Moving toward a sustainable ecological science: don't let data go to waste!

Timothée Poisot, Dominique Gravel

Nov. 2012

Intro

Claude Bernard (Bernard 1864) wrote that "art is *me*; science is *us*". This sentence has two meaning. First, the altruism of scientists is worth more than the self-indulgence of mid-nineteenth century Parisian art scene. Second, and we will keep this one in mind, creativity and insights come from individuals, but validation and rigor are reached through collective efforts, cross- validation, and peerage. Given enough time, the conclusions reached and validated by the efforts of many will take prominence over individualities, and this (as far as Bernard is concerned), is what science is about. With the technology available to a modern scientist, one should expect that the dissolution of *me* would be accelerated, and that several scientists should be able to cast a critical eye on data, and use this collective effort to draw robust conclusions.

In molecular evolution, there exists a large number of databases (GenBank, EMBL, SwissProt, and many more) in which information can be retrieved. This values (and allows) a new type of scientific research: building over the raw material of others, it is now possible to identify new phenomenon or evaluate the generality of previously studied ones. The job of these scientists is not to *make* data, neither to *stole* them, it's rather to gather them and, most of all, look at them in a different way. This would not be possible, if not for the existence of public, free, online repositories. It's impossible to be as enthusiastic when looking at current practices in ecology. Apart from a few, non-specific initiatives (*DataDryad*), or small-scale initiatives which are not always properly maintained (*Interaction Web Networks Database*), there is no data sharing culture among ecologists.

TODO (Reichman, Jones, and Schildhauer 2011) *DataONE* as a signal that some organisations are ready to invest time and money

In this paper, using the example of ecological networks, we will argue that improving our data sharing practices will improve both the science, and the reputation of the scientists. We will illustrate how simple steps can be taken to greatly improve the situation, and how we can encourage the practice of data-sharing at different levels.

Why we morally must

- Most data underlying published ecological research is generated by publicly- or charitably-funded researchers

TODO Find some examples of data sharing and availability policies from funding bodies

- It allows reproducibility of the science, which is supposed to be the rule

Using journals to publish scientific information should not only serve the purpose of disseminating an interesting discussion of data: it should maximize the ability of other researchers to replicate, and thus both validate and expand, results. It's interesting to see that, while editors and referees alike are very careful about the way the *Materials & Methods* parts of a paper are worded, it's extremely rare to recieve any comment about the data availability. This can cause problems at all steps of the life of a paper. How can a paper describing a new method be adequately reviewed if data are not available? How can you be sure that you are correctly applying a method if you can't reproduce the results? The movement of *reproducible research* advocates that a paper can be self-contained, *i.e.* be not only the text, but also the data, and the computer code to reproduce the figures. Even without going to such lengths, releasing data and computer code alongside a paper should be viewed as an ethical decision. Barnes (Barnes 2010) made the point that computer code is good enough to be shared, even though researchers are not professional programmers.

- It will fight bad authorship practices, people hitch-hicking on other people's work
- Data are costly (time and money) to acquire, acquiring new instead of using old ones is wasteful

(Heidorn 2008) dark data, there is already enough material to answer some pending questions

[@Wicherts2006] surveyed the field of psychology, and showed that asking for the raw data often doesn't result in a successful data sharing outcome, even after 6 months of repeated inquiries. Authors can claim to have 'lost' the data, can be extremely slow to reply, can ignore emails, the given contact email address may be invalid and difficult to find the 'current' contact address. Authors also die, and sadly this can result in the loss of valuable scientific data unless it has been accessibly and discoverably archived elsewhere. Ultimately, authors can also flat out refuse to give the data.

Why is it beneficial for the one who collected data

- A proxy to your science: data are a mean for people to get familiar with what you do

(Ince, Hatton, and Graham-Cumming 2012) improves reproducibility and adequate communication of your results

(Vandewalle 2012) showed that sharing computer code improved the scientific impact

(Piwowar, Day, and Fridsma 2007) Sharing detailed research data is associated with increased citation rate for your papers

- It stimulates collaboration and creativity
- A measure of your productivity

How we technically can

- End the rule of Excel: JSON schemes or XML to represent context-rich data
- FigShare and other projects: data can have a DOI and be cited/shared
- Local databases but linked globally: APIs and programmatic access

How it should be encouraged

- Journal policies, and referee expertise

Journals are in the best position to make things move (Vision 2010), because a scientist career depend on getting its papers accepted. Although when possible, a bottom-up approach sould always be prefered, some journals are now asking the authors to deposit their ecological data in a public repository (Fairbairn 2011 ; Whitlock et al. 2010). This is mandatory for sequences in all journals (*GenBank*), and archiving of all data in TreeBase, DataDryad, or FigShare is becomming a common practice. The referees are, however, rarely asked to evaluate if the adequate data are released (*e.g.* network metrics and summary statistics instead of full networks), and even more rarely given access to the data during the evaluation process.

Ecological journals have policies in place

- Evaluation for funding

Conclusion

In the last two years, there were an important number of media outbursts, and public indignation, about the role of science and scientific conduct, which may all have been avoided if the practice of putting data publicly online was widespread. The so-called *climategate* (Jasanoff 2010) could have been largely averted if all data were made public in the earlier days of the affair, as it was later clearly demonstrated that the apparent lack of transparency eroded public trust in scientists (Leiserowitz et al. 2010 ; Ravetz 2011). Even more recently, the controversy over a study on the carcinogenicity of GM maize (Séralini et al. 2012) was thickened by the refusal of both sides (Monsanto and the french research group) to release the full data, in addition to many undisclosed conflicts of interests (Meldolesi 2012).

List of possible boxes

- The story of the BCI data
- What we could tell about network biogeography with public data?

Barnes, Nick. 2010. "Publish your computer code: it is good enough." *Nature* 467: 753. doi:10.1038/467753a.

Bernard, C. 1864. *Introduction à l'étude de la médecine expérimentale.*

Fairbairn, Daphne J. 2011. "The advent of mandatory data archiving." *Evolution* 65: 1–2. doi:10.1111/j.1558-5646.2010.01182.x.

Heidorn, P. Bryan. 2008. "Shedding Light on the Dark Data in the Long Tail of Science." *Library Trends* 57: 280–299. doi:10.1353/lib.0.0036.

Ince, Darrel C., Leslie Hatton, and John Graham-Cumming. 2012. "The case for open computer programs." *Nature* 482: 485–488. doi:10.1038/nature10836.

Jasanoff, Sheila. 2010. "Testing time for climate science." *Science.*

Leiserowitz, Anthony, Edward W. Maibach, Connie Roser-Renouf, Nicholas Smith, and Erica Dawson. 2010. "Climategate, public opinion, and the loss of trust." *Social Science Research Network.*

Meldolesi, Anna. 2012. “Media leaps on French study claiming GM maize carcinogenicity.” *Nature Biotechnology* 30: 1018.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. 2007. “Sharing detailed research data is associated with increased citation rate.” *PloS one* 2: 308. doi:10.1371/journal.pone.0000308.

Ravetz, J. R. 2011. “’Climategate’ and the maturing of post-normal science.” *Futures*.

Reichman, O. J., Matthew B. Jones, and Mark P. Schildhauer. 2011. “Challenges and opportunities of open data in ecology.” *Science* 331: 703–5. doi:10.1126/science.1197962.

Séralini, Gilles-Eric, Emilie Clair, Robin Mesnage, Steeve Gress, Nicolas Defarge, Manuela Malatesta, Didier Hennequin, and Joël Spiroux de Vendômois. 2012. “Long term toxicity of a Roundup herbicide and a Roundup-tolerant genetically modified maize.” *Food and chemical toxicology* 50: 4221–31. doi:10.1016/j.fct.2012.08.005.

Vandewalle, Patrick. 2012. “Code Sharing is Associated with Research Impact in Image Processing.” *Computing in Science and Engineering*: 1–5.

Vision, Todd J. 2010. “Open Data and the Social Contract of Scientific Publishing.” *BioScience* 60: 330–331. doi:10.1525/bio.2010.60.5.2.

Whitlock, Michael C., Mark A. McPeek, Mark D. Rausher, Loren Rieseberg, and Allen J. Moore. 2010. “Data archiving.” *The American naturalist* 175: 145–6. doi:10.1086/650340.