Data Sharing & Standards

Emma Dunne | APW 2023 | Toolkit Day 1

Paleontology & ‘Big Data’

- Large data compilations in have opened up—and continue to inspire—vast new research areas
    - **Analytical/Quantitative paleobiology**

Image: University of Chicago

Paleontology & ‘Big Data’

- Large data compilations in have opened up—and continue to inspire—vast new research areas
    - **Analytical/Quantitative paleobiology**
- Advances in tools to handle and analyse these data
- Required several changes to research protocols
    - Data management, sharing, and citation

Image: University of Chicago

Open science

Open science

- Transparent and accessible knowledge
- Shared and developed through collaborative networks
- Open to all levels of society
- Examples:
    - Open access publishing
    - Science communication
    - Data and code sharing

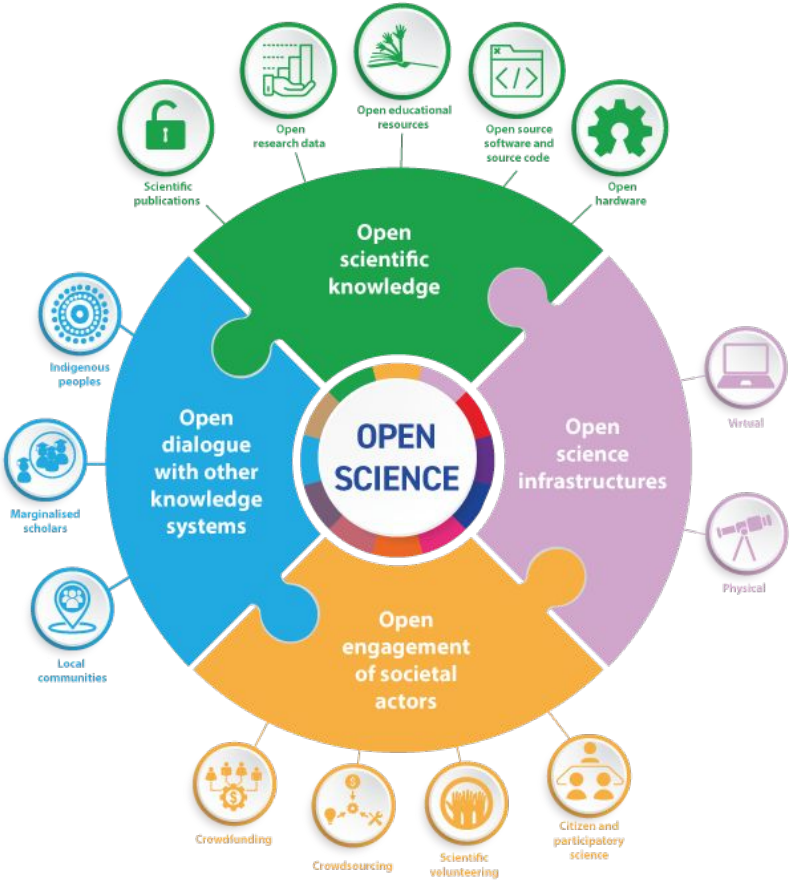


Image: Pillars of Open Science, UNESCO

Open data

- Openly accessible, exploitable, editable and shared by anyone for any purpose
- Open Data Handbook requires that the data be:

  A. Legally open = open license

  - Licensed under an open license (e.g. Creative Commons CC0)

  B. Technically open

  - Accessible and at no extra cost

Image: data.europa.eu

Benefits of open data

Benefits of open data

Open data is good for research and researchers:

- Reproducibility of studies
- Transparency — greater research integrity
- Increased accessibility of resources
- Expansion of ideas and research opportunities
- Increased engagement (within and outside of academia)
- Can even improve citations (Maitner *et al.* 2023)

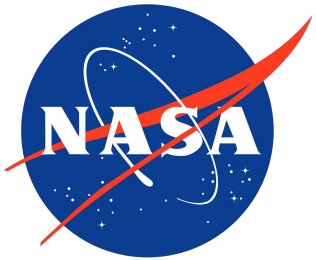


Image: Open data sources

Reluctance to share data & code

92% of publications in Agricultural and Biological Sciences fail to share code (in comparison, only 49% fail to share data) (PLOS 2023)

95% of ecology and evolution publications since 2010 don't share their code (Maitner *et al.* 2023)

- Unfamiliarity with best sharing practices
- Insecurity about code quality
- Fears of misuse
- Excess preparation costs

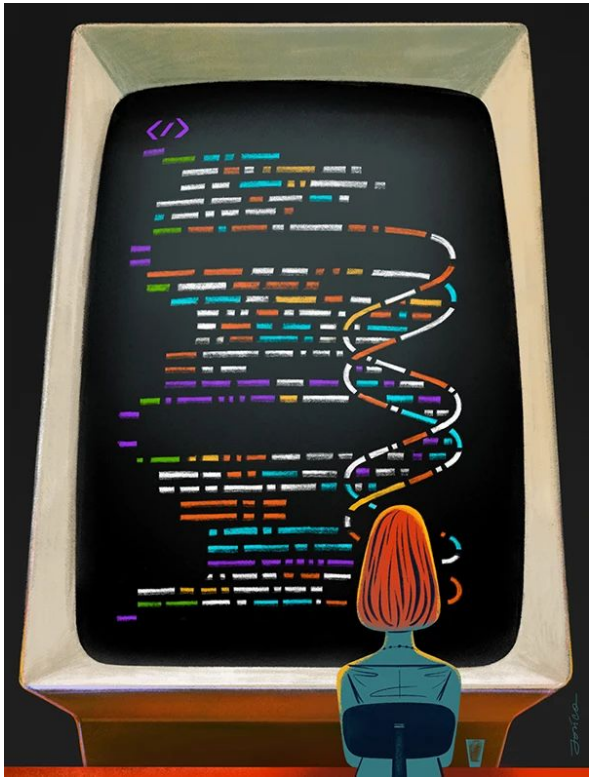


Image: Paweł Jońca

Reluctance to share data & code

92% of publications in Agricultural and Biological Sciences fail to share code (in comparison, only 49% fail to share data) (PLOS 2023)

95% of ecology and evolution publications since 2010 don't share their code (Maitner *et al.* 2023)

- **Unfamiliarity with best sharing practices**
- **Insecurity about code quality**
- Fears of misuse
- **Excess preparation costs**

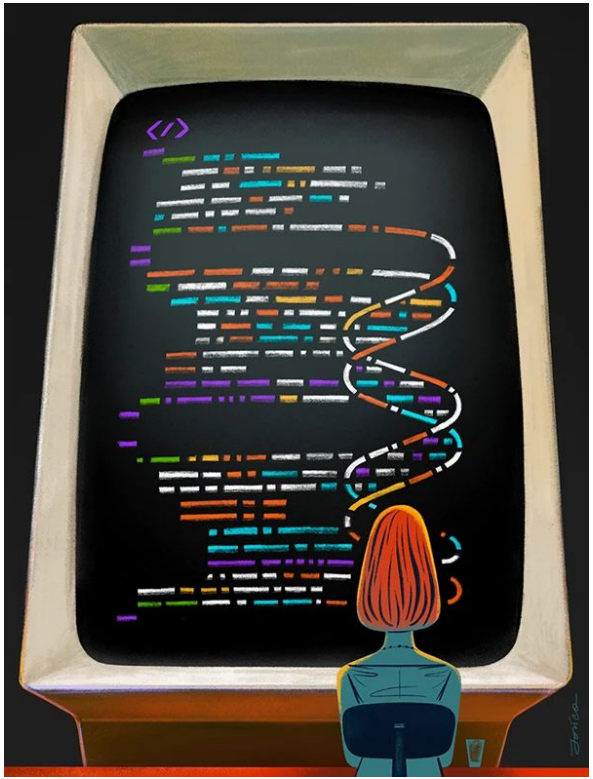


Image: Paweł Jońca

Data & code sharing

- Requires adherence to certain standards
- FAIRsharing = resource on data and metadata standards, inter-related to databases and data policies
- Many different repositories to choose from
- **DOI** = a **d**igital **o**bject **i**dentifier to track digital/physical/abstract items

Images: Data sharing repositories logos

Activity

Data & code sharing in recent (2010–present) analytical paleobiology papers

Record results here:
menti.com
code: **1835 9838**

Preprints

- A preprint is an openly available scientific manuscript that an author uploads to a public server prior to peer review
- Assigned a DOI
- Examples:
  - *EcoEvoRxiv*
  - *bioRxiv*
  - *EarthArXiv*
  - *OSF Preprints*

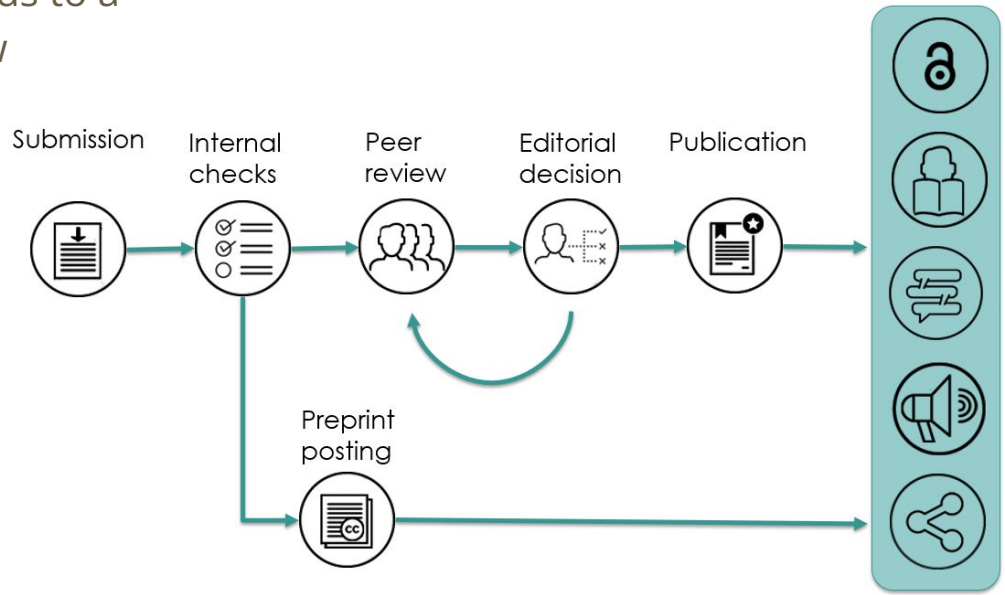


Image: PLOS Blog

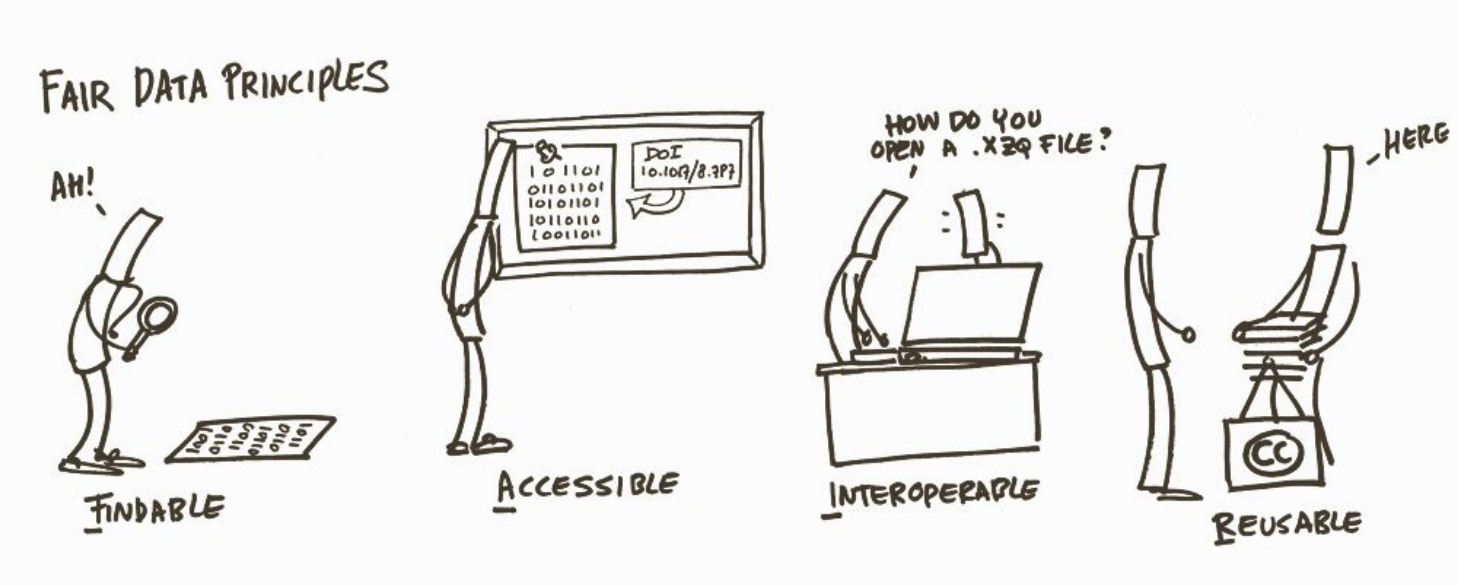
Open data standards

- Reusable agreements that help researchers and organisations to publish, access, share and use better quality data
  - Individuals and teams
  - Museums, universities, etc.
- Biodiversity Information Standards (TDWG)
  - *“promotes standards and guidelines for the recording and exchange of data about organisms”*

Biodiversity Information Standards
TDWG

FAIR Guiding Principles

- Developed to enhance data discovery and reuse (Wilkinson *et al.* 2016)



TRUST Principles

- Developed to demonstrate the trustworthiness of digital repositories (Lin *et al.* 2020)
- *"Repositories must earn the trust of the communities they intend to serve and demonstrate that they are reliable and capable of appropriately managing the data they hold"*

CARE Principles of Indigenous Data Governance

- Promote the ethical use and reuse of Indigenous data (Carroll *et al.* 2020)
- Developed by the International Indigenous Data Sovereignty Interest Group
- Complement the FAIR Guiding Principles

Research integrity

- Several flavours of **Questionable Research Practices** in the statistical analysis of data and the presentation of the results (e.g. ***P*-hacking**)
- In ecology and evolution (Fraser *et al.* 2018):
  - *"64% of surveyed researchers reported they had at least once failed to report results because they were not statistically significant" (**Cherry picking**)*
  - *"51% had reported an unexpected finding as though it had been hypothesised from the start" (**HARKing**)*

Image: Dirk-Jan Hoek

Pruitt data scandal

- Jonathan Pruitt, behavioural ecologist at McMaster University, Canada
- Resigned after 2+ years of allegations of **data irregularities** (Viglione, 2020)
- Numerous **retractions** (17 on last count, amounting to 900+ citations) (Kozlov, 2022)
- Pruitt blames *"mistakes in data management"*
- Students, (former) lab members and collaborators still dealing with the fallout

Image: Bernard Dupont (CC BY-SA 2.0)

Tanis data scandal

- Robert de Palma, paleobiologist at the University of Manchester, UK
- Accused of **faking data** (Price, 2022)
    - *"plotted line graphs and figures [in the] paper contain numerous irregularities"*
- The raw, machine-produced data underlying the analyses is not publicly available
    - Analyst died years prior to publication
- **Race to publish** before others (During *et al.* 2022)

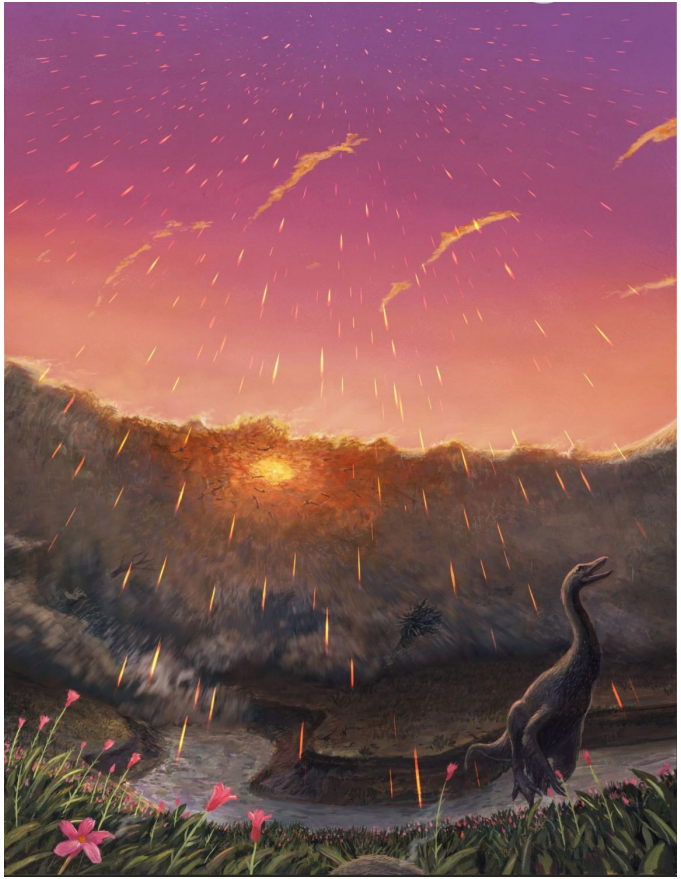


Image: Joschua Knüppe

Drivers of unethical behaviour

- Pressure to publish – “publish or perish“ culture (Raja & Dunne, 2022)
- Financial incentives
- Lack of oversight (limited or no consequences for misconduct)
- Poor research culture
- Competitive environment
- Lack of training or awareness
- Personal and emotional factors

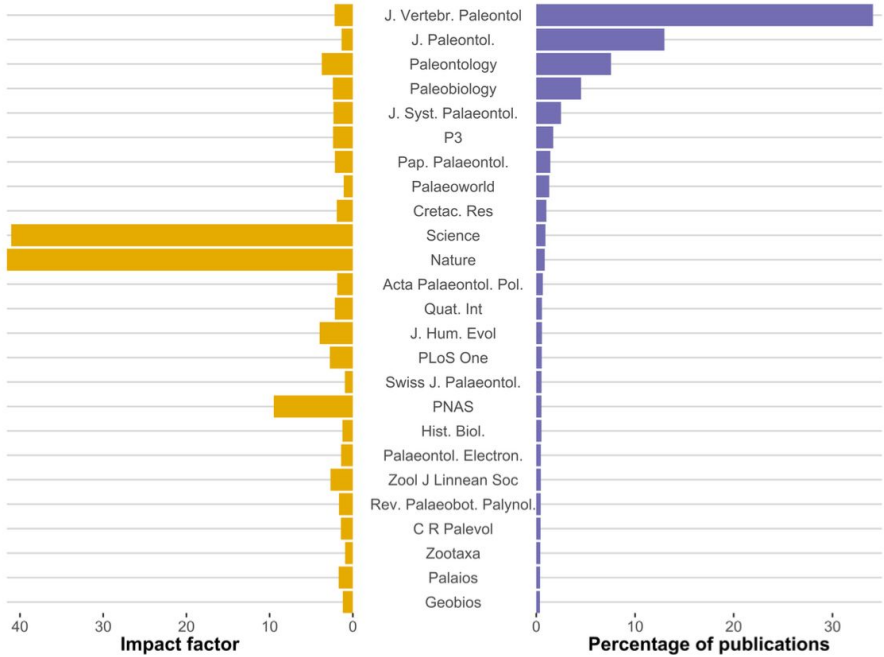


Image: Raja & Dunne (2022)

Data & code sharing in paleobiology

- Paleobiology lags behind other fields (Dillon *et al.* 2023)
- Several data standard initiatives launched:
  - Paleo Data Working Group
  - Enabling FAIR Data project (for Earth, Space, and Environmental Science)
- More and more paleobiologists are using large datasets and code in their analyses
  - Training opportunities & resources
  - Interoperability & future-proofing