Open Data: Myths and Realities

Deborah Fitchett and Erin-Talia Skinner
Library, Teaching and Learning, Lincoln University, Christchurch, New Zealand

Introduction

We both believe that data from publically funded research should be published openly, as something the taxpayers have a right to, and as a way to promote the advancement of science. But we know researchers have some important concerns about publishing data, so we want to respond to those. We're using a question-and-answer format and hope this will be useful both to researchers considering publishing their own data, and support staff who may need to answer similar questions from researchers they work with.

1. It won't benefit me, so why should I bother?

Question: Look, all this talk about publishing datasets and making them openly available is all well and good but it won't help me. I conduct my research, I publish my paper and I'm done. I got what I came for and that's the end of it.

Answer: But you also want to be cited, right?

Question: Well… yes…..

Answer: So publish in a data archive that gives it a DOI and people can cite the data as well as the paper.

Even better, a study by Heather Piwowar and Todd Vision[1] of 11,000 papers based on microarray data shows that papers where associated datasets were published had a 9% citation advantage. Actually a number of studies in different fields[2-5] are looking at this citation advantage, but the reason I'm citing this one here – that's because the authors publish their data.

2. Someone might scoop me if I post my data.

Question: I know for a fact that there are pharmaceutical researchers in the United States, Denmark and Japan who are conducting research in the same area as me. This could be as big as aspirin! Why on earth would I want to publish my data online? They could see what I'm doing and beat me to the patent office!

Answer: Usually researchers publish the data at the same time as the paper. By the time that's published, you've got a big headstart on any follow-up research.

But where a patent's at stake, you can deposit the data in a data archive for proof that you were there first, but embargo its public release for long enough to secure the patent.

3. No-one cares about this data.

Question: I carry out research into the mating habits of the Amazon Horned Frog, specifically those living on the left bank of the Zamora River in the Morona Santiago province of

Ecuador. I'm one of three researchers in the world who care about this frog. Trust me; no one else is going to care about this research so there's no reason for me to make it available.

Answer: What about the researcher comparing mating habits of frogs worldwide? Or the researcher in the future comparing how this frog's habits have changed over time? Or the researcher working on other aspects of the Zamora River ecosystem? Or researchers in a field you never even considered.?

In 1993 a psychiatrist was researching panic disorders by injecting sleeping patients with caffeine – as you do. And a couple of their patients reported olfactory hallucinations. Now, the researchers couldn't care less about hallucinations, but they mentioned it in a letter to the editor – and that got picked up and influenced several studies about migraines[6].

4. Someone might see my mistakes.

Question: Look, I've got a chance at a permanent position at my university. I don't want to publish my data in an open repository. I might have made a mistake in one of my studies and something like that could cost me the position.

Answer: So wouldn't you want someone to spot the mistake early while you can still fix it and not years down the track after you've staked your reputation on dozens of papers? Not to mention the real-world consequences.

If researchers routinely published their data along with their conclusions, the fraudulent link between vaccinations and autism[7] would never have got off the ground. The blunder in an Excel spreadsheet that made austerity in an economic crisis look like a good idea[8] could have been spotted right away.

So send your data with your paper to be peer-reviewed. Then publishing it will give your readers confidence in the reliability of your results.

5. I can't release my data due to ethical or legal concerns.

Question: My colleagues and I have conducted a multi-year study into the efficacy of ways to help people with treatment-resistant affective disorders. This data includes a great deal of personal information about the subjects.. And what if someone misinterprets the data and sues us when their treatment goes wrong? It would be an ethical and legal minefield!

Answer: Social and health science researchers have developed lots of ways to preserve confidentiality. You could aggregate the data, like Statistics New Zealand does with the census data. Or use randomized data-distortion techniques. Or strip identifying data entirely.

As for legal concerns, MetService just puts a legal disclaimer on their website saying you use their data at your own risk[9], so you can't sue them if an unexpected storm ruins your boating trip. Or your boat.

6. Open data isn't appropriate for my field.

Question: I'm a psychology professor and I mostly conduct studies on the use of cognitive-behavioural therapy techniques. These are all small studies. Open data publication might be worthwhile in fields that produce large datasets but researchers in my field just don't do that.

Answer: Name a field and researchers in it are publishing their data. As a small sample, my slide shows some data archives from broad to narrow:

- Dryad, accepting data from every field but especially strong in biology and medicine;
- ICPSR for political and social science;
- Data.Govt.NZ (and governmental data sites worldwide);
- PANGAEA for earth and environmental science;
- ARDA for religion;
- NZSSDS for New Zealand social science survey data;
- VADS for visual arts;
- the NASA Exoplanet Archive;
- SAMHDA within the mental health field;
- PacMARS for the Pacific Marine Arctic region;
- Data Sharing for Demographic Research;
- GloWbE for corpus linguistics; and
- FlyBase for drosophilaphiles.

And fig**share** provides a platform specialising in negative data. So whether or not you can publish about your insignificant findings, you can publish the dataset here and make it citable. And yes, there are already at least 18 results there when I search for 'cognitive behavioral therapy'.

It's true that publishing data is extremely useful for large datasets. Lots of places are doing that and hopefully more will join them! But it's also extremely useful for small datasets, because if you pull together enough small datasets they become a large dataset.

7. I'm too busy and it's too hard.

Question: I have research to conduct, classes to teach, post-grads to supervise, meetings with business people interested in commercialising my research, departmental politics to navigate and my twin 18-month-old daughters are teething. I do not have time to do this and anyway, I have no idea how to do it so please go away.

Answer: There are more and more tools all the time devoted to making it easier for you to publish your data.

University of Edinburgh has great set of easy file naming conventions.[10] The University of California Curation Center has developed a plugin for Excel that helps you make sure your data is labelled and tidy before you upload it.[11]

When it comes to choosing a data archive, your journal or funder might already name a preferred repository. If not, Re3Data[12] and DataBib[13] both let you search for

appropriate research data repositories in your field. And those data archives will each take you through step by step the upload process. This also acts as a final checklist of the information you need to provide along with the data.

This labelling and documenting your data for publication might take some time. But it can also save time by making it easy for new members of your research group to follow it without an induction. Or risking the kind of mistake that happens if you don't induct them.

8. My institution doesn't reward data publication.

Question: This all sounds well and good but the fact is, my institution does not reward data publication. It doesn't even support it. Why should I care about publishing data when my institution doesn't?

Answer: Hands up everyone whose institution just might be interested in something that could give its published research a 9% citation advantage?

Your institution mightn't be paying attention now, but they will soon. As for support, data publication is on the radar of every academic and research library in the country. But we have to put our time where you put your priorities. So we need to know that this is important to you. Talk to your research office and talk to your library: tell them you want to know more. Talk to your colleagues, your deans, your research directors, and tell them why your institution needs to be developing policy, support structures, and rewards.

And if at first they're not convinced - show them the data.

Bibliography

[1] Piwowar, H. & Vision, T.J. (2013). Data Reuse and the Open Data Citation Advantage. *PeerJ PrePrints* 1:e1v1. doi:10.7287/peerj.preprints.1v1

[2] Diepenbroek, M. (n.d.). Data Sharing Effect on Article Citation Rate in Paleoceanography - KomFor. *KomFor*. Retrieved June 28, 2013, from http://www.komfor.net/blog/unbenanntemitteilung

[3] Dorch, B. (2012). On the Citation Advantage of linking to data. Retrieved from http://hprints.org/hprints-00714715

[4] Gleditsch, N. P., & Strand, H. (2003). Posting Your Data: Will You Be Scooped or Will You Be Famous? *International Studies Perspectives*, *4*(1), 72–107. doi:10.1111/1528-3577.04105

[5] Henneken, E. A., & Accomazzi, A. (2011). *Linking to Data - Effect on Citation Rates in Astronomy* (arXiv e-print No. 1111.3618). Retrieved from http://arxiv.org/abs/1111.3618

[6] Koerth-Baker, M., (2011). "Caffeine hallucinations: Why "Letters to the Editor" matter in science", *BoingBoing*, http://boingboing.net/2011/10/05/caffeine-hallucinations-why-letters-to-the-editor-matter-in-science.html

[7] Godlee, F., Smith, J., & Marcovitch, H. (2011). Wakefield's article linking MMR vaccine and autism was fraudulent. *BMJ*, *342*(jan05 1), c7452–c7452. doi:10.1136/bmj.c7452

[8] Herndon, T., Ash, M., & Pollin, R. (2013). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Political Economy Research Institute Working Paper Series, 322.

[9] MetService. (2013) Terms of Use of MetService websites. Retrieved from http://about.metservice.com/about/about-this-site/terms-of-use/

[10] Thompson, A. (2005, June 27). Records management guidance: standard naming conventions for electronic records. Retrieved July 1, 2013, from http://www.recordsmanagement.ed.ac.uk/InfoStaff/RMstaff/RMprojects/PP/FileNameRules/Rules.htm

[11] University of California Curation Center. (2012). DataUp: Describe, Manage and Share Your Data. Retrieved July 1, 2013, from http://dataup.cdlib.org/

[12] Registry of Research Data Repositories. (n.d.). Retrieved July 1, 2013, from http://www.re3data.org/

[13] Databib | Research Data Repositories. (n.d.). Retrieved July 1, 2013, from http://databib.org/