Open, Reliable and Transparent Data

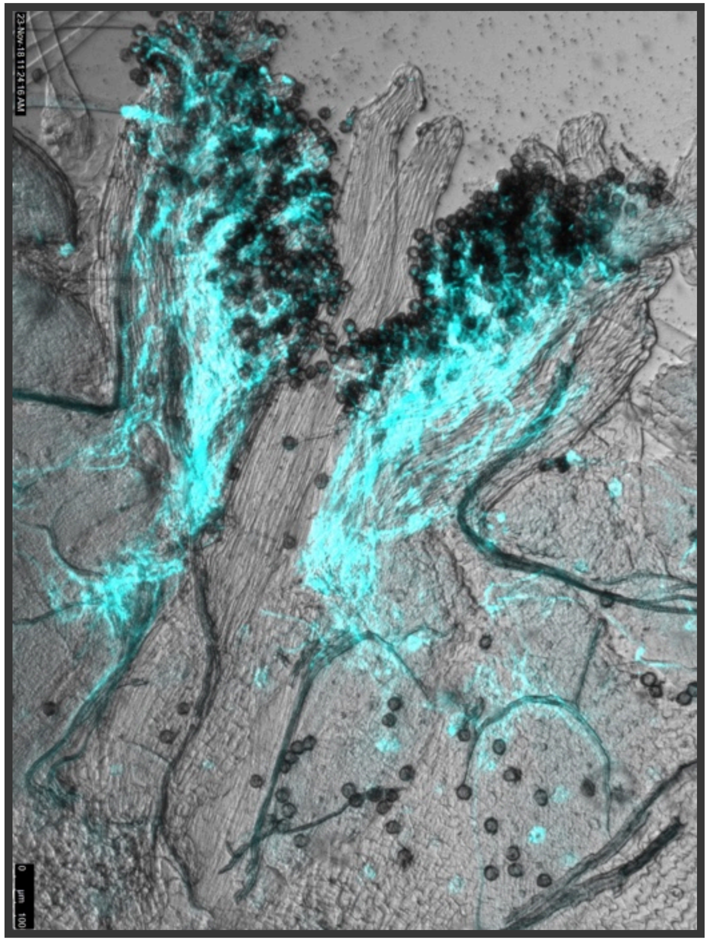
Iain R. Moodie

Stockholm Mini-symposium

2024-02-28

A brief annecdote

Sexual selection in plants



Pollen tubes interacting with pistil tissues - Jeanne Tonnabel

- Bateman gradients in angiosperms
  - N = 2 (in 2021)
- Project goal
  - Conduct a meta-analysis 🤔
- Find datasets that could be re-analysed in this new context
- Combine into a meta-analysis to test predictions

Sexual selection in plants

- Initial search
  - N=2167 😊
- After sorting
  - N=30 🥹
- After trying to source data
  - N=9 😐

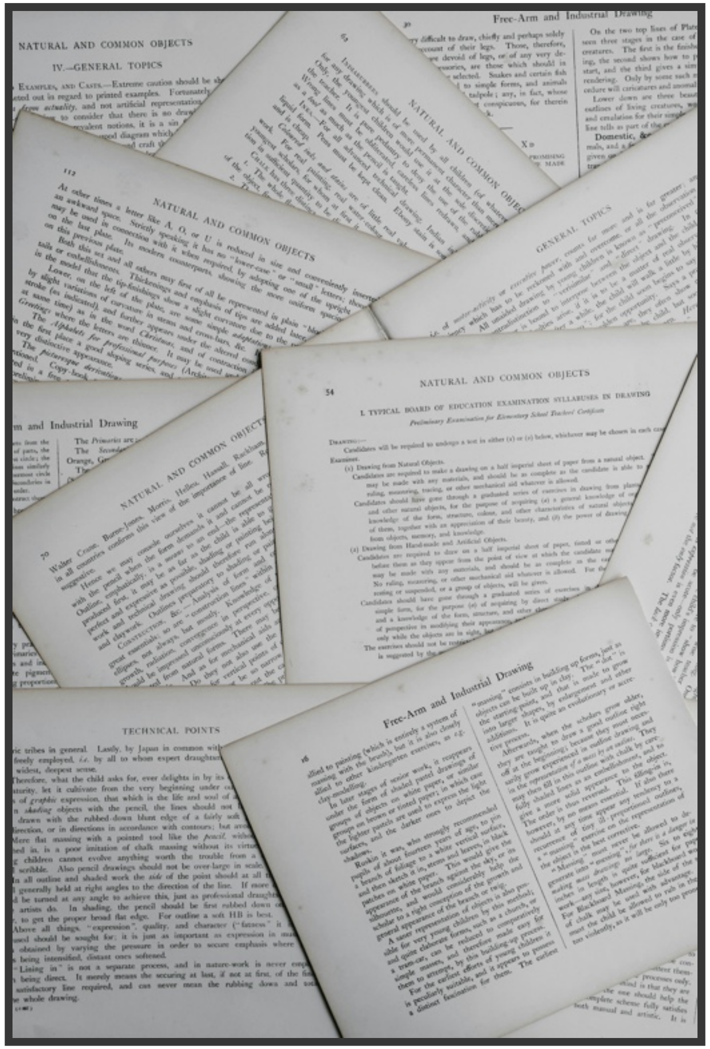


Photo by Annie Spratt

Datasets we couldn’t use

- Data not archived
  - No way to contact author
  - No response to contact
  - Data had been lost
  - Not willing to share data

- Data archived
  - Inaccessible
  - Incomplete
  - Incomprehensible

Lost from science

Exxon Valdez oil spill 1989

- 40.8 million litres of crude oil spilled
- Settlement funds from Exxon used for research and monitoring the impacts of the spill
- Between 1989 and 2010, 419 projects were funded
- In 2012, NCEAS tried to compile all historic datasets
- **70% were unrecoverable**

Lost from science

Transparency in research

Opaque research

- Publication bias
  - Not all research is published
- Incomplete or insufficiently detailed methods
- Selective reporting in results
  - Confirmation bias
  - “HARKing”
  - “P-hacking”
- Unaccessible underlying data

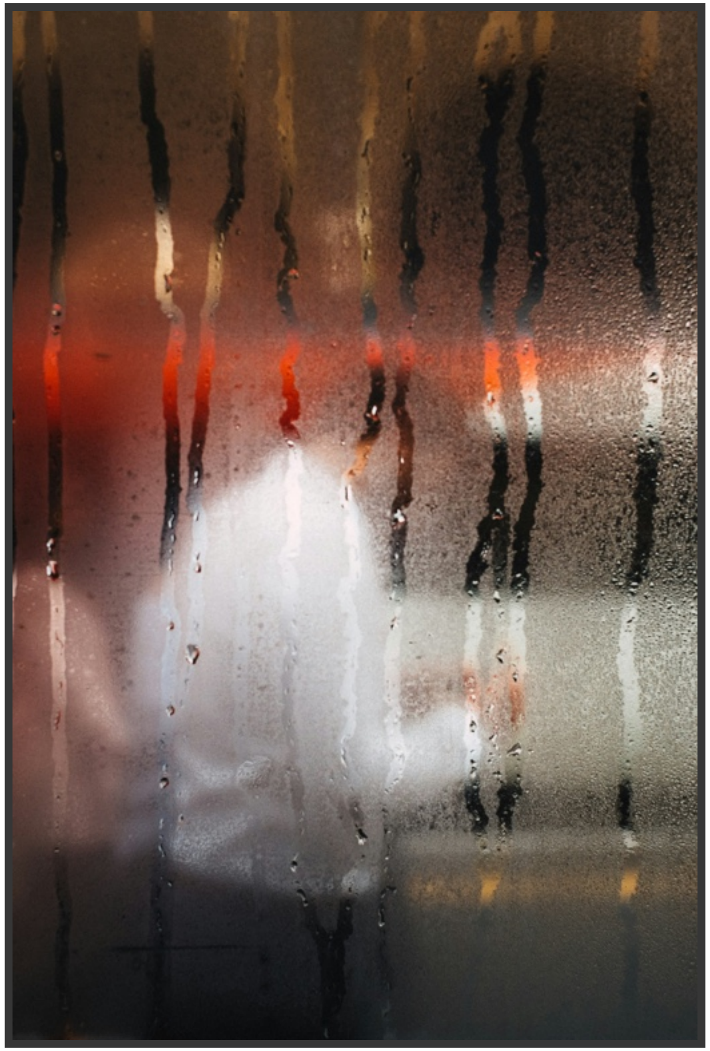


Photo by Clem Onojeghuo

Opaque research limits science

- Harder to replicate or re-use methods
- Harder to build upon to progress the field
- Harder to interpret results
- Harder to trust the conclusions

Photo by Karl Hedin

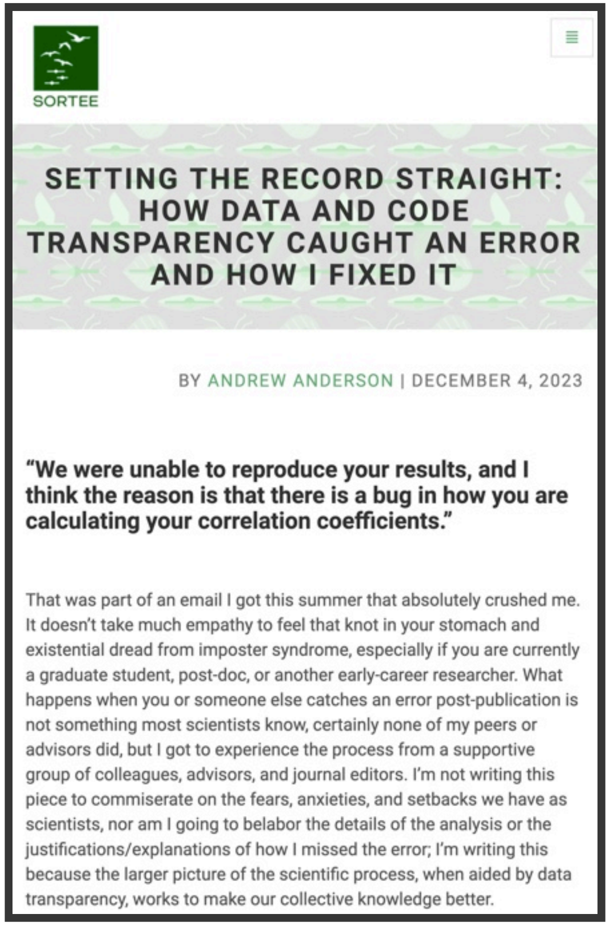
Open, Reliable and Transparent Science

Open, Reliable and Transparent Data

And why you should care about it.

Reproducible and reliable results

- Promotes accountability and trust
- Mistakes can be corrected[1]
- Analytical decisions can be justified
- Scientific misconduct can't hide



SORTEE

SETTING THE RECORD STRAIGHT: HOW DATA AND CODE TRANSPARENCY CAUGHT AN ERROR AND HOW I FIXED IT

BY ANDREW ANDERSON | DECEMBER 4, 2023

"We were unable to reproduce your results, and I think the reason is that there is a bug in how you are calculating your correlation coefficients."

That was part of an email I got this summer that absolutely crushed me. It doesn't take much empathy to feel that knot in your stomach and existential dread from imposter syndrome, especially if you are currently a graduate student, post-doc, or another early-career researcher. What happens when you or someone else catches an error post-publication is not something most scientists know, certainly none of my peers or advisors did, but I got to experience the process from a supportive group of colleagues, advisors, and journal editors. I'm not writing this piece to commiserate on the fears, anxieties, and setbacks we have as scientists, nor am I going to belabor the details of the analysis or the justifications/explanations of how I missed the error; I'm writing this because the larger picture of the scientific process, when aided by data transparency, works to make our collective knowledge better.

New questions & new methods

Photo by Monika Manenti

- Built upon more effectively
  - Deeper understanding of data & analysis
  - Used to develop new tools/methods/protocols
  - E.g. Bumpus 1899
- Viewed in a new light
  - Beyond the original paper
  - Paradigm shifts
- Analysed using the latest methods
- Meta analysis

More accurate meta-analysis

- Easy extraction of accurate data
    - No need to extract from figure
    - Reduces ambiguity and error
- Go beyond the results section
    - Helps reduce bias from selective reporting
    - Capture the full picture of the study
- Extends the life the dataset
    - Can always be accessed

Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8**, 777–784 doi: 10.1111/2041-210X.12758

COMMENTARY

Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting

Katharina Gerstner*[,1,2], David Moreno-Mateos[3,4], Jessica Gurevitch[5], Michael Beckmann[2], Stephan Kambach[6,7], Holly P. Jones[8] and Ralf Seppelt[2,9]

[1] iDiv – German Centre for Integrative Biodiversity Research, 04103 Leipzig, Germany; [2] Department Computational Landscape Ecology, UFZ – Helmholtz Centre for Environmental Research, 04318 Leipzig, Germany; [3] Basque Centre for Climate Change – BC3, 48940 Leioa, Basque Country, Spain; [4] IKERBASQUE, Basque Foundation for Science, 48008 Bilbao, Basque Country, Spain; [5] Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, USA; [6] Institute for Biology, Martin-Luther-University Halle-Wittenberg, 06099 Halle (Saale), Germany; [7] Department Community Ecology, UFZ – Helmholtz Centre for Environmental Research, 06120 Halle (Saale), Germany; [8] Department of Biological Sciences and Institute for the Study of the Environment, Sustainability, and Energy, Northern Illinois University, DeKalb, IL 60115, USA; and [9] Institute of Geoscience & Geography, Martin-Luther-University Halle-Wittenberg, 06099 Halle (Saale), Germany

Summary

1. Ecological and evolutionary research increasingly uses quantitative synthesis of primary research studies (meta-analysis) for answering fundamental questions, informing environmental policy and summarizing results for decision makers.

2. Knowing how meta-analysis works is important for researchers so that their research can have broader impact. Meta-analytic thinking encourages scientists to see single primary research studies as substantial contributions to a larger picture.

3. To facilitate inclusion in a meta-analysis, relevant primary research studies must be found and basic information about the methods and results must be thoroughly, clearly and transparently reported. While many published papers provide this information, it is common for essential data to be omitted, leading to study exclusion from meta-analyses.

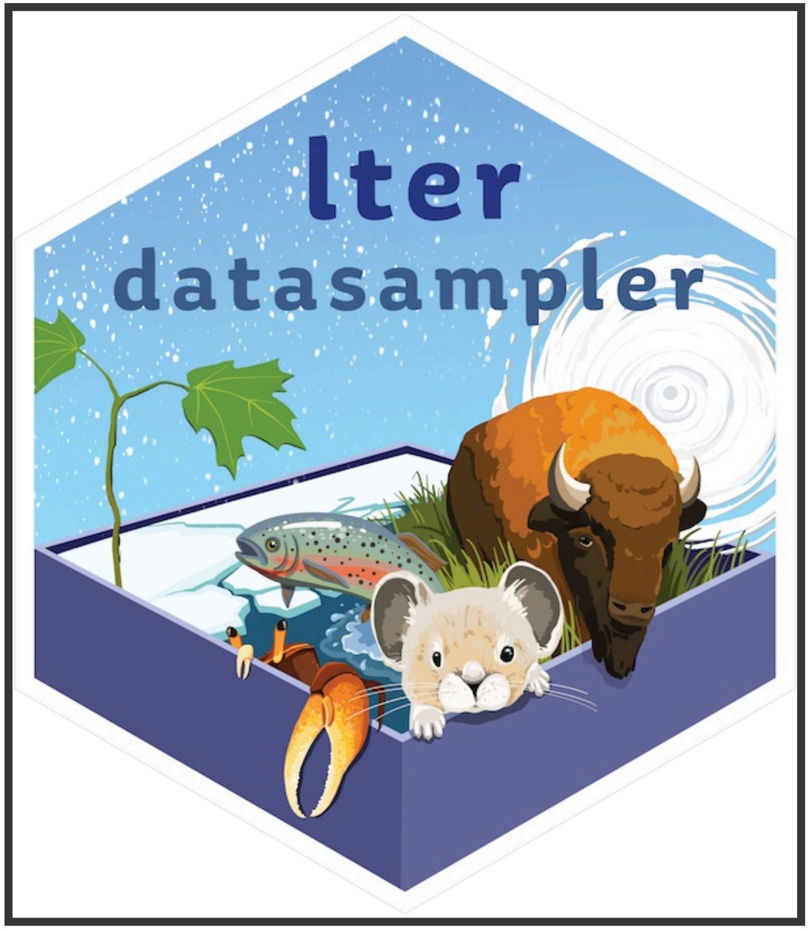
4. We provide guidelines for correctly reporting basic data needed from primary studies in ecology and evolutionary biology so that they can be included in meta-analyses, together with examples that show how data should be reported to enable calculation and analysis of effect sizes, and how data should be made accessible.

5. These guidelines are important for reporting research results in general, whether or not results are included in subsequent meta-analyses, because they are necessary for the interpretation and assessment of study outcomes. Increased implementation of these guidelines by authors, editors and publishers, and reinforcement by funders, will foster higher quality and more inclusive syntheses, further the goals of transparency and reproducibility in science, and improve the quality and value of primary research studies.

Key-words: data reporting, meta-analysis, open science, quantitative review, reproducibility, research synthesis, transparency

Gerstner et al. (2017) *Methods in Ecology and Evolution*

Learning and teaching



Long Term Ecological Research program (LTER) Datasets

- Teaching students using example datasets
  - Real biological “quirks”
  - Real scenarios
  - Can teach good practises from the start
- Learning and understanding new methods
  - Complexity can be broken down
  - Walkthrough when code also available

Benefits for the data archiver

- Increased exposure, reach, and trustworthiness
- Citation advantage (+25%) [1]
- Your own best collaborator
  - Data is clean and ready to use
  - Well annotated
  - Cannot be lost

Photo by Anton

Reducing research loss & waste

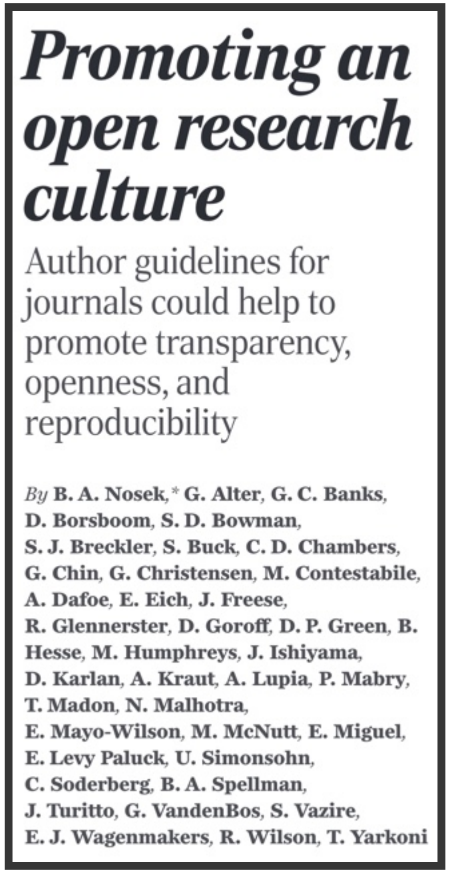
Roche et al. (2015) *PLOS Biology*

- Removes need for duplicated data collection effort
  - Time/location/event dependant data
  - Research animal use
- Reduces cost of research

How are things going?

Transparency and Openness Promotion (TOP) guidelines

- “A set of standards applied to journals to measure their alignment with open scientific principles”
    - Specific guidance on data transparency:
        - Level 3: open data + peer review of dataset and analysis
        - Level 2: open data in trusted repository
        - Level 1: mandatory data statement
- >5000 journals are signatories
- Field specific advice for ecology and evolution



Promoting an open research culture

Author guidelines for journals could help to promote transparency, openness, and reproducibility

By **B. A. Nosek,* G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni**

Nosek et al. 2015

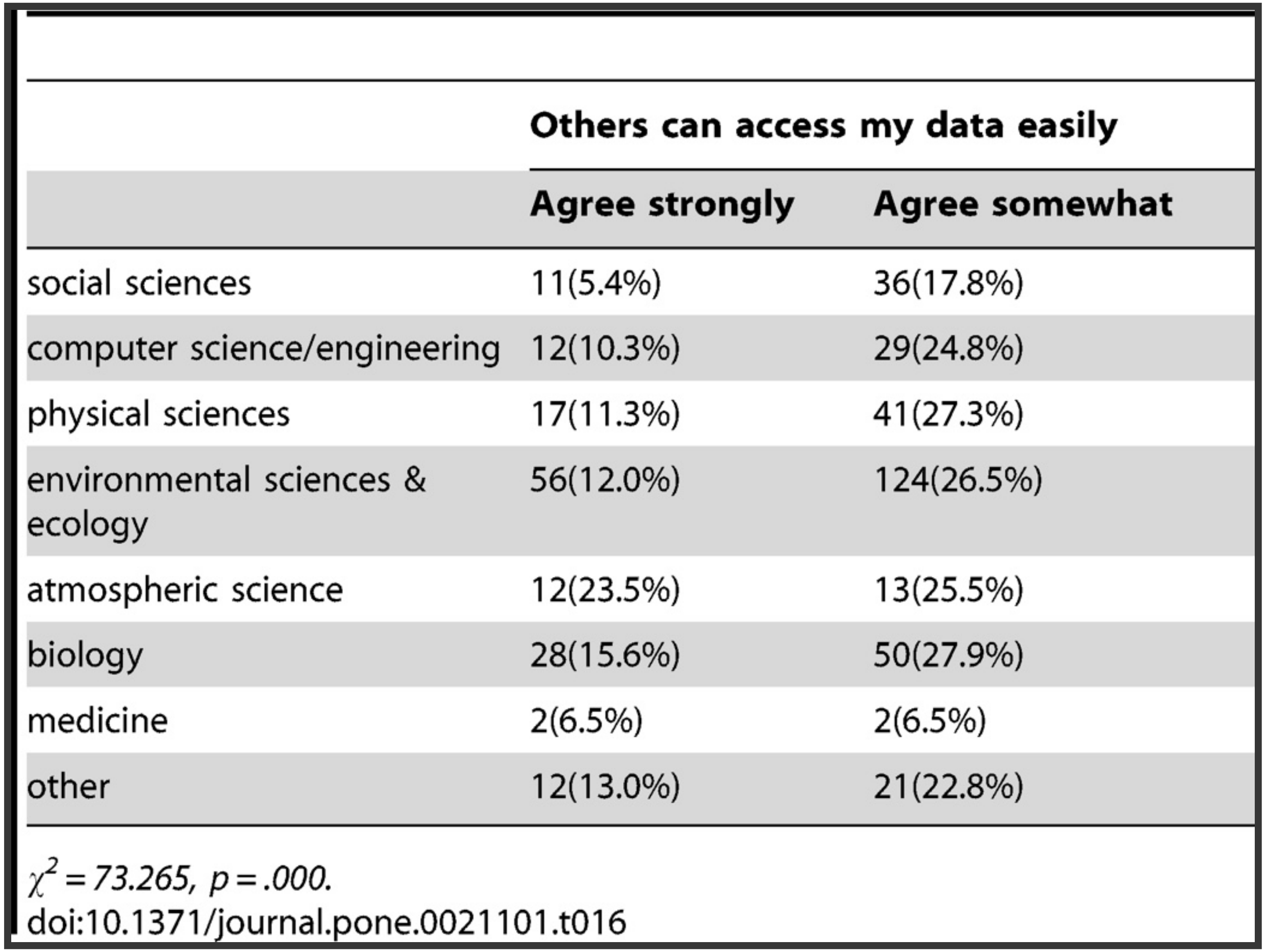
Top down pressure

- Journals
  - Mandated archiving has become “the norm”
- Funding sources
  - Open access requirements extending to datasets
- Institutions
  - To help staff meet requirements of the above

Community driven approaches

- Positive attitudes towards data transparency are common
  - 95% of scientists in ecology and evolution think that data should be publically archived (Whitlock et al. 2010)
- Lack of data transparency is seen as a problem
  - 67% of scientists think that lack of access to data is a major impediemnt to progress in science (Tenopir et al. 2011)

How well are we doing?



	Others can access my data easily	
	Agree strongly	**Agree somewhat**
social sciences	11(5.4%)	36(17.8%)
computer science/engineering	12(10.3%)	29(24.8%)
physical sciences	17(11.3%)	41(27.3%)
environmental sciences & ecology	56(12.0%)	124(26.5%)
atmospheric science	12(23.5%)	13(25.5%)
biology	28(15.6%)	50(27.9%)
medicine	2(6.5%)	2(6.5%)
other	12(13.0%)	21(22.8%)

$\chi^2 = 73.265$, $p = .000$.
doi:10.1371/journal.pone.0021101.t016

Tenopir et al. 2011

How well are we doing?

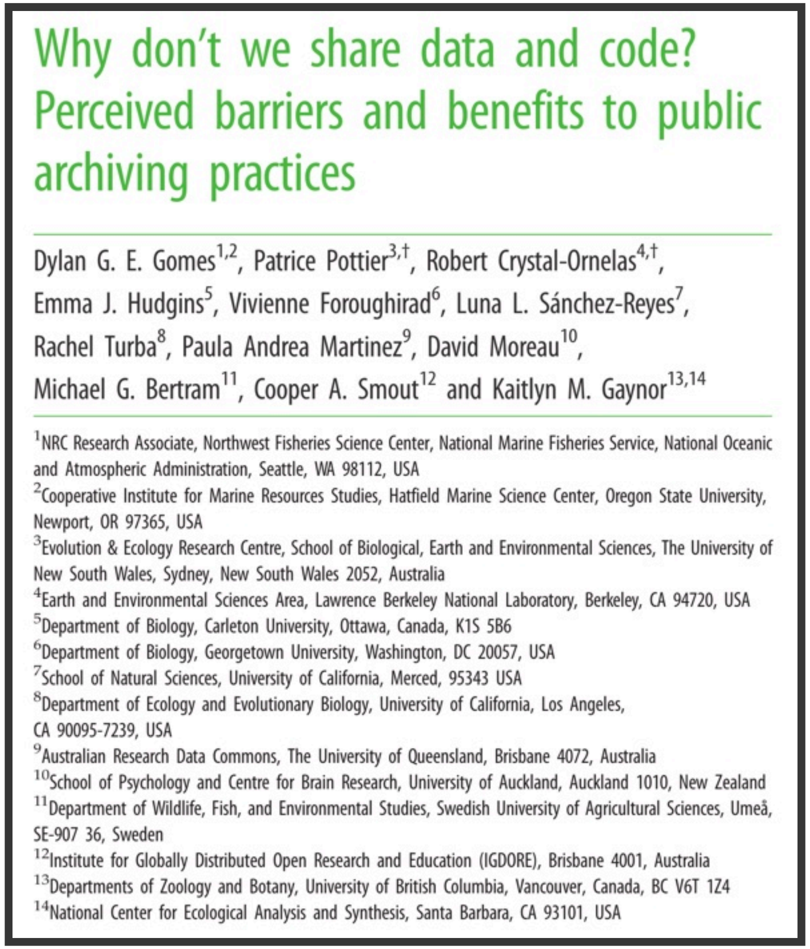
Published without sufficient data to replicate:

- 89% (N=18) of micro-array gene expression analyses (Ioannidis et al. 2009)
- 35% (N=19) of population genetic studies (Gilbert et al. 2012)
- 64% (N=100) of non-molecular eco/evo studies in journals that **mandate data archiving** (Roche et al. 2015)

Photo by Steven Wright

How do we improve things?

- Why we don’t share data?
  - Knowledge barriers
  - Re-use concerns
  - Disincentives
- How to work towards data transparency



Why don’t we share data and code? Perceived barriers and benefits to public archiving practices

Dylan G. E. Gomes[1,2], Patrice Pottier[3,†], Robert Crystal-Ornelas[4,†], Emma J. Hudgins[5], Vivienne Foroughirad[6], Luna L. Sánchez-Reyes[7], Rachel Turba[8], Paula Andrea Martinez[9], David Moreau[10], Michael G. Bertram[11], Cooper A. Smout[12] and Kaitlyn M. Gaynor[13,14]

[1]NRC Research Associate, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA 98112, USA
[2]Cooperative Institute for Marine Resources Studies, Hatfield Marine Science Center, Oregon State University, Newport, OR 97365, USA
[3]Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, The University of New South Wales, Sydney, New South Wales 2052, Australia
[4]Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[5]Department of Biology, Carleton University, Ottawa, Canada, K1S 5B6
[6]Department of Biology, Georgetown University, Washington, DC 20057, USA
[7]School of Natural Sciences, University of California, Merced, 95343 USA
[8]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095-7239, USA
[9]Australian Research Data Commons, The University of Queensland, Brisbane 4072, Australia
[10]School of Psychology and Centre for Brain Research, University of Auckland, Auckland 1010, New Zealand
[11]Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, Umeå, SE-907 36, Sweden
[12]Institute for Globally Distributed Open Research and Education (IGDORE), Brisbane 4001, Australia
[13]Departments of Zoology and Botany, University of British Columbia, Vancouver, Canada, BC V6T 1Z4
[14]National Center for Ecological Analysis and Synthesis, Santa Barbara, CA 93101, USA

Knowledge barriers

What’s the process?

- Do not know how to share data effectively
  - Which online data repository to use?
  - What format to share data in?

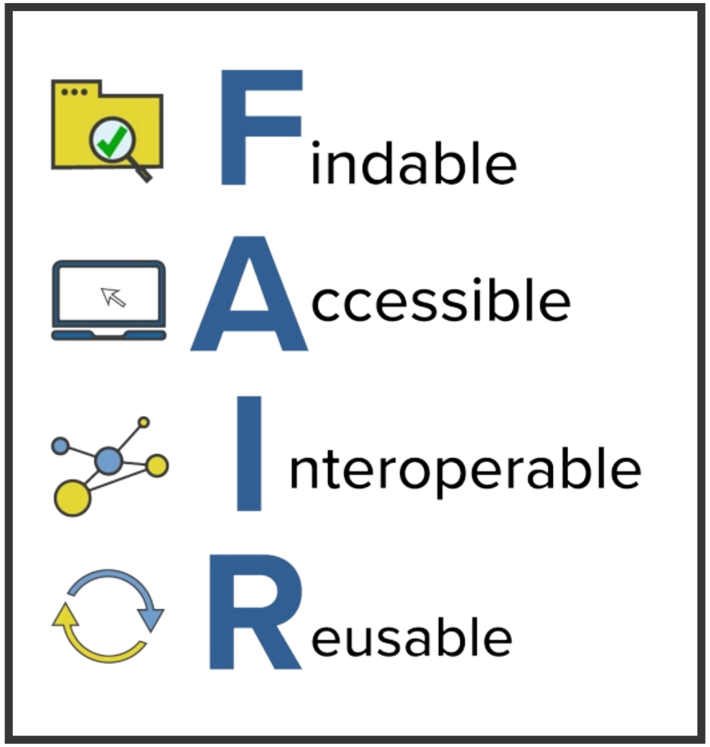
What’s the process?

- Online guides and primers
  - British Ecological Society “Guides to Better Science”
  - UKRN Primers
  - SORTEE (coming soon)

British Ecological Society Primer Series

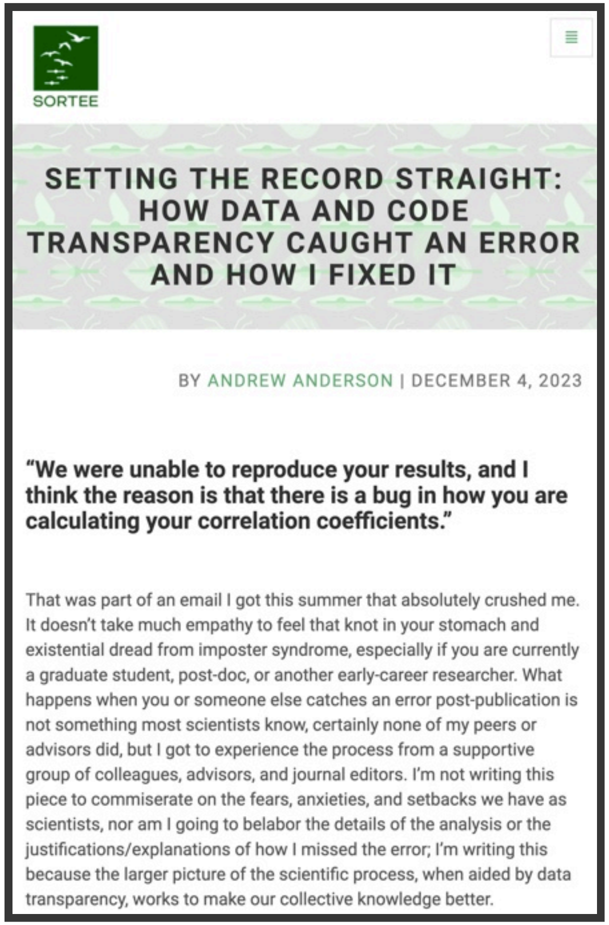
What’s the process?

- Institutional libraries
  - Often under-utilised advice and guidance
- FAIR templates and guides
- Any data is better than no data!
  - Learn by doing



The FAIR Principles

Insecurities



SORTEE

SETTING THE RECORD STRAIGHT: HOW DATA AND CODE TRANSPARENCY CAUGHT AN ERROR AND HOW I FIXED IT

BY ANDREW ANDERSON | DECEMBER 4, 2023

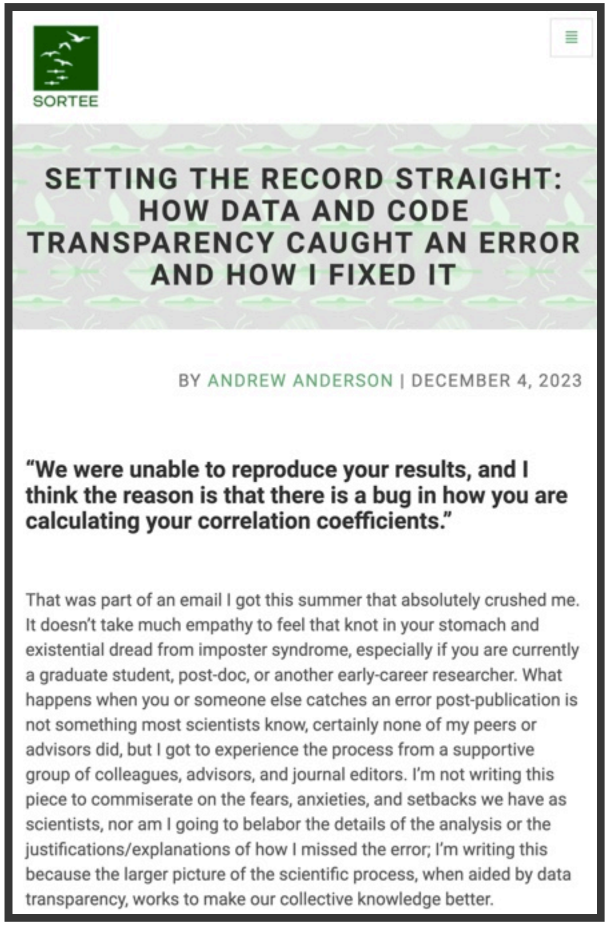
“We were unable to reproduce your results, and I think the reason is that there is a bug in how you are calculating your correlation coefficients.”

That was part of an email I got this summer that absolutely crushed me. It doesn’t take much empathy to feel that knot in your stomach and existential dread from imposter syndrome, especially if you are currently a graduate student, post-doc, or another early-career researcher. What happens when you or someone else catches an error post-publication is not something most scientists know, certainly none of my peers or advisors did, but I got to experience the process from a supportive group of colleagues, advisors, and journal editors. I’m not writing this piece to commiserate on the fears, anxieties, and setbacks we have as scientists, nor am I going to belabor the details of the analysis or the justifications/explanations of how I missed the error; I’m writing this because the larger picture of the scientific process, when aided by data transparency, works to make our collective knowledge better.

Blog post by Andrew Anderson

- Early career researchers can feel especially vulnerable
- Fear, insecurity and embarressment are powerful emotions

Insecurities



SORTEE

SETTING THE RECORD STRAIGHT: HOW DATA AND CODE TRANSPARENCY CAUGHT AN ERROR AND HOW I FIXED IT

BY ANDREW ANDERSON | DECEMBER 4, 2023

“We were unable to reproduce your results, and I think the reason is that there is a bug in how you are calculating your correlation coefficients.”

That was part of an email I got this summer that absolutely crushed me. It doesn’t take much empathy to feel that knot in your stomach and existential dread from imposter syndrome, especially if you are currently a graduate student, post-doc, or another early-career researcher. What happens when you or someone else catches an error post-publication is not something most scientists know, certainly none of my peers or advisors did, but I got to experience the process from a supportive group of colleagues, advisors, and journal editors. I’m not writing this piece to commiserate on the fears, anxieties, and setbacks we have as scientists, nor am I going to belabor the details of the analysis or the justifications/explanations of how I missed the error; I’m writing this because the larger picture of the scientific process, when aided by data transparency, works to make our collective knowledge better.

Blog post by Andrew Anderson

- Share before publication
  - Lab meetings or data review sessions
  - Pre-print (private or open)
- Data being hard to understand is bigger issue
- Culture that prioritises learning over citisism

Don’t see value in their data

- Too niche
- Too small
- Why would someone be interested?

Photo by Diego PH

Don’t see value in their data

- Highly subjective
- Hard to predict future use
- + all other benefits

Photo by Diego PH

Re-use concerns

Misinterpretation

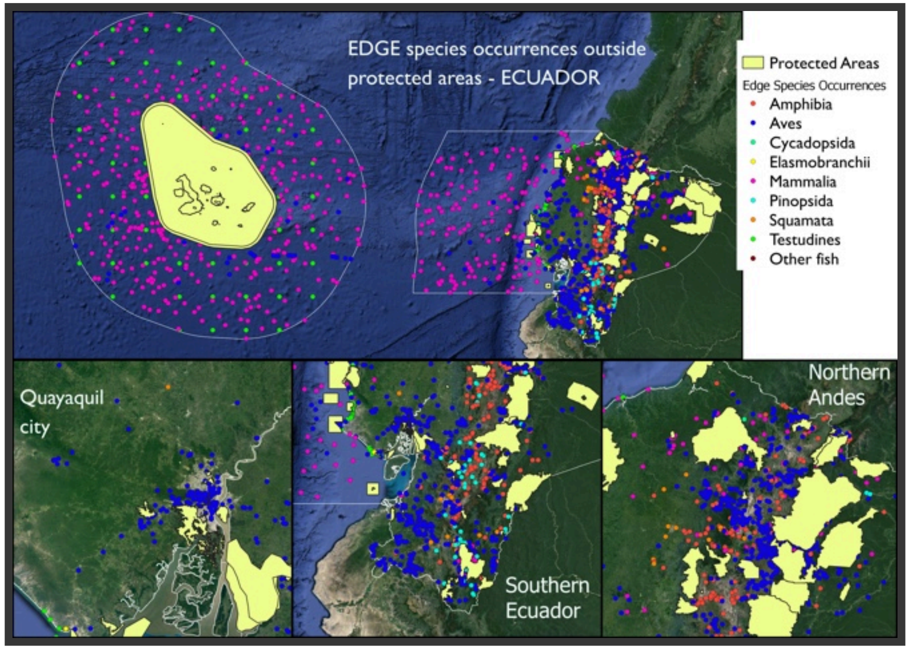
- Fear of inappropriate use
  - Lack of familiarity with particular dataset
  - Miss crucial details and draw misleading conclusions

Misinterpretation

- High quality metadata
    - Peer review
- Contactable
- Not a unique problem to data

Sensitive information

- Dual use problem
- Weigh up benefits and costs
- Ethical (and legal) implications
- Sharing limited subset
  - Species example guidelines: Chapman 2020



GBIF

Disincentives

Scooping

Fear of:

- A researcher performs an analysis on publicly shared data that the original data collected had not done yet
  - Being “scooped”
- Reduced collaborations
- Loss of future publications
  - Metric used to assess performance

Photo by Saher Suthriwala

Scooping

Less likely than you would imagine:

- Ideas are plentiful
- Original collectors in best position to act
- Most analyses by original authors on published data happen within 2 years[1]
- Most analysis by other researchers peak at 5 years^1

Photo by Saher Suthriwala

Scooping

- Pre-print to “claim”
- If major concern:
  - restrictions on use of data can be made
  - embargo periods
- Change in mindset to see data as a valuable contribution

Photo by Saher Suthriwala

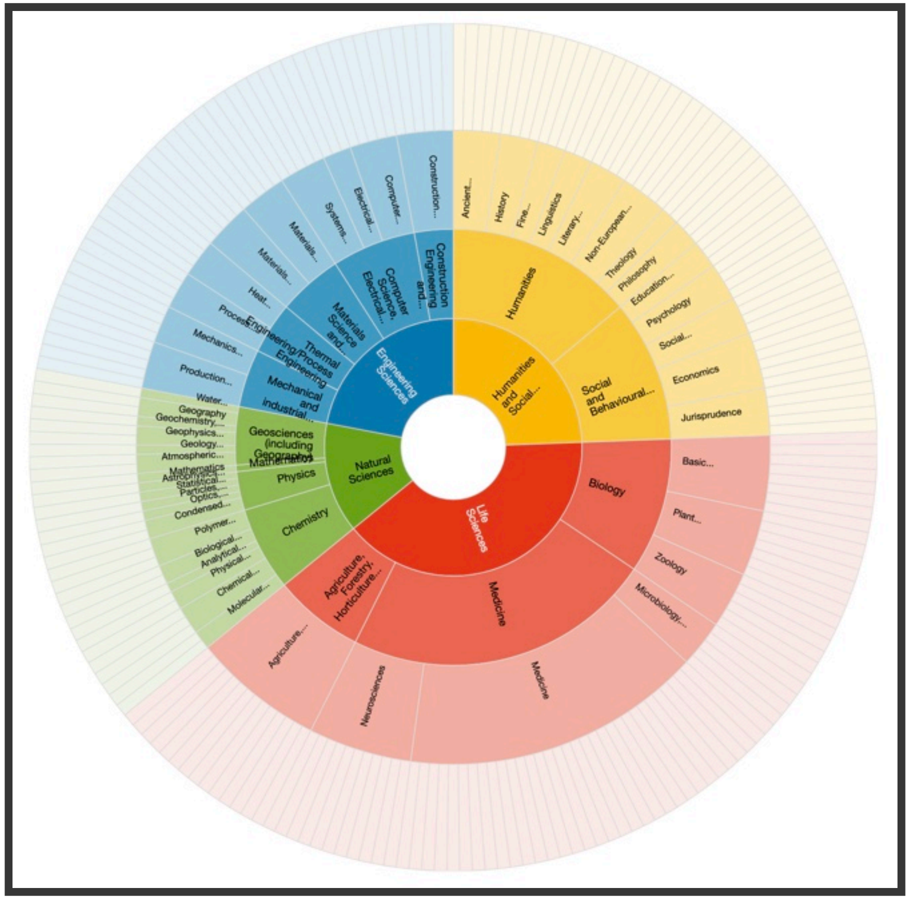
How to work towards data transparency

1. Plan to publish your data!

- What data needs to be recorded?
- What metadata might be needed?
- How raw/cleaned should my data be?
- Talk with collaborators early about plans

2. Identify an appropriate repository

- Field specific
- Data type specific
- Journal preferences
- Good starting place: re3data.org



Subjects covered by re3data.org

3. Make a nice README file

- One or more plain text files that describe the data in detail
- Write early!
- Check repository guidelines
- Document your data

4. Pre-peer-review peer-review

- Ask a colleague to look through your README and dataset
  - Data/code review sessions
  - Can they make sense of it?

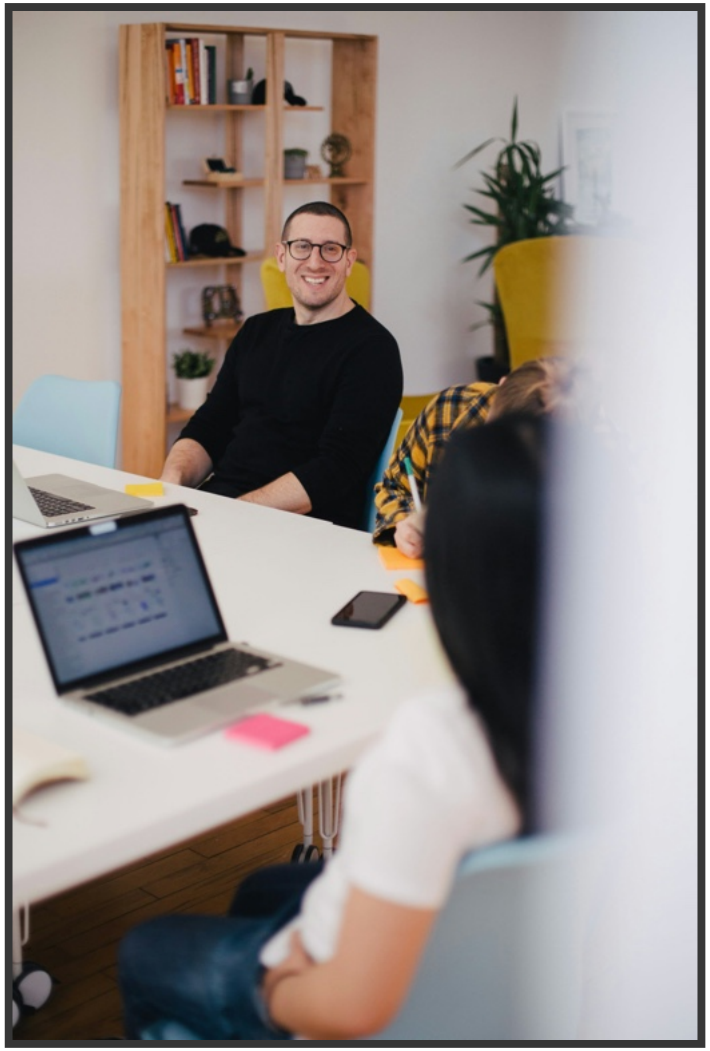


Photo by Jason Goodman

5. Publish your data

- Make sure it has a citable DOI
- Cite your data in your publication!
- Talk about it with your colleagues

Thank you for listening

- Slides & references:
  - irmoodie.com/slides/datatransparency-stockholm-2024
- Want to learn more:
  - www.sortee.org
- Contact me:
  - iain.moodie@biol.lu.se or irmoodie.com
- Questions?