SPARC Europe

Briefing Paper

The Open Data Citation Advantage

Introduction

The impact of scientific research continues to be evaluated by mechanisms that rely on the citation rates of published literature, so researchers, funders and research-performing organisations (RPOs) are all understandably keen to increase citations. SPARC Europe ran the Open Access Citation Advantage Service, which tracked literature on whether or not there is a citation advantage for Open Access articles. This is now so well established that they have stopped tracking the evidence. There is now increasing evidence that open practice with data also has a positive effect in a range of domains so widely spread that those who disagree with the effect would need to find conclusive evidence of domains where it does not hold. Several studies have shown that papers where the underlying data is available receive more citations over a longer period, and help generate further research. Open practice with data boosts citations to papers it underlies, providing a clear payoff to researchers for their data sharing effort. Moreover, open data practice increases the likelihood that the data itself will be cited as a scholarly output. The situation is complex; 'open practice' with data does not always simply mean 'open data', and other factors can also increase citation counts. In some domains, open data and publication linking is the norm and so no comparative studies can be done between those that do and do not make data available. The evidence is strong enough, however, to warrant effort from funders and RPOs to support researchers in open data practice. This briefing reviews a sample of the evidence, and recommends closer attention to potential links between the domains where these benefits have been found, and their prior investment in research data infrastructure.

Sharing microarray data and its effect on citation

Two studies have compared citation rates for publications that share underlying microarray data. Piwowar, Day & Fridsma (2007) were one of the first groups to identify a strong correlation between article citation rates and the associated publication of open data resources. The study focused on research conducted in the field of oncology, comparing cancer clinical trial studies which generated gene expression microarray data. Citation frequency was compared between studies which made their microarray data available and those which did not. As this type of data is difficult and expensive to generate, it was considered a good candidate to demonstrate the value of open data publication. Since this was one of the earliest studies to identify a strong correlation between the open publication of data and increased article citation frequency, it has been regularly cited by open data advocates. Most particularly, the arresting claim that:

> *"Publicly available data was significantly ($p = 0.006$) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression."*

The study recognised that there are a broad array of possible factors contributing to the frequency of article citation. An effort was made to account for some of these by controlling for journal impact factor, the age of the publication and whether or not an American author was included in the study. Despite mitigating for these elements, the study still suffered from two primary weaknesses – a small sample size (85 studies) and no allowance for the impact on citation rate of the number of collaborators contributing to a study.

Piwowar and Vision's (2013) follow-up study aimed to address these concerns. By continuing to focus on research generating gene expression microarray data, but broadening the range of research away from simply cancer clinical trials, it increased sample size dramatically (from 85 to 10,555), extended the timeframe under consideration and added extra covariates. The measured effect was smaller (9%, 95% confidence interval 5% - 13%) but still marked, and there is a greater degree of confidence in the results.

Some felt that, whilst the effect demonstrated by Piwowar & Vision was genuine, it was specific to a specialised field and unlikely to translate to other research domains. Studies in astronomy, social sciences, international relations and other domains have since shown otherwise. We summarise a sample here.

The effect of data sharing on citation rates in astrophysics and astronomy

Three studies have examined the comparative citation rates of publications with and without data links in astrophysics and astronomy. Each of these studies was based on information from the Astrophysics Data System (ADS), a compilation of three bibliographic databases. The ADS allows search results to be limited to articles with data links. Across the studies, a citation advantage of between 20-50% has been observed for articles that link to associated data.

Henneken & Accomazzi (2011) analysed articles published in four journals between 1995 and 2000. They restricted the subject matter of the articles in their analysis to ensure a reasonable comparison. They demonstrated a 20% citation advantage over 10 years for articles with links to open data. The authors questioned whether this citation advantage could be attributed to other factors such as open access publication or links from data centres, but found that the dataset analysed was homogenous for these other publication attributes, so concluded that the effect observed was real. They illustrate their findings in this graph

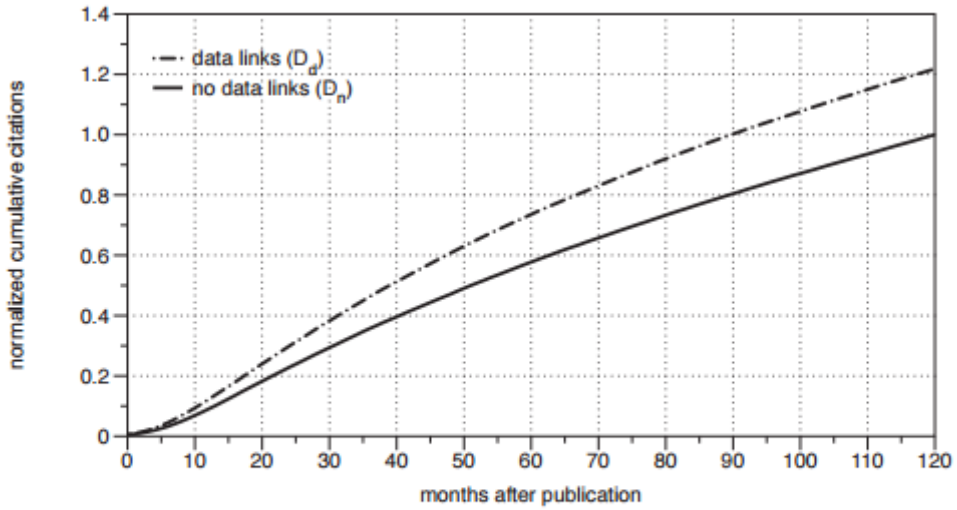

Figure 3. The cumulative citation distributions for data sets D_d and D_n. The citation counts have been normalized by the total number of citations for articles without data links, 10 years after publication.

Dorch (2012) meanwhile restricted his initial enquiry to papers published in the Astrophysics Journal between 2000-2010 and demonstrated a 26% advantage per paper per year; this increased to 55% for the final period (2009-2010) under examination. Dorch also recognised that other biases may be at play, such as paper length, co-authorship and the data sharing culture of the field. He performed a basic enquiry into papers related to telescope data in an attempt to address the potential field bias. Comparing papers using the keyword "The Very Large Telescope" showed that the 500 most recent papers with data links received on average 50% more citations than the 500 most recent papers without data links.

A further study (Drachen, Ellegaard, Larsen and Dorch, 2016) looks at the same journals as Henneken & Accomazzi in the period 2000-2014. In line with the previous Dorch study, they observed an increase in citations, particularly in the later period – 25% overall, 40% for 2009-2014.

International Relations

The effect doesn't just exist in the sciences. Gleditch (et al) looked at international relations, a subfield of political science, and collected citation data for 430 articles in the Journal of Peace Research for the 10 year period from 1991-2001. For each they identified whether the article offered data in any form, either through appendices, URL, or contact addresses. Control variables included author gender, length and age of article, whether co-authored or not, and the degree to which it included formal multivariate analysis. Their analysis shows a doubling of citations where data is made available, even after correction for the control variables. This increases to tripling from 1998, when the journal's data replication policy was introduced.

Social Sciences in the USA

A wide-ranging study of the effect of data sharing in US social science research (Pienta, Alter & Lyle, 2010) looked at a slightly different question – the extent to which data sharing increases use of the underlying data by leading to more publications. It examined thousands of studies funded by NSF and NIH between 1985 and 2001, and distinguished between no sharing (papers which made no reference to availability of data), formal sharing (data placed in disciplinary or institutional archive and linked to from paper) and informal sharing (such as statements about availability of data from a paper's author.) The analysis shows that formal sharing of data results in a 150% increase in the number of publications per dataset with a weaker effect for informal sharing of 75%.

Disciplinary dimensions of the citation advantage

The examples reviewed in this briefing offer evidence of a significant correlation between article citations and access to underlying data via domain or institutional repositories. Similar effects are seen for code sharing (Vanderwalle 2012.) Evidence that the citing authors are actually influenced by data access is more limited. Where it has been shown (e.g. Piwowar and Vision) the indication is that only a minority of the additional citations are from papers that reuse the data. But quality is more important than quantity here, and in two senses. Firstly, data sharing evidently pays off in terms of generating additional research which may generate citations to the original papers, the data or both. Secondly, authors cite papers for a number of reasons including the perceived quality of the article they are citing, and there is some evidence that data availability boosts reader perceptions of article quality.

Disciplinary factors may be a key issue; the evidence for a citation benefit is discipline-specific and data sharing and reuse practices are well known to vary across disciplines (Borgman, 2012; Tenopir et al, 2015), and even at the sub-disciplinary level (Cragin et al 2010; Lyon et al 2010). Nonetheless, it exists to some degree in every discipline which has been examined so far.
Studies so far have concentrated on domains with some degree of investment in data sharing infrastructure and data policies or community norms that promote data access and reuse. For some, however, that change has occurred relatively recently as technology has made data sharing feasible. Work such as Gleditch et al has influenced practice in its field and promoted a culture of replication and change in journal policy regarding data availability. Characteristics of research domains, such as collaboration, predominance of quantitative or interpretive approaches and common procedures (Becher and Trowler, 2001) all have an effect. Tam (2017) and Jacobs (2015) examine these issues in geography and social sciences respectively, and Borgman (2012) considers the effect of collaboration.

This suggests a need to tailor support and advocacy measures to discipline characteristics. For example, research projects that are highly collaborative and cross-disciplinary may have more realistic expectations of a citation advantage from data sharing. Lone cultural anthropologists, on the other hand, may be more receptive to benefits of data management tools that support rich contextualisation, and whose use in turn stimulates collaboration and potentially wider sharing and reuse. All need infrastructure to make data sharing possible and some need further evidence to persuade them of the desirability of change. It is critical to continue tracking and sharing the evidence to demonstrate tangible benefits and stimulate more people to share data.

References

Becher, T., & Trowler, P. (2001). *Academic Tribes And Territories: Intellectual Enquiry and the Culture of Disciplines.* McGraw-Hill Education (UK).

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology,* 63(6), 1059–1078. doi:10.1002/asi.22634

Cragin, Melissa H, Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. Philosophical transactions. Series A, *Mathematical, physical, and engineering sciences,* 368(1926), 4023–38. doi:10.1098/rsta.2010.0165

Dorch Bertil. On the Citation Advantage of linking to data: *Astrophysics*. 2012. <hprints-00714715v2>https://hal-hprints.archives-ouvertes.fr/hprints-00714715v2

Drachen, T.M. et al., (2016). Sharing data increases citations. *LIBER Quarterly*. 26(2), pp.67–82. DOI: http://doi.org/10.18352/lq.10149http://arxiv.org/abs/1111.3618

Gleditsch, N. P., Metelits, C., & Strand, H. (2003). Posting your data: Will you be scooped or will you be famous. *International Studies Perspectives*, *4*(1), 89-97.

Henneken, Edwin A. & Accomazzi Alberto (2011) Linking to Data – Effect on Citation Rates in Astronomy. http://arxiv.org/abs/1111.3618

Jacoby, JoAnn. "Share and share alike? Data-sharing practices in different disciplinary domains." Social Science Libraries: Interdisciplinary Collections, Services, *Networks* 144 (2010): 79.

Lyon, L., Rusbridge, C., Neilson, C., & Whyte, A. (2010). *DCC SCARP: disciplinary approaches to sharing, curation, reuse and preservation - final report.* Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

Pienta, Amy M.; Alter, George C.; Lyle, Jared A. (2010) *The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data* http://hdl.handle.net/2027.42/78307

Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2(3): e308. doi:10.1371/journal.pone.0000308

Piwowar HA, Vision TJ. (2013) Data reuse and the open data citation advantage. *PeerJ* 1:e175 https://doi.org/10.7717/peerj.175

Tam, Winnie. (2016) *Discipline and Research Data in Geography.* Unpublished PhD Thesis. Centre for Information Management, Loughborough University.

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. (2015) Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS ONE* 10(8): e0134826. doi:10.1371/journal.pone.0134826

Vandewalle, P. *Computing in Science & Engineering,* 2012 - scitation.aip.org

* * *

SPARC Europe would like to thank DCC for carrying out this work

www.dcc.ac.uk