On the Citation Advantage of linking to data

Bertil Dorch

Discussion paper

On the Citation Advantage of linked data: Astrophysics

S. B. F. Dorch
The Royal Library, Copenhagen, Denmark
First version July 1 2012

Abstract

This paper present some indications of the existence of a Citation Advantage related to linking to data, using astrophysics as a case. Using simple measures, I find that the Citation Advantage presently (at the least since 2009 and in *The Astrophysical Journal*) amounts to papers with links to data receiving on the average 50% more citations per paper per year, than the papers without links to data. A similar study by other authors should a cummulative effect after several years amounting to 20%. Hence, a Data Sharing Citation Advantage seems inevitable.

1. Introduction

Demands for Data Management Plans are currently surfacing among research funding agencies, and infrastructures and policies are arising aiming at both archiving, structuring and distributing research data resulting from scholarly ventures within practically all fields of science. Hence, while scientists are increasingly being evaluated and funded according to quantitative measures, e.g. bibliometrics, it is timely to ask whether there is a Citation Advantage related to the activity of linking, i.e. sharing data, e.g. similar to the postulated Citation Advantage related to Open Access (jf. Kurtz et al. 2005, Eysenbach 2006, Kurtz et al. 2007, Swan 2010).

A study by Piwowar et al. (2007) based on the citation history of cancer microarray clinical trial publications suggests that sharing research data is indeed associated with an increased citation rate: Publicly available data was associated with a significant increase in citations, independent of other factors.

I present here a simple bibliometric study of publications within the field of astrophysics to investigate a possible increased citation impact resulting from linking to data, using the The SAO/NASA Astrophysics Data System, henceforth ADS (cf. Kurtz et al 2000).

2. Dataset and method

ADS is developed by NASA (The National Aeronautics and Space Administration), was launched in 1992, is hosted by the High Energy Astrophysics Division at the Harvard-Smithsonian Center for Astrophysics.

ADS is an online publication database of over 8 million astronomy and physics papers from both peer reviewed and non-peer reviewed sources, e.g. conference proceedings and books. Abstracts are available free online for almost all articles, and full scanned articles are available in GIF and PDF for older articles. New articles have links to electronic versions hosted at the journal's webpage.

ADS receives abstracts or tables of contents from approximately two hundred journal sources. The service may receive data referring to the same article from multiple sources, and creates one bibliographic reference based on the most accurate data from each source. For electronic articles, a list of the references

given at the end of the article is easily extracted. For scanned articles, reference extraction relies on OCR. The ADS can also list the citations for each paper in the database.

In the cases of the major astronomy journals - the nucleus being *Astrophysical Journal*, *Astronomical Journal*, *Astronomy and Astrophysics*, *Publications of the Astronomical Society of the Pacific* and *Monthly Notices of the Royal Astronomical Society* - the coverage is complete and these journals account for approximately two-thirds of the papers in the database. Many astronomers regard ADS their primary seach engine for searching for literature and citation analyses. In fact, tests show that ADS exceeds commercial databases such as Web of Science, both in number of relevant publications and in terms of citational coverage. The ADS search engine is tailor-made for searching astronomical abstracts. The database can be queried for author names, astronomical object names, title words, abstract text, and results can be filtered according to a number of criteria (cf. Eichhorn 2000).

For each publication record in ADS, a number of links are possible, including data links to on-line data, e.g. at other data centers. Links of this type are abbreviated "D" aka. D-links (cf. Accomazzi & Eichhorn 2004, Eichhorn *et al.* 2007 and Accomazzi 2011). Therefore, it is possible to limit ADS search results to publications with or without D-links. Other possible limits include specific journals, keywords, publication years etc. Furthermore, when returning citation data, it is possible to limit to non-author citations and peer reviewed citations only.

In this study, I limited the enquiry to the journal to *The Astrophysical Journal*, henceforth *ApJ*, a leading relatively high Impact Factor journal, often considered the top astrophysics journal for subject specific research. Additionally, I limit the study to papers published in *ApJ* in the 11-year period in the current millenium 2000 - 2010: In that period, the *ApJ* published in total 2457 to 3671 papers per year. On the average 829 papers per year have D-links, i.e. about 28% of the papers in ApJ in the chosen period, cf. Fig. 1.

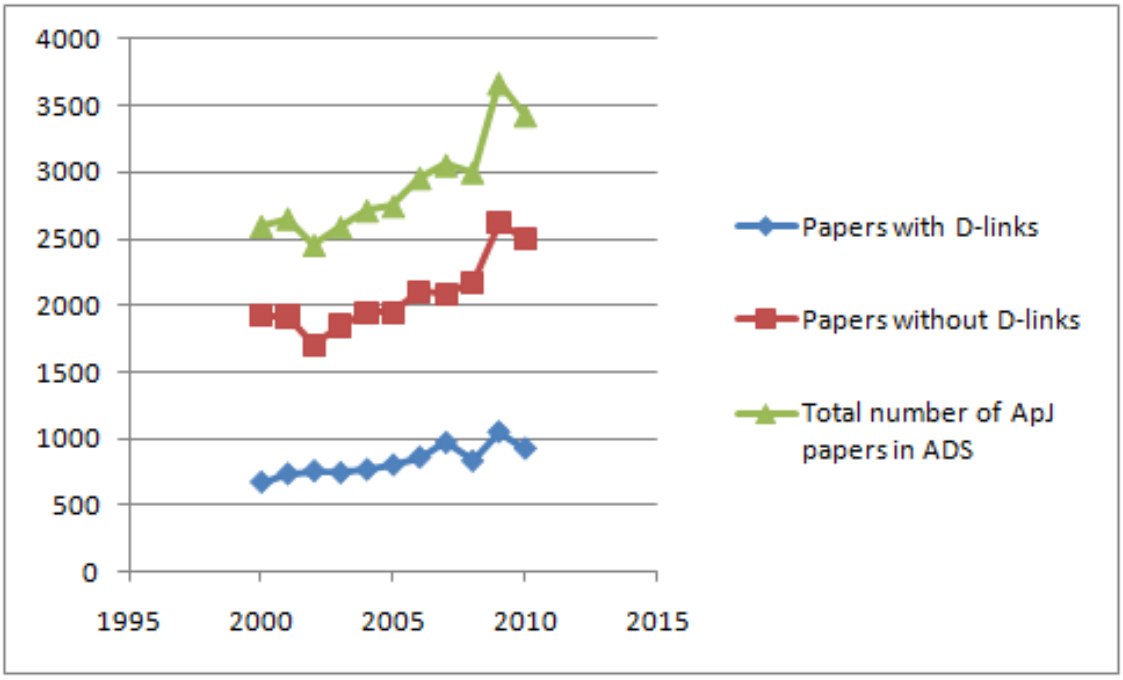

Fig. 1. The number of papers in The Astrophysical Journal 2000 to 2010 as a function of the year of publication as registered in ADS. Upper curve (green): Total number of papers. Middle curve (red): Papers without links to data. Lower curve (blue): Papers with links to data.

3. Results and discussion

The examined papers in the dataset with D-links received in total 29475 citations per year on average, whereas the papers without links to data received correspondingly more citations in total per year, namely 59995 citations per year, amounting to a total of 89470 citations per year, on average, rounded to integers.

However, as mentioned there being fewer papers with links to data, it turns out, that these papers on the average actually received more citations per paper. I.e. during the period the D-link papers in total on average receive 26% more citations per paper per year, than the papers without links to data (on average about 36 refereed citations per paper, excluding self-citations). Lately, i.e. since 2009, that fraction is more like 55% more citations per paper, cf. Fig. 2.

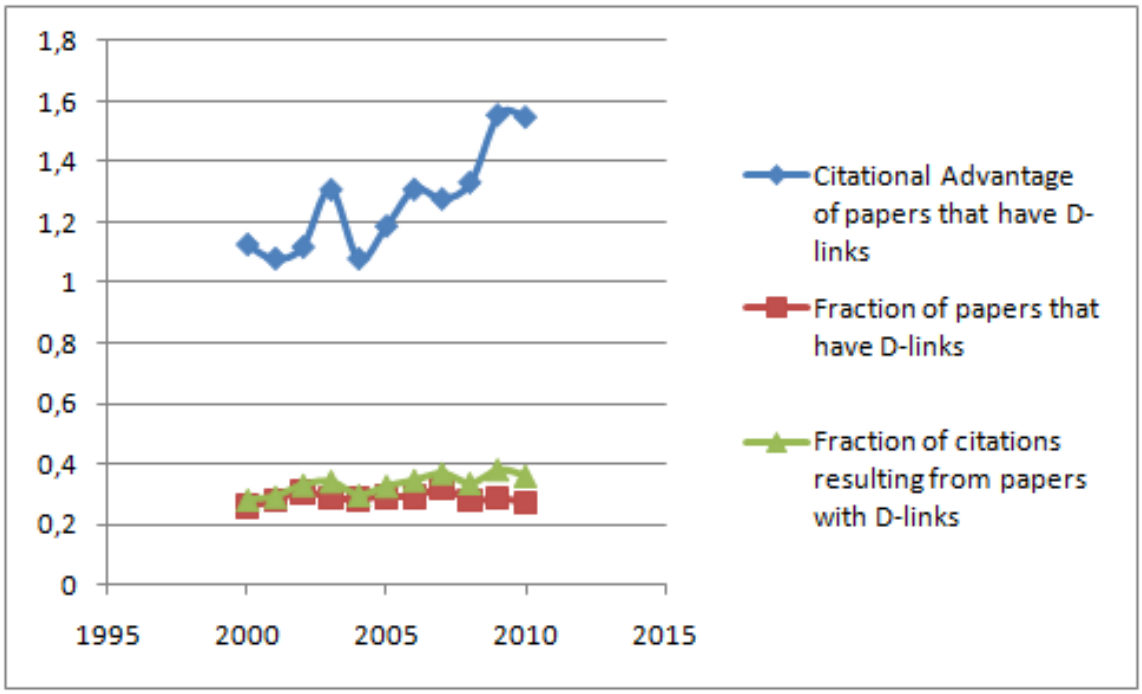

Fig. 2. Upper curve (blue): The Citational Advantage of papers that links to data as a function of the year of publication as registered in ADS (defined as the ratio of the average number of citations per year to papers with links to data, and the average number of citations per year to papers without such links). Middle curve (green): The fraction of the total number of citations that results from papers with links to data (average being 0,33). Lower curve (red): The fraction of papers that actually have links to data (average being 0,28).

This rather simple study indicates a clear tendency for papers with links to data to received more citations per year, on the average, that papers that do not link to data. However, there are several biases that could be looked further into, in more extensive studies. E.g. whether longer papers, papers with more authors etc. display generically difference citation patterns. Also of potential importanc could be whether some sub-subjects that “naturally” link to data already have a higher citation impact than other fields (that could e.g. papers based on space missions or telescope data as opposed to theoretical papers) - that is, whether field normalization should be performed.

However, a simple ADS search of refereed publications with non-self, refereed citations and keyword = "VLT" (The Very Large Telescope) shows that the 500 most recent papers (now not limiting to *ApJ* or publication year 2000-2010) with D-links receive about 50% more citations on (simple) average than the

latest 500 without links to data. I.e. within this more narrow field, of papers seemingly related to the VLT, it also seems that there inherently is a Citation Advantage.

Henneken & Accomazzi (2011) performed a similar analysis to the above, restricting their dataset to publications with at least 3 keywords in common,m based on a set of 50 keywords: They looking at cummulative citations to astrophysical papers after a 10 year period an report a 20% increase in citation count for articles with D-links, compared to those without. All in all, evidence is mounting that linking to data, and hence enabling sharing - if not in fact directly yielding open and free access to data - does indeed yield academic and scholarly incentives to those who do so, and this evidence is thereby also giving some merit to initiatives aiming at furthering the development of data infrastructures, services and policies.

Acknowledgements

This research has made use of NASA's Astrophysics Data System Bibliographic Services.

References

Accomazzi, A. & Eichhorn, G. (2004) "Publishing Links to Astronomical Data On-line". In Astronomical Data Analysis Software and Systems (ADASS) XIII, Proceedings of the conference held 12-15 October, 2003 in Strasbourg, France. Edited by Francois Ochsenbein, Mark G. Allen and Daniel Egret. ASP Conference Proceedings, Vol. 314. San Francisco: Astronomical Society of the Pacific, 2004., p.181

Accomazzi, A. (2011) "Linking Literature and Data: Status Report and Future Efforts". In Future Professional Communication in Astronomy II, Astrophysics and Space Science Proceedings, Volume 1. ISBN 978-1-4419-8368-8. Springer Science+Business Media, LLC, 2011, p. 135

Eichhorn, G., Kurtz, M.J., Accomazzi, A., Grant, C.S. & Murray, S.S. (2000) "The NASA Astrophysics Data System: The search engine and its user interface". Astronomy and Astrophysics Supplement, v.143, p.61-83

Eichhorn, G. Accomazzi, A., Grant, C.S., Kurtz, M.J., Thompson, D.M. & Murray, S.S. (2007) "Access to the literature and connection to on-line data". Bulletin of the Astronomical Society of India, Vol. 35, p. 717-725

Eysenbach G (2006) "Citation Advantage of Open Access Articles". PLoS Biol 4(5): e157. doi:10.1371/journal.pbio.0040157

Henneken, E.A. & Accomazzi, A. (2011) "Linking to Data - Effect on Citation Rates in Astronomy". To appear in the proceedings of ADASS XXI: http://arxiv.org/abs/1111.3618

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C.S., Murray, S.S. & Watson, J.M. (2000) "The NASA Astrophysics Data System: Overview". Astronomy and Astrophysics Supplement, v.143, p.41-59

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E. et al. (2005) "Effect of use and access on citations". Information Processing & Management, 41(6), 1395-1402.

Kurtz, M. J., & Henneken, E. A. (2007). "Open access does not increase citations for research articles Astrophysical Journal": http://arxiv.org/abs/0709.0896

Piwowar, H.A., Day, R.S. & Fridsma, D.B. (2007) "Sharing Detailed Research Data Is Associated with Increased Citation Rate". PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

Swan, Alma (2010) "The Open Access citation advantage: Studies and results to date.": http://eprints.soton.ac.uk/id/eprint/268516