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

Timothée Poisot, Ross Mounce, Dominique Gravel

Nov. 2012

Introduction

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 organizations 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, and data citation (http://datacite.org/whycitedata) to encourage and reward sharing.

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

some UK-based funder policies: - BBSRC (http://www.bbsrc.ac.uk/organisation/policies/position/policy/data-sharing-policy.aspx) cites OECD report principles that "Publicly-funded research data are a public good, produced in the public interest" and "Publicly-funded research data should be openly available to the maximum extent possible" the policy itself: "The value of data often depends on timeliness" "it is expected that timely release would generally be no later than the release through publication of the main findings" applicants for BBSRC funding, as with NSF, NERC, and many other funders applicants *must* provide a "statement on data sharing" in grant proposals which will be assessed.

- NERC (http://www.nerc.ac.uk/research/sites/data/policy2011.asp) "All the environmental data held by the NERC Environmental Data Centres will normally be made openly available to any person or any organisation who requests them." "All those who use data provided by NERC are required to acknowledge the source of the data" "All applications for NERC funding must include an outline Data Management Plan" "The outline data management plan will be evaluated as part of the standard NERC grant assessment process. All successful applications will be required to produce a detailed data management plan in conjunction with the appropriate NERC data centre." "... Those funded by NERC who do not meet these requirements risk having award payments withheld or becoming ineligible for future funding from NERC"
- 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 receive 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

Also, add a point about the quantification of authorship, some refs to cite here: http://aidanmkeith.wordpress.com/2012/11/06/tiered-authorship-as-a-simple-quantifiable-and-greyscale-measure-of-contribution/

- 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

(Wicherts et al. 2006) 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 that is increasingly being appreciated and encouraged by research funder agencies, as an example: the NSF (US) Grant Proposal Guidelines for 2013 have renamed the 'Publications' section to 'Products' specifically to make it clear that they apppreciate research products that "include, but are not limited to, publications, data sets, software, patents, and copyrights" (http://nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_sigchanges.jsp). Published datasets are now truly creditworthy, first class research objects (http://www.force11.org/white_paper) in the eyes of many funders.

How we technically can

Data representation

Except when they are deposited into large-scale databases, such as the ones we previously mentioned, data usually live (in various states of dormancy) on the hard drives of researchers. These data are usually formatted in the way where they were used to produce the few figures used in the published account, which is to say mostly as a spreadsheet, or a raw text file (Akmon et al. 2011). JSON / XML

Data sharing

- Local databases but linked globally: APIs and programmatic access

An important obstacle is that maintaining a global database requires funding on a scale which is orders of magnitude higher than what most grants will cover. The other solution, building on an increased use of strict data specification, is to link several local databases through APIs. A potential research output for working groups should thus be to design a strong data specification, and to publish it for other researchers to adopt. In the ecological sciences, there are now publications outlets focused only on methodological papers (*Methods in Ecology and Evolution*, and to some extent *BMC Bioinformatics*), and several other journals have sections for methodological papers. The conception of a data specification can thus be valorised as a research output in the form of a publication, which is accounted for by funding and tenure committees.

- FigShare and other projects: data can have a DOI and be cited/shared

Data freedom!

- Apply appropriate Creative Commons licenses or waivers to digitally available research data to remove legal barriers to re-use and prevent legal ambiguity over what type of re-use is allowed by the authors. We echo (Hrynaszkiewicz and Cockerill 2012) that the CCO waiver is best for factual non-copyrightable data (e.g. measurements) and that the Creative Commons Attribution license (CC BY) is best for copyrightable data such as photographs. Both these licenses are in accordance with the Panton Principles (http://pantonprinciples.org/) for open data in science. See (Hagedorn et al. 2011) for further details, including an explanation of the pitfalls of more restrictive Creative Commons licenses.

How it should be encouraged

The role of journals

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 should always be preferred, editors have in their hand a formidable lever to modify our collective behavior. 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 becoming 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. In practice, authors are still free to release summary statistics instead of raw data, which allows to reproduce the paper, but not to confirm the validity of the approach.

Journal-led mandates cannot be the only solution used. When compliance with journal stipulations are retrospectively checked, even clinical trials data compliance (Prayle, Hurley, and Smyth 2012) and *GenBank* archiving of data are not universally adhered to, even in the 'best' journals of highest reputation (Noor, Zimmerman, and Teeter 2006). Journals must take care that data archiving mandates are enforced and not just 'rhetoric', be it through increased editorial control, or by asking the referees to evaluate the data sharing plans.

Ecological journals have policies in place

The role of funding agencies

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?

Akmon, Dharma, Ann Zimmerman, Morgan Daniels, and Margaret Hedstrom. 2011. "The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs." *Archival Science* 11: 329–348. doi:10.1007/s10502-011-9151-4.

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.

Hagedorn, Gregor, Daniel Mietchen, Robert Morris, Donat Agosti, Lyubomir Penev, Walter Berendsohn, and Donald Hobern. 2011. "Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information." *ZooKeys* 150: 127–149.

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.

Hrynaszkiewicz, Iain, and Matthew Cockerill. 2012. "Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals." *BMC Research Notes* 5: 494.

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.

Noor, Mohamed A. F., Katherine J. Zimmerman, and Katherine C. Teeter. 2006. "Data Sharing: How Much Doesn't Get Submitted to GenBank?." *PLoS Biol* 4: 228.

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.

Prayle, Andrew P., Matthew N. Hurley, and Alan R. Smyth. 2012. "Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study." *BMJ* 344.

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.

Wicherts, Jelte M., Denny Borsboom, Judith Kats, and Dylan Molenaar. 2006. "The poor availability of psychological research data for reanalysis." *American Psychologist* 61: 726–728.