What should scientists have to say about where the data's at?

Alex O. Holcombe

Blog post published September 23, 2010 in *Alex Holcombe's blog*
https://doi.org/10.59350/nnecz-w9n35

Abstract

I have argued before that scientists must do more to make available the original data behind their articles. Any scientific claim should ideally be verifiable, via scrutiny of the original evidence. Currently, the vast majority of scientists never share their data.

Keywords

Cognitive Neuroscience (Subfield), Open Access

Recommended citation

Holcombe, A. O. (2010, September 23). What should scientists have to say about where the data's at?. *Alex Holcombe's Blog*. https://doi.org/10.59350/nnecz-w9n35

Copyright

Copyright © Alex O. Holcombe 2010. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

I have argued before that scientists must do more to make available the original data behind their articles.

Any scientific claim should ideally be verifiable, via scrutiny of the original evidence. Currently, the vast majority of scientists never share their data. They also know it can be very difficult for others to 'check their work'. When combined with the heavy pressure many scientists feel to publish rapidly and frequently, sloppy practices develop. A few even engage in outright fraud. Actually, usually everything is done right, but the inability to verify this means that the few bad apples undermine confidence in even the good science.

I think most scientists already agree that making original data available is a good thing. There are situations where it is not feasible or there are ethical issues associated with sharing the data (as can occur when it is difficult to anonymize data from humans), and cases where more publications are expected to result from a hard-won data set and perhaps scientists should have the right to hold the data back. Even excepting these problematic cases, however, very few scientists have been publishing their data. Of course, before the internet it was nearly impossible to publish original data, and so standard practice was to not do it, and this never really changed. Almost everyone has kept doing what they always did- publish the results of the data analysis but not the data itself.

Even before internet publication was possible, many journals and professional societies have had the policy that the data should be made available to other scientists 'upon request'. Sadly, scientists do not seem to be living up to even this standard of sending the data to another scientist when it is requested. In a recent study published in PLoS ONE, the authors asked other PLoS authors for the original data behind their PLoS-published results. Of the ten authors asked, only one provided the original data set, even after they were repeatedly reminded that PLoS journals explicitly require authors to share their data!

We have got to take steps to reduce the harmful hypocrisy evident here, and make more of science verifiable. I have suggested to some people that PLoS ONE and other journals require authors to complete a 'data availability statement'. Basically, authors would have to indicate whether the data associated with the science they are reporting will be available at the time of publication. If not, why not. If so, where it will be available. Perhaps this doesn't sound like much, and indeed it wouldn't impose any substantive burden. They'd essentially be doing little more than confirming how they plan to comply with already-existing PLoS policy.

Nevertheless, I think it would amount to a big push, because I suspect presently the vast majority of authors willfully ignore the issue or never even think about it. By forcing authors to write something about it, they'd have to confront the issue. I contend this would 1) push many authors who otherwise would not bother to archive their data to do so, and 2) nudge many authors who have not even thought about the issue to start thinking about ways they can archive their data, at least for future articles.

There is some feeling at PLoS that PLoS would have some responsibility to police the expectation of data availability, just as PLoS feels a responsibility to follow up on allegations of scientific

misconduct associated with its publications. This complicates the issue, and I am not sure how far that should go.

If PLoS journals or other journals were to require a 'data availability statement', how exactly should the requirement be worded? Remember that in the case of PLoS journals, we're talking about something that would apply to multiple disciplines, so we can't get into the details of what's expected or feasible in various subfields, what internet repositories are suitable, etc. So I'm thinking of something simple like this:

> *PLoS policy is that data should, if practical, be made publicly available at the time of publication. Many authors do this via their institutional repository or internet-accessible databases specific to certain scientific areas. For more discussion of the possibilities and the privacy and other issues that prevent publication of some datasets, see [insert link to long discussion and list of resources]*
>
> *1. Are the original data that you analyzed to support the claims of your article currently publicly available? YES/NO*
> *2. Will these data be publicly available when your article is published? YES/NO*
> *3. If you answered NO to 2, why will the data not be available?*
> *4. If you answered YES to 1 and/or 2, where are/will be the data available?*
>
> *Your answers to the above questions will appear in a 'data availability' section associated with your article.*

I certainly hope someone can think of a better way to wording the above! Additionally, you may have objections to the substance of these questions or to the proposed policy in general. I'd love to hear them.