Showing our working

Transparency for trustworthy science

Richard J. Acton

2023-04-03

## Why show our working?

We all remember being told to show our working in school, usually in maths 🧮. Why were we asked to do this? So that someone else can follow our reasoning, step-by-step, and see (with a little effort) for themselves if they found our reasons sound, or at least partially so. This is one of the fundamental motivating factors for open science, sharing our data and methods so that others can assess our conclusions for themselves with all of the same information. This is core to the corrective mechanism that drives the scientific progress, you can’t progress if you can’t spot mistakes, gaps, and misunderstandings; importantly, you can’t spot these if you can’t see our working. Ideally anyone can ask: “how do you know what you think you know?” and we can provide a detailed and compelling answer that anyone can challenge and, with some effort, check for themselves. Trust in our conclusions is, rightly, derived from the transparency and accountability of our processes. This applies both within the scientific community and to our relationship with the public.

The number and complexity of the steps that take us from our starting point to our conclusions in modern science has grown as the depth of our understanding of the world has increased. This has made it harder, as a practical matter, to show all of our working from start to finish. Doing so however is no less important now than it was ever been, if anything the complexity of modern science makes it more important than ever. The length of the story that now has to be told to get from basic assumptions to conclusions is quite long. In many cases there is also a great deal of context needed to understand some of the questions we now tackle. This can present a significant communication challenge when interacting with the public and even specialists in other disciplines. The effort necessary to asses the strength of others’ conclusions has risen, this makes the importance of explanatory clarity greater than ever. One of the factors which makes completeness of description challenging is that It is rarely one person, or even one research group, that is responsible for the full chain of steps that produce a modern research paper, especially ‘prestigious’ ones with lots experiments often making use of varied methods. Thus, there is no single person with full insight into the granular details of every experiment and method used in many modern papers. Consequently, ensuring that every detail needed for reproducible work can be a significant coordination challenge among co-authors.

So how are we, the scientific community doing at this task of showing our work? Unfortunately not as well as you might hope.

## Are we any good at showing our working? How hard can it be?

To start with, across disciplines our work is getting harder to read. It has become laden with ‘science-ese’ or general scientific jargon, or so conclude Plaven-Sigray and co-authors in a 2017 paper in eLIFE “The readability of scientific texts is decreasing over time” (Plavén-Sigray et al. 2017). This does not help the general accessibility of our work to colleagues, students, science journalists or the public. Nor does it appear to driven by specific technical jargon which is a useful communication shorthand, it’s apparently mostly the addition of seemingly superfluous polysyllabic obfuscationalisms presumably so that we can show-off our erudition 😉.

Richard F. Harris’s 2017 book “Rigor mortis: how sloppy science creates worthless cures, crushes hope, and wastes billions” (Harris 2017) & Stuart Ritchie’s 2020 book “Science fictions: exposing fraud, bias, negligence and hype in science” (Ritchie 2020) called popular attention to issues of reproducibility especially in the life sciences. In 2021 a series of papers was published by the ‘[Reproducibility Project: Cancer Biology](https://www.cos.io/rpcb)’ summarising the results of efforts spanning 8 years to reproduce the findings of 193 experiments from 53 prominent papers in field of cancer biology. Their results were not particularly reassuring. **0 of the 193 experiments were described in sufficient detail for the project team to design protocols to repeat them** 😅, yes none of them could be repeated without additional information from the original researchers. Of the 193 experiments the researchers were eventually able to repeat 50, where they were able to get some additional information from the original researchers, 32% of whom did not respond helpfully or at all to their inquiries. Between 40% and 80% of these 50 experiments, were successfully replicated, depending on how the criteria for successful replication were applied. Reproducing biological work is genuinely a difficult task, lab work requires significant skill that is often hard to fully codify in protocols. Biology has a lot of inherent variability even when you are going to great lengths to get the same starting conditions there are sometimes factors that it is difficult to control. Analysing your data though, reproducing that should be easy right 😎? It’s all on the computer you can just run the same thing again right 🥺? Alas it is not as simple as you might think 🤦. It can take several months to reproduce a bioinformatic analysis in a published paper if it is possible at all, see: “Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome” (Garijo et al. 2013). (Significant strides have been made in reproducible analysis since that paper was written but I’ve not found a more recent attempt to quantify the problem.)

Why is this the case? What makes reproducibility hard and what can we do to make it better?

## How can we do better?

As the data outputs manager for the [Human Developmental Biology Initiative (HDBI)](https://www.hdbi.org) It’s my role to make it easier for the other scientists in HDBI to make their data available to others, and easier for them to make both their experiments and analyses reproducible.

### What’s needed for data & methods to be open and reproducible?

If you begin looking into the area of reproducibility you will, before too long, encounter the somewhat nebulous term ‘metadata’. *Meta* is from the Greek meaning something like “higher” or “beyond”, metadata, therefore is data about data. So what sorts of things are metadata? Who generated it? When? With what machine? With what settings on that machine? Why? What properties did the samples measured have? Species, type(s) of cell, stage of development? What where the sample preparation steps? The answers to all these questions and more can be considered metadata. This abundance of possible properties that could be recorded leads to the following questions:

* How do we decide which pieces of metadata we need for a given experiment?
* How do we store and organise them nicely so people and computers can read them?

The answers to these questions are complex and context dependent but in general scientific communities have to get together and decide for themselves what they need to know and how best to represent that information by creating community standards that they can agree to adhere to.

This has varying degrees of success…. 🤷 

Fortunately, There are some general principles that we can adhere to when designing domain specific metadata standards and things it is sensible for everyone to consider when sharing their data. These are the principles of ‘linked data’ where we can used the ‘resource description framework’ to devise a suitable representation of our metadata. This is not always the most practical format of metadata for day-to-day use but if you can automatically translate your working format to a linked data format then others can use a shared set of tools to combine data from your domain with data from theirs. In this way components common to different domains can be re-used e.g. an agreed set of standard terms to refer to certain classes of things like cell-types or sequencing technologies so that we don’t use different words for the same thing. For more on metadata standards checkout [the section in my book](https://hdbi.gitlab.io/data-management/hdbi-data-resource/01-what-constitutes-data.html#sec-metadata).

One of the ways of thinking about making our data available for others to use, so that they can not only use it in their own work, but also check existing conclusions is based around the acronym FAIR:

#### FAIR (Findable, Accessible, Interoperable, Re-usable)

* Findable
	+ Has a unique identifier that can be looked up in a database, plus some associated terms so you can find the id with a search.
* Accessible
	+ If I’ve got the ID I can download a copy, or figure out who to ask for permission to download a copy if there are e.g. privacy restrictions.
* Interoperable
	+ It’s in a file format I can read (without expensive proprietary software).
	+ It’s described in standard terminology so I can easily search for it and connect it with data about the same/related things.
* Re-usable
	+ Where is came from/how it was generated and other attributes are well documented according to community standards.
	+ It’s licensed so it can be used.

Reproducing lab work 👨‍🔬 and computational work 👨‍💻 analogises quite well to cooking 👨‍🍳, a recipe is comprised of a list of ingredients (Data, Materials & Reagents), a set of steps to follow (Code, Protocols), and descriptions of the environment in which the food is cooked. e.g. The type and temperature of the oven (Compute environment, Lab environment) and additional information that helps me find, contextualise and appropriately use the recipe (metadata): Some recipes are overly vague and some highly specific, this might depend on the difficultly and stakes of getting a tasty or at least edible result. Science can be like high stakes cooking (Think, a meal for diplomats from two countries with a tense relationship who also happen to be restaurant critics with a bunch of mutually incompatible food allergies and religious dietary restrictions) so the recipes have to be good, *really good* 😅.

Here’s a table fleshing out the analogy with some small examples of the sorts of information that can fall into these categories as they apply to these different disciplines:

Cooking 🧑‍🍳

Lab Work 🧑‍🔬

Computational Analysis 🧑‍💻

Inputs

Ingredients

❌ 7oz Flour (vague)

✅ 200g Plain Wheat Flour (all-purpose/550/55/0)

Materials & Reagents

❌ HeLa cells

✅ UKBi001-A

Data

❌ Human Genome

✅ Homo sapiens (NCBI:txid9606) genome (Ensembl 109, GRCh38.p13)

Process

Cooking Instructions

❌ Bake at medium temperature until bronzed (vague)

✅ Bake at 190C for 35mins if using a fan oven

Protocols

❌ Gel electrophoresis (vague)

✅ Agarose gel electrophoresis (0.5%) TBE buffer, 120V, ethidium bromide, 10kb ladder from…

Code

❌ random-script-from-email.R

✅ git repository

Environment

Kitchen Conditions

Ambient temperature, pressure, and humidity (What’s the boiling point of water in your kitchen?)

Lab Environment

Ambient temperature, pressure, and humidity (What’s the boiling point of water in your kitchen Lab?)

How you wash you glassware (no really this has affected the reproducbility of experiments)

Compute Environment

OS 🐧/🪟/🍎

🩹 R v4.2.0

✅ Environment management tools & Containers renv.lock

Context

Discernment

History, Culture, & Origins of a Dish 🌍

Allergens 😵

Appropriate pairings 🍷

Metadata

❌ Who: Steve, When: Tuesday (vague)

✅ Who: Steven Stickler (ORCID: 0000-1234-1234-1234), When: 2023-03-30 15:34 (UTC+0)

Metadata (same as Lab +)

Who: PGP fingerprint: 96C2 0929 FA88 DD89 9270

When: commit hash: 954086fdf13d8…

When a protocol can’t quite capture your bench work well enough that someone else could do your experiment if they read it, then you can take the approach of [JoVE](https://www.jove.com/) (The Journal of Visualised Experiments). If that’s a bit much the less formal approach is available, anyone with a smartphone or action camera can film their experimental work upload it to figshare and get a DOI to reference in a protocol on [protocols.io](https://www.protocols.io/) or in a paper. (Enlist the help of a student an alarming fraction of them are surprisingly capable videographers thanks to social media.)

What does a ‘computational environment’ mean? When doing an analysis you don’t re-implement all steps from scratch you use existing tools to perform many calculations, these in turn use other tools creating a ‘tree’ of ‘dependencies’. The way these tools work can change if the software gets updates so to re-run your analysis exactly I need to know not just the steps that you took but the versions of the tools you were using and the versions of the tools your tools where using, and so on. It’s tools all the way down. Fortunately there are tools for taking inventory of all the versions of all the tools that you’ve used, sharing this list and even re-creating the same computational environment from these inventories. Checkout [the section in my book](https://hdbi.gitlab.io/data-management/hdbi-data-resource/04-working-with-data.html#sec-reproducible-computational-analysis) to learn more about this.

## How can we encourage the adoption of these practices?

So why are we not working more reproducibly already? It’s quite hard to do in certain cases often because tooling and automations have not caught up to make it easier. It’s also not yet a norm to which we expect one-another to conform in the scientific community, either when we review others’ work or when we have our own reviewed. In his article Five selfish reasons to work reproducibly (Markowetz 2015) Florian Markowetz lays out some excellent reasons to get ahead of the curve on working reproducibly.

The laboriousness of recording and providing this level of detail can be a major impediment to researchers actually sharing their processes if they don’t feel that doing so will be time well spent. So we should ask what can be automated, what tools, practices and standard procedures can scientists adopt to make the provision of sufficiently detailed structured information a part of their workflow that does not get in their way, and if anything makes their lives easier?

### Crafting a ‘pit of sucess’

Make FAIR data and reproducibility the default and the expectation in every area, such that if you ‘go with the flow’ your work will be FAIR and reproducible. Not everyone will have the time, inclination or incentive to strive relentlessly towards a ‘pinnacle of excellence’ so raise to floor not the ceiling and construct a ‘pit of success’ into which we can all trip without trying too hard. We can do our best to make it easy & useful to do so but this may not be quite enough to get up over the hump in all cases. So in some cases we may have to take a slightly more “Nice paper/grant/dataset you’ve got there ’be a shame if you couldn’t publish/fund/generate it, unless…” approach.

Here is my advice for people in various roles for using a mixture of carrots 🥕 and sticks📏 to improve FAIR data practices and working reproducibility.

* Core facility Staff
	+ If you run/work in Microscopy, flow cytometry, bioinformatics, proteomics, sequencing, etc. facilities develop policies which make good metadata annotations a condition of researchers using your facilities and getting access to the data once it has been generated. Make it as frictionless as possible so as not to put them off, and showcase how useful it is to be able to call on well structured and anotated datasets.
* Human Resources
	+ Make proper data stewardship part of the on-boarding and leaving process. You shouldn’t be able to clear HR when you are leaving if your data is not in a suitable state to hand over to others.
* Software Developers & IT Staff
	+ Build tools and systems which make the ingestion, annotation, sharing, accessing, and processing of data as intuitive, seamless and integrated a process as possible with open-source tools.
	+ (Don’t develop a proprietary platform or product to try and solve these issue or people like me will tell others not to use it as the incentives don’t line up with the degree of interoperability and portability needed in science. Go with an open-source business model that is compatible with the openness required by scientific process.)
	+ Package your software so that it can easily be included in portable reproducible environments like Docker simply and declaratively.
* Peer Reviewers
	+ When you review things ask questions about reproducibility, and FAIR data. This is where the expectation of higher standards in this area can begin to be set. (Also if you get asked to review stuff a lot and would like to reduce incoming requests suddenly becoming a reproducibility nut might lighten your load 😉). If you are asking these questions in your reviews of papers it may clue editors in that academics now expect this and shape their decisions on what to put out for review in the first place.
* Journal Editors
	+ If you are a journal editor and you are deciding between many good submissions make this a criterion for what you choose.
* Grant Reviewers
	+ If you review grants ask about peoples plans for making their data FAIR and their analyses reproducible. This will get them thinking about it well in advance and hopefully planning for it, especially if they think it might be the difference between getting funded or not.
* Press
	+ If you are a member of the press or a populariser of science, report favourably on publications that show their work and skeptically of those that don’t.
	+ Ask questions about reproducibility and openness when interviewing scientist and push university PR departments about these qualities in the papers they choose to make press releases about.
* Public *(also applies if you fall into any of the other categories)*
	+ Ask your elected representative or the relevant minister (At time of writing [Michelle Donelan](https://www.gov.uk/government/people/michelle-donelan) Secretary of State for Science, Innovation and Technology) why the research councils aren’t holding their grant awardees to a higher standard on reproducibility and FAIR data so that the best use can be made of public research funds?

## Where Can I Start / Learn More?

I’ve written a short ebook as a resource for HDBI members “[Data: Inception to Publication & Beyond](https://hdbi.gitlab.io/data-management/hdbi-data-resource/)” this is directed at a more technical audience but aims to be written in an accessible style. It features many links to external resources to learn more about a given topic in various media. It’s a living document and I’m working on some new material for it, I’m always looking for feedback, comments and suggestions for improvement from any readers.

To get a sense of what it covers here’s a table of contents:

1. What Constitutes Data?
2. When Should I Generate Data?
3. How to Store Your Data
4. Working With Data
5. When to Publish Data
6. Where to Publish Data
7. What Data to Publish
8. How to License Your Data
9. How to Manage References

Garijo, Daniel, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E. Bourne, and Yolanda Gil. 2013. “Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome.” Edited by Christos A. Ouzounis. *PLoS ONE* 8 (11): e80278. <https://doi.org/10.1371/journal.pone.0080278>.

Harris, Richard F. 2017. *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions*. New York: Basic Books.

Markowetz, Florian. 2015. “Five Selfish Reasons to Work Reproducibly.” *Genome Biology* 16 (1). <https://doi.org/10.1186/s13059-015-0850-7>.

Plavén-Sigray, Pontus, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. “The Readability of Scientific Texts Is Decreasing over Time.” *eLife* 6 (September). <https://doi.org/10.7554/elife.27725>.

Ritchie, Stuart. 2020. *Science Fictions: Exposing Fraud, Bias, Negligence and Hype in Science*. London: The Bodley Head.