EMBARGOED until Oct 1st 2013: 7 am EST; 12 midday UK time

PeerJ

Your Peers, Your Science
Academic Publishing is Evolving

Scientists who share data publicly receive more citations

A new study finds that papers with data shared in public gene expression archives received increased numbers of citations for at least five years. The large size of the study allowed the researchers to exclude confounding factors that have plagued prior studies of the effect and to spot a trend of increasing dataset reuse over time. The findings will be important in persuading scientists that they can benefit directly from publicly sharing their data.

The study, which adds to growing evidence for an open data citation benefit across different scientific fields, is entitled "Data reuse and the open citation advantage". It was conducted by Dr. Heather Piwowar of Duke University and Dr. Todd Vision of the University of North Carolina at Chapel Hill, and published today in PeerJ, a peer reviewed open access journal in which all articles are freely available to everyone (https://PeerJ.com).

The study examined citations to over ten thousand articles that generated new gene expression data, a quarter of which had data publicly archived in the GEO and ArrayExpress repositories. Papers with publicly available data received about 9% more citations overall, with the difference increasing over time. The researchers concluded that much of this citation difference was due to actual data reuse.

"Professional advancement in science is still highly dependent on how well your paper gets cited, even in a field like genomics where the data underlying that paper may have far more scientific impact over the long term." said Dr. Vision, a biologist affiliated with the National Evolutionary Synthesis Center and the Dryad Digital Repository. "Until the happy day when hiring and promotion committees catch up with how to value data sharing for its own sake, it is comforting to know that scientists can still receive credit for data sharing in a currency that counts."

The researchers also mined the full text of articles for references to dataset identifiers in order to study trends in data reuse directly. They took the unusual step of discussing the obstacles they encountered in the paper. Dr. Piwowar, at the time of the study a postdoc with the DataONE project, said "We need more open and cohesive infrastructure to support collecting evidence about the process and products of science. This evidence is needed to inform important policy decisions. For example, data archiving requirements, infrastructure, and education should be informed by evidence about how data is and is not reused."

The mined references revealed that scientists generally stopped publishing papers using their own datasets within two years, while other scientists continued to reuse their data for at least

six years. It also showed that data reuse is on the rise. “Not only were the number of reuse papers higher”, says Dr. Piwowar, “but analyses from 2002 to 2004 were reusing only one or two datasets, while a quarter of the studies by 2010 were using three or more.”

###

EMBARGOED until Oct 1st 2013: 7 am EST; 12 midday UK time

Link to the PDF of this Press Release: http://bit.ly/PeerJPiwowar

Link to the Press Preview of the Original Article (this link should only be used BEFORE the embargo ends): http://static.peerj.com/press/previews/2013/10/175.pdf (note: this is an author proof and so may change slightly before publication)

Link to the Published Version of the article (quote this link in your story – the link will ONLY work after the embargo lifts): https://peerj.com/articles/175 - your readers will be able to **freely** access this article at this URL.

PeerJ encourages authors to publish the full peer reviews, and author rebuttals, for their article. For the purposes of due diligence by the Press, we can provide these materials as a PDF (and they will be published alongside the final article). Please contact us at press@peerj.com to request a copy of the reviews.

Citation to the article: Piwowar HA, Vision TJ. (2013) Data reuse and the open data citation advantage. PeerJ 1:e175 http://dx.doi.org/10.7717/peerj.175

Other Information: The raw data behind this study are publicly available in the Dryad Digital Repository at http://doi.org/10.5061/dryad.781pv. This link will only work after Oct 1st

Funding: This study was funded by U.S. National Science Foundation grants to the DataONE (OCI-0830944) and Dryad (DBI-0743720) projects, and a Discovery grant to Michael Whitlock from the Natural Sciences and Engineering Research Council of Canada.

###

About PeerJ

PeerJ is an Open Access publisher of peer reviewed articles, which offers researchers a lifetime membership, for a single low price, giving them the ability to openly publish all future articles for free. The launch of PeerJ occurred on February 12th, 2013. PeerJ is based in San Francisco, CA and London, UK and can be accessed at https://peerj.com/.

###

Media Contacts

For the Authors:

Dr Heather Piwowar
Email: hpiwowar@gmail.com

For PeerJ:

press@peerj.com
https://peerj.com/about/press/

###

Abstract (from the article)

Background. Attribution to the original contributor upon reuse of published data is important both as a reward for data creators and to document the provenance of research findings. Previous studies have found that papers with publicly available datasets receive a higher number of citations than similar studies without available data. However, few previous analyses have had the statistical power to control for the many variables known to predict citation rate, which has led to uncertain estimates of the "citation benefit". Furthermore, little is known about patterns in data reuse over time and across datasets.

Method and Results. Here, we look at citation rates while controlling for many known citation predictors and investigate the variability of data reuse. In a multivariate regression on 10,555 studies that created gene expression microarray data, we found that studies that made data available in a public repository received 9% (95% confidence interval: 5% to 13%) more citations than similar studies for which the data was not made available. Date of publication, journal impact factor, open access status, number of authors, first and last author publication history, corresponding author country, institution citation history, and study topic were included as covariates. The citation benefit varied with date of dataset deposition: a citation benefit was most clear for papers published in 2004 and 2005, at about 30%. Authors published most papers using their own datasets within two years of their first publication on the dataset, whereas data reuse papers published by third-party investigators continued to accumulate for at least six years. To study patterns of data reuse directly, we compiled 9,724 instances of third party data reuse via mention of GEO or ArrayExpress accession numbers in the full text of papers. The level of third-party data use was high: for 100 datasets deposited in year 0, we estimated that 40 papers in PubMed reused a dataset by year 2, 100 by year 4, and more than 150 data reuse papers had been published by year 5. Data reuse was distributed across a broad base of datasets: a very

conservative estimate found that 20% of the datasets deposited between 2003 and 2007 had been reused at least once by third parties.

Conclusion. After accounting for other factors affecting citation rate, we find a robust citation benefit from open data, although a smaller one than previously reported. We conclude there is a direct effect of third-party data reuse that persists for years beyond the time when researchers have published most of the papers reusing their own data. Other factors that may also contribute to the citation benefit are considered. We further conclude that, at least for gene expression microarray data, a substantial fraction of archived datasets are reused, and that the intensity of dataset reuse has been steadily increasing since 2003.