Title: Data reuse and scholarly reward: understanding practice and building infrastructure

Authors: Todd Vision[1], Heather Piwowar[2]

Abstract

Recently introduced funding agency policies seek to increase the availability of data from individual published studies for reuse by the research community at large. The success of such policies can be measured both by data *input* ("is useful data being made available?") and research *output* ("are these data being reused by others?"). A key determinant of data input is the extent to which data producers receive adequate professional credit for making data available. One of us (HP) previously reported a large citation difference for published microarray studies with and without data available in a public repository. Analysis of a much larger sample, with more covariates, provides a more reliable estimate of this citation boost, as well as additional insights into patterns of reuse and how the availability of data affects publication impact. A more recent study tracking the reuse of 100 datasets from each of ten different primary data repositories reveals large variation in patterns of reuse and citation. Our findings (a) illuminate ways in which the reuses of archived data tend to differ in purpose from that of the original producers; (b) inform data archiving policy, such as how long data embargoes need to be in order to protect the proprietary interests of producers; (c) and allow us to answer the vexing question of what the return on investment is for data archiving. In conducting these studies, we have become aware of gaps in data citation practice and infrastructure that limit the extent to which researchers receive credit for their contributions. We describe early efforts to bake good data citation and usage tracking into cyberinfrastructure as part of DataONE, the Data Observation Network for Earth. Finally, we introduce total-impact, a tool that allows researchers to track the diverse impacts of all their research outputs, including data, and empowers them to be recognized for their scholarly work on their own terms.

Software and data availability:

- Research software and data: https://github.com/hpiwowar (CCZero for data where possible, MIT for code)
- Dryad - new BSD license: http://code.google.com/p/dryad/
- DataONE - Apache license: http://www.dataone.org/developer-resources
- total-impact - MIT license: https://github.com/total-impact/

[1] Dept. of Biology, University of North Carolina, Chapel Hill, NC

[2] DataONE and the National Evolutionary Synthesis Center, Dept of Biology, Duke University, Durham NC

PeerJ PrePrints | https://peerj.com/preprints/14 | v1 received: 29 Apr 2013, published: 29 Apr 2013, doi: 10.7287/peerj.preprints.14