March 2005)

doi 10.1136/bmj.38422.611736.E0

Advanced Center for Genome Technology, Department of Botany and Microbiology, University of Oklahoma, 101 David L. Boren Blvd, Rm. 2025, Norman, OK 73019, USA
Jonathan D Wren *research scientist*
Correspondence to: J D Wren Jonathan.Wren@OU.edu

BMJ: first published as 10.1136/bmj.38422.611736.E0 on 12 April 2005. Downloaded from http://www.bmj.com/ on 20 April 2024 by guest. Protected by copyright.