/706/648/479

One size doesn't fit all

Jacopo Bertolotti
University of Exeter, j.bertolotti@exeter.ac.uk

I have found over and over that in any discussion about if and how to share data, and how much to share, a large fraction of the debate boils down to people coming from different disciplines, with different approaches to the everyday working of scientific research, and thus different problems that need to be solved. Maybe using a simple example can be useful to avoid misunderstandings, so let's consider the case where in our paper an important piece of information is summarized as a spectrum.

One possibility is to present the spectrum as a nicely readable plot, but not to include any data. Proponents of this approach defend this position noticing that if the plot is well formatted it is possible to extract the data directly from the plot (there is software that does that for you), so there is no need to explicitly provide the data.

Personally I tend to disagree with this position, as the data retrieved from the plot is never of very high quality.

A low-effort but high gain approach is to just provide the data used for that plot. As the plot itself should be well formatted and captioned there is no need to carefully curate the data itself. Anyone able to understand what the spectrum means are also able to distinguish which of the columns is the wavelength or frequency and which one is the transmission or fluorescence and so on. And since all the units used are already in the plot, and the way the spectrum was obtained is described in the method section of the paper, there is no need for any annotation.

Personally I find that the amount of work needed to do this level of data sharing is minimal, and can really help people who want to use your results as a stepping stone to new science. This basic level of data sharing has practically no drawback, and I cannot think of any reason for people not to do it. Actually I think that this data should be available from the journal web page simply by clicking on the plot.

It is beyond this basic level of data sharing that things get complicated. If my spectrum is actually the average of many measurements, should I share all the measurements or just the data in the plot?

Should I share the script I used to collect the data? Should I share the data before I made a background subtraction? What if the data where the background was not subtracted was simply never saved? And what about all those spectra that didn't really work - should we share them too? I don't think there is a "one size fits all" kind of answer to these questions, and each sub-discipline will have different reasons to answer one way or another to each of these questions.

I work in the field of optics and photonics, where it is common to replicate previously published measurements as an intermediate step to your own original results (although the "replication" part is seldom advertised). It is also a field where people can argue bitterly on how to interpret the data,

but arguments about the data itself are much more rare. As a consequence the main reason for sharing data is to allow other scientists to better understand what we did.

In contrast, a large experiment like Atlas is not easy to reproduce independently, and analyzing the data produced is a highly specialized job, making it very unlikely that someone is able to both produce and analyze the data. Given this separation of specializations it makes sense that the people producing the data will make them available to anyone able and willing to analyze them. In most cases the only people truly able to analyze them are anyway already part of the collaboration, so there is no risk of getting scooped. But even here Atlas does not share "all" the data. Actually a lot of the data is never saved, as people spent many years developing smart triggers that allow them to save only those event that has any chance of containing something interesting. All events where nothing happens are just discarded and nobody will ever see them. This is not a problem, because the way this selection is made is open and well known among the specialists. As everyone agrees that this is the right way to select the data, the fact that the data is selected is a non-problem.

But in many biomedical disciplines the experiments (although not even remotely as massive as Atlas) are large, complex, full of subtleties, difficult or expensive to reproduce and yet carried out by relatively small groups in direct competition with each other. In such an environment it is difficult to all agree on a "best" way to do things like they can do at LHC, LIGO etc, and cross-checking each other becomes a very taxing and unrewarding task. It's not a surprise that people in such disciplines are the most vocal in asking that really all the data are shared. For them it is not just a way to help understanding this or that paper, it is a way to try to spot the frauds before too much damage is done. And I totally agree that what they ask makes total sense in their own context.

The danger I see is that what makes sense in genetics will be pushed on to people working in high energy physics, or what works there will be pushed on people working in optics, in a misguided quest for a "one size fit all" answer to what can only be a nuanced and highly differentiated problem.