Of carrots and sticks

Jens Kattge, Sandra Díaz and Christian Wirth

Journals and funders increasingly require public archiving of the data that support publications. We argue that this mandate is necessary, but not sufficient: more incentives for data sharing are needed.

With the digital revolution, data exchange has become easy in a technical sense. As a result, data that were originally collected for one specific purpose can now be used in different contexts, to answer new scientific questions. Data sharing offers prospects for progress, and not only for data-intensive sciences like remote-sensing. Scientific domains that are typically dominated by numerous small data sets — such as ecology, biodiversity or medicine — stand to benefit.

Historically, data sharing was limited by the absence of centralized easily accessible archives for scientific data. Data were usually stored at the research institutions where they were produced, and formats ranged from a centralized institutional digital repository to hand-written field notebooks. Upon publication, the underlying data sets were typically shared via bilateral communication between researchers, and mostly used solely to repeat a given study and verify its results. The originality of these publications might be particular. In such informal structures, metadata are often lacking and data are prone to rapid loss of information content, which makes reuse difficult.

As the focus of much research is shifting towards larger-scale questions — such as global biodiversity scenarios, worldwide organisational specialization patterns and continental pandemics — and data collection is becoming ever more efficient, the approach to storing and disseminating data needs to change.

We argue that data sharing is already rewarded with recognition, influence and collaborations, but stronger incentives in terms of citations are overdue. Only if the full scientific value of generating and disseminating data is acknowledged, will data sharing become the integral part of the scientific system that it needs to be.

### Three stumbling blocks

Constraints to data sharing are now more social than technical. Scientists tend to embrace the opportunity of sharing data in bilateral contexts, but they are often reluctant to release their data to the broader scientific community. The reasons for this are manifold, but we feel that three are particularly prominent.

First, high-quality data are hard to obtain. Researchers should expect their fieldwork to yield one or several primary publications. The originality of these publications might be jeopardized if the data are widely available before they are published, or if individual data sets are amalgamated in large collective synthesis publications. The reluctance towards making hard-earned data publicly available is understandable — at least as long as the measurements have not been sufficiently exploited by those who obtained them.

Second, data are context dependent. Without appropriate contextual information, for example metadata regarding locations, methods and shortcomings, they can easily be misinterpreted and misused. Opening data sets to other researchers means that the circumstances of collection — known by those who performed the measurements — need to be carefully recorded and communicated.

The third factor is related to the previous one: significant effort is often necessary to prepare the data for reuse before they can be made available to the scientific community. In addition to contextualising data, formats may have to be adjusted and a point of contact may be necessary in case there are questions.

In order to overcome these obstacles to voluntary data sharing, public archiving has been made mandatory by many publishers and funding agencies. However, this approach — using a proverbial stick to encourage data sharing — has not (yet) led to a broad cultural change in researchers’ actions. We therefore advocate complementary incentives — a carrot that will supplement the stick.

### Big data, many references

There are already incentives for sharing high-quality data accompanied by all the relevant contextual information. Benefits to those who readily make their data available include the facilitation of new collaborations and the development of their professional network; researchers can enhance their own original data set by allowing others to access it and contribute; data sharing can also result in joint publications with other groups who use the data, and it may yield data publications and hence citations beyond those of the original papers.

These benefits have perhaps not been cultivated as much as they could be, but improvements are under way. Examples include the development of domain-specific data repositories, which explicitly support networking opportunities (see Box 1). But collaborations and networks tend to form

---

**Box 1 | Types and examples of data depositories.**

**Generic data depositories**, such as PANGAEA (www.pangaea.de) or DRYAD (http://datadryad.org), compile data sets from a wide range of scientific domains. They guarantee long-term availability of contributed data sets, can ensure the presence of appropriate metadata to some extent and provide an opportunity to make data widely visible and accessible.

**Domain-specific data repositories**, such as the FLUXNET (http://fluxnet.ornl.gov) database for micro-meteorological eddy-covariance measurements or the TRY (www.try-db.org) database for plant traits, cover a narrower range of data, but in turn offer more intense data curation and networking opportunities. For example, FLUXNET has developed standardized data curation workflows allowing globally integrated analyses of ecosystem-atmosphere exchange across all measurement sites; the TRY database facilitates outlier detection, duplicate identification and gap-filling of missing data; GBIF, the Global Biodiversity Facility (www.gbif.org) compiles occurrence data for all kinds of species and has developed highly efficient algorithms to identify and potentially correct mistakes in geo-locations.
Open code for open science?

Steve M. Easterbrook

Open source software is often seen as a path to reproducibility in computational science. In practice there are many obstacles, even when the code is freely available, but open source policies should at least lead to better quality code.

Poor code quality is endemic, and not just in scientific computation. It is always tempting to build something ‘quick and dirty’, under the assumption that it can be cleaned up later. This is especially true at the cutting edge of a field — why invest time writing beautifully engineered code from the outset, if you’re not sure that what you’re trying to do is even possible?

In software engineering, this is known as technical debt: by deferring issues such as code readability and maintainability, a debt is created that someone in the future might have to pay, in the extra effort needed to re-run or modify the code. The point of the metaphor is not that debt is bad per se. After all, we frequently incur debt to obtain something of immediate value, for example, using a mortgage to buy a house. The point is that such debts have to be managed carefully, to prevent them spiralling out of control.

Open source policies in scholarly journals can help here. If journals ask for open code, they create a strong incentive for authors to clean up the code each time a paper is produced, rather than deferring such tasks indefinitely. As a second order effect, such policies should encourage more scientists to take the opportunity to improve their software-building skills, through courses such as Software Carpentry (http://software-carpentry.org/).