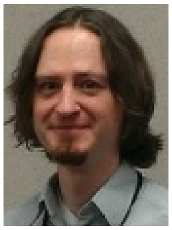
Global-level data sets may be more highly cited than most journal articles.

blogs.lse.ac.uk/impactofsocialsciences/2014/05/15/global-level-data-sets-highly-cited/

5/15/2014

Scientists can be reluctant to share data because of the need to publish journal articles and receive recognition. But what if the data sets were actually a better way of getting credit for your work? **Chris Belter** *measured the impact of a few openly accessible data sets and compared to journal articles in his field. His results provide hard evidence that the production, archival, and sharing of data may actually be a more effective way to contribute to the advancement of scientific knowledge.*



As with so many things, scientific research has been transformed by the rise of big data. These days, scientific discoveries happen through computer screens at least as often as they do through microscopes. In fact, many of the most significant recent discoveries like climate change and the Higgs boson were only possible because of big data.

As a result, there's been a lot of talk in the scientific community about how to deal with all of this data. Many in the community have been pushing to make that data openly accessible to everyone. They argue that sharing data is necessary to replicate and verify published research, to allow others in the community to more easily benefit from and build on others' research, and to foster a culture of openness and transparency in science.

But despite these arguments, scientists have, for the most part, been reluctant to share their data. Part of the problem is that researchers are not usually taught how to describe and format their data sets in ways that make them usable by others. But more fundamentally, scientists are reluctant to share their data because they are worried about not receiving credit for it. With the rise of publication-based assessments like the RAE, scientists have come under increasing pressure to publish journal articles with their data to advance in their careers. Sharing their data with the world not only increases the risk that others might benefit from that data without crediting them for collecting it, but also takes time and energy away from their production of new journal articles.

So, to try and persuade scientists to share their data, and to give credit to those scientists who are already doing so, I attempted to measure the impact that a few openly accessible data sets have had on scientific research. In my recent paper in Plos One, I analyzed the impact that three freely available oceanographic data sets curated by the US National Oceanographic Data Center have had on oceanographic research by using citations as a measure of impact. Since scientific assessments like the RAE increasingly use citations to journal articles for this purpose, I wanted to do the same for data sets.

I did the analysis in two phases. First, I did a series of searches in several journal article databases for articles that mention each of these data sets. The number of results from each search gave me a rough citation count for each data set in each database. Second, I delved a bit more deeply into the citations to one particular data set, the World Ocean Atlas and World Ocean Database, to try and learn more about when, where, and how this data set has been used in scientific research. Finally, I compared the data sets' citation counts to those of journal articles in oceanography to see how the data sets stacked up against that traditional measure of scientific research.

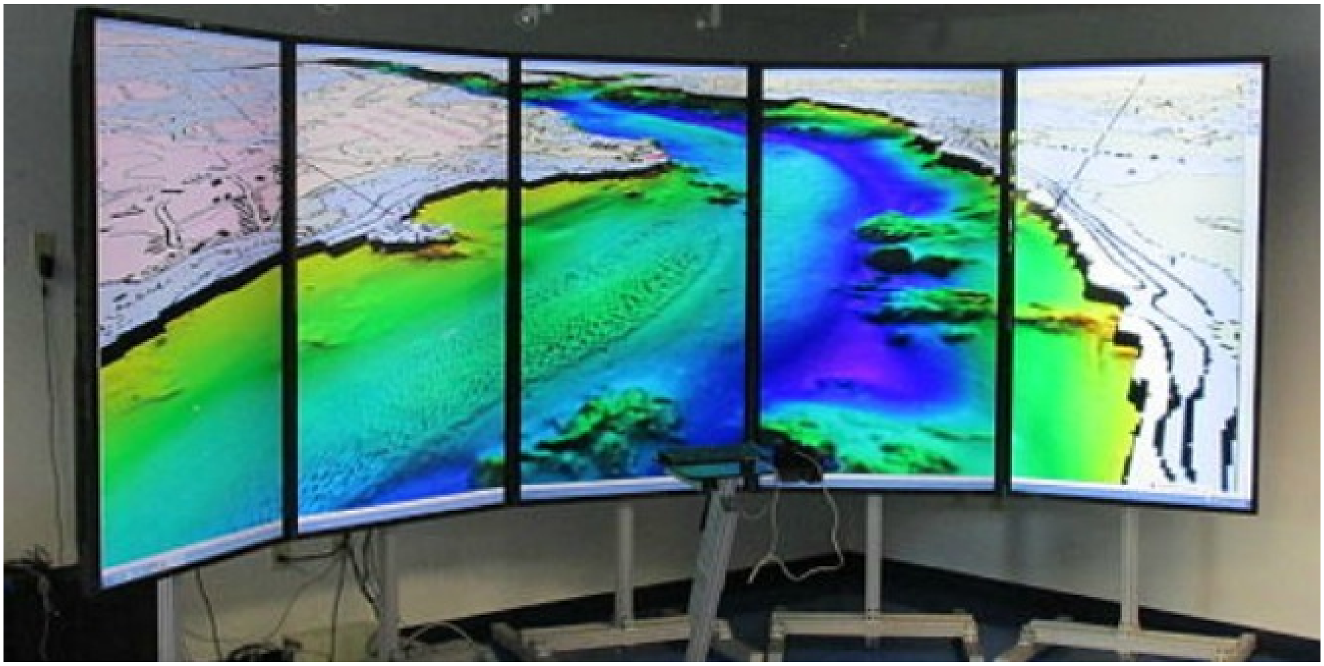


Image credit: John Phelan (Wikimedia, CC BY-SA)

My results suggest that all three data sets are more highly cited than most journal articles. Each data set has probably been cited more often than 99% of the journal articles in oceanography that were published during the same years as the data sets. One data set in particular, the World Ocean Atlas and World Ocean Database, has been cited or referenced in over 8,500 journal articles since it was first released in 1982. To put that into perspective, this data set has a citation count over six times higher than any single journal article in oceanography from 1982 to the present. True to its name, the World Ocean Atlas and World Ocean Database data set not only has global data, but also has global use. Numerous citations to the data set have come from every continent and every country in the world that produces substantial amounts of scientific research.

As I see it, these findings are important for three reasons. First, they provide hard evidence that journal articles may not necessarily be the only, or even the best, way to contribute to scientific knowledge. If these data sets are anything to go by, the production, archival, and sharing of data may actually be more important to advancing our understanding of the world than most journal articles are. My hope is that this evidence will start to convince scientists that sharing their data is not just a good thing to do, but can also be a significant contribution to science in its own right.

Second, I think these results have implications for how we measure the impact of scientific research. Although citations and other forms of bibliometric analysis can be useful to support the peer review process, they can often lead people to think that journal articles are the only things that count because they are the only things that can be counted. My analyses suggest that these data sets, and by extension the activities necessary to make them available and usable, can not only be counted, but can also count for more than most journal articles.

Finally, these results also suggest an opportunity for data repositories. The data sets I analyzed are unusual because each one is actually a collection of smaller data sets gathered by different researchers at different times and in different places. These smaller data sets were then standardized, quality controlled, and merged to form the global data sets in my analysis. The high citation counts I found for these data sets may have been the direct result of these processes. After all, global data sets are much more useful to the broader community than local ones. If this is the case, then other data repositories like Genbank, Dryad, and ICPSR have the opportunity to create global-level data sets in other fields by combining the individual- or local-level data sets they already have. This might also encourage researchers to put more data into these repositories, since the more data such a global-level data set includes, the more useful it becomes.

The creation of such global-level data sets would benefit everyone. It would make science more transparent and reproducible because everyone would have access to the underlying data. It would enable scientists, and data repositories, to show the impact of their data on future scientific research. And it would allow everyone to access and build on the data gathered by everyone else – which, after all, is what science is all about.

Note: This article gives the views of the authors, 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

Chris Belter *is a public services librarian at the US National Oceanic and Atmospheric Administration. His research focuses on the how bibliometric analysis techniques can expand and enhance the services provided by libraries and librarians.*