

Information in practice

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

Jonathan D Wren

Editorial by Suber

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 2005;330:1128-31

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, although highly incomplete in its coverage, is none the less biased towards publications of higher popular demand.

Introduction

The lower cost of electronic publication and dissemination provided by the internet in combination with potential increases in 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.¹⁻⁴ Recently, the National Institutes of Health announced its intention to require open access publication of all its funded research.^{5,6}

While some scientific publications may not have been published in an open access journal, they may, none the less, be openly accessible to the public on non-journal websites.⁷ A better understanding is needed of how commonly scientific publications are shared online, what types of publications are shared, and whether or not this is changing.

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

A program was written in Visual Basic .NET to read Medline records and access the Google application programming interface (API), also available in Visual Basic .NET, enabling queries to be sent to Google in an automated manner (see bmj.com).

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).⁸ All journals had articles indexed in Medline dating back at least to 1994 and were subscription based.

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 prefer them because they can be printed out without loss of formatting. 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. However, though DOIs are recorded by PubMed, they are not provided in the distributed version used to obtain article information, and there is still variance among and within journals regarding the inclusion of DOIs within reprints (PDFs).

I therefore had to design highly restrictive queries to send to Google. The first query term was the rarest of the authors' last names. The second query term was the rarest word found within the authors' affiliation



This is the abridged version of an article that was posted on bmj.com on 12 April 2005: <http://bmj.com/cgi/doi/10.1136/bmj.38422.611736.E0>

field. Thirdly, I used the title in quotes so that only exact matches would be returned.

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.

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>.” The end result was a list of journal articles indexed by Google and freely available online at non-journal websites (see bmj.com for details).

Benchmarking query recall

I chose the *Journal of Biological Chemistry* (*J Biol Chem*) to assess 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. 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.

Search engines are not comprehensive in their indexing of web accessible documents.⁹ 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. 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 is found).

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 (see bmj.com). 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. Six documents returned either a blank page or a “404 not found” error. 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 (fig 1). 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 ($P < 0.02$), 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). 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,

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)

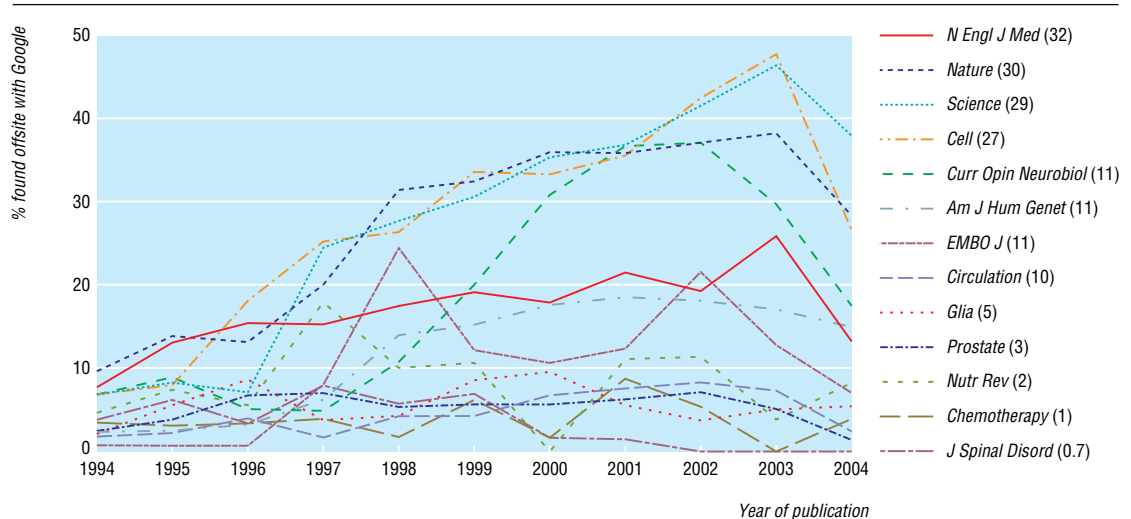


Fig 1 Subscription based journal articles locatable with Google at non-journal websites, with approximate impact factors for 2002 in parentheses. No articles were found in Medline for *J Spinal Disord* from 2002-3

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 (see bmj.com).

I also examined file sharing for four open access or delayed open access journals (fig 2). The free availability of these articles could obviate the need to share them on non-journal websites. On the other hand, the free availability of articles might encourage them to be copied and shared.¹⁰ I used a *t* test 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.

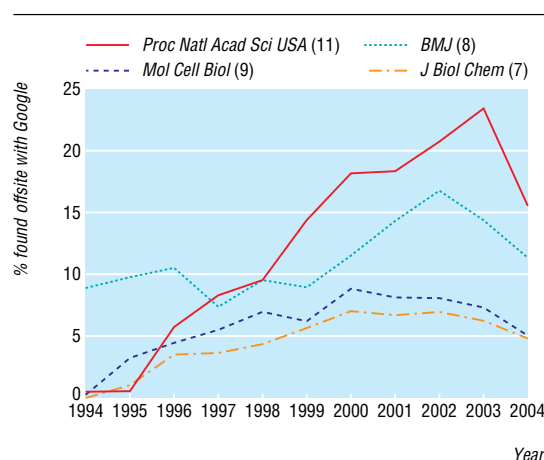


Fig 2 Open access journal articles locatable with Google at non-journal websites

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,^{11 12} 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 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.

One weakness of this study is that it is difficult to assess the true fraction of journal articles accessible at non-journal websites because of incomplete search engine indexing⁹; the reported numbers almost certainly underestimate the real numbers. This incomplete indexing is not specific to Google. I also found a similar performance with Yahoo and MetaCrawler. The relatively low proportion of indexed articles may be due partly to difficulties searching PDF content. New search engines specifically for academics, such as Google Scholar, should help researchers to locate these full text articles with greater precision, although incomplete web page indexing will probably remain an issue.

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

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.¹³⁻¹⁵

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: JW 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 Shatill SJ. Open access, yes! Open excess, no! *Blood* 2004;103:3257.
- 2 Plutchak TS. Embracing open access. *J Med Libr Assoc* 2004;92:1-3.
- 3 Graczynski MR, Moses L. Open access publishing—panacea or Trojan horse? *Med Sci Monit* 2004;10:ED1-3.
- 4 Kaiser J. Scientific publishing. Seeking advice on “open access,” NIH gets an earful. *Science* 2004;305:764.
- 5 Roehr B. NIH moves towards open access. *BMJ* 2004;329:590.
- 6 Enhanced public access to NIH research information. <http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-04-064.html> (accessed 4 April 2005).
- 7 Dufva M. Open access will deter illegal file-sharing. *Nature* 2003;426:15.
- 8 ISI journal citation reports. www.isinet.com (accessed 4 April 2005).
- 9 Lawrence S, Giles CL. Searching the world wide web. *Science* 1998;280:98-100.
- 10 Suber P. *SPARC Open access newsletter*. 2004. www.earlham.edu/~peters/fos/newsletter/01-02-04.htm#manycopy (accessed 4 April 2005).
- 11 Lawrence S. Free online availability substantially increases a paper's impact. *Nature* 2001;411:521.
- 12 Perneger TV. Relation between online “hit counts” and subsequent citations: prospective study of research papers in the BMJ. *BMJ* 2004;329:546-7.
- 13 Spinellis D. The decay and failures of web references. *Commun ACM* 2003;46:71-7.
- 14 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.
- 15 Wren JD. 404 not found: the stability and persistence of URLs published in Medline. *Bioinformatics* 2004;20:668-72.

(Accepted 11 March 2005)

*Tips on ...***Breaking bad news**

Nothing tests our communication skills so much as breaking bad news. Such conversations can be extremely emotional for both doctor and patient. The right words said in the right way make a huge difference. Here are some tips:

- Always read the patient's clinical notes, the test results, in detail. Make a mental note of the patient's resuscitation status and past communications
- Speak to the nurse in charge of the patient and ask him or her to be present during the conversation
- Ensure privacy; try handing your bleep over to someone else
- Arranging the conversation in advance generally gives a better outcome. Ensure that appropriate family members or carers are present
- Introduce yourself. Asking questions—such as “What do you understand about your problems so far?”—will give you clues to the patient's ideas, concerns, and expectations
- Avoid jargon. Give information slowly and clearly, making sure that the patient has time to understand
- The crucial point in the conversation is the “bad news” itself. The way you phrase this depends on how the patient has responded so far in the conversation. Most will have an idea as to what's coming next. Explain the situation in a simple, unambiguous way and let the information sink in

- Discuss further options (therapeutic or palliative) and make a plan for the future. Remember to give hope. Information leaflets, Macmillan services, support groups, etc, play a vital role
- In concluding the conversation, ensure that everyone has understood the diagnosis, and the plan. Your job is not complete until you document everything in the patient's notes and fill out the necessary referrals
- Watching and learning from seniors handling these situations helps immensely. Ask for feedback from nurses.

Chinmay Patvardhan *senior house officer in medicine, University Hospital of North Tees, Stockton*
(chcinmayp77@hotmail.com)

We welcome articles up to 600 words on topics such as *A memorable patient, A paper that changed my practice, My most unfortunate mistake*, or any other piece conveying instruction, pathos, or humour. Please submit the article on <http://submit.bmj.com> Permission is needed from the patient or a relative if an identifiable patient is referred to. We also welcome contributions for “Endpieces,” consisting of quotations of up to 80 words (but most are considerably shorter) from any source, ancient or modern, which have appealed to the reader.