Data Sharing Effect on Article Citation Rate in Paleoceanography

Jonathan R. L. Sears

American Geophysical Union, Washington, DC, USA

IN53B-1628

ABSTRACT

The validation of scientific results requires reproducible methods and data. Often, however, data sets supporting research articles are not openly accessible and interlinked. This analysis tests whether open sharing and linking of supporting data through the PANGAEA® data library measurably increases the citation rate of articles published between 1993 and 2010 in the journal ***Paleoceanography*** as reported in the Thomson Reuters Web of Science database. The 12.85% (171) of articles with publicly available supporting data sets received 19.94% (8,056) of the aggregate citations (40,409). Publicly available data were thus significantly (p=0.007, 95% confidence interval) associated with about 35% more citations per article than the average of all articles sampled over the 18-year study period (1,331), and the increase is fairly consistent over time (14 of 18 years). This relationship between openly available, curated data and increased citation rate may incentivize researchers to share their data.

BACKGROUND AND OBJECTIVES

Areas such as the physical sciences increasingly encourage open sharing of supporting data, however a literature search did not show such a comparison in the geosciences to date. The present bibliometric analysis was therefore conducted to test whether open sharing of supporting data measurably increases article citation rate in the area of paleoceanography, as a "data-driven science", based on a sample of articles published between 1993 and 2010 in the journal *Paleoceanography* (published by AGU, ISSN 0883-8305). The journal is considered a high-impact journal in its field, ranking highest for impact factor among the Paleontology category since 1995, and third in the Oceanography category (Thomson Reuters, 2010).

PANGAEA® (Publishing Network for Geoscientific and Environmental Data) data library (http://www.pangaea.de) includes over a half a million data sets. It is open to any project or individual scientist to archive and publish data, with a strong following within the paleoceanography community (Diepenbroek, 2011). AGU is one of several publishers cooperating with PANGAEA to establish web service-enabled reciprocal linking between articles and their supporting data sets.

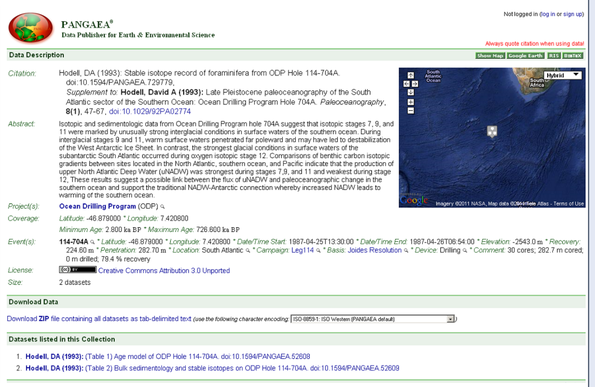

Figure 1: PANGAEA metadata record showing (a) Formal Citation to Data set including DOI ; (b) Parent article citation and link via DOI; and (c) Geospatial coordinates; and (d) Rich added metadata – parameters, date/time, sensors etc.

PROCEDURE AND ANALYSIS

Noting that the relatively long cited half-life* of *Paleoceanography* articles overall is 8.6 years based on Journal Citation Reports (JCR)®, 2009 data (Thomson Reuters, 2010), it was decided to examine articles published over a span of 18 years, from 1993 to 2010. [* The cited half-life is defined as "the number of publication years from the current year which account for 50% of the current citations received", or "the median age of the articles that were cited in the JCR year".]

A total of 171 articles published between 1993 and 2010 in the journal *Paleoceanography* were identified as having supporting data archived in the PANGAEA data library (Figure 1). Those articles were then searched in Thomson Reuters Web of Science (WoS) bibliographic database, and the number of citations to each article, date of publication, and corresponding author's country of origin were recorded (Figure 2). Searches were conducted over a 2-day period and double-checked to ensure that the counts represented an accurate snapshot of citations at a fixed point in time, i.e. that no additional citations were added during the study. Citation counts to the article set were also made via CrossRef (www.crossref.org) and Scopus (www.scopus.com) for comparison, and as a check for reasonableness of each result. The control group was the entire set of articles published in *Paleoceanography* between 1993 and 2010 (1,331). Thus the comparison is with the total population 1993-2010; in other words the comparison is not "with data vs. without data" but rather "with data vs. with or without data".

Results were tabulated in an Excel spreadsheet (Supplement 1) and output as a plot of citations/year (Figure 3).

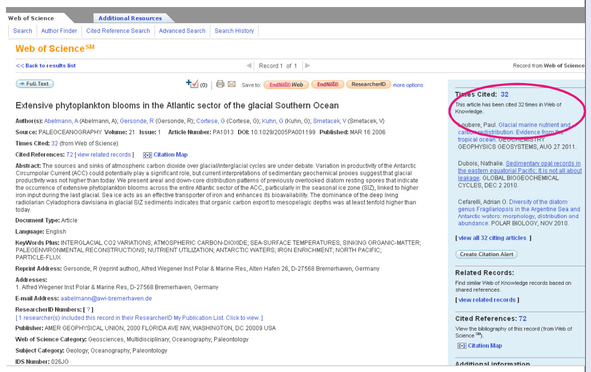

Figure 2. Web of Science record from Paleoceanography showing number of citing articles (circled in red)

RESULTS

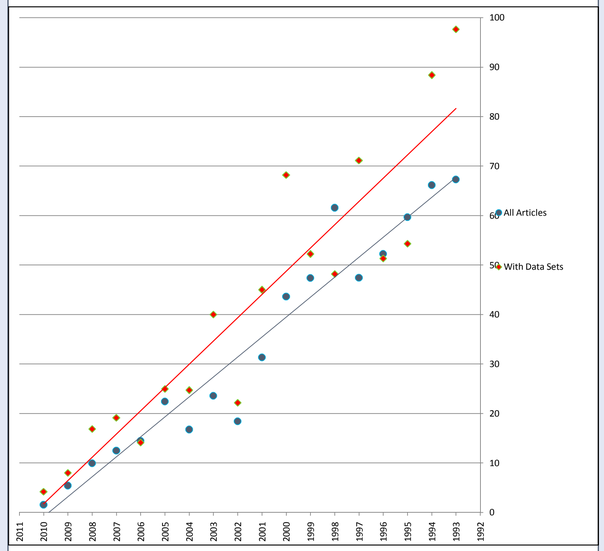

Figure 3. Plot of mean citations by year to articles with data sets vs. citations to all articles in *Paleoceanography* between 1993 and 2010. Trendline: With Data Sets: $R^2 = 0.845$; All Articles: $R^2 = 0.936$

Data sets in the study were almost always deposited with PANGAEA at the time of publication of the parent article, and citations naturally increase over time. Mean citations to articles with data were shown to be higher than those without data in 14 out of the 18 study years .

The citation counts seen in Web of Science, Crossref, and Scopus bibliographic databases are fairly consistent overall, bearing in mind the different universes of coverage and selection criteria. Some small differences may be due to source journals, policy on indexing comments and corrections, etc., with Scopus showing slightly more citations overall per article than Web of Science, and Crossref consistently at about ⅓ to ½ of the Web of Science numbers.

The 12.85% (171) of articles with publicly available supporting data sets received 19.94% (8,056) of the aggregate citations (40,409), or about 35% more citations per article than the average of all articles sampled over the 18-year study period (1,331). Thus, publicly available data was significantly (p = 0.007, 95% confidence interval) associated with a 35% increase in citations .

CONCLUSIONS

This study demonstrates a connection between the availability of supporting data sets and the citation counts of their parent articles, and although correlation does not equal causation, it is possible to speculate about likely reasons behind the increase in citations. Firstly there is the increased exposure among interested groups, not only directly via PANGAEA data library search clients, but also through syndication and dissemination of metadata among numerous registries and data centers and through various networking arrangements. A researcher who locates relevant data via any of these channels is able to link directly to the primary article. It is also clear that a community of paleoceanographers has become primary users of these data services. The correlation between publicly available data and increased literature impact in this study may further motivate researchers to share their data

REFERENCES

Aalbersberg, I. J., and Kähler , O. (2011), Supporting Science through the Interoperability of Data and Articles, *D-Lib Mag.*, 17(1/2), doi:10.1045/january2011-aalbersberg

Baker, D. N., C. E. Barton, W. K. Peterson, and P. Fox (2008), Informatics and the 2007-2008 Electronic Geophysical Year, *Eos Trans AGU*, 89(48) 485-486, doi:10.1029/2008EO480001

Costello, M. J. (2009), Motivating online publication of data, *BioScience*, 59(5), 418-427, doi:10.1525/bio.2009.59.5.9

Diepenbroek (2011), PANGAEA ® and the ICSU World Data System (WDS) - towards a global system of data publishers and data libraries, *Geophysical Research Abstracts*, Vol. 13, EGU2011-4859, EGU General Assembly 2011, http://meetingorganizer.copernicus.org/EGU2011/EGU2011-4859.pdf

Kleiner, K., Data on demand (2011), *Nature Climate Change*, 1, 10-12, doi:10.1038/nclimate1057

Lee, K., and L. Bero (2006), Ethics: Increasing accountability, *Nature*, doi:10.1038/nature05007

Lowry, R., E. Urban, and P. Pissierssens (2009), A New Approach to Data Publication in Ocean Sciences, *Eos Trans. AGU*, 90(50), doi:10.1029/2009EO500004.

National Science Foundation (2011), Dissemination and Sharing of Research Results, http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Parsons, M.A., R. Duerr, and J.-B. Minster (2010), Data citation and peer review, *Eos Trans AGU*, 91(34) 297-298, doi:10.1029/2010EO340001

Piwowar, H. A., R. S. Day, and D. B. Fridsma (2007), Sharing detailed reseach data is associated with increased citation rate. *PLoS ONE*, 2(3) e308, doi:10.1371/journal.pone.0000308

Thomson Reuters (2010). 2009 Journal Citation Reports ®.

CITATION

Sears, J.R.L. (2011), Data sharing effect on article citation rate in paleoceanography, Abstract IN53B-1628, presented at 2011 Fall Meeting, AGU, San Francisco, Calif., 5-9 Dec.

CONTACT

Jon Sears jsears@agu.org +1-202-777-7525