Data Citation and Sharing: What’s in it for me?

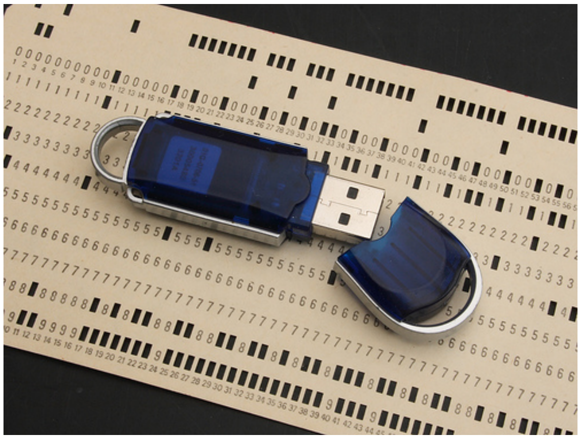
1/7/2014

Research funders, data managers, librarians, journal editors and researchers themselves are calling for a change in the culture of research to ensure formal data citation is the norm, rather than the exception. **Sarah Callaghan** *looks at the reasons for and against a more fluid data environment and finds that as well as being good for science, data sharing is also good for the scientist.*

“I’m all for the free sharing of information, provided it’s them sharing their information with us.” – Archchancellor Ridcully, Unseen University, Ankh-Morpork (Unseen Academicals p. 166).

Substitute the word “data” (or “code” or “methodology” or “workflows” or…) for “information” in the above quote and you’ve got a sentiment that a lot of researchers share, though maybe not in quite such a blunt way. You’d find it very hard to argue that data sharing isn’t difficult, time consuming, expensive, and generally not part of scientific practice. Conversely, you’d find it even harder to argue that data shouldn’t be shared.

Let’s get the reasons for sharing out of the way first.



1. Science is all about reproducibility – if someone else can’t reproduce your results, then your conclusions are invalid, and therefore the science doesn’t work. For a lot of scientific domains, reproducing results means using the original data collected, which means having access to it in the first place, which means sharing.
2. Data sharing cuts down on academic fraud. It’s hard work fabricating datasets (I know this from personal experience, having spent most of my PhD trying to simulate synthetic rain fields that looked anything like the real ones…), and having other people using your data means that they’re more likely to notice if something seems a bit wrong (which is also useful for error corrections).
3. Data sharing saves time and money. If a dataset already exists to test your hypothesis, why spend the effort and the money to collect an entirely new one?
4. Data sharing improves the transparency of the research process. If the data’s available to anyone who wants it, then you can’t be accused of hiding evidence about a controversial topic (like climate change).

All good things, right? So, the question then becomes: why don’t researchers share their data as a matter of course?

There are lots of reasons, which have been collated by lots of other researchers, ranging from fear of getting scooped (it happened to me!), to worry that others will find errors in the dataset or use it to misrepresent a key finding, to a simple and understandable desire to squeeze all possible research benefit out of the data before making it public. In a time of increasing competition and decreasing science budgets, hoarding data might make the difference between getting a grant and not, and therefore building a career as a scientist, or not.

Still, more and more research funders (including the UK Research Councils and the USA’s National Science Foundation) are becoming more interested in how the data produced by their funding is being managed, and have issued policies about this, including making statements about providing the resources needed for researchers to properly manage and share their data. Compliance, however, is patchy. Some academic fields (like my own, Earth Sciences) are very well supplied with long running and well-established national and international data centres, which serve their communities by offering services in data curation and archiving, and metadata creation and management. Other researchers aren't nearly as lucky, and face the situation of having to share their data via ftp sites, or departmental webpages, if at all.

Yet, as well as sharing data being good for science, it appears that data sharing is also good for the scientist. Piwowar et al (2007), showed that, for a sample of 85 cancer microarray clinical trial publications, 48% of trials with publicly available microarray data received 85% of the aggregate citations. A similar study (Piwowar and Vision, 2013) showed that of 10,555 studies that created gene expression microarray data, studies that made data available in a public repository received 9% more citations than similar studies for which the data was not made available. It remains to be seen whether this holds true for other research fields, as not that many research domains have an established norm of citing data, though anecdote seems to suggest it might be.

Data sharing can be done on a very informal basis, but that informality sometimes acts as a disincentive to researchers to share their data. Often sharing agreements are predicated on a quid pro quo arrangement, for example: "You can use my data, but I get co-author credit on any resulting publications" which requires trust on both sides.

In order to encourage data sharing, and the proper acknowledgement of data use, data citation is being promoted by a wide range of academics including research funders, data managers, librarians, journal editors and researchers themselves. Changing the culture of research so that formal data citation is the norm, rather than the exception, is a slow process, though it has been gaining momentum rapidly over recent years. Important players such as Thompson Reuters are developing tools and services built around data citation, including their Data Citation Index and research funders are providing guidance on how and what to cite.

The main premise in data citation is that data is a research output on par with other scientific publications and should be treated as such. Citation also gives the data producer formal attribution and credit, as well as allowing the tracking of the dataset's impact on the field. Data citation helps to meet the need for reproducibility by providing information about the dataset and its location, as well as ensuring that the cited dataset is archived and managed properly for the long term as part of the scientific record.

Page from alchemic treatise of Ramon Llull (Beginning of the 16th century)

So, after a lot of words, to answer the question asked in the title: Data sharing and citation is good for the data producer and user because it allows datasets to be found and reused easily, it encourages the proper archiving and curation of data that is important to the scientific record, and it provides rewards to the data producer in the form of data citations, and increased rates of paper citations where the data is shared.

Fundamentally though, we share our data because we want to be good scientists and we want to improve the level of public knowledge. If data isn't shared, we might as well be doing alchemy, not science!

Note: This article gives the views of the author, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Sarah Callaghan *is a Senior Researcher and Project Manager for the British Atmospheric Data Centre . She blogs about data citation, data publication and women in science at Citing Bytes and can be found on Twitter at @sorcha_ni.*