Open Research Data (ORD) and data sharing: a scientist’s perspective

Florian Altermatt
My research

Modeling/ synthesis

Laboratory experiments

Outdoor experiments

Field studies (e.g., monitoring)

Realism

Causality
Motivation/reasons for data sharing
Motivation/reasons for data sharing

An open mind on open data

The move to make scientific findings transparent can be a major boon to research, but it can be tricky to embrace the change.

By Virginia Gewin

In a movement building steady momentum to make research data, software code, and experimental methods publicly available and transparent, a spirit of openness is gaining traction in the scientific community, and it is the only way, say advocates, to address a crisis in science whereby too few findings are successfully reproduced. Furthermore, they say, it is the best way for researchers to gather the range of observations that are necessary to speed up discoveries or to identify large-scale trends. The open-data shift poses a conundrum for junior researchers, who are carving out their niche. On the one hand, the drive to share is gathering official steam. Since 2013, global scientific bodies — including the European Commission, the US Office of Science and Technology Policy and the Global Research Council — have begun to back policies that support increased access to research.

On the other hand, scientists disagree about how much and when they should share data, and they debate whether sharing is more likely to accelerate science and make it more robust, or to introduce vulnerabilities and problems. As more journals and funders adopt data-sharing requirements, and as a growing number of enthusiasts call for more openness, junior researchers must find their place between adopters and those who continue to hold out, even as they strive to launch their own careers.

One key challenge facing young scientists is how to be open without becoming scientifically vulnerable. They must determine the risk of jeopardizing a job offer or a collaboration proposal from those who are wary of — or unfamiliar with — open science. And they must learn how to capitalize on the movement’s benefits, such as opportunities for more citations and a way to build a reputation.
Motivation/reasons for data sharing

- Open science
- Reproducibility
- Visibility
- Credibility
- Role model for my students and postdocs
- Teaching
- Journal’s requirement
- Funding Agency’s requirement

“spirit of openness”

internal forcing

external forcing
Types of data shared – Matrices

Interaction data (e.g., plant-insect interactions)

Simple matrices or data frames

Usually as .csv or .txt files plus metadata

Pearse & Altermatt 2013  *Ecology Letters*
Carrara et al. 2012  *PNAS*
Types of data shared – Videos

- Videos too big to be shared.
- Simple matrices or data frames of processed data.
- Usually as .csv or .txt files plus metadata

Fronhofer & Altermatt 2015 *Nature Communications*
Pennekamp et al. 2018 *Nature*
Types of data shared – Code, protocols, sequence data

- Code
- Lab protocols
- Sequence data

Specific file formats

Usually shared on existing databases (GitHub, GenBank, NCBI etc.) or in supplements of papers

Altermatt et al. 2015 *Methods in Ecology and Evolution*
Deiner et al. 2017 *Nature Communications*
Possibilities/Opportunities of ORD data sharing

- BioRxiv
- Dryad
- GitHub

- April 2012: Nature Precedings folds after posting nearly 2300 manuscripts.
- April 2013: PeerJ Preprints started by for-profit publisher.
- November 2013: BioRxiv launched by Cold Spring Harbor Laboratory.
- February 2016: ASAPbio holds first meeting to promote preprints in life sciences.
- Early 2017: U.K. Medical Research Council, Wellcome Trust, HHMI, NIH endorse preprints in proposals.

Kaiser 2017 Science
Dryad (www.dryad.org) – a data repository

Dryad as a generic data repository:

> community consensus (many use it)

> searchable

> data get a DOI
GitHub (www.github.com) – a developer/sharing platform

Learn Git and GitHub without any code!
Using the Hello World guide, you’ll create a repository, start a branch, write comments, and open a pull request.

Read the guide
Start a project

Discover interesting projects and people to populate your personal news feed.
Your news feed helps you keep up with recent activity on repositories you watch and people you follow.

Explore GitHub
Advantages of ORD data sharing

- Forces lab members to structure datasets
- Visibility
- Track record / credits for data use
- Availability for meta-analyses
- Outsourcing the long-term storage of data
Time investment for data sharing

1–2 hours of uploading

versus

virtually endless

Costs of ORD data sharing

0–120 CHF

versus

permanent part time data scientist working for the group
Challenges/obstacles of ORD data sharing

- Big data (videos with >10 TB per experiment)
- Time-series / incremental datasets, when to publish?
- How to handle “re-analyses”
- Costs of producing data vs. benefits of sharing them
- Scooping
- “Sensitive” data (e.g., monitoring programs), 3rd party data
- Short-term protection of early career scientists (PhD students, Postdocs)
Do we always do ORD data sharing? -> No…

- Worry of being scooped

- Long term datasets -> how is investment rewarded?

- Worry of miss-use

- Only partial rights on data (e.g., to use, but not to publish)

- Other priorities / laziness…
Thanks!  Questions?

Roman Alther, Francesco Carrara, Kristy Deiner, Emanuel Fronhofer, Pravin Ganesanandamoorthy, Andrea Giometto, Isabelle Gounand, Eric Harvey, Katharina Kaelin, Chelsea Little, Elvira Mächler, Martina Ramel, Fabienne Santschi, Mat Seymour

Funding: